Universidade Federal do Piauí
Centro de Ciências da Natureza
Programa de Pós-Graduação em Ciência da Computação

Aprimorando a regulação de serviços assistenciais a partir da triagem automática de solicitações

Hugo de Oliveira Cordeiro

Teresina-PI, Março de 2017

Hugo de Oliveira Cordeiro

Aprimorando a regulação de serviços assistenciais a partir da triagem automática de solicitações

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da UFPI (área de concentração: Sistemas de Computação), como parte dos requisitos necessários para a obtenção do Título de Mestre em Ciência da Computação.

Universidade Federal do Piauí – UFPI

Centro de Ciências da Natureza

Programa de Pós-Graduação em Ciência da Computação

Orientador: Pedro de Alcântara dos Santos Neto

Teresina-PI

Março de 2017

Hugo de Oliveira Cordeiro
Aprimorando a regulação de serviços assistenciais a partir da triagem automática de solicitações/ Hugo de Oliveira Cordeiro. – Teresina-PI, Março de 2017-
117 p. : il. (algumas color.) ; 30 cm.

Orientador: Pedro de Alcântara dos Santos Neto

Dissertação (Mestrado) – Universidade Federal do Piauí – UFPI
Centro de Ciências da Natureza
Programa de Pós-Graduação em Ciência da Computação, Março de 2017.

1. Regulação de Solicitações de Serviços Assistenciais. 2. Inteligência Artificial. 3. Descoberta de Conhecimento em Bancos de Dados. 4. Mineração de Dados. 5. Computação Aplicada. I. Pedro de Alcântara dos Santos Neto. II. Universidade Universidade Federal do Piauí III. Aprimorando a regulação de serviços assistenciais a partir da triagem de solicitações

CDU 02:141:005.7

Hugo de Oliveira Cordeiro

Aprimorando a regulação de serviços assistenciais a partir da triagem automática de solicitações

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da UFPI (área de concentração: Sistemas de Computação), como parte dos requisitos necessários para a obtenção do Título de Mestre em Ciência da Computação.

Trabalho aprovado. Teresina-PI, 29 de Março de 2017:

Pedro de Alcântara dos Santos Neto
Orientador

Cleber Zanchettin
Convidado 1

Ricardo de Andrade Lira Rabêlo
Convidado 2

Vinícius Ponte Machado
Convidado 3

Teresina-PI
Março de 2017

"Computers have promised us a fountain of wisdom
but delivered a flood of data."
(A frustrated MIS executive)

Resumo

Evitar o desperdício de recursos é um dos maiores desafios para a melhoria do serviço de atenção à saúde no Brasil. No âmbito das operadoras de planos de saúde brasileiras, destaca-se a regulação de solicitações de serviços assistenciais como mecanismo de combate ao desperdício oriundo de erro médico, fraude e abuso. Tradicionalmente a avaliação das solicitações é um processo manual e, considerando o volume de solicitações, torna-se um mecanismo de manutenção cara e não escalável que atrasa o acesso do paciente ao serviço além de causar desgaste na interação entre operadoras de planos de saúde e prestadores de serviços. Este trabalho propõe a triagem automática de solicitações para otimizar o processo de regulação. Para isso, técnicas de mineração de dados foram utilizadas para a construção de modelos preditivos a fim de responder automaticamente solicitações que tiverem a probabilidade de autorização igual ou superior a um fator de confiança escolhido pela operadora de planos de saúde, encaminhando o restante das solicitações para a avaliação manual. A abordagem proposta foi avaliada em sete bases de dados disponíveis para o estudo, contemplando seis operadoras de planos de saúde públicas e uma privada. Nos experimentos realizados foi possível autorizar automaticamente entre 20% e 90% das solicitações mediante a variação do fator de confiança. Esses resultados mostram a viabilidade da utilização da triagem proposta para reduzir a carga de trabalho de avaliadores humanos de acordo com a necessidade das operadoras de planos de saúde, o que pode tornar o processo mais rápido, escalável e mais barato.

Palavras-chaves: Mineração de Dados, Planos de Saúde, Regulação de Solicitações de Serviços Assistenciais.

Abstract

Avoiding resources waste is one of the greatest challenges for improving health care services in Brazil. Within the scope of Brazilian health plan providers, services requests prior authorization is highlighted as a mechanism to inhibit wastage originated from medical mistakes, fraud, and abuse. Traditionally the requests evaluation is a manual process and considering the requests volume it becomes an expensive and nonscalable mechanism which delay patient access to service besides causing friction between health plan providers and health care providers. This work proposes the requests triage to optimize the prior authorization process. Thereunto, data mining techniques were used to create predictive models to automatically answer requests that present the authorization probability greater or equals to a confidence factor chosen by the health plan provider, forwarding the remainder of the requests to manual evaluation. The proposed approach was assessed in seven databases available for the study, comprising of six public health plan providers and one private. In the performed experiments it was possible to authorize automatically from 20% to 90% of the requests by varying the confidence factor. These results demonstrate the viability of using this approach to reduce human evaluation workload accordingly to health plans providers necessities, which can make the process faster, scalable and cheaper.

Keywords: Data Mining, Health Plan Providers, Healthcare Prior Authorization.

Lista de ilustrações

Figura 1 – Quantidade de beneficiários associados a OPSs no período de 2007 a 2014 (ANS, 2015a). 2
Figura 2 – Margem de lucro das OPSs no período de 2007 a 2014 (ANS, 2015a). . 2
Figura 3 – Despesas de OPSs, por tipo, no período de 2007 a 2014 (ANS, 2015a). 3
Figura 4 – Processo de regulação de solicitações de serviços assistenciais. 4
Figura 5 – Taxonomia dos métodos de mineração de dados, baseado no trabalho de Maimon (MAIMON; ROKACH, 2005), com destaque no foco deste trabalho 17
Figura 6 – Histograma da quantidade de exemplos no conjunto de testes (eixo das ordenadas) para a estimativa da probabilidade para o rótulo "positivo" (eixo das abcissas). 24
Figura 7 – Processo de regulação com a triagem automática. 27
Figura 8 – Etapas da metodologia para implantação da triagem automática em OPSs. 28
Figura 9 – Estratificação e balanceamento das bases de dados. 35
Figura 10 – Treinamento dos algoritmos a partir dos dados estratificados e balanceados. 35
Figura 11 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados I com o algoritmo C4.5. 38
Figura 12 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados II com o algoritmo C4.5. 38
Figura 13 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados IV com o algoritmo C4.5. 39
Figura 14 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados V com o algoritmo C4.5. 39
Figura 15 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados VI com o algoritmo C4.5. 40
Figura 16 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados VII com o algoritmo C4.5. 40
Figura 17 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados I com o algoritmo *Ripper*. 42
Figura 18 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados II com o algoritmo *Ripper*. 43
Figura 19 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados IV com o algoritmo *Ripper*. 43
Figura 20 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados V com o algoritmo *Ripper*. 44

Figura 21 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados VI com o algoritmo *Ripper*. 44
Figura 22 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados VII com o algoritmo *Ripper*. 45
Figura 23 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados I com o algoritmo *NaiveBayes*. 46
Figura 24 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados II com o algoritmo *NaiveBayes*. 47
Figura 25 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados IV com o algoritmo *NaiveBayes*. 47
Figura 26 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados V com o algoritmo *NaiveBayes*. 48
Figura 27 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados VI com o algoritmo *NaiveBayes*. 48
Figura 28 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados VII com o algoritmo *NaiveBayes*. 49
Figura 29 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados I com o algoritmo *Random Forest*. 52
Figura 30 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados II com o algoritmo *Random Forest*. 52
Figura 31 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados IV com o algoritmo *Random Forest*. 53
Figura 32 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados V com o algoritmo *Random Forest*. 53
Figura 33 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados VI com o algoritmo *Random Forest*. 54
Figura 34 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados VII com o algoritmo *NaiveBayes*. 54
Figura 35 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados I com o algoritmo SVM. 57
Figura 36 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados II com o algoritmo SVM. 57
Figura 37 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados IV com o algoritmo SVM. 58
Figura 38 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados V com o algoritmo SVM. 58
Figura 39 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados VI com o algoritmo SVM. 59

Figura 40 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados VII com o algoritmo SVM. 59
Figura 41 – Fator de Confiança de 10,10% para a combinação entre a Base de Dados I, a técnica *Random Oversampling* e o algoritmo *Naive Bayes*. 63
Figura 42 – Fator de Confiança de 98,30% para a combinação entre a Base de Dados I, a técnica *Random Oversampling* e o algoritmo *Naive Bayes*. 64
Figura 43 – Fator de Confiança de 80,50% para a combinação entre a Base de Dados II, a técnica *MetaCost* e o algoritmo *Random Forest*. 65
Figura 44 – Fator de Confiança de 100,00% para a combinação entre a Base de Dados II, a técnica *MetaCost* e o algoritmo *Random Forest*. 66
Figura 45 – Fator de Confiança de 38,80% para a combinação entre a Base de Dados IV, a técnica *Random Oversampling* e o algoritmo *SVM*. 67
Figura 46 – Fator de Confiança de 86,60% para a combinação entre a Base de Dados IV, a técnica *Random Oversampling* e o algoritmo *SVM*. 68
Figura 47 – Fator de Confiança de 11,00% para a combinação entre a Base de Dados V, a técnica *Random Oversampling* e o algoritmo *Naive Bayes*. 69
Figura 48 – Fator de Confiança de 99,90% para a combinação entre a Base de Dados V, a técnica *Random Oversampling* e o algoritmo *Naive Bayes*. 70
Figura 49 – Fator de Confiança de 17,90% para a combinação entre a Base de Dados VI, a técnica *Random Oversampling* e o algoritmo *Naive Bayes*. 71
Figura 50 – Fator de Confiança de 99,10% para a combinação entre a Base de Dados VI, a técnica *Random Oversampling* e o algoritmo *Naive Bayes*. 72
Figura 51 – Fator de Confiança de 41,20% para a combinação entre a Base de Dados VII, a técnica *Random Oversampling* e o algoritmo *Naive Bayes*. 73
Figura 52 – Fator de Confiança de 79,20% para a combinação entre a Base de Dados VII, a técnica *Random Oversampling* e o algoritmo *Naive Bayes*. 74

Lista de tabelas

Tabela 1 – Matriz de Confusão. 21
Tabela 2 – Características das bases de dados (BD) utilizadas. 32
Tabela 3 – Exemplos de atributos removidos durante a seleção manual. 33
Tabela 4 – Criação dos atributos “Idade” a partir do atributo “Data Nascimento”. 34
Tabela 5 – Características das bases de dados (BD) após a seleção de solicitações e a seleção manual de atributos realizada por especialistas da empresa parceira. 36
Tabela 6 – Características das Bases de Dados (BD) utilizadas após a construção de atributos e seleção automática de atributos. 36
Tabela 7 – Fatores de confiança mínimos e máximos, por base de dados, para o classificador C4.5. 41
Tabela 8 – Fatores de confiança mínimos e máximos, por base de dados, para o classificador *Ripper*. 46
Tabela 9 – Fatores de confiança mínimos e máximos, por base de dados, para o classificador *Naive Bayes* - Sem Tratamento e SMOTE. 50
Tabela 10 – Fatores de confiança mínimos e máximos, por base de dados, para o classificador *Naive Bayes* - *Random Oversampling* e *MetaCost*. 51
Tabela 11 – Fatores de confiança mínimos e máximos, por base de dados, para o classificador *Random Forest* - Sem Tratamento e SMOTE. 55
Tabela 12 – Fatores de confiança mínimos e máximos, por base de dados, para o classificador *Random Forest* - *Random Oversampling* e *MetaCost*. . . . 56
Tabela 13 – Fatores de confiança mínimos e máximos, por base de dados, para o classificador SVM. 60
Tabela 14 – Fatores de confiança para a combinação entre a Base de Dados I, a técnica *Random Oversampling* e o algoritmo *Naive Bayes*. 62
Tabela 15 – Fatores de confiança para a combinação entre a Base de Dados II, a técnica *MetaCost* e o algoritmo *Random Forest*. 63
Tabela 16 – Fatores de confiança para a combinação entre a Base de Dados IV, a técnica *Random Oversampling* e o algoritmo *SVM*. 65
Tabela 17 – Fatores de confiança para a combinação entre a Base de Dados V, a técnica *Random Oversampling* e o algoritmo *Naive Bayes*. 66
Tabela 18 – Fatores de confiança para a combinação entre a Base de Dados VI, a técnica *Random Oversampling* e o algoritmo *Naive Bayes*. 68
Tabela 19 – Fatores de confiança para a combinação entre a Base de Dados VII, a técnica *Random Oversampling* e o algoritmo *Naive Bayes*. 69

Lista de abreviaturas e siglas

ANS	Agência Nacional de Saúde Suplementar
BD	Base de Dados
DCBD	Descoberta de Conhecimento em Bancos de Dados
FC	Fator de Confiança
OPS	Operadora de Planos de Saúde
SUS	Sistema Único de Saúde
TVN	Taxa de Verdadeiros Negativos
TVN	Taxa de Verdadeiros Positivos
VPN	Valor Preditivo Negativo
VPN	Valor Preditivo Positivo

Sumário

1 INTRODUÇÃO 1
1.1 Definição do Problema 3
1.2 Justificativa 4
1.3 Objetivos 5
1.4 Visão Geral da Proposta 5
1.5 Contribuições 6
1.6 Estrutura do Trabalho 7

2 TRABALHOS RELACIONADOS 9
2.1 Mapeamento Sistemático de Estudos 9
2.2 Trabalhos não incluídos no Mapeamento Sistemático de Estudos . . 12
2.3 Considerações Finais 14

3 DESCOBERTA DE CONHECIMENTO EM BANCOS DE DADOS 15
3.1 Pré-processamento 15
3.2 Mineração de dados 16
3.2.1 Aprendizado de máquina 17
3.2.1.1 Terminologia e definições 18
3.2.1.2 Tipos de aprendizado 19
3.2.1.3 Paradigmas de aprendizado 19
3.2.1.3.1 Paradigma bayesiano 20
3.2.1.3.2 Paradigma baseado em árvores de decisão 20
3.2.1.3.3 Paradigma estatístico 21
3.2.1.4 Avaliação de Algoritmos 21
3.3 Técnicas de Tratamento de Desbalanceamento 22
3.4 Validação cruzada 23
3.5 Fator de Confiança 23

4 TRIAGEM AUTOMÁTICA DE SOLICITAÇÕES DE SERVIÇOS ASSISTENCIAIS EM SAÚDE 25
4.1 Metodologia para a implantação da triagem automática em OPSs . 28
4.1.1 Definição do escopo 28
4.1.2 Construção dos modelos preditivos 29
4.1.3 Análise do fator de confiança 29
4.1.4 Avaliação do impacto 29

5 AVALIAÇÃO DA TRIAGEM AUTOMÁTICA . . . 31
5.1 Questões e métricas . . . 31
5.2 Execução do Experimento . . . 31
5.2.1 Definição do escopo . . . 32
5.2.2 Construção dos modelos preditivos . . . 32
5.2.2.1 Pré-processamento . . . 33
5.2.2.2 Mineração de dados . . . 34
5.2.2.2.1 Resultados . . . 36
5.2.3 Análise do fator de confiança . . . 37
5.2.3.1 Resultados . . . 37
5.2.3.1.1 C4.5 . . . 37
5.2.3.1.2 *Ripper* . . . 42
5.2.3.1.3 *NaiveBayes* . . . 45
5.2.3.1.4 *Random Forest* . . . 50
5.2.3.1.5 SVM . . . 55
5.2.3.1.6 Considerações . . . 60
5.2.4 Avaliação do impacto . . . 61
5.2.4.1 Base de Dados I . . . 62
5.2.4.2 Base de Dados II . . . 62
5.2.4.3 Base de Dados IV . . . 64
5.2.4.4 Base de Dados V . . . 65
5.2.4.5 Base de Dados VI . . . 67
5.2.4.6 Base de Dados VII . . . 69
5.3 Ameaças à validade . . . 70
5.3.1 Validade externa . . . 71
5.3.2 Validade Interna . . . 72

6 CONCLUSÕES . . . 75
6.1 Limitações . . . 76
6.2 Continuidade da Pesquisa . . . 76

REFERÊNCIAS . . . 79

APÊNDICES 83

APÊNDICE A – ATRIBUTOS UTILIZADOS POR BASE DE DADOS . . . 85

APÊNDICE B – RESULTADOS . . . 87

1 Introdução

A saúde é algo primordial para a vida humana estando diretamente relacionada à visão de futuro e planos de vida de qualquer pessoa (JAMIESON, 1995). Melhores condições de saúde constituem um fator central para felicidade e bem estar, além de ter uma importante contribuição para o progresso econômico, uma vez que populações saudáveis vivem mais, são mais produtivas e significam um gasto menor para os governos (WHO, 2010a).

No Brasil, a assistência à saúde é um direito constitucional (BRASIL, 1988) e apresenta dois subsistemas: um público e um privado. No subsistema público, a assistência à saúde é entregue gratuitamente a cidadãos brasileiros e estrangeiros por meio do Sistema Único de Saúde (SUS), de responsabilidade governamental; enquanto o segundo subsistema engloba as Operadoras de Planos de Saúde (OPS) privadas, que prestam um serviço complementar supervisionado pela Agência Nacional de Saúde Suplementar (ANS). Segundo a ANS, o mercado complementar constitui-se em um dos maiores sistemas privados de saúde do mundo (ANS, 2015b). Isso verifica-se em números absolutos sob a análise da abrangência de atendimento com 1.370 OPSs prestando assistência médica a cerca de 70 milhões de beneficiários (ANS, 2015a). Além disso, dados da ANS (ANS, 2015a) indicam um crescimento, desde o ano de 2001, do número de beneficiários associados às OPSs. A Figura 1 ilustra esse crescimento. Em números relativos, no entanto, essa abrangência representa a cobertura de somente cerca de 30% da população brasileira, e está concentrada na região sudeste do país, onde mais de 60% das OPSs se encontram e mais de 65% dos contratos são mantidos (ANS, 2015a). Essas circunstâncias mostram que ainda há uma significativa parcela da população brasileira que não é atendida e que há potencial de crescimento para o mercado de saúde privado brasileiro.

Contudo, para continuar a crescer e incluir a parcela da população brasileira ainda não atendida, as OPSs precisam enfrentar os baixos lucros. É possível observar na Figura 2 uma margem de lucro média inferior a 1% no período de 2007 a 2014, segundo a ANS (ANS, 2015a). O crescimento na quantidade de beneficiários trouxe consigo, naturalmente, o aumento das despesas assistenciais (i.e. despesas resultantes da utilização da cobertura oferecida). A Figura 3 ilustra esse tipo de despesa como o principal fator dos baixos lucros, evidenciando a importância da gestão de custos assistenciais para que as OPSs mantenham o seu bem estar financeiro e, dessa maneira, consigam investir em crescimento e melhoria do serviço prestado.

Somado a isso, em uma estimativa conservadora, entre 20% e 40% dos recursos de saúde são desperdiçados e com a redução desse desperdício pode-se aumentar significativa-

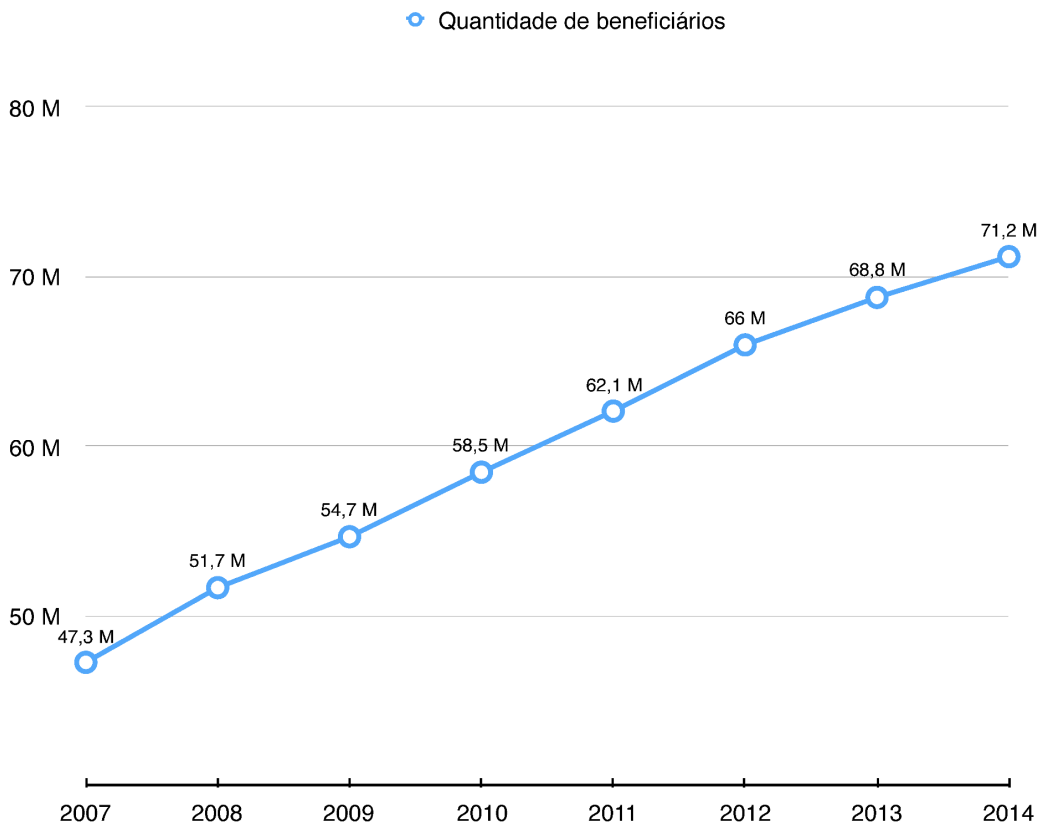

Figura 1 – Quantidade de beneficiários associados a OPSs no período de 2007 a 2014 (ANS, 2015a).

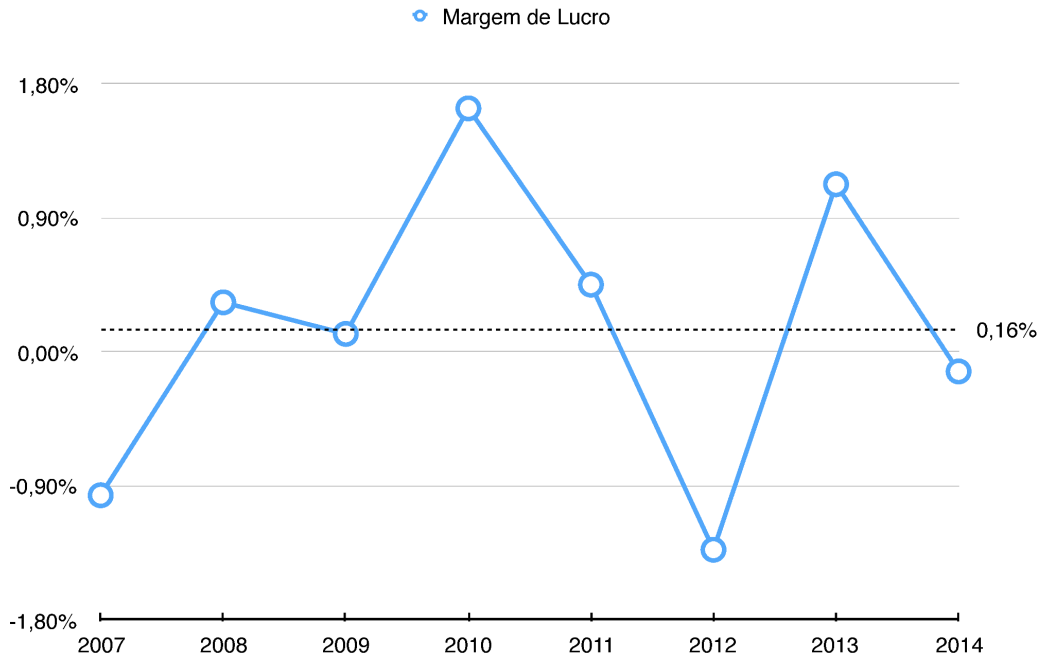

Figura 2 – Margem de lucro das OPSs no período de 2007 a 2014 (ANS, 2015a).

mente a capacidade de os sistemas de saúde fornecerem serviços de qualidade e melhorarem a saúde (WHO, 2010b). Esse desperdício pode ocorrer de várias maneiras, dentre as quais destaca-se a realização de assistência inadequada (i.e. que estão em discordância em

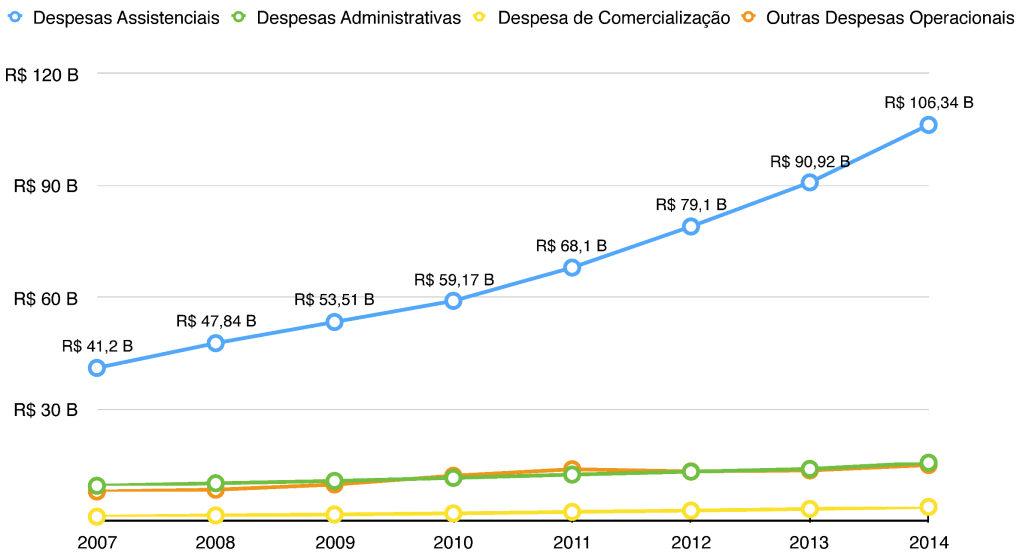

Figura 3 – Despesas de OPSs, por tipo, no período de 2007 a 2014 (ANS, 2015a).

relação aos protocolos estabelecidos por órgãos reguladores), pois além de causar incômodo aos pacientes, que são submetidos a exames e procedimentos desnecessários, eleva os custos assistenciais. Nesse aspecto, a implementação de métodos, técnicas e propostas para otimizar o uso de recursos torna-se essencial para o bom funcionamento das OPSs. Enfatiza-se a regulação de solicitações de serviços assistenciais para evitar o desperdício e mau uso dos recursos na prestação do atendimento médico aos beneficiários.

1.1 Definição do Problema

A regulação de solicitações de serviços assistenciais (doravante denominada "regulação") é uma área chave em um plano de saúde. Seu principal objetivo é tentar garantir o acesso à alternativa assistencial mais adequada à necessidade do beneficiário, evitando o desperdício oriundo de erro médico, fraude e abuso. Parte do processo envolvido na atividade de regulação é descrito a seguir: (1) durante o atendimento a um paciente, o profissional de saúde avalia a necessidade de realização de procedimentos para auxiliar no diagnóstico ou tratamento do beneficiário; (2) os procedimentos são solicitados via serviço disponibilizado pela OPS; (3) a solicitação então é direcionada à área de regulação que verifica a conformidade com os padrões e com os protocolos clínicos estabelecidos; (4) caso a análise seja positiva, a realização dos procedimentos é autorizada, (5) ou então negada, com o detalhamento dos motivos para negação dos itens solicitados. A Figura 4 ilustra esse processo.

Para o beneficiário, a regulação proporciona qualidade do serviço utilizado, uma vez que busca a melhor alternativa para a sua condição clínica, enquanto que para as OPSs, a utilização da regulação pode diminuir as despesas assistenciais, pois evita a realização

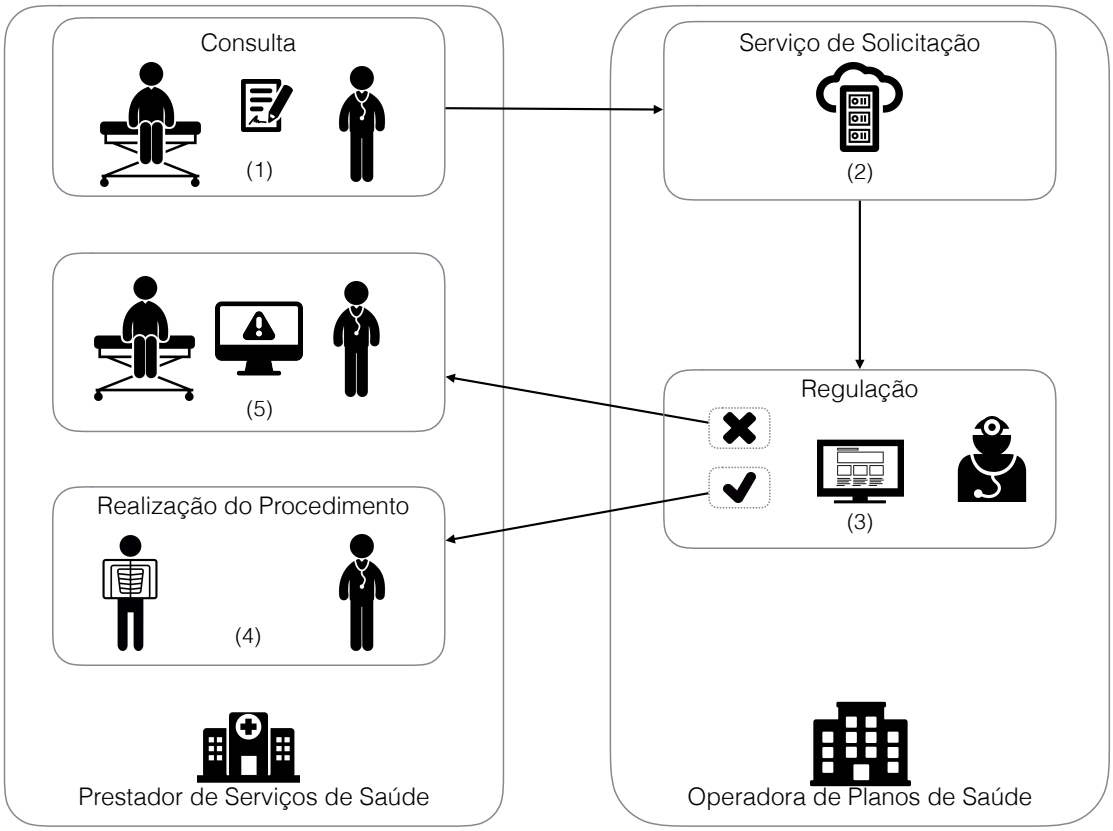

Figura 4 – Processo de regulação de solicitações de serviços assistenciais.

de serviços inadequados (desperdício assistencial).

1.2 Justificativa

Por se tratar de um processo que envolve uma decisão humana, a regulação pode ser incômoda para alguns beneficiários, que têm o acesso ao serviço atrasado pela análise da solicitação. Somado a isso, discordâncias sobre a aplicabilidade de determinado procedimento para a situação do beneficiário, podem causar conflitos com médicos e prestadores de serviço, além de possivelmente atrasar a resposta definitiva da solicitação. O cerne da questão, entretanto, está no investimento necessário para o estabelecimento e manutenção desse mecanismo de controle, custando para algumas OPSs mais de dez vezes o valor economizado diretamente no processo. Mas, por outro lado, a sua ausência é ainda mais onerosa, uma vez que, quando percebida por profissionais de saúde mal intencionados, permite a prática da fraude e do abuso assistencial para o benefício financeiro dos envolvidos.

O processo de regulação tradicional é oneroso para todos os atores envolvidos (é caro para as OPSs e incômodo para prestadores de serviços e beneficiários) e existem poucos trabalhos que se debruçam sobre os dados disponíveis nos sistemas de informação

utilizados para auxiliar a avaliação das solicitações. Devido a essa observação, foi escolhido como problema alvo a regulação em OPSs brasileiras. Acredita-se que existam informações relevantes implícitas nesses dados que podem ser utilizadas para automação do processo.

A escolha do mercado brasileiro de saúde privada deu-se pelas suas proporções absolutas, uma das maiores do mundo e também pelo potencial de crescimento da área, uma vez que abrange, aproximadamente, somente 30% da população brasileira. Desta maneira, soluções que permitam uma redução de gastos das operadoras de plano de saúde podem viabilizar a popularização dos planos privados e preencher a lacuna que o SUS deixou na atenção secundária à saúde (i.e. diagnóstico e tratamento médico especializado).

1.3 Objetivos

O objetivo principal deste trabalho consiste na proposição de um mecanismo para triagem automática de solicitações de serviços assistenciais em saúde (doravante denominada "triagem automática") em OPSs brasileiras, visando reduzir a carga de trabalho no processo tradicional de regulação, porém, sem comprometer a qualidade da assistência. Além do objetivo principal, pretende-se alcançar os seguintes objetivos específicos:

1. Avaliar quais técnicas de mineração de dados podem ser mais efetivas para automação da regulação;

2. Utilizar diferentes técnicas de tratamento de classes desbalanceadas para verificar como elas auxiliam na solução do problema;

3. Estimar o impacto da triagem automática no nível de redução da carga de trabalho no processo tradicional de regulação;

4. Avaliar a viabilidade da utilização da triagem automática para autorização de solicitações em saúde.

1.4 Visão Geral da Proposta

Na maioria das OPSs no Brasil o processo tradicional de regulação conta com o auxílio de sistemas de informação. Tais sistemas atuam na intermediação das solicitações e possuem os dados de quem solicitou, por que solicitou, o que solicitou, por quem foi analisado, quando e qual a resposta. Esses dados podem ser usados para a criação de modelos preditivos utilizando técnicas de mineração de dados.

Este trabalho propõe minimizar o trabalho humano na regulação por meio da automação da regulação a partir da aplicação de técnicas de mineração de dados para a criação de modelos capazes de fornecer a probabilidade de uma nova solicitação ser autorizada ou negada baseados em dados anteriores. A triagem automática é realizada por um autômato que avalia as solicitações e autoriza automaticamente aquelas que apresentam uma probabilidade para a autorização igual ou superior a um valor definido pela OPS. As demais solicitações são encaminhadas para a avaliação humana tradicional.

O método para introdução da triagem automática proposto neste trabalho é composto por quatro etapas: definição do escopo de atuação, construção dos modelos preditivos, análise do fator de confiança para customizar a automação e avaliação do impacto dessa triagem no processo de regulação. A primeira etapa consiste em definir precisamente a área de atuação, uma vez que uma OPS pode fornecer planos com diferentes segmentações de atendimento (hospitalar, ambulatorial e odontológico). Diferentes planos requerem regulações com características distintas, dessa maneira, a etapa de definição do escopo contempla a escolha do produto que deve receber a triagem automática. A construção dos modelos preditivos é realizada por meio de técnicas de mineração de dados. Após os tratamentos iniciais, a base de dados do produto escolhido é submetida aos algoritmos de mineração de dados para criar os modelos de predição. Tais modelos fornecem uma sugestão de resultado mas com um nível de certeza, aqui denominado de fator de confiança. Assim, uma das etapas é fazer um estudo sobre a variação de tal fator, visando entender como sua manipulação poderia impactar na triagem, analisando tanto a possível redução de trabalho manual de avaliação das solicitações (via automação da resposta), quanto da sua efetiva qualidade (acerto x erro) dentro do processo. Por fim, a avaliação do impacto da triagem automática é realizada a partir de conversas com especialistas, utilizando os resultados associados à exploração dos vários fatores de confiança trabalhados na etapa anterior, para se definir quais níveis de confiança poderiam ser usados dentro de um ambiente real.

1.5 Contribuições

Como contribuições do trabalho, destacam-se:

1. Definição de um método para triagem automática de solicitações em saúde;
2. Avaliação do método em diferentes contextos;
3. Identificação de fatores que possam dificultar a adoção do método em contextos reais.

1.6 Estrutura do Trabalho

O restante deste trabalho está organizado da seguinte maneira: no Capítulo 2 são discutidos trabalhos que inspiraram o desenvolvimento da pesquisa; no Capítulo 3 é apresentado o processo de DCBD e também os algoritmos utilizados neste trabalho; a triagem automática é apresentada no Capítulo 4, no qual são detalhadas as suas etapas; no Capítulo 5 são apresentados os resultados obtidos nos experimentos da triagem proposta; e por fim, no Capítulo 6 são discutidas as conclusões, limitações e trabalhos futuros da pesquisa realizada.

2 Trabalhos Relacionados

Não existem muitos trabalhos na literatura abordando a utilização da computação no apoio a regulação de solicitações de serviços assistenciais. Apesar dos primeiros trabalhos sobre o tema datarem de mais de vinte anos atrás (FARMAN; FARAG; YEAP, 1992) (BOUHADDOU et al., 1993), somente nos últimos dez anos houve uma maior exploração do assunto. Ainda assim, a utilização de técnicas computacionais no auxílio ao processo de regulação ainda é uma promessa (TERRY, 2015b) (TERRY, 2015a). Este capítulo discute estudos encontrados que abordam esse tema e outros trabalhos relevantes para o desenvolvimento da pesquisa apresentada.

2.1 Mapeamento Sistemático de Estudos

A primeira abordagem realizada para se obter os trabalhos relacionados ao tema deste trabalho foi a realização de um Mapeamento Sistemático de Estudos (MSE) (KITCHENHAM; CHARTERS, 2007) com o objetivo de identificar o estado da arte na literatura relativo a utilização de técnicas computacionais no suporte ao processo de regulação de solicitações de serviços assistenciais.

No contexto da pesquisa, estudos podem ser caracterizados como primários ou secundários. Por estudos primários, entende-se a condução de estudos que visem caracterizar uma determinada abordagem em uso dentro de um contexto específico (MAFRA; TRAVASSOS, 2006). Estudos secundários, por sua vez, buscam identificar, avaliar e interpretar resultados relevantes a um determinado tópico de pesquisa, fenômeno de interesse ou questão de pesquisa (KITCHENHAM, 2004).

Um MSE é um estudo secundário que exibe, em um alto nível de granularidade, evidências de um determinado domínio. Isso permite a identificação de abordagens exploradas do domínio em questão, sendo útil para, dentre outras coisas: indicar áreas onde é necessário a condução de estudos primários ou áreas em que um estudo secundário é mais apropriado (KITCHENHAM; CHARTERS, 2007).

A pesquisa por estudos primários foi guiada com o intuito de elencar as principais técnicas computacionais utilizadas no auxílio ao processo de regulação de solicitações de procedimentos médicos, quais áreas da medicina as utilizam e qual o impacto da adoção.

Os portais de busca escolhidos para pesquisa foram: *Scopus*, *Compendex* e *Web Of Science*, pois eles cobrem as bases de dados consideradas mais importantes das Engenharias e das Ciências Exatas/Aplicadas: *IEEE Xplore*, *ACM Digital Library*, *Elsevier*, e *Springer*.

No total, foram encontrados 165 trabalhos. Um pré-processamento foi realizado para excluir trabalhos duplicados e, com isso, obtiveram-se 130 candidatos a estudos primários. Após a análise destes trabalhos, incluindo a análise de referências (*snowballing*), somente três trabalhos foram avaliados como relevantes aos objetivos do MSE.

BOUHADDOU et al. (1993) apresentaram a implementação do *Iliad*, um sistema especialista baseado em regras para a indicação de três procedimentos cirúrgicos: colecistectomia, remoção de catarata, artroscopia do joelho. A base de conhecimento foi construída baseada na literatura médica sobre a aplicabilidade dos procedimentos e representa um consenso entre os recursos utilizados e especialistas de uma OPS norte americana e do Centro Médico da Universidade de Utah. O sistema foi utilizado no serviço de regulação da companhia em questão que selecionou treze cirurgiões de quatro clínicas parceiras para um experimento que durou de abril a dezembro de 1992. Estes cirurgiões preenchiam um formulário sobre a situação do paciente e enviavam para a companhia via fax e eram inseridos no sistema para se obter a reposta sobre a indicação cirúrgica e um relatório justificando a posição adotada era enviado para o prestador de serviços. Os resultados obtidos mostraram: confiabilidade nas respostas do sistema, apresentando consenso entre OPS e prestadores de serviço; agilidade pela utilização do fax para a troca de informações; e aumento significativo do nível de documentação para a avaliação e determinação da aplicabilidade dos procedimentos antes da sua execução.

CARROLL et al. (2006) avaliaram a efetividade do *SmartPA*, um sistema automatizado de regulação baseado em regras, para a diminuição do uso e dos gastos de inibidores seletivos da COX-2, uma alternativa mais cara aos medicamentos antiinflamatórios não-esteróides, mas menos tóxica ao sistema gastrointestinal. O sistema investiga o histórico médico e de utilização de fármacos para determinar o acesso ao medicamento, caso o acesso seja negado, o paciente e o médico podem submeter a solicitação manualmente para uma avaliação adicional. A pesquisa foi realizada no programa de saúde social *Medicaid* do estado do Missouri, Estados Unidos da América, na modalidade *fee-for-service* onde os prestadores de serviço recebem da OPS pagamento por cada procedimento executado (assemelhando-se ao sistema brasileiro) e engloba o período de doze meses anteriores e posteriores à implementação do sistema. Um outro estado norte americano, que não possui o mecanismo de regulação de solicitações de serviços médicos, foi utilizado como parâmetro de controle. Os resultados mostram que o sistema obteve sucesso no controle do uso e dos gastos de inibidores seletivos da COX-2, havendo também diminuição em ambos os casos, enquanto aumentaram substancialmente no estado de controle. Esta diminuição foi observada principalmente entre pacientes com baixo risco de problemas gastrointestinais, o que significou um aumento na utilização de medicamentos antiinflamatórios não-esteróides tradicionais, que não foi observado em fármacos protetores do sistema gastrointestinal.

LUNDEEN et al. (2013) analisaram o tempo de acesso, conformidade com regime

de profilaxia e acurácia da projeção da dose obtidas com um sistema Web de regulação de solicitações de serviços médicos para o antibiótico palivizumab, no programa de saúde social *Medicaid* do estado da Carolina do Norte, nos Estados Unidos América. Para isso, uma retrospectiva histórica de todas as solicitações de palivizumab nas temporadas de 2010/2011 e 2011/2012 do vírus sincicial respiratório, analisando o tempo de acesso, número de doses aprovadas, aplicadas, data da administração e dosagem utilizada. Esta foi a primeira aplicação Web para regulação da utilização do palivizumab nos Estados Unidos América, e funcionalidades como aprovação automática e carregamento de arquivos foram efetivas na diminuição em 3,7 dias, em média, no tempo de acesso ao medicamento. A conformidade com o regime de profilaxia foi um pouco maior quando comparados com o Registro de Resultados dos pacientes do programa de saúde, sendo esse um ponto importante, uma vez que a não conformidade pode resultar em hospitalização de pacientes e o desperdício das doses aplicadas. Por fim, a projeção da dosagem parece prever com acurácia a quantia necessária para cada paciente.

Foi observado que o auxílio da tecnologia da informação tem sido pouco explorado para melhorar o processo de regulação. A área de farmácia é a que mais tem buscado esse apoio, principalmente para fornecer o medicamento correto com melhor custo/benefício para os segurados e para dar celeridade ao processo, economizando tempo e dinheiro de todos os envolvidos. Foi identificado, também, que os trabalhos possuem um foco bem específico, analisando somente a regulação de um medicamento, ou grupo de medicamentos específicos, sendo necessária uma investigação mais profunda para determinar as razões dessa abordagem.

Nos trabalhos selecionados foi percebido que as abordagens atuais concentram-se em transferir o conhecimento do médico regulador para um sistema computacional que possa julgar as solicitações realizadas. Vale ressaltar que os sistemas utilizados buscam a automação da autorização de pedidos, que representam a maior parte dos pedidos.

Acredita-se que a pouca quantidade de estudos encontrados dá-se pelo fato do processo de regulação não ter sido utilizado em larga escala até pouco tempo atrás. Especificamente falando dos Estados Unidos América, onde a maioria dos trabalhos retornados pela busca foi originado, o processo de regulação só se tornou mais abrangente após mudanças ocorridas durante o governo do presidente Obama, que esteve no poder durante 2009 e 2017. Essas mudanças permitiram a expansão do número de segurados sujeitos ao processo de regulação de prescrição de medicamentos, tendo havido um aumento na quantidade de trabalhos cujo foco é a eficácia do processo de regulação. Dessa maneira, a utilização de técnicas computacionais para auxiliar o processo de regulação de prescrição de medicamentos ainda está em andamento, enquanto outras áreas ainda não utilizam esta abordagem.

2.2 Trabalhos não incluídos no Mapeamento Sistemático de Estudos

Outros trabalhos importantes para o desenvolvimento desta pesquisa não foram achados no mapeamento sistemático executado, pois não estavam escritos em inglês (ou usavam nomenclaturas diferentes[1]) ou não abordavam especificamente a regulação de solicitações de serviços médicos. Estes trabalhos são apresentados a seguir.

Em ARAúJO; SANTANA; SANTOS NETO (2016) propuseram uma metodologia baseada em técnicas de pré-processamento para melhorar a qualidade das informações presentes na base de dados de procedimentos odontológicos de uma OPS e com isso utilizá-las no aprendizado do processo de regulação de solicitações de serviços médicos/odontológicos. A abordagem proposta possui oito etapas (remoção de dados éticos e legais, seleção manual de atributos, seleção automática de atributos, tratamento de valores desconhecidos, transformação de dados, balanceamento, treinamento e avaliação dos classificadores) e foi realizada por um especialista de domínio e analista de dados. Foram escolhidos três algoritmos para o aprendizado do processo de regulação: C4.5, *Naive Bayes* e *Multilayer Perceptron*; que foram avaliados utilizando a técnica *10-fold cross-validation* e as seguintes métricas: precisão, *recall*, acurácia, *f-measure*, área sobre a curva ROC e índice Kappa. Com o algoritmo C4.5 foi possível obter uma acurácia superior a 91%, demonstrando que o processo de regulação de solicitações de serviços médicos/odontológicos pode ser aprendido por algoritmos de aprendizagem de máquina, desde que se utilize técnicas de pré-processamento para a melhoria da qualidade dos dados.

Alguns dos trabalhos inspiradores para esta pesquisa focam na detecção de fraude e abuso após a realização da assistência, avaliando o mérito do pagamento destes aos prestadores de serviço ou reembolso de beneficiários.

ORTEGA; FIGUEROA; RUZ (2006) propuseram um sistema de detecção de fraude baseado em redes neurais na solicitação de benefícios requisitados a uma empresa de seguro de saúde chilena. Os dados iniciais para a pesquisa consistiam em dois grupos distintos, o primeiro contendo 169 solicitações abusivas extremamente bem documentadas pela companhia e o último contendo 500,000 solicitações do período de 2001 a 2003 classificadas como “aprovada”, “rejeitada” ou “reduzida”. Para a criação dos modelos os dados de solicitações foram agrupadas em quatro grupos: afiliados, profissionais, empregadores e solicitação; mantendo a proporção entre casos fraudulentos/abusivos e normais. Para cada um desses grupos um comitê de dez *perceptrons* multi camadas é treinado mensalmente para assinalar a probabilidade de novas amostras como abusivas/fraudulentas ou normais, dessa maneira auditores da companhia podem concentrar esforços nas solicitações sus-

[1] Por exemplo: a utilização de *Medical Claims* no lugar de *Prior Authorization*

peitas. O sistema proposto conseguiu uma detecção de aproximadamente 75% dos casos fraudulentos/abusivos por mês e adiantou a descoberta desses comportamentos em seis meses.

KOSE; GOKTURK; KILIC (2015) implementaram e avaliaram um *framework* interativo (eFAD, *electronic fraud and abuse detection*), para detectar casos fraudulentos e abusivos independentemente dos atores ou bens envolvidos e uma estrutra para identificar novos tipos fraude e abusos nas requisições de pagamentos e reembolsos por serviço. O trabalho foi motivado pela insatisfação dos especialistas de domínio na pouca participação no processo tradicional de aprendizado de máquina, resumido a auxiliar os engenheiros de conhecimento, dessa maneira o trabalho desenvolve um sistema de suporte à decisão baseado em aprendizado de máquina interativo para identificar casos suspeitos de fraude e abuso e direciona-los para análise do especialista. O *framework* é composto por quatro componentes: incialmente o conhecimento do especialista a respeito do objetivo e das hipóteses é incorporado através de *storyboards*; atores, bens e atributos são extraídos destes artefatos em um segundo momento para a criação de dois *data warehouses*; o terceiro componente avalia os escores de riscos para os atores envolvidos em uma dada requisição; e o quarto componente consiste em uma ferramenta de visualização destas informações que permite a interação (e mudança de parâmetros) com o especialista. O eFAD foi avaliado experimentalmente com dados reais de uma companhia de seguros da Turquia que compreendiam 100,000 beneficiários e 845,247 requisições do período de 2008 a 2011. O *framework* em questão conseguiu níveis de acurácia entre 70.8% e 89.6% para vários comportamentos fraudulentos e abusivos, resultado considerado encorajador para sua utilização.

HILLERMAN; CARVALHO; REIS (2015) analisaram nas requisições de pagamentos e reembolsos de prestadores de serviço individuais para identificar comportamentos suspeitos, mais especificamente o comportamento conhecimento como “Dia Impossível” que se configura quando há uma quantidade de atendimentos tão alta que seria inviável a realização de todos por um único profissional. A base de dados analisada correspondia ao ano de 2013 e continha aproximadamente um milhão de pagamentos para cerca de 13,000 prestadores de serviço individuais e 350,000 beneficiários. Após um tratamento inicial, os dados foram submetidos ao algoritmo *k-means* com um modelo encontrado experimentalmente de 4-grupos. Dentre os agrupamentos criados, percebeu-se a presença dos suspeitos iniciais de praticarem “Dias Impossíveis” no grupo de número três e outros prestadores com altos fatores de risco foram reunidos no grupo de número quatro. Baseado nesse *framework* é possível criar controles automáticos para sinalizar novas requisições suspeitas de fraude e direciona-las para uma análise mais minuciosa.

2.3 Considerações Finais

Este capítulo apresentou estudos relevantes para o desenvolvimento da pesquisa atual obtidos por meio de um MSE referente à utilização de técnicas computacionais no auxílio do processo de regulação de solicitações de procedimentos em saúde e de buscas sobre a detecção de fraude e abuso no âmbito das OPSs. Apesar da detecção de fraude e abuso no pagamento de serviços médicos ser bastante explorado na literatura, não foi possível identificar o mesmo esforço no auxílio ao processo de regulação, que é o foco deste trabalho e principal item de destaque. Vale ressaltar também a característica comum nas abordagens encontradas de apenas sinalizar os possíveis casos fraudulentos e abusivos, deixando a decisão final para o especialista.

3 Descoberta de Conhecimento em Bancos de Dados

A Descoberta de Conhecimento em Bancos de Dados (DCBD) é o processo de coletar, limpar, processar, analisar e obter informações úteis e relevantes a partir de um conjunto de dados (AGGARWAL, 2015) e surgiu para auxiliar na extração de conhecimento do grande volume de dados da era digital (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Técnicas de DCBD são indispensáveis no aspecto econômico, uma vez que as empresas analisam dados para obter vantagens estratégicas, melhorar a eficiência, e valorizar o serviço prestado, e também no aspecto científico, se considerarmos que os dados são a matéria prima básica para a construção de teorias e modelos (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). A DCBD pode ser divida em três estágios (AGGARWAL, 2015):

Coleta de Dados: Estágio inicial no qual os dados contidos no banco são selecionados e minimamente organizados para o processo de descoberta;

Pré-processamento: Os dados oriundos da coleta ainda podem apresentar atributos incompletos, mal concebidos e até mesmo irrelevantes para o problema tratado, por isso nesse estágio os dados são transformados em um formato adequado para viabilizar o processamento e a aplicação das técnicas analíticas;

Mineração de Dados: Por fim, os dados são analisados por meio da aplicação de diferentes estratégias e técnicas com o intuito de produzir um modelo para auxiliar o enfrentamento do problema em questão.

3.1 Pré-processamento

O pré-processamento tem como objetivo o entendimento do problema e a seleção, limpeza e transformação dos dados. Essa etapa é de grande importância para a remoção de quaisquer características que possam afetar a qualidade dos dados e impactar negativamente na mineração. O pré-processamento realizado no presente trabalho contém os seguintes momentos: seleção de dados, seleção manual de atributos, construção de atributos, seleção automática de atributos, estratificação e balanceamento.

Seleção de dados: Os dados coletados são analisados e uma nova seleção é realizada para remover eventuais inconsistências;

Seleção manual de atributos: O número de atributos utilizados passa por uma primeira redução que exclui aqueles de baixa qualidade ou que não representam informações úteis para o problema abordado;

Construção de atributos: Novos atributos são construídos para representar informações contidas indiretamente nos dados;

Seleção automática de atributos: Algoritmos são utilizados para calcular a importância de cada atributo no processo de mineração e remover aqueles que não contribuírem para o processo;

Estratificação e balanceamento: Por último, os dados são separados em conjuntos de treinamento e de teste. Os dados de treinamento passam por um último tratamento para diminuir a diferença entre a quantidade de exemplos das classes que serão analisadas.

3.2 Mineração de dados

Antes dos algoritmos utilizados neste trabalho serem apresentados, é interessante expor algumas características dos paradigmas de mineração de dados. Existem vários métodos de mineração de dados, utilizados para diferentes propósitos. A Figura 5 apresenta uma taxonomia da classificação destes métodos segundo a intenção do processo de descoberta e conhecimento.

Na Figura 5 pode-se distinguir os dois principais métodos de mineração de dados: verificação e descoberta (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996) (MAIMON; ROKACH, 2005).

Métodos de verificação lidam com a avaliação de uma hipótese proposta por uma fonte externa (e.g. um usuário especialista). Esse ramo não é muito utilizado na mineração de dados uma vez que a maioria dos problemas atacados nesse campo está focada na descoberta de novas hipóteses (MAIMON; ROKACH, 2005).

Já os métodos de descoberta são aqueles que automaticamente identificam padrões nos dados, podendo ainda ser subdivididos em: métodos descritivos e métodos preditivos. Métodos descritivos são orientados a interpretação dos dados e focam na compreensão de suas relações (e.g. visualização). Métodos preditivos buscam a construção de modelos comportamentais para antecipar valores de novas amostras apresentadas ao sistema. Vale ressaltar que a categorização entre os objetivos de predição e descrição não é exclusiva (i.e. métodos preditivos também podem ser descritivos, e vice versa) (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

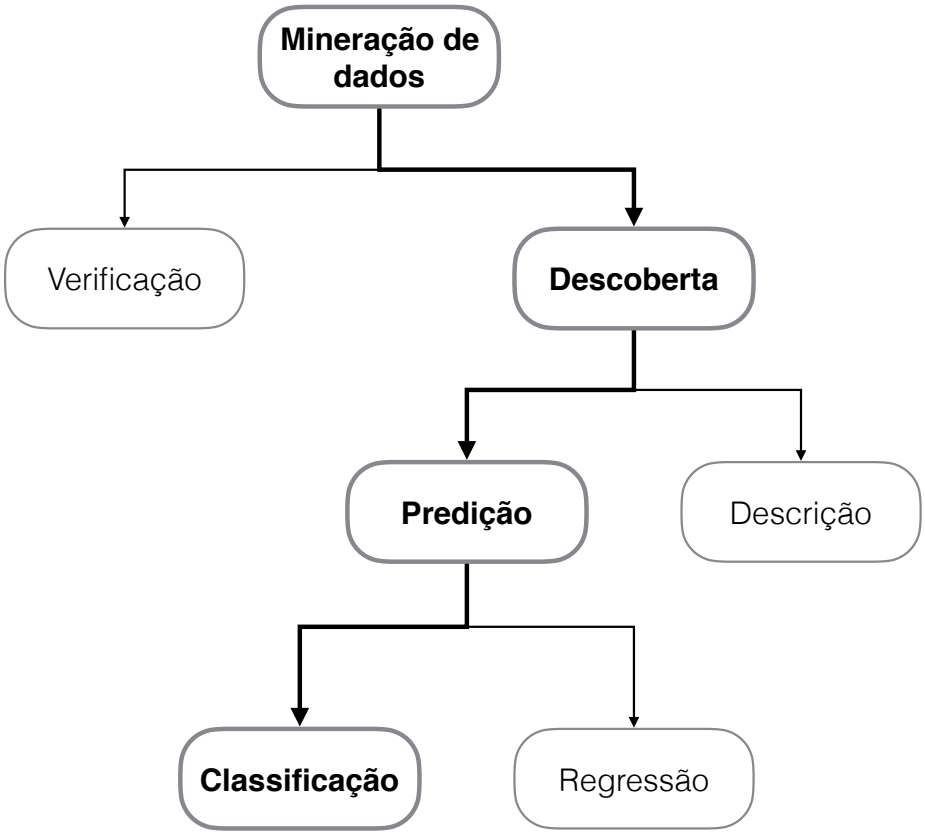

Figura 5 – Taxonomia dos métodos de mineração de dados, baseado no trabalho de Maimon (MAIMON; ROKACH, 2005), com destaque no foco deste trabalho

Uma vez que, neste trabalho, os dados de regulação já existiam e o interesse principal era prever o comportamento futuro para novas requisições de procedimentos, optou-se por trabalhar com o ramo descoberta. Além disso, como a mineração de dados pode ser descrita também como a aplicação de algoritmos de aprendizado de máquina em grandes bancos de dados (ALPAYDIN, 2010), a próxima subseção discute a teoria e algumas técnicas de aprendizado de máquina.

3.2.1 Aprendizado de máquina

Para se resolver problemas em um computador faz-se necessário o emprego de um algoritmo, i.e. uma sequência de instruções que devem ser seguidas para, dada uma entrada conhecida, obter-se uma saída esperada. Alguns problemas têm essa transformação bem descritas, como a ordenação de uma lista de números. Porém, existem problemas em que, apesar de possuir entradas conhecidas, não é sabido uma sequência de instruções que devem ser seguidas para alcançar a saída esperada, por exemplo, a classificação de mensagens eletrônicas como *SPAM*. A falta de conhecimento na construção de algoritmos para resolver esses problemas é superada pela grande quantidade de dados disponíveis. Como solução, apresentam-se exemplos de entradas, e.g. mensagens eletrônicas, e suas respectivas

saídas esperadas para que um computador (máquina) possa inferir as características que contribuem para a solução do problema (ALPAYDIN, 2010).

Com o desenvolvimento da informática, a quantidade de dados disponíveis esta se expandindo e o acesso a exemplos de entradas para diversos problemas foi facilitado. Companhias podem guardar, em seus servidores, dados detalhados de cada transação que ocorra em seu processo de negócio e eventualmente disponibilizar na internet para que outras pessoas possam utilizá-los. Analisando esses dados, as empresas esperam extrair informações que possam ser utilizadas para obter vantagens competitivas e identificar tendências. Apesar de não ser possível identificar o processo real de geração de informações, é comum a construção de aproximações satisfatórias, sendo esse o nicho do aprendizado de máquina (ALPAYDIN, 2010).

Desde 1980, técnicas de aprendizado de máquina vêm ganhando um papel mais central nas pesquisas de inteligência artificial e cresceu bastante em interesse entre os pesquisadores. Em 1983, SIMON definiu aprendizado como qualquer mudança em um sistema que permita a melhora de sua performance em uma segunda execução de uma mesma tarefa ou em uma nova tarefa sob uma mesma população. Nesse sentido, um sistema pode mudar adquirindo novo conhecimento de uma fonte externa, ou pode se modificar para utilizar o seu atual conhecimento de maneira mais eficiente (SHAVLIK; DIETTERICH, 1990). Dessa maneira, o aprendizado de máquina pode ser definido como métodos computacionais que usam experiência para melhorar a sua performance ou fazer predições precisas (MOHRI; ROSTAMIZADEH; TALWALKAR, 2012).

3.2.1.1 Terminologia e definições

Para um melhor entendimento dos algoritmos de aprendizado de máquina, primeiro é necessário um alinhamento de definições e terminologia. Os termos mais utilizados neste trabalho são:

Exemplo: Item, ou instância, de dados utilizado;

Características: Atributos associados a um determinado item, geralmente apresentado como uma tupla, ou vetor, de atributos;

Rótulo: Categoria (em problemas de classificação) ou valor real (em problemas de regressão) associado a um determinado item;

Dados: Conjunto de exemplos utilizado no processo de aprendizado de máquina. Geralmente, esse conjunto é particionado em dois grupos:

- **Dados de Treinamento:** Exemplos utilizados para treinar um modelo modelo paramétrico a partir de um algoritmo de aprendizagem;

Dados de Teste: Exemplos utilizados para avaliar o modelo criado, sendo esses separados dos dados de treinamento e estão indisponíveis na fase de aprendizagem;

3.2.1.2 Tipos de aprendizado

Existem várias maneiras de se classificar algoritmos de aprendizado de máquina. Considerando os diferentes cenários para tipos de dados de treinamento disponíveis temos a classificação pelos estilos de aprendizagem, que são (MOHRI; ROSTAMIZADEH; TALWALKAR, 2012):

Aprendizado supervisionado: O algoritmo recebe exclusivamente um conjunto de exemplos rotulados como dados de treinamento e faz suposições para novas amostras. Como a classificação de todas as instâncias é conhecida, avaliar a performance do algoritmo é simples;

Aprendizado não supervisionado: O algoritmo recebe exclusivamente um conjunto de exemplos não rotulados como dados de treinamento. Uma vez que não há instâncias rotuladas, pode ser difícil quantificar a avaliação da performance desse tipo de algoritmo;

Aprendizado semi-supervisionado: O algoritmo recebe um conjunto de dados com exemplos rotulados e não rotulados como dados de treinamento, e faz suposições para novas amostras. Esse tipo de aprendizado é mais utilizado em problemas nos quais os dados são de fácil acesso porém caros de se obter.

Uma vez que, para este trabalho, há uma grande disponibilidade de dados já rotulados, o aprendizado supervisionado é destacado. Nesse tipo de aprendizado, um problema é solucionado ao aprender o mapeamento de uma dada entrada X em uma saída Y (ALPAYDIN, 2010).

3.2.1.3 Paradigmas de aprendizado

Uma outra maneira de se classificar algoritmos de aprendizado de máquina é em relação aos paradigmas de aprendizado. Com o intuito de representar diferentes paradigmas de aprendizado, e manter consistência com as pesquisas anteriores do grupo, este trabalho utiliza algoritmos de três paradigmas: bayesiano, baseado em árvores de decisão e estatístico. Estes paradigmas são explorados nas próximas subseções, juntamente com os algoritmos utilizados.

3.2.1.3.1 Paradigma bayesiano

No paradigma bayesiano a aprendizagem acontece por inferência probabilística. Essa abordagem procura identificar todas as variáveis relevantes nos dados de treinamento para construir um modelo probabilístico de suas interações e a inferência é realizada por meio do cálculo das probabilidades condicionadas aos atributos da instância de validação ou teste. Vale destacar que novas instâncias podem ser classificadas por meio da combinação de múltiplas hipóteses, ponderadas por suas probabilidades (MITCHELL, 1997). Este trabalho utiliza o algoritmo *Naive Bayes* do paradigma bayesiano, que assume que as variáveis são independentes para determinada classe. Apesar dessa premissa ser geralmente uma suposição ruim, esse algoritmo tem resultados práticos comparáveis a outros classificadores mais sofisticados.

3.2.1.3.2 Paradigma baseado em árvores de decisão

Algoritmos desse paradigma utilizam a estrutura de árvores para representar o conhecimento e para tomar decisões, por meio de regras bem definidas, sobre qual a classe de uma dada instância (QUINLAN, 1987). A partir dos dados de treinamento, uma árvore é criada de maneira que cada nó interno contém um teste e o resultado desse teste é utilizado para definir por qual ramo o processo de decisão deve seguir até atingir uma folha da árvore que contém um rótulo para a instância em questão. Este trabalho utiliza três algoritmos do paradigma baseado em árvores de decisão: C4.5, *Ripper*, e *Random Forest*.

O algoritmo C4.5 é, provavelmente, o algoritmo mais popular no aprendizado de máquina, e para muitas aplicações as árvores produzidas por esse algoritmo são pequenas e precisas, resultando em rotulações rápidas e confiáveis (MURTHY; KASIF; SALZBERG, 1994). Esse algoritmo usa os atributos com maior razão de ganho para criar os nós internos da árvore, até que as instâncias estejam separadas pelos rótulos nos nós folhas.

Outro algoritmo utilizado, o *Ripper*, examina as classes incrementalmente. Para cada rótulo, o algoritmo busca um conjunto de regras que contemple todos os exemplos antes de partir para a próxima classe (COHEN, 1995).

Fruto da pesquisa de combinação de algoritmos, o *Random Forest* é a combinação entre árvores de decisão em que cada árvore depende dos valores de um vetor randômico amostrado independentemente e com a mesma distribuição para todas as árvores (BREIMAN, 2001). Entre as vantagens da sua utilização, destaca-se uma maior resistência a ruídos que dificultam o processo de aprendizagem.

3.2.1.3.3 Paradigma estatístico

Esse paradigma se baseia nos campos da estatística e da análise funcional para encontrar uma aproximação matemática da função que relaciona o comportamento preditivo entre as caraterísticas das observações com o rótulo apresentado (FRIEDMAN; HASTIE; TIBSHIRANI, 2001). Nessa abordagem os dados são considerados como pontos em um espaço euclidiano e o objetivo é obter uma função aproximada para todos os pontos de uma região do espaço em questão sendo possível encontrar o rótulo para dados desconhecidos no momento da criação do modelo. Este trabalho utiliza o algoritmo SVM *Support Vector Machine* do paradigma estatístico, que busca construir um hiper-plano (ou conjunto de hiper-planos) que separe com a maior distância exemplos de classes diferentes.

3.2.1.4 Avaliação de Algoritmos

A corretude de uma classificação pode ser avaliada por meio da computação do número de exemplos corretamente identificados da classe de interesse (verdadeiros positivos), do número de exemplos corretamente identificados que não pertencem a classe de interresse (verdadeiros negativos), e do número de exemplos que foram incorretamente atribuídos a classe de interesse (falso positivos) ou que não foram identificados como exemplos da classe de interesse (falso negativo). Estes quatro valores constituem a chamada Matriz de Confusão, ilustrada na Tabela 1 para uma situação de classificação binária (SOKOLOVA; LAPALME, 2009).

Tabela 1 – Matriz de Confusão.

		Predição	
		Positiva	Negativa
Exemplo	Positivo	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Negativo	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Existe, na literatura, uma diversidade de métricas calculadas a partir desses valores para se avaliar a performance de um dado classificador. Neste trabalho, são utilizadas as seguintes métricas: Taxa de Verdadeiros Positivos, Taxa de Verdadeiros Negativos, Valor Preditivo Positivo e Valor Preditivo Negativo.

A Taxa de Verdadeiro Positivo (também conhecida como *recall*, ou *hit rate*, ou *sensitivity*), Equação 3.1, representa a proporção de exemplos da classe de interesse corretamente classificados dentre todos os exemplos da classe de interesse. O Valor Preditivo Positivo (também conhecido como *precision*), Equação 3.2, representa a proporção de exemplos da classe de interesse corretamente classificados dentre todos os exemplos

atribuídos à classe de interesse.

$$TVP = \frac{VP}{VP + FN} \times 100 \tag{3.1}$$

$$VPP = \frac{VP}{VP + FP} \times 100 \tag{3.2}$$

A Taxa de Verdadeiro Negativo (também conhecida como *specificity*), Equação 3.3, representa a proporção de exemplos corretamente identificados como não pertencentes a classe de interesse dentre todos os exemplos que não pertecem à classe de interesse. O Valor Preditivo Negativo, Equação 3.4, representa a proporção de exemplos corretamente identificados como não pertencentes à classe de interesse dentre todos os exemplos não atribuídos à classe de interesse.

$$TVN = \frac{VN}{VN + FP} \tag{3.3}$$

$$VPN = \frac{VN}{VN + FN} \tag{3.4}$$

3.3 Técnicas de Tratamento de Desbalanceamento

A utilização do aprendizado de máquina no problema abordado por este trabalho é afetado pelo desbalanceamento inerente à detecção de fraude e abuso. Em situações reais, bases de dados dessa temática costumam possuir uma grande diferença entre a quantidade de exemplos rotulados como válidos e inválidos, sendo a quantidade de exemplos válidos superior à de exemplos inválidos. Para tratar essa questão, este trabalho utiliza as técnicas de *Random Oversampling*, SMOTE (do ingês *Synthetic Minority Over-sampling Technique*) (CHAWLA et al., 2002) e *Meta Cost* (DOMINGOS, 1999); onde as duas primeiras atuam nos dados de treinamento e a última atua sobre os algoritmos empregados. Essas técnicas são descritas a seguir.

***Random Oversampling*:** Aumenta o número de exemplos do rótulo minoritário (inválidos) nos subconjuntos de treinamento por meio da replicação aleatória dos exemplos existentes;

SMOTE: Adiciona informação ao subconjunto de treinamento por meio da criação de informações sintéticas, combinando um exemplo com seus vizinhos, i.e. exemplos com características e rótulos semelhantes;

Meta Cost a maioria dos classificadores desenvolvidos no campo do aprendizado de máquina assume que todos os erros resultam em um mesmo custo, o que raramente

é o caso nos problemas da DCBD. Para atenuar ou eliminar esse problema a técnica *MetaCost* converte classificadores baseados em erro por meio de uma matriz de custos utilizada para re-rotular os dados de treinamento com a categoria que apresentar o menor custo para suas características.

3.4 Validação cruzada

É uma técnica baseada em dados para estimar a acurácia de um modelo matemático (KOHAVI, 1995). Na prática, geralmente a quantidade de dados rotulados é pequena, e separar exemplos exclusivamente para validação pode deixar um número insuficiente de dados para treinamento (MOHRI; ROSTAMIZADEH; TALWALKAR, 2012). Dessa maneira, na validação cruzada, o conjunto de dados é particionado em subconjuntos, em que alguns desses subconjuntos são utilizados para treinamento e o restante é empregado na validação do modelo.

Usualmente, em aplicações no aprendizado de máquina, o método mais utilizado de validação cruzada é o *k-fold* onde k assume um valor entre 5 e 10 (MOHRI; ROSTAMIZADEH; TALWALKAR, 2012). Na validação cruzada *k-fold*, também conhecida como método da rotação, o conjunto de dados D é particionado em k subgrupos (*folds*) $D_1, D_2, \ldots, D_k$ de tamanhos aproximadamente iguais. O algoritmo de aprendizado é executado k vezes; a cada iteração $t \in \{1, 2, \ldots, k\}$, $D \setminus D_t$ é utilizado para criar um modelo que é testado com D_t. A estimativa de acurácia da validação cruzada é igual ao número total de classificações corretas, divido pelo número de instâncias existentes (KOHAVI, 1995).

3.5 Fator de Confiança

Considerando o problema de classificação discutido anteriormente, os modelos criados podem fornecer saídas contínuas (e.g. uma estimativa da probabilidade dos rótulos de novos exemplos) ou discretas (e.g. um único rótulo discreto para cada novo exemplo) (FAWCETT, 2006). Para a execução dos experimentos foi utilizada a ferramenta WEKA (WEKA, 2016) (do inglês *Waikato Environment for Knowledge Analysis*) e as implementações dos algoritmos escolhidos criam modelos capazes de fornecer saídas contínuas. Este trabalho utiliza um parâmetro, chamado de fator de confiança, para definir o valor mínimo para a estimativa da probabilidade do rótulo positivo.

A Figura 6 é utilizada com o intuito de exemplificar o conceito acima para uma classificação binária entre os rótulos “negativo” e “positivo”. O gráfico apresentado consiste em um histograma da quantidade de exemplos no conjunto de testes (eixo das ordenadas) para a estimativa da probabilidade para o rótulo “positivo” (eixo das abcissas).

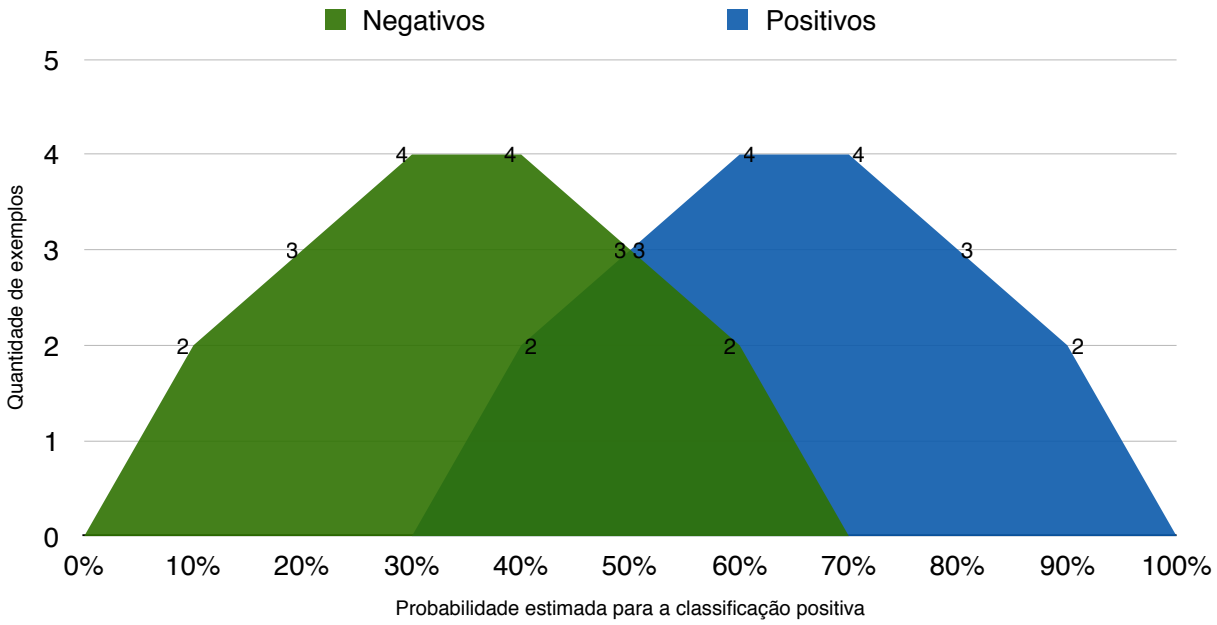

Figura 6 – Histograma da quantidade de exemplos no conjunto de testes (eixo das ordenadas) para a estimativa da probabilidade para o rótulo "positivo" (eixo das abcissas).

No exemplo da figura, com 36 exemplos particionados igualmente entre positivos e negativos, um fator de confiança de 50% levaria a correta classificação de 16 exemplos positivos e a incorreta classificação de 5 exemplos negativos. Com um fator de confiança de 70% metade dos exemplos positivos seriam corretamente classificados e nenhum exemplo negativo seria incorretamente classificado. Com a introdução desse parâmetro espera-se dar liberdade para a OPS configurar a classificação de novas instâncias de acordo com o contexto e o impacto da sua utilização é explicado no próximo capítulo.

4 Triagem Automática de Solicitações de Serviços Assistenciais em Saúde

A Triagem Automática de Solicitações de Serviços Assistenciais em Saúde, utilizada aqui somente como Triagem Automática, é uma proposição deste trabalho que visa avaliar solicitações em saúde, respondendo de forma automática àquelas que forem consideradas seguras, e repassando as demais para a avaliação de reguladores humanos, com o intuito de reduzir a carga de trabalho manual nos centros de regulação.

A ideia da triagem de solicitações iniciou em um trabalho de pesquisa anterior (ARAúJO; SANTANA; SANTOS NETO, 2016), que teve como principal objetivo a avaliação do uso de técnicas de aprendizado de máquina, via DCBD, na regulação de solicitações odontológicas. Os resultados obtidos indicaram que tais técnicas podem ser bastante efetivas em tal contexto, com um nível de precisão bastante alto (acima de 90%).

O trabalho iniciado foi então expandido para o contexto médico. Foram realizados alguns experimentos com o mesmo objetivo do trabalho citado anteriormente, porém utilizando bases de dados de planos de saúde médicos. Os resultados obtidos indicaram um bom nível de resposta, porém, com uma quantidade considerável de erros quando se envolve eventos de natureza médica. Por conta disso, procurou-se propor uma alternativa que ao mesmo tempo pudesse reduzir o trabalho dos reguladores humanos, mas que não comprometesse a saúde dos pacientes envolvidos.

É importante ressaltar que as solicitações que chegam a uma central de regulação são feitas por profissionais em saúde, sejam médicos ou dentistas, que são os responsáveis diretos por toda e qualquer questão associada à realização dos tratamentos/procedimentos/exames requisitados. Uma central de regulação deve apenas analisar tais solicitações para eventualmente negar aquelas que apresentem alguma forma de incorreção. Por sinal, várias operadoras simplesmente não possuem um setor de regulação, o que significa, na prática, que tudo que for solicitado por um profissional em saúde é aprovado, uma vez que não há avaliação da pertinência do pedido, pois a solicitação é meramente um registro do que será realizado. Ou seja, a responsabilidade técnica do que será realizado com o paciente é do solicitante, que deve ser devidamente registrado nos conselhos profissionais de área.

Os erros que podem acontecer dentro do processo de triagem automática são de duas naturezas diferentes:

- Uma resposta positiva para um evento que deveria ter uma resposta negativa (erro 1). Nesse caso, algo que deveria ser negado seria aprovado pelo mecanismo de triagem automática. Isso significa que a OPS arcará com os custos de algo que não deveria ser

feito. Dependendo do que for autorizado indevidamente, o custo pode ser considerável. Porém, via de regra, uma autorização esporádica errada não tende a ser algo que leve uma OPS a um estado de alerta.

- Uma resposta negativa para um evento que deveria ter uma resposta positiva (erro 2). Nesse caso, algo que deveria ser autorizado seria negado pelo mecanismo de triagem automática. Esses casos são bem mais complicados que o anterior. Tal erro pode gerar consequências desastrosas a uma operadora, pois uma negativa como essa pode levar um paciente para uma situação crítica, culminando com sua morte.

Analisando os tipos de erro acima, torna-se claro que uma abordagem para triagem automática deve conter algum mecanismo de impedir a ocorrência do erro 2, mas, que mesmo assim, torne-se atrativa sob o aspecto da redução do trabalho humano na avaliação de solicitações.

Ao se aprofundar sobre o problema, foi notado algo que é inerente à regulação: existe um grande desbalanceamento entre as classes autorizado e negado. Nas bases em que foram feitos experimentos, era comum que o número de itens autorizados fosse muito maior, variando de 70% até bases com cerca de 98% de autorizações. Essa constatação foi muito importante para a definição de uma proposta para o mecanismo de triagem automática. A ideia base foi usar a resposta automática, obtida via modelos classificadores gerados com técnicas de aprendizagem de máquina, combinados com o fator de confiança de tal resposta, somado ainda ao tipo da resposta (autorizada x negada). Tudo isso junto pode ser harmonizado para que seja criada uma abordagem que reduza o trabalho humano e que mesmo assim mantenha a qualidade das respostas, evitando a negação daquilo que deveria ser autorizado.

O processo de regulação com a triagem automática é descrito a seguir: (1) durante o atendimento a um paciente, o profissional de saúde avalia a necessidade de procedimentos para auxiliar no diagnóstico ou tratamento do beneficiário; (2) os procedimentos são solicitados via serviço disponibilizado pela OPS; (3) um autômato avalia as solicitações e emite sua resposta; caso seja sim e o fator de confiança da resposta seja superior ao configurado, a solicitação é automaticamente autorizada; (4) em caso contrário, a solicitação é direcionada a um regulador que verifica a conformidade com os padrões e com os protocolos clínicos estabelecidos; (5) se a análise for positiva, a realização dos procedimentos é autorizada. A Figura 7 ilustra esse processo.

Conforme destacado anteriormente, a grande mudança no processo de regulação, com a introdução da triagem automática, é a obtenção da resposta de um autômato. Essa resposta pode ser autorizar ou negar. No mecanismo de triagem automática proposto neste trabalho, apenas as solicitações com a resposta autorizadas devem ser respondidas de forma automática, evitando a negação automática do acesso do paciente à alternativa assistencial

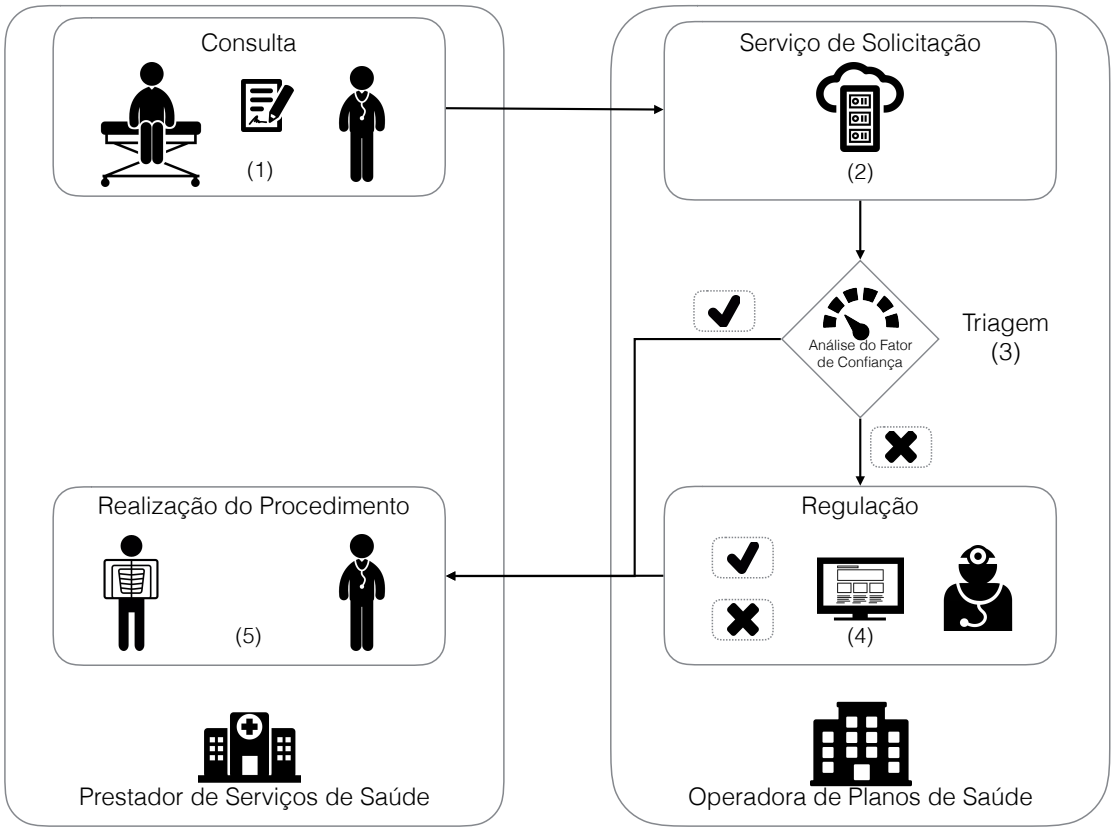

Figura 7 – Processo de regulação com a triagem automática.

proposta pelo médico. Além disso, para que sejam respondidas automaticamente, o fator de confiança da resposta, gerado pelo autômato, deve estar acima de um valor pré-definido pela OPS. Esse valor pode mudar ao longo do tempo, de acordo com o contexto da OPS, uma vez que, quando a situação da operadora estiver em equilíbrio, pode-se baixar o fator de confiança, fazendo com que mais solicitações sejam respondidas automaticamente. Por outro lado, pode-se configurar um fator de confiança elevado (próximo a 100%), o que irá causar um maior encaminhamento para reguladores humanos. Tal proposta garante que todas as eventuais respostas negativas tenham passado pela avaliação de uma figura humana e o uso do fator de confiança permite que seja configurado o nível de automação desejado.

Este trabalho, além de propor essa abordagem para a triagem automática, também apresenta uma metodologia para sua implantação em contextos reais. Tal metodologia é descrita nas próximas seções.

4.1 Metodologia para a implantação da triagem automática em OPSs

A Figura 8 apresenta as etapas da metodologia para implantação da triagem automática em OPSs, que são descritas logo em seguida.

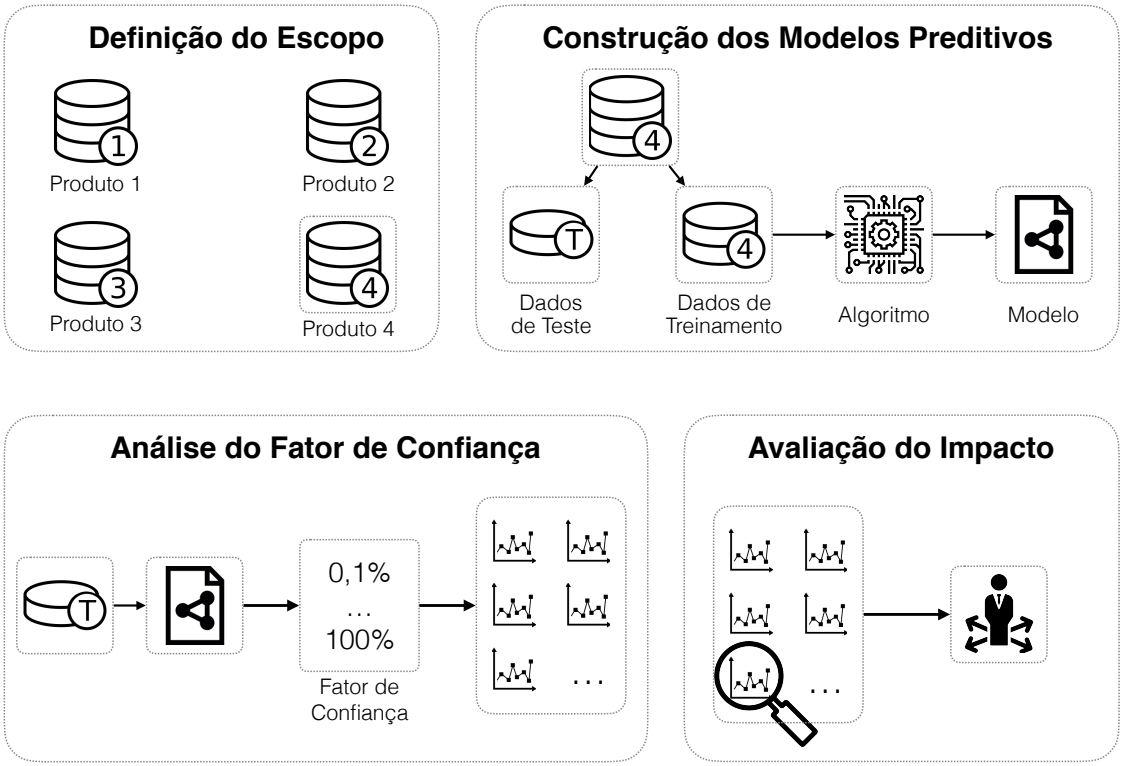

Figura 8 – Etapas da metodologia para implantação da triagem automática em OPSs.

4.1.1 Definição do escopo

Uma OPS pode oferecer diferentes produtos para o mercado. A diferença entre os produtos de uma operadora referem-se basicamente à segmentação, que pode ser ambulatorial (consulta, exames ou tratamentos), hospitalar (internação, urgência ou eletiva) e odontológica. Cada um desses produtos possui mecanismo de funcionamento diferente e requer mecanismos de regulação distintos, uma vez que envolvem procedimentos, tempo de resposta e volume de solicitações diferentes.

Dessa maneira, esta etapa tem como objetivo a seleção do item assistencial a ter a triagem automática implantada.

4.1.2 Construção dos modelos preditivos

A construção dos modelos preditivos é realizada por meio de técnicas de mineração de dados. A princípio, a base de dados do produto selecionado é submetida à fase de Pré-Processamento, na qual os dados são selecionados e tratados para remover informações ou características que possam afetar negativamente, ou que não contribuem para as próximas fases da mineração de dados. Após os tratamentos iniciais, os dados são particionados em conjuntos de treinamento e de teste, e os conjuntos de treinamento são submetidos a análise em buscas de padrões para criar os modelos preditivos. Nesta etapa deve-se gerar diferentes modelos, utilizando-se de diferentes técnicas, para se avaliar qual delas é a mais eficiente para o problema a ser tratado.

Ao final desta etapa um modelo preditivo é criado para o item assistencial indicado, que nada mais é do que o autômato responsável por responder as solicitações encaminhadas a uma OPS.

4.1.3 Análise do fator de confiança

O fator de confiança é um ponto importante da triagem automática e representa qual a probabilidade mínima para uma solicitação ser autorizada automaticamente. A etapa de análise do fator de confiança compreende a realização de experimentos para aferir a capacidade de identificação de solicitações a serem autorizadas e negadas, nos diferentes modelos preditivos.

Para cada modelo construído é realizada a validação cruzada *10-fold* com diferentes fatores de confiança. Em cada experimento, as solicitações do conjunto de teste são apresentadas ao modelo criado pelo classificador que atribui uma probabilidade para aquela solicitação ser autorizada ou negada.

A análise se dá pela variação do fator de confiança (a partir de 0,1% até 100%), na qual para cada valor são medidas as taxas de verdadeiros negativos e verdadeiros positivos em busca de uma faixa de valores que proporcione o aumento gradual no reconhecimento de solicitações inadequadas (verdadeiros negativos) sem comprometer o reconhecimento das solicitações adequadas (verdadeiros positivos).

4.1.4 Avaliação do impacto

Por fim, a avaliação do impacto da triagem automática é realizada apreciando-se os diferentes resultados obtidos com a análise dos fatores de confiança.

Tal análise é feita com o apoio de especialistas da OPS, levando-se em consideração a potencial redução da carga de trabalho no processo tradicional da regulação, conseguida

pela autorização automática de parte das solicitações, juntamente com a avaliação das principais métricas obtidas com o modelo preditivo utilizado.

Em um cenário real, as solicitações que apresentassem a probabilidade para autorização maior ou igual ao fator de confiança seriam autorizadas automaticamente, caso contrário seriam encaminhadas para a regulação humana. Assim, com um fator de confiança de 0% todas as solicitações deveriam ser autorizadas automaticamente e do outro lado, com um fator de confiança de 100%, todas as solicitações deveriam ser encaminhadas para a regulação humana.

A principal questão a ser abordada na avaliação do impacto é o erro envolvido na autorização automática (desperdício assistencial) e no direcionamento para a regulação humana (aumento do trabalho manual). Deve-se avaliar se o uso da triagem automática realmente traz ganhos e, se trouxer, qual o modelo a ser usado e com qual fator de confiança. Uma vez definidos tais aspectos, a implantação efetiva em um software de uma operadora pode ser feita a partir da invocação de uma função que avalia a solicitação, executando o modelo preditivo com as entradas informadas, para então gerar um resultado, que será considerado o resultado da regulação. Caso o fator de confiança esteja acima do configurado, a regulação será aprovada; em caso contrário ela continuará aguardando por uma avaliação humana.

5 Avaliação da Triagem Automática

A abordagem proposta neste trabalho necessita de uma avaliação para que seja possível mensurar se existem ganhos associados. O propósito desta avaliação é inferir se o uso da triagem automática no processo de regulação tradicional pode diminuir a quantidade de trabalho realizada pelos reguladores humanos, sem aumentar significativamente o desperdício assistencial. Para isso, foram analisadas várias bases de dados de diferentes serviços, na forma de uma avaliação em retrospectiva, por meio da aplicação da metodologia proposta para implantação da triagem automática. O intuito é de identificar o nível de resultados alcançados em vários contextos de OPSs.

5.1 Questões e métricas

É importante destacar que, devido ao alto grau de desbalanceamento, o ponto principal da avaliação é encontrar o menor erro envolvido na autorização automática e no encaminhamento de solicitações para o processo tradicional de regulação. Dessa maneira, os experimentos foram realizados com o objetivo de responder as seguintes perguntas:

- Qual o percentual de solicitações que pode ser autorizado automaticamente e qual o erro envolvido no processo? A métrica de valor preditivo positivo foi utilizada para responder essa pergunta.
- Qual o percentual de solicitações que podem ser encaminhadas para a regulação tradicional e qual o erro envolvido no processo? A métrica de valor preditivo negativo foi utilizada para responder essa pergunta.

5.2 Execução do Experimento

Os dados utilizados nesta avaliação foram obtidos de uma empresa que oferece serviço de apoio à gestão de planos de saúde. As sete bases de dados utilizadas cobrem procedimentos médicos e odontológicos, realizados nos âmbitos ambulatorial e hospitalar, e foram submetidas individualmente à metodologia descrita neste trabalho.

A Tabela 2 apresenta algumas características das bases utilizadas. Um ponto comum a todas as bases de dados, e que dificulta o processo de descoberta de conhecimento, é a grande disparidade entre a quantidade de solicitações autorizadas e negadas. Enquanto a base de dados mais balanceada apresenta uma razão de cerca de 70% de solicitações

autorizadas e cerca de 30% negadas, a maioria das bases apresenta uma razão de mais de 95% de solicitações autorizadas e menos de 5% negadas.

Tabela 2 – Características das bases de dados (BD) utilizadas.

BD	Fins Lucrativos	Plano	Atendimento	Autorizados/Não Autorizados
I	SIM	Médico	Hospitalar	72,24% / 27,76%
II	NÃO	Médico	Ambulatorial	99,76% / 0,24%
III	NÃO	Médico	Ambulatorial	100,00% / 0,00%
IV	NÃO	Médico	Ambulatorial	99,13% / 0,87%
V	NÃO	Médico	Hospitalar	97,76% / 2,24%
VI	NÃO	Médico	Hospitalar	97,50% / 2,50%
VII	NÃO	Odontológico	Ambulatorial	96,77% / 3,23%

5.2.1 Definição do escopo

A etapa definição do escopo foi resolvida pela disponibilidade das bases de dados. Cada base estava associada a um dado produto, que possuía uma cobertura e uma segmentação específica.

5.2.2 Construção dos modelos preditivos

A construção dos modelos preditivos é realizada por meio de técnicas de Mineração de Dados. A princípio, as bases de dados disponíveis são submetidas à fase de pré-processamento, na qual os dados são selecionados e tratados para remover informações ou características que possam afetar negativamente as próximas fases. Após os tratamentos iniciais, na fase de mineração de dados, os dados são algoritmicamente analisados em busca de padrões para criar os modelos preditivos.

Para a execução dos algoritmos relacionados à fase de mineração foi utilizada a ferramenta WEKA (WEKA, 2016) por ser amplamente utilizada e aceita na literatura de descoberta de conhecimento em bancos de dados, além do seu fácil manuseio e possibilidade de utilização programática. Os experimentos foram codificados com o intuito de facilitar a re-execução e parametrização, para isso foi utilizado o ambiente de desenvolvimento gráfico Eclipse (ECLIPSE, 2016) e a linguagem Java (JAVA, 2016), devido à sua compatibilidade com a ferramenta WEKA. Com excessão do algoritmo SVM, que foi configurado para fornecer saídas contínuas por meio da opção “Construir Modelos Logísticos”, os algoritmos foram executados com seus parâmetros padrão (i.e. não foi realizado o ajuste de parâmetros, também conhecido como *tunning*).

5.2.2.1 Pré-processamento

A fase de pré-processamento tem como objetivo o entendimento do problema e a seleção, limpeza e transformação dos dados. Tendo em vista o desbalanceamento presente na base de dados original, essa etapa é de grande importância para que se possa remover quaisquer características que possam afetar mais ainda a qualidade dos dados e impactar negativamente na etapa de mineração. O pré-processamento realizado no presente trabalho contém os seguintes momentos: seleção de solicitações, seleção manual de atributos, construção de atributos, seleção automática de atributos, estratificação e balanceamento.

Com o intuito de melhorar a qualidade dos dados disponíveis para a mineração de dados, os especialistas da empresa parceira determinaram um período específico para a seleção de solicitações. Esta delimitação é essencial para a remoção de inconsistências nas bases de dados por mudanças nos padrões ou protocolos assistenciais estabelecidos.

Na seleção manual de atributos, o número de atributos utilizados para representar a regulação de um procedimento passa por uma primeira redução. Nem todos os atributos contidos na base de dados, em entidades envolvidas no processo, possuem qualidade suficiente para serem utilizados ou representam informações úteis para a regulação. A Tabela 3 apresenta exemplos de atributos removidos durante a seleção manual. Cinco grupos de atributos foram removidos: sigilosos e/ou privados, i.e. dados que possibilitem identificar o segurado (atributo “CPF”); irrelevantes, i.e. identificadores internos do sistema da OPS (atributo “ID”); preenchidos com valor padrão (atributo “Desconto”); duplicados (atributos “Valor” e “Total”); e poluídos, i.e. preenchidos fora do padrão esperado (atributo “Idade $>$ 130 anos”).

Tabela 3 – Exemplos de atributos removidos durante a seleção manual.

ID	Segurado	Idade	Valor	Desconto	Total
1	João	Novo	10.0	0.0	10.0
2	Maria	Nova	50.0	0.0	50.0
3	Alberto	Velho	20.0	0.0	20.0
4	Catarina	Menor	30.0	0.0	30.0

Após a seleção manual de atributos, novos atributos são construídos para representar informações contidas indiretamente nos dados. Os atributos construídos descrevem o histórico de procedimentos realizados por um dado beneficiário e também constituem informações mais detalhadas sobre o procedimento solicitado e sobre a data de realização. Um exemplo da construção de atributos pode ser visto na Tabela 4, na qual o atributo “Data Nascimento” é transformado em “Idade”.

Seguidamente, foi utilizado o método de ganho de informação para atribuir um

Tabela 4 – Criação dos atributos “Idade” a partir do atributo “Data Nascimento”.

Data Nascimento	Idade
28/09/1975	41
04/07/2005	11
01/01/1988	29

valor para a importância de cada atributo no processo de decisão e aqueles que possuíam um ganho igual a zero foram removidos das bases de dados. Com isso, somente os dados que podem efetivamente contribuir para o processo de triagem são utilizados para a mineração de dados.

Os dados resultantes da seleção automática de atributos precisam ser particionados em subconjuntos para treinamento e testes. Essa separação é de suma importância para a criação e avaliação dos modelos gerados. São necessários muitos exemplos de solicitações para o treinamento dos algoritmos de classificação e, ao mesmo tempo, é indispensável reservar uma boa quantidade de dados para aferir a qualidade do modelo criado. Neste trabalho o conjunto original de dados é estratificado aleatoriamente em dez subconjuntos de mesmo tamanho que são utilizados no método de validação cruzada *10-fold*, para a análise do fator de confiança.

Para finalizar o pré-processamento, as técnicas de *Random Oversampling* e SMOTE são utilizadas para criar subconjuntos de treinamento balanceados. Enquanto a *Random Oversampling* aumenta o número de exemplos de solicitações não autorizadas nos subconjuntos de treinamento, por meio da replicação aleatória dos exemplos existentes, a SMOTE busca adicionar informação ao subconjunto por meio da criação de informações sintéticas combinando um dado exemplo e seus vizinhos. O processo de estratificação e balanceamento é ilustrado na Figura 9.

5.2.2.2 Mineração de dados

É nesta fase que os modelos de triagem são criados. O escopo da classificação neste trabalho consiste em separar novas solicitações em autorizadas ou negadas. Os seguintes algoritmos foram empregados nesta etapa: C4.5, *Ripper*, *Random Forest*, *Naive Bayes* e SVM. Tais algoritmos foram selecionados por conta de terem tido bons resultados nas pesquisas iniciais executadas pelo grupo.

A Figura 10 ilustra o treinamento dos algoritmos para a criação dos modelos preditivos. Este processo é repetido a fim de gerar dez modelos para cada combinação entre algoritmo e técnica de tratamento de desbalanceamento. Esses modelos são utilizados para calcular os resultados a partir do método de validação cruzada *10-fold*.

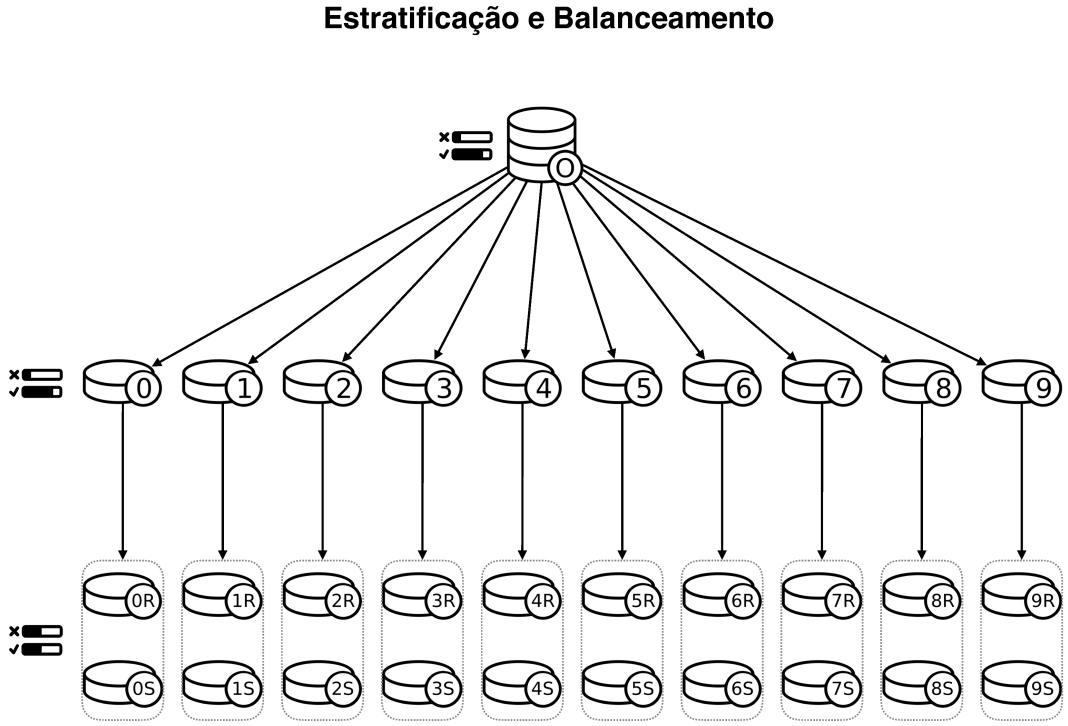

Figura 9 – Estratificação e balanceamento das bases de dados.

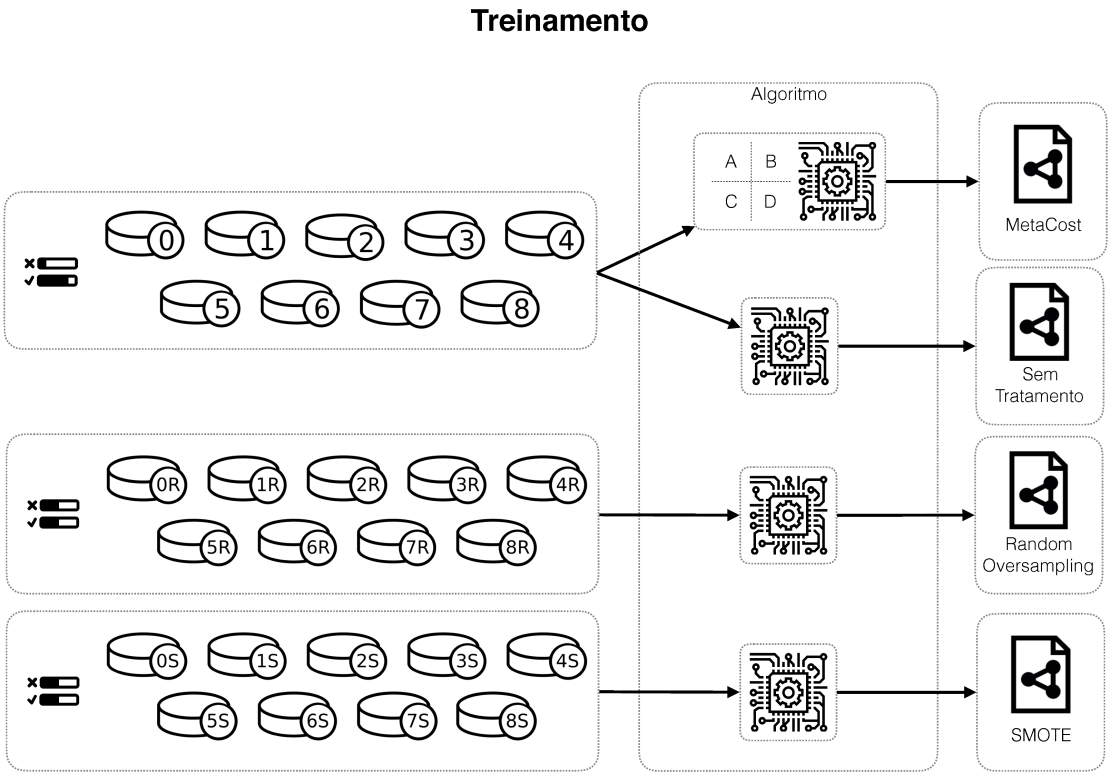

Figura 10 – Treinamento dos algoritmos a partir dos dados estratificados e balanceados.

5.2.2.2.1 Resultados

A Tabela 5 apresenta as características das bases obtidas para o estudo após a seleção de solicitações e a seleção manual de atributos realizadas por especialistas da empresa parceira. Essas características são consideradas como as características iniciais para o trabalho. A Base de Dados III foi excluída por não conter nenhuma solicitação negada, fato esse que impede seu uso para fins de aprendizado.

Tabela 5 – Características das bases de dados (BD) após a seleção de solicitações e a seleção manual de atributos realizada por especialistas da empresa parceira.

BD	Período	Procedimentos	Autorizados	Negados	Atributos (A)
I	11/2012 a 07/2016	14.548	72,24%	27,76%	12
II	07/2014 a 05/2016	16.587	99,76%	0,24%	17
III	09/2007 a 03/2016	486	100,00%	0,00%	12
IV	12/2009 a 07/2016	45.021	98,65%	1,35%	07
V	07/2014 a 05/2016	16.943	97,76%	2,24%	12
VI	08/2007 a 04/2016	4.086	97,50%	2,50%	11
VII	08/2007 a 04/2016	55.787	94,12%	5,88%	11

A Tabela 6 ilustra as características das bases de dados após a construção de atributos e de seleção automática de atributos. A construção (C) de atributos gerou dezenove novos atributos que representam o histórico de procedimentos do segurado (14 atributos), detalhes sobre o procedimento executado (3 atributos) e por fim sobre a data de execução (2 atributos)[1]. Por fim, o ganho de informação de cada atributo foi calculado para cada base de dados e os atributos cujo ganho de informação é igual a zero foram removidos (E). Por se tratar de múltiplas bases de dados, os atributos finais não foram coincidentes em todas as situações e um relatório detalhado por base pode ser encontrado no Apêndice A.

Tabela 6 – Características das Bases de Dados (BD) utilizadas após a construção de atributos e seleção automática de atributos.

BD	Procedimentos	Autorizados	Negados	Atributos (A + C - E)
I	14.548	72,24%	27,76%	12 + 19 - 11
II	16.587	99,76%	0,24%	17 + 20 - 14
III	—	—	—	—
IV	45.021	98,65%	1,35%	07 + 19 - 8
V	9.866	97,76%	2,24%	12 + 19 - 12
VI	4.026	97,50%	2,50%	11 + 19 - 12
VII	55.787	94,12%	5,88%	11 + 19 - 13

[1] O atributo adicional da Base de Dados II indica se o segurado já foi submetido à cirurgia e foi construído a partir da Base de Dados V pois as duas estão relacionadas.

Por fim, as bases de dados foram aleatoriamente estratificadas em dez subconjuntos e cada subconjunto deu origem a mais dois, por meio da criação (SMOTE) e duplicação (*Random Oversampling*) de instâncias negadas para que os novos subconjuntos apresentassem a mesma quantidade de solicitações autorizadas e negadas para o treinamento dos algoritmos de classificação.

5.2.3 Análise do fator de confiança

Vários experimentos foram realizados para cada combinação entre base de dados e algoritmo, visando explorar toda a faixa de valores de fator confiança para a identificação dos melhores resultados.

5.2.3.1 Resultados

Para o objetivo deste trabalho é essencial que a identificação de solicitações negadas seja maximizada, contudo, mantendo um número razoável na identificação de solicitações autorizadas. Foram estipulados arbitrariamente os valores mínimos de 50% para a Taxa de Verdadeiros Negativos (TVN, detecção de solicitações negadas) e de 25% para a Taxa de Verdadeiros Positivos (TVP, detecção de solicitações autorizadas) para um resultado ser considerado razoável. A seguir os resultados são apresentados divididos por algoritmo e técnica de tratamento de desbalanceamento para cada base de dados.

5.2.3.1.1 C4.5

As Figuras 11, 12, 13, 14, 15 e 16 ilustram o comportamento da TVN e da TVP causado pela alteração do fator de confiança e a Tabela 7 exibe as faixas de valores para o fator de confiança nas quais os requisitos mínimos foram atendidos (as bases de dados que não apresentaram uma faixa de fatores de confiança dentro dos valores mínimos definidos não foram incluídas na tabela).

Nos experimentos sem o tratamento para o desbalanceamento somente a Base de Dados I obteve um resultado satisfatório, sendo possível determinar uma faixa de fatores de confiança de 19,60% o que permite escolher gradativamente entre autorizar automaticamente mais ou menos solicitações.

A utilização da técnica SMOTE diminuiu timidamente a TVN máxima para a Base de Dados I mas em contrapartida aumentou bastante a faixa de fatores de confiança, o que concede maior poder de configuração da triagem. A utilização desta técnica melhorou os resultados das bases mais desbalanceadas, mas em graus diferentes: as Bases de Dados II, VI e VI conseguiram atingir os mínimos estipulados mas os valores máximos não chegam

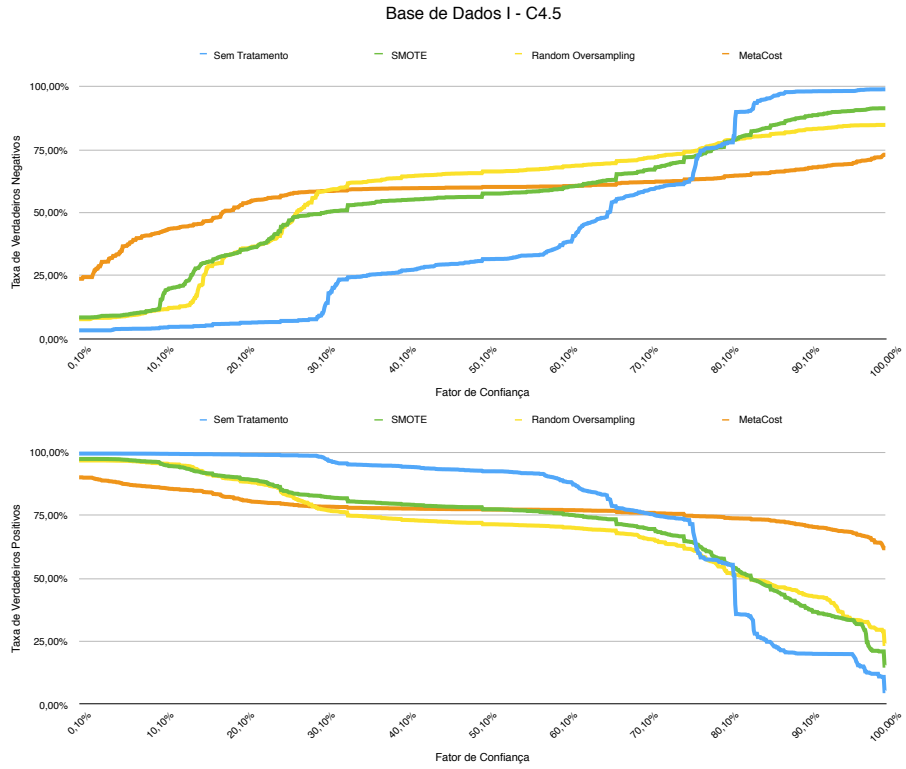

Figura 11 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados I com o algoritmo C4.5.

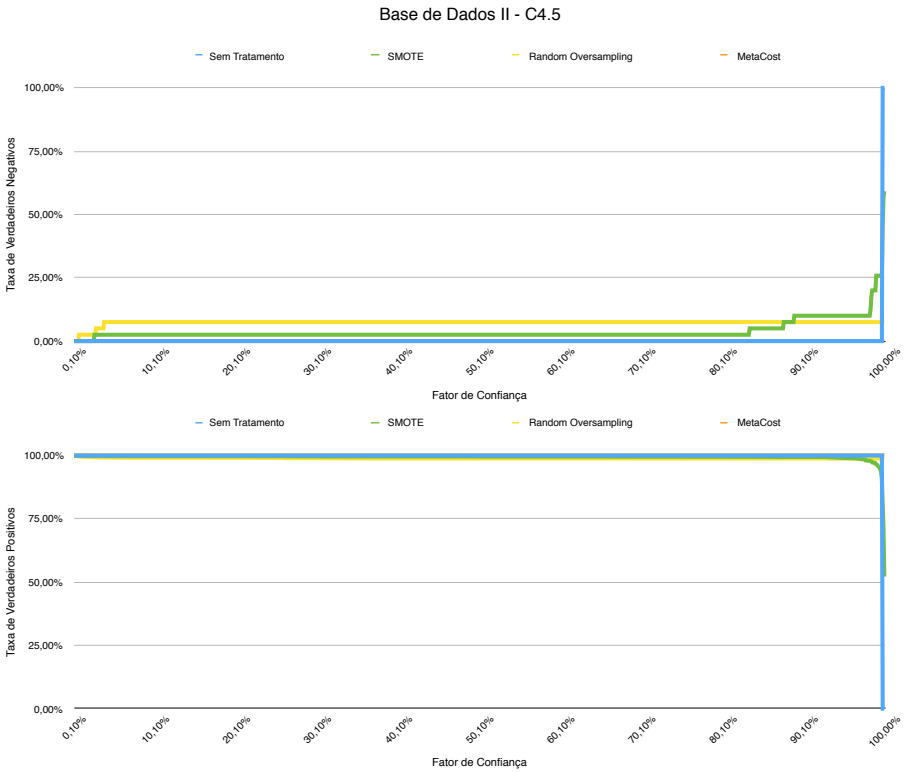

Figura 12 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados II com o algoritmo C4.5.

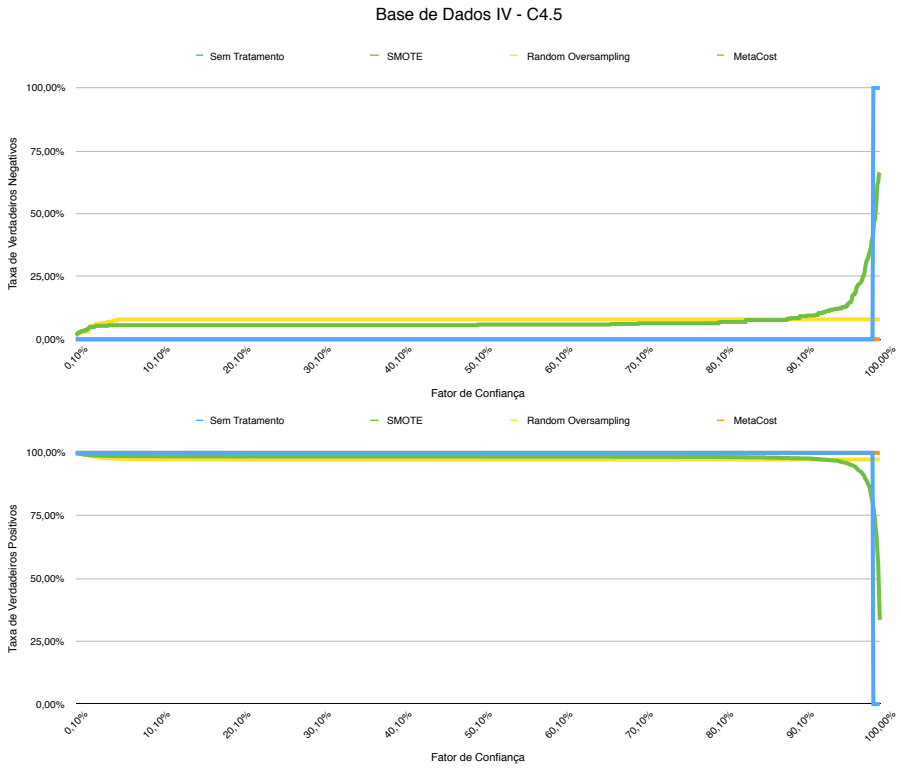

Figura 13 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados IV com o algoritmo C4.5.

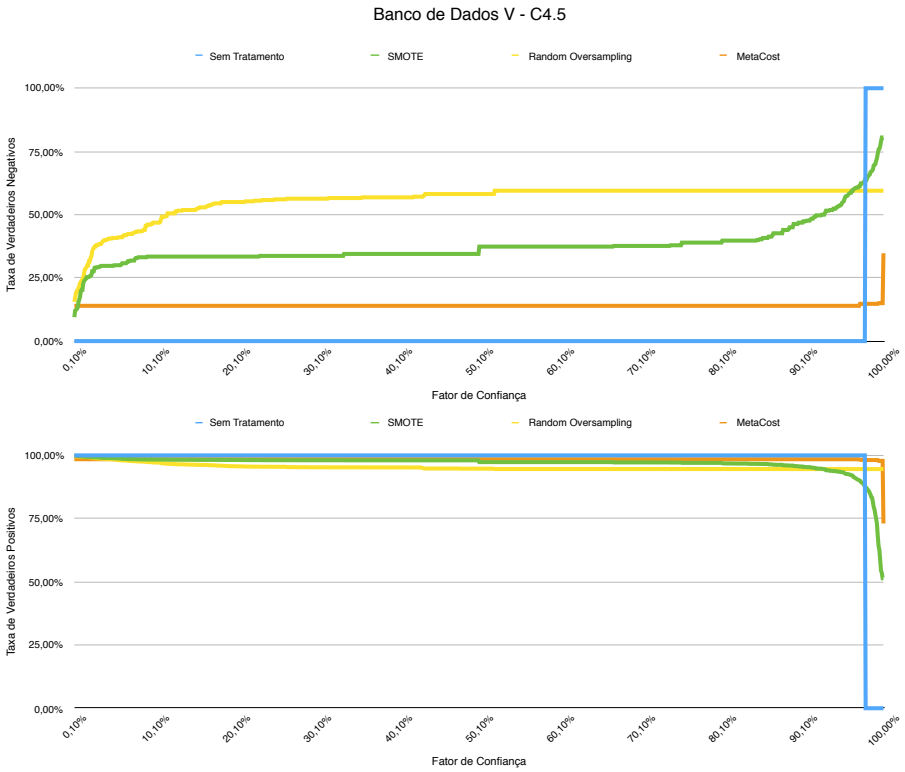

Figura 14 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados V com o algoritmo C4.5.

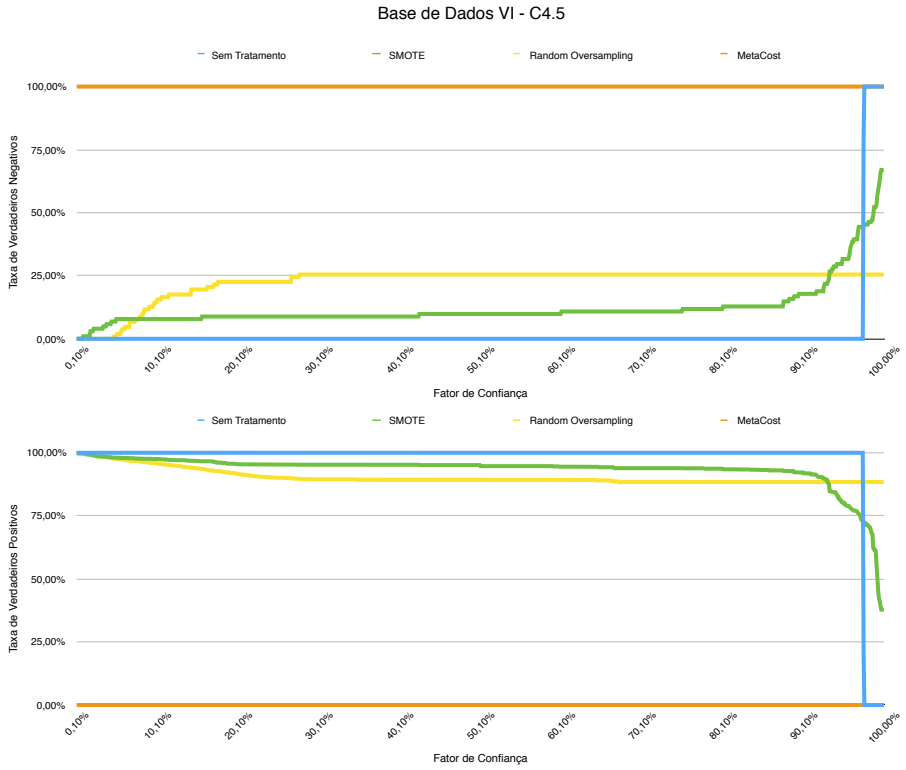

Figura 15 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados VI com o algoritmo C4.5.

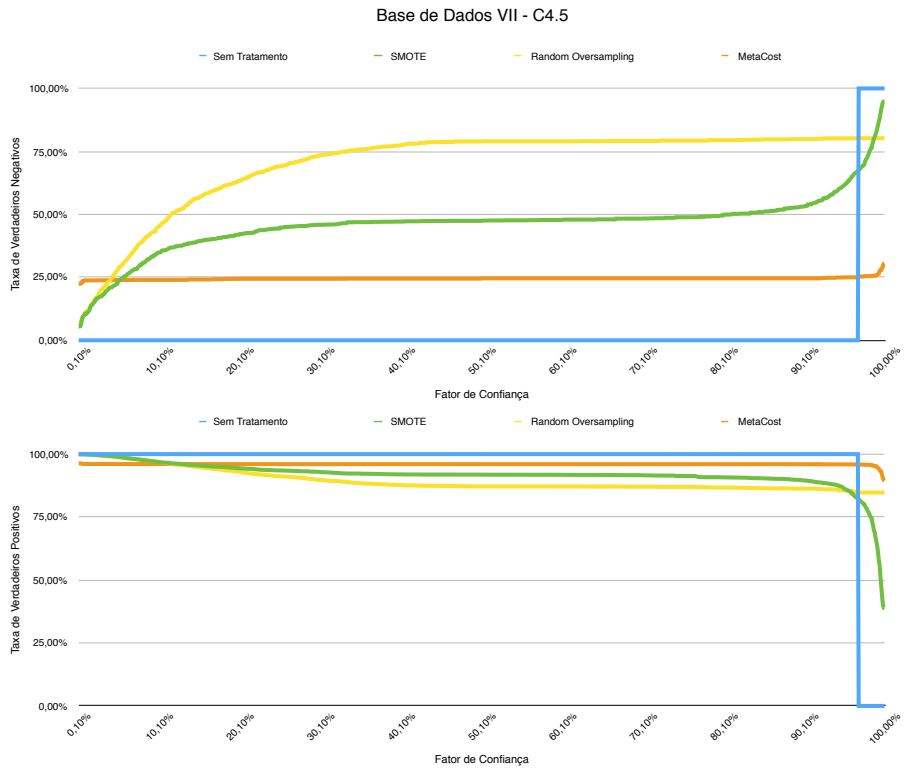

Figura 16 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados VII com o algoritmo C4.5.

Tabela 7 – Fatores de confiança mínimos e máximos, por base de dados, para o classificador C4.5.

BD	Fator de Confiança	TVN	TVP	Diferença
		Sem Tratamento		
I	65,60%	50,11%	81,52%	19,60%
	85,20%	95,02%	25,65%	
		SMOTE		
I	30,60%	50,06%	82,36%	67,00%
	97,69%	90,84%	28,00%	
II	99,90%	56,67%	70,98%	0,10%
	100,00%	59,17%	52,20%	
IV	99,50%	51,03%	69,64%	0,50%
	100,00%	66,31%	33,56%	
V	92,40%	50,00%	94,66%	7,60%
	100,00%	80,53%	51,40%	
VI	98,80%	52,18%	61,62%	1,20%
	100,00%	66,91%	37,80%	
VII	81,00%	50,08%	90,73%	19,00%
	100,00%	94,84%	39,01%	
		Random Oversampling		
I	27,10%	50,38%	81,34%	72,70%
	99,80%	84,80%	28,96%	
V	11,60%	50,53%	96,68%	88,40%
	100,00%	59,47%	94,59%	
VII	11,60%	50,08%	96,15%	88,40%
	100,00%	80,23%	84,76%	
		MetaCost		
I	17,90%	50,26%	82,55%	82,10%
	100,00%	72,96%	61,99%	

a configurar bons resultados; já as Bases de Dados V e VII conseguiram altas TVN mas uma faixa de fatores de confiança inferior ao resultado da Base de Dados I.

Assim como a aplicação da técnica SMOTE, a utilização do *Random Oversampling* diminuiu a TVN máxima para a Base de Dados I mas em contrapartida aumentou a faixa de fatores de confiança. Este resultado se repetiu para as Bases de Dados V e VII, mas a utilização da técnica em questão não garantiu os valores mínimos definidos nas Bases de Dados II, IV e VI.

A técnica *MetaCost* não atingiu os valores mínimos estipulados para as bases de dados mais desbalanceadas mas garantiu para a Base de Dados I o resultado mais equilibrado em relação as métricas de TVN e TVP além da maior faixa de fatores de confiança.

5.2.3.1.2 *Ripper*

As Figuras 17, 18, 19, 20, 21 e 22 ilustram o comportamento da TVN e da TVP causado pela alteração do fator de confiança e a Tabela 8 exibe as faixas de valores para o fator de confiança nas quais os requisitos mínimos foram atendidos (as bases de dados que não apresentaram uma faixa de fatores de confiança dentro dos valores mínimos definidos não foram incluídas na tabela).

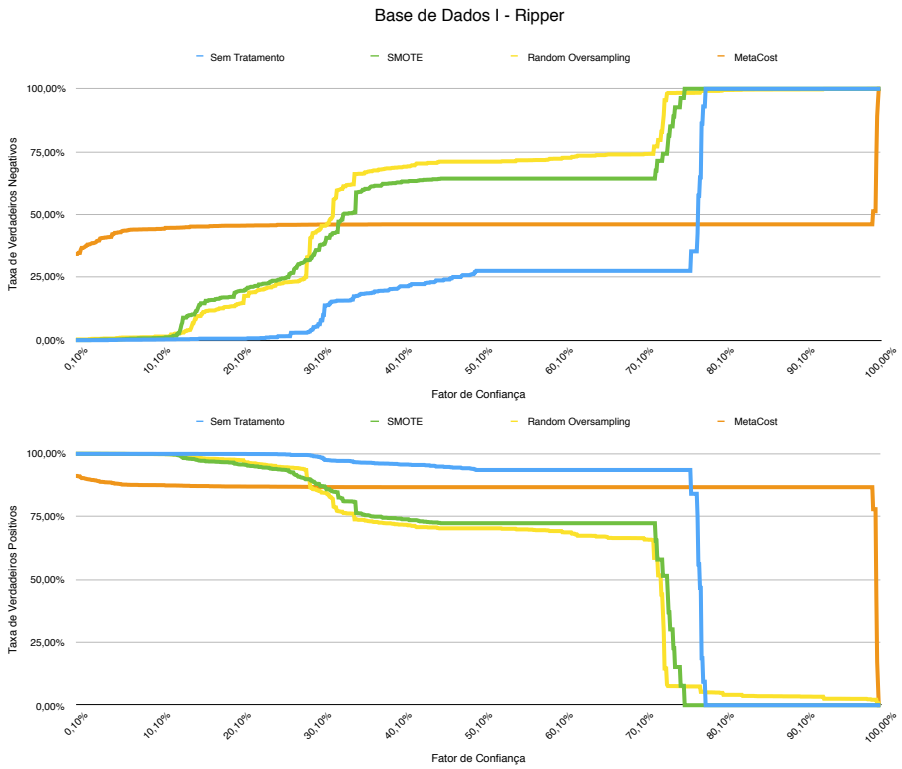

Figura 17 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados I com o algoritmo *Ripper*.

Assim como nos experimentos com o algoritmo C4.5, a aplicação do algoritmo *Ripper* sem a utilização de técnicas para o tratamento de desbalanceamento só produziu uma faixa de fatores de confiança dentre os valores mínimos estipulados para as métricas de TVN e TVP na Base de Dados I, ainda assim a faixa encontrada é muito curta para este ser considerado um bom resultado.

A utilização da técnica SMOTE melhorou substancialmente os resultados para a Base de Dados I, aumentando a TVN máxima e a faixa de fatores de confiança, apesar de diminuir a TVP no limite superior do Fator de Confiança. Em relação às bases de dados mais desbalanceadas, somente a Base de Dados VII obteve resultados razoáveis com a técnica SMOTE.

A utilização do *Random Oversampling* obteve resultados similares ao da técnica SMOTE, apresentando uma leve melhora no valor máximo da TVN. Já na Base de Dados

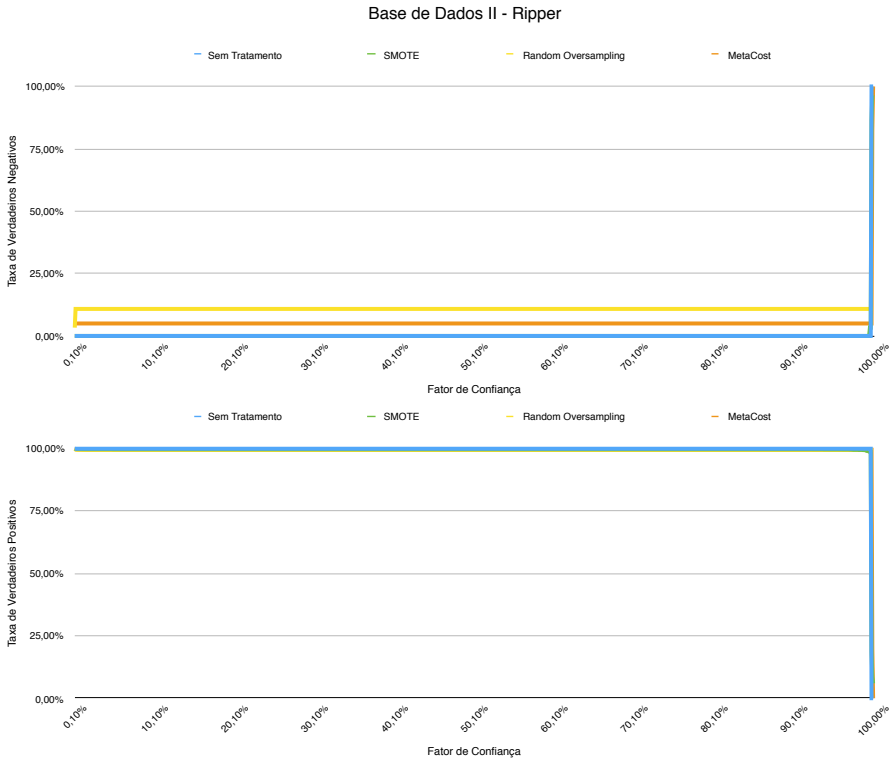

Figura 18 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados II com o algoritmo *Ripper*.

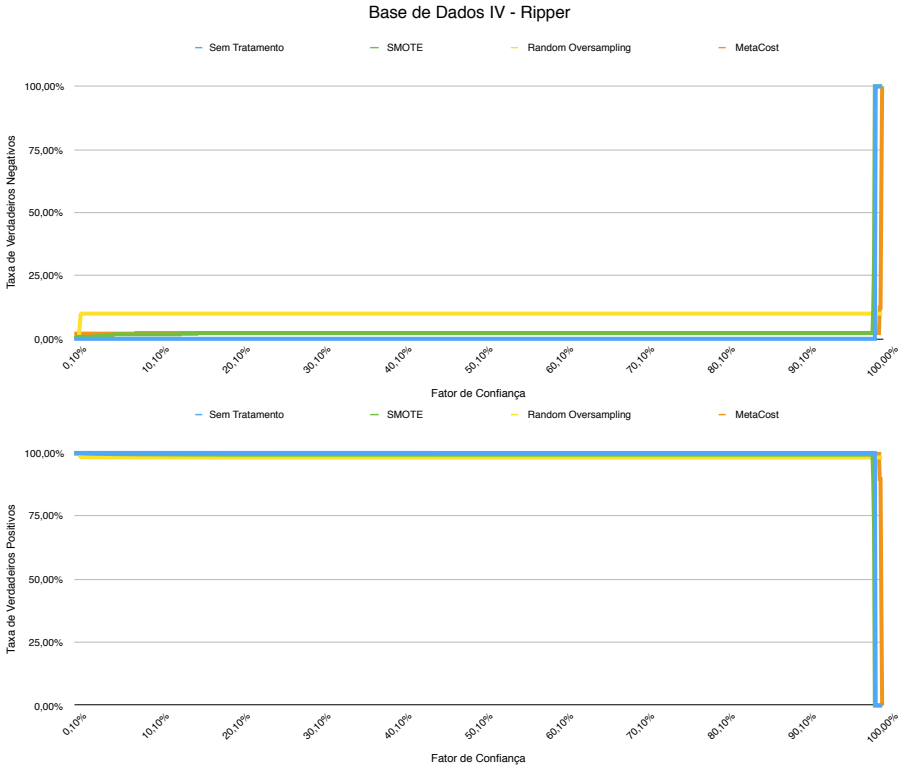

Figura 19 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados IV com o algoritmo *Ripper*.

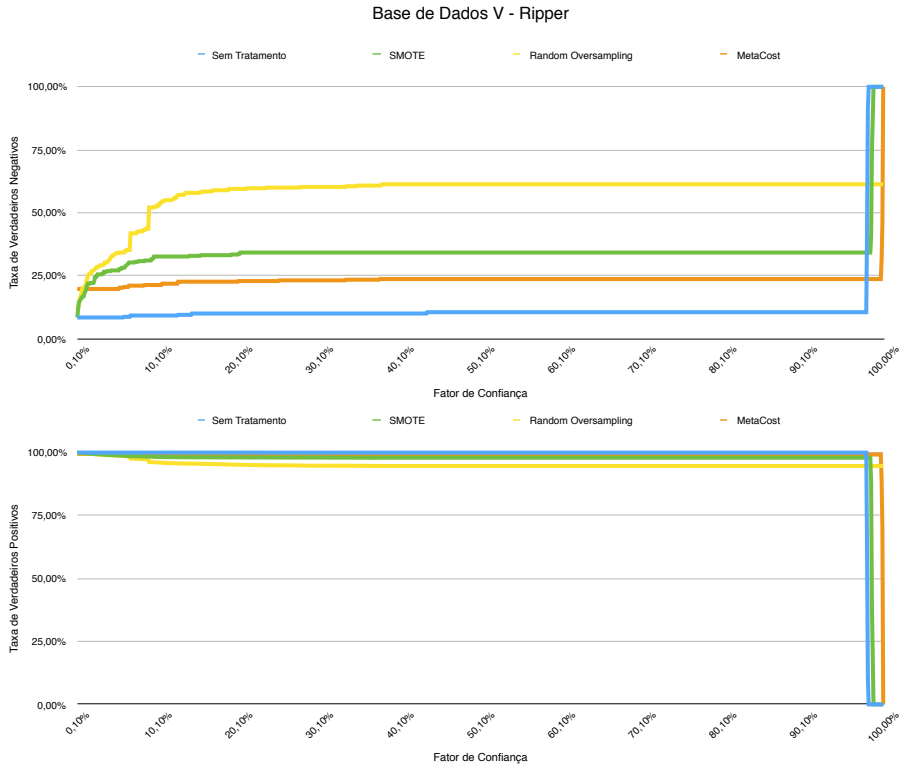

Figura 20 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados V com o algoritmo *Ripper*.

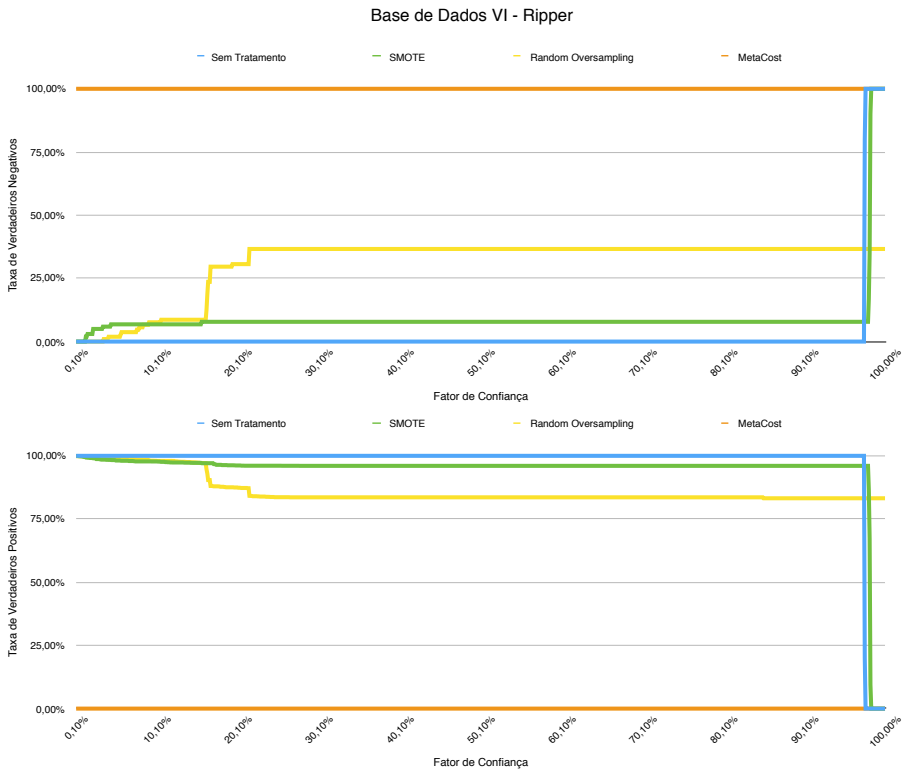

Figura 21 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados VI com o algoritmo *Ripper*.

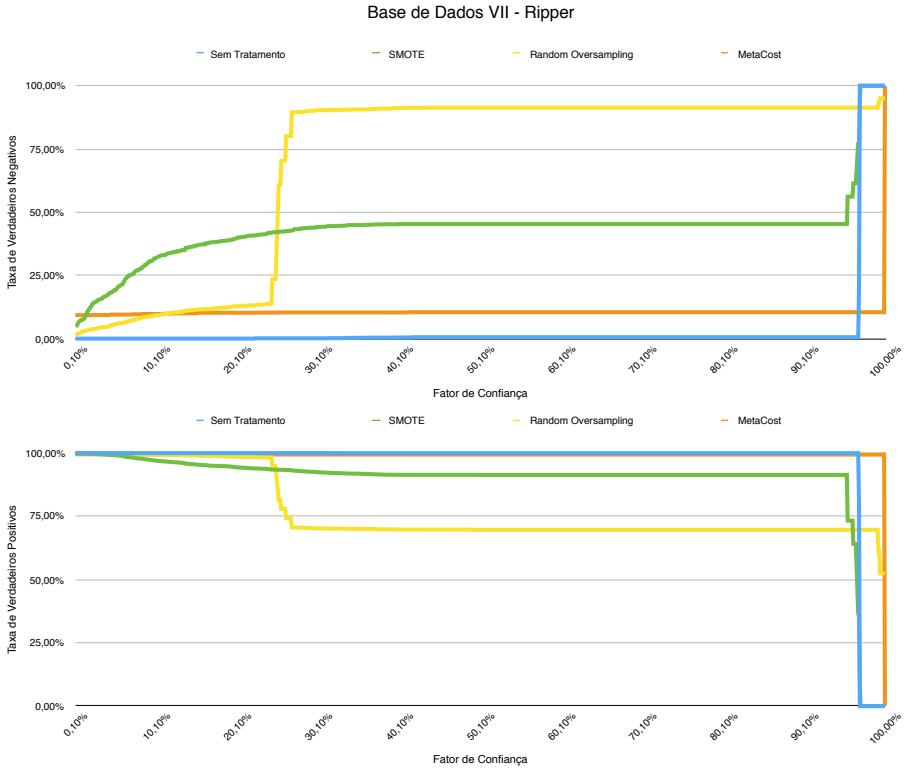

Figura 22 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados VII com o algoritmo *Ripper*.

VII houve uma considerável melhora tanto nos valores de TVN e TVP, quanto na faixa de fatores de confiança identificada. Com o *Random Oversampling* também foi possível obter uma grande faixa de fatores de confiança para a Base de Dados V, na qual, apesar do valor máximo de TVN não ser muito alto, os valores da TVP são próximos a 95%.

A técnica *MetaCost* não obteve resultados para as bases de dados mais desbalanceadas e nem garantiu para a Base de Dados I uma grande faixa de fatores de confiança.

5.2.3.1.3 *NaiveBayes*

As Figuras 23, 24, 25, 26, 27 e 28 ilustram o comportamento da TVN e da TVP causado pela alteração do fator de confiança e as Tabelas 9 e 10 exibem as faixas de valores para o fator de confiança nas quais os requisitos mínimos foram atendidos (as bases de dados que não apresentaram uma faixa de fatores de confiança dentro dos valores mínimos definidos não foram incluídas na tabela).

Com o algoritmo *Naive Bayes* foi possível determinar uma faixa de fatores de confiança para todos as bases de dados sem a utilização de nenhuma técnica de tratamento para o desbalanceamento e, com exceção da Base de Dados II, em todos os casos a TVN foi superior a 85%. Porém, somente na Base de Dados I foi possível definir uma faixa de fatores de confiança grande o suficiente para permitir uma escolha gradativa entre

Tabela 8 – Fatores de confiança mínimos e máximos, por base de dados, para o classificador *Ripper*.

BD	Fator de Confiança	TVN	TVP	Diferença
		Sem Tratamento		
I	77,40%	57,48%	56,01%	0,30%
	77,70%	64,73%	46,68%	
		SMOTE		
I	33,30%	50,11%	81,16%	40,90%
	74,20%	85,02%	30,17%	
VII	95,40%	56,17%	73,32%	1,40%
	96,80%	77,00%	36,71%	
		Random Oversampling		
I	32,10%	55,96%	78,93%	41,00%
	73,10%	89,53%	29,20%	
V	9,00%	52,11%	96,23%	91,00%
	100,00%	61,32%	94,64%	
VII	25,10%	51,52%	84,76%	74,90%
	100,00%	95,06%	52,49%	
		MetaCost		
I	99,10%	51,27%	77,97%	0,40%
	99,50%	72,75%	43,38%	

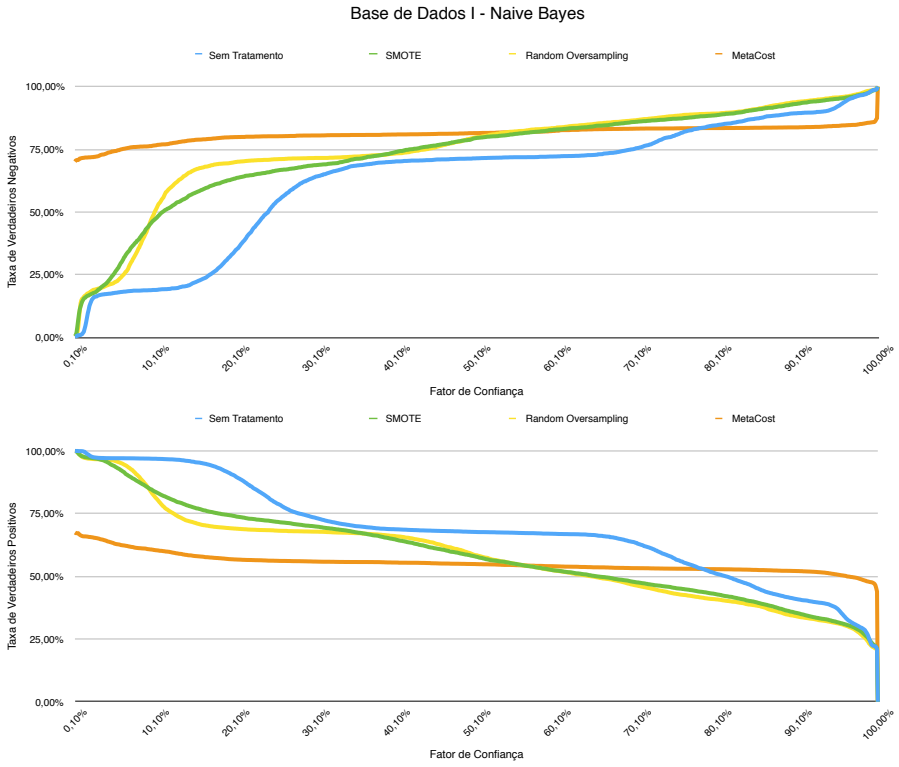

Figura 23 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados I com o algoritmo *NaiveBayes*.

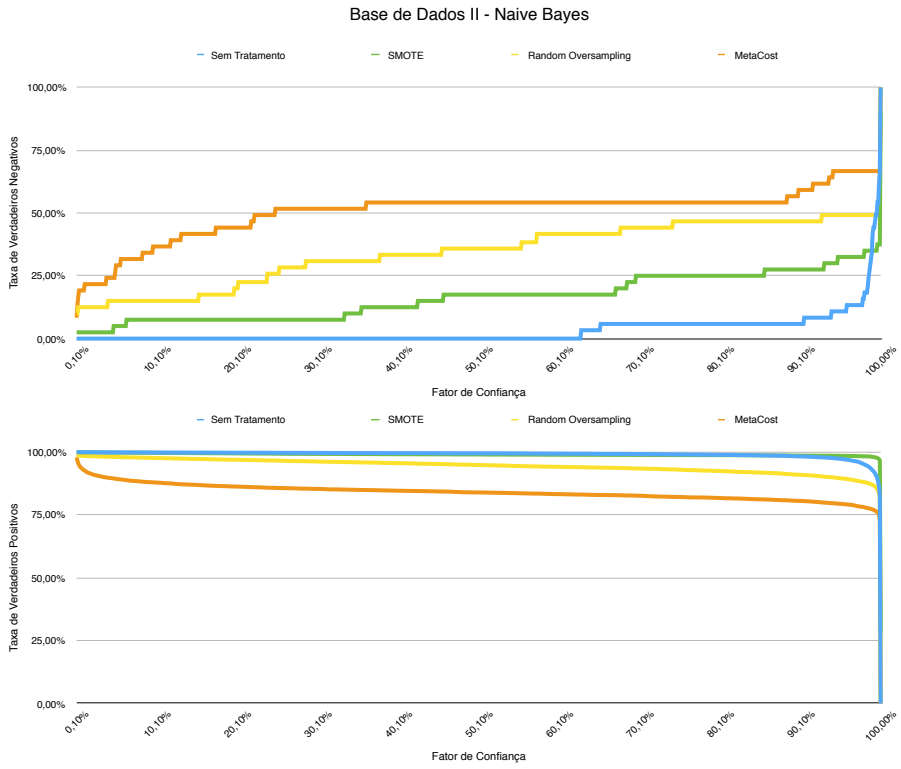

Figura 24 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados II com o algoritmo *NaiveBayes*.

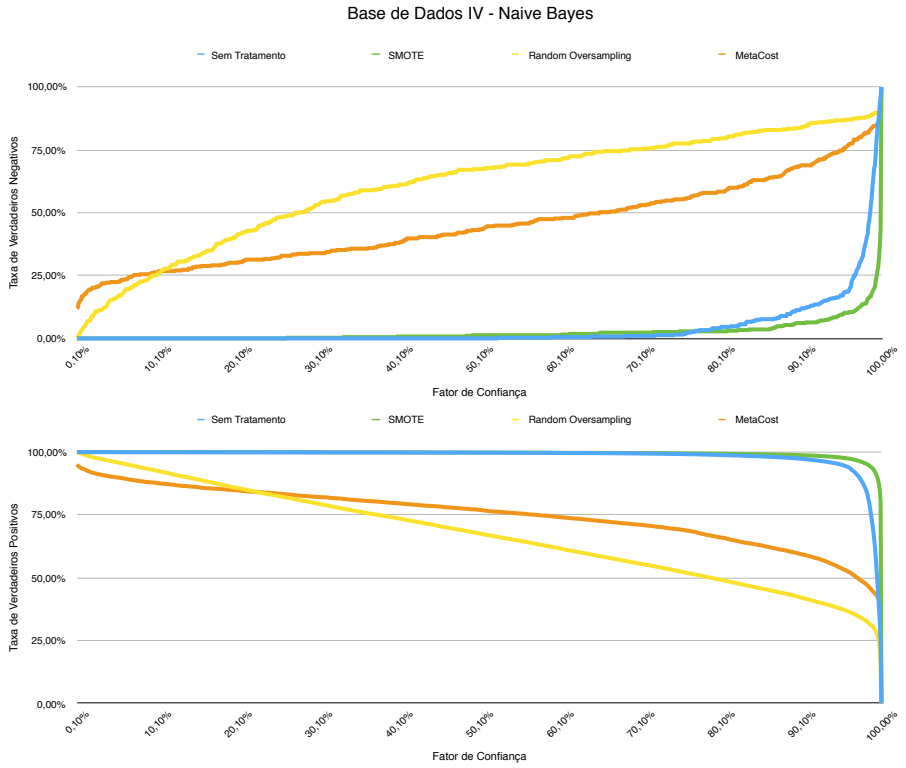

Figura 25 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados IV com o algoritmo *NaiveBayes*.

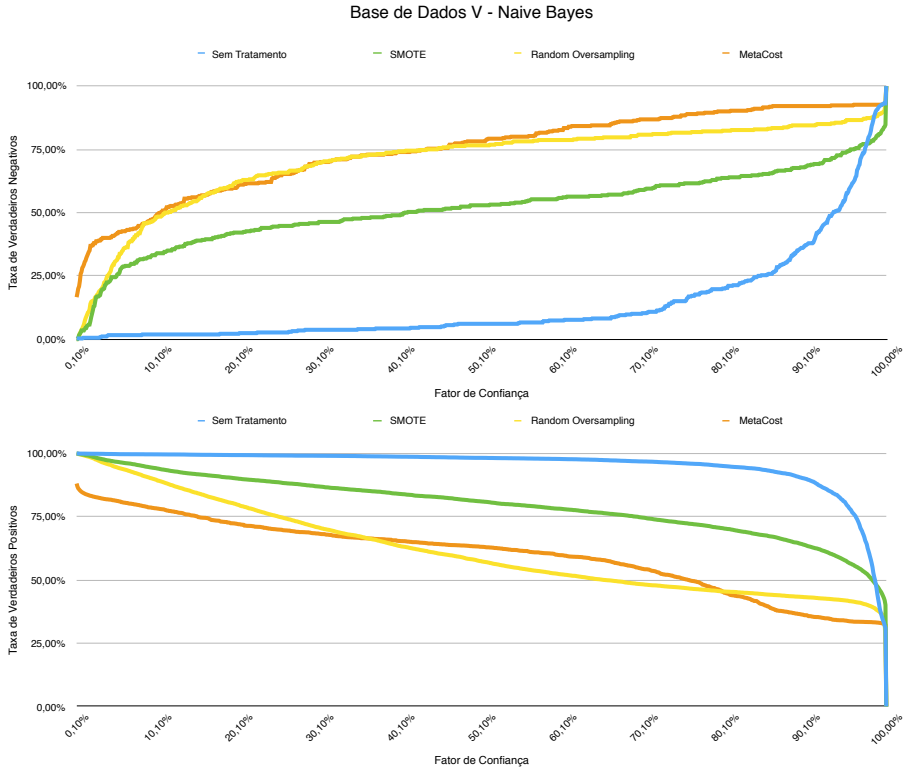

Figura 26 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados V com o algoritmo *NaiveBayes*.

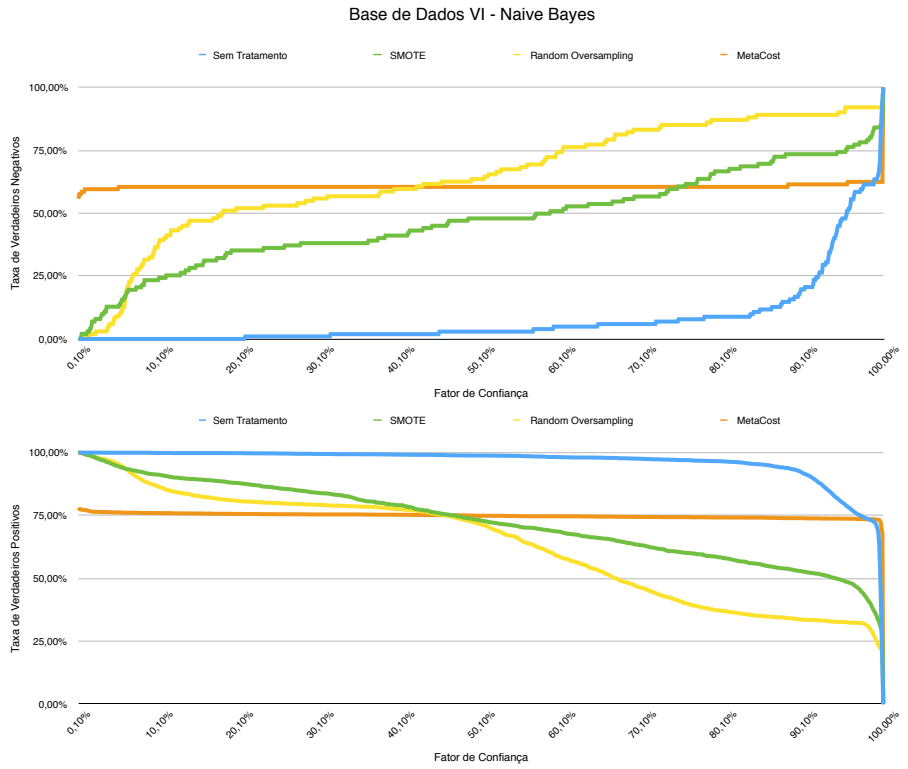

Figura 27 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados VI com o algoritmo *NaiveBayes*.

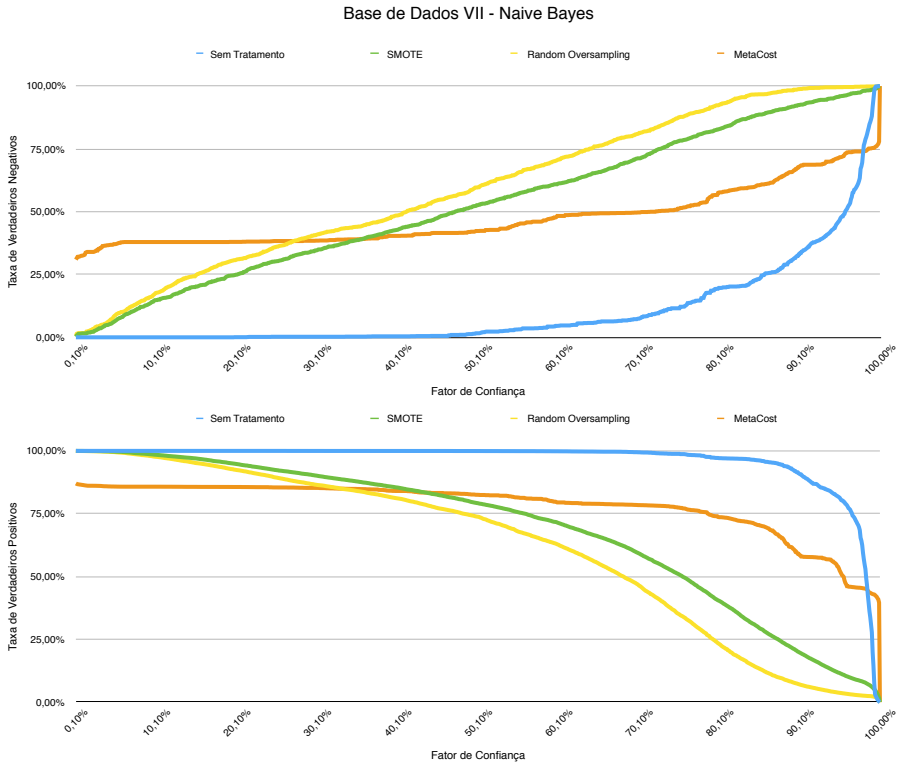

Figura 28 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados VII com o algoritmo *NaiveBayes*.

autorizar automaticamente mais ou menos solicitações.

A utilização da técnica SMOTE não mudou o desempenho na Base de Dados I, enquanto diminuiu o desempenho nas duas bases de dados mais desbalanceadas (Bases de Dados II e IV), que não atingiram os valores mínimos das métricas de TVN e TVP. Para as Bases de Dados V e VI houve uma redução no valor máximo da TVN mas também houve um aumento substancial da faixa de fatores de confiança. Para a Base de Dados VII houve uma melhora tímida no valor máximo da TVN, juntamente com um aumento substancial da faixa de fatores de confiança.

A utilização do *Random Oversampling* obteve resultados similares ao da técnica SMOTE porém um aumento significativo na maioria das faixas de fatores de confiança excetuando-se a Base de Dados II. Essas faixas são suficiente para permitir a escolha gradativa entre a autorização automática em massa ou a detecção do máximo de solicitações inadequadas possíveis.

A técnica *MetaCost* conseguiu aumentar a faixa de Fatores Confiança para todas as bases de dados quando comparadas às faixas do experimento sem a utilização de técnicas para o tratamento do desbalanceamento. A sua utilização também diminuiu as diferenças entre TVP e TVN equilibrando os resultados sem alterar drasticamente, com exceção da Base de Dados VI, os valores máximos de TVN.

Tabela 9 – Fatores de confiança mínimos e máximos, por base de dados, para o classificador *Naive Bayes* - Sem Tratamento e SMOTE.

BD	Fator de Confiança	TVN	TVP	Diferença
		Sem Tratamento		
I	24,20%	50,38%	81,31%	74,70%
	98,90%	97,57%	25,38%	
II	99,60%	54,17%	89,32%	0,30%
	99,90%	66,67%	82,57%	
IV	98,70%	53,57%	75,15%	1,10%
	99,80%	92,88%	31,81%	
V	93,50%	50,26%	84,20%	6,40%
	99,90%	93,16%	31,11%	
VI	95,50%	50,82%	78,69%	4,30%
	99,80%	90,00%	27,13%	
VII	95,70%	50,14%	78,97%	3,30%
	99,00%	88,01%	27,81%	
		SMOTE		
I	11,00%	50,01%	82,22%	74,70%
	98,80%	97,52%	25,35%	
V	40,60%	50,00%	83,87%	59,30%
	99,90%	84,74%	40,15%	
VI	58,70%	50,82%	69,00%	41,20%
	99,90%	85,09%	27,26%	
VII	47,30%	50,14%	81,09%	40,00%
	87,30%	90,45%	25,11%	

5.2.3.1.4 *Random Forest*

As Figuras 29, 30, 31, 32, 33 e 34 ilustram o comportamento da TVN e da TVP causado pela alteração do fator de confiança e as Tabelas 11 e 12 exibem as faixas de valores para o fator de confiança nas quais os requisitos mínimos foram atendidos (as bases de dados que não apresentaram uma faixa de fatores de confiança dentro dos valores mínimos definidos não foram incluídas na tabela).

Assim como nos experimentos do algoritmo *Naive Bayes*, com o algoritmo *Random Forest* foi possível determinar uma faixa de fatores de confiança para todos as bases de dados sem a utilização de nenhuma técnica de tratamento para o desbalanceamento. Porém somente as Bases de Dados I, V e VII obtiveram valores para TVN maiores que 80% e faixas de fatores de confiança grandes o suficiente para permitir uma escolha gradativa entre autorizar automaticamente mais ou menos solicitações.

Para a Base de Dados I, a utilização da técnica SMOTE não apresentou melhorias significativas. Já para as Bases de Dados II, IV e VI houve um leve aumento da faixa de fatores de confiança, porém continuaram muito pequenas para para permitir uma escolha

Tabela 10 – Fatores de confiança mínimos e máximos, por base de dados, para o classificador *Naive Bayes* - *Random Oversampling* e *MetaCost*.

BD	Fator de Confiança	TVN	TVP	Diferença
		Random Oversampling		
I	10,10%	50,58%	81,46%	88,20%
	98,30%	97,75%	25,72%	
II	99,70%	51,67%	84,54%	0,20%
	99,90%	60,00%	81,41%	
IV	27,90%	50,24%	80,79%	71,70%
	99,60%	89,78%	25,98%	
V	11,00%	50,00%	88,41%	88,90%
	99,90%	93,68%	27,05%	
VI	17,90%	50,00%	81,42%	81,20%
	99,10%	92,09%	25,50%	
VII	41,20%	50,03%	80,32%	38,00%
	79,20%	91,56%	25,18%	
		MetaCost		
I	0,10%	70,09%	67,46%	99,80%
	99,90%	87,42%	43,97%	
II	24,80%	51,67%	85,79%	75,10%
	99,90%	66,67%	72,98%	
IV	65,00%	50,22%	72,59%	34,90%
	99,90%	88,03%	37,19%	
V	10,40%	50,00%	78,07%	89,50%
	99,90%	93,16%	32,26%	
VI	0,10%	55,73%	78,11%	99,80%
	99,90%	62,45%	67,72%	
VII	72,20%	50,14%	78,19%	27,70%
	99,90%	77,90%	39,60%	

gradativa entre autorizar automaticamente mais ou menos solicitações. Para as Bases de Dados V e VII, entretanto, essa melhora na faixa de fatores de confiança foi substancial apesar de não melhorar significativamente os valores máximos da TVN.

A utilização do *Random Oversampling* apresentou resultados semelhantes a utilização da técnicas SMOTE em relação aos valores máximos e mínimos das TVN e TVP porém, para as Bases de Dados I, V e VII o aumento nas faixas de fatores de confiança foi superior permitindo uma transição mais gradual entre as situações extremas.

A técnica *MetaCost* conseguiu aumentar a faixa de Fatores Confiança para todas as bases de dados. Para as Bases de Dados V e VII este aumento foi substancial; a Base de Dados I apresentou um leve aumento e o mesmo foi identificado nas Bases de Dados II, IV e VI, mas as faixas dessas bases continuou muito curta para permitir a transição gradual entre as situações extremas. Em relação aos valores máximos das TVN e TVP, não houve grandes alterações nas bases de dados menos desbalanceadas (Bases de Dados

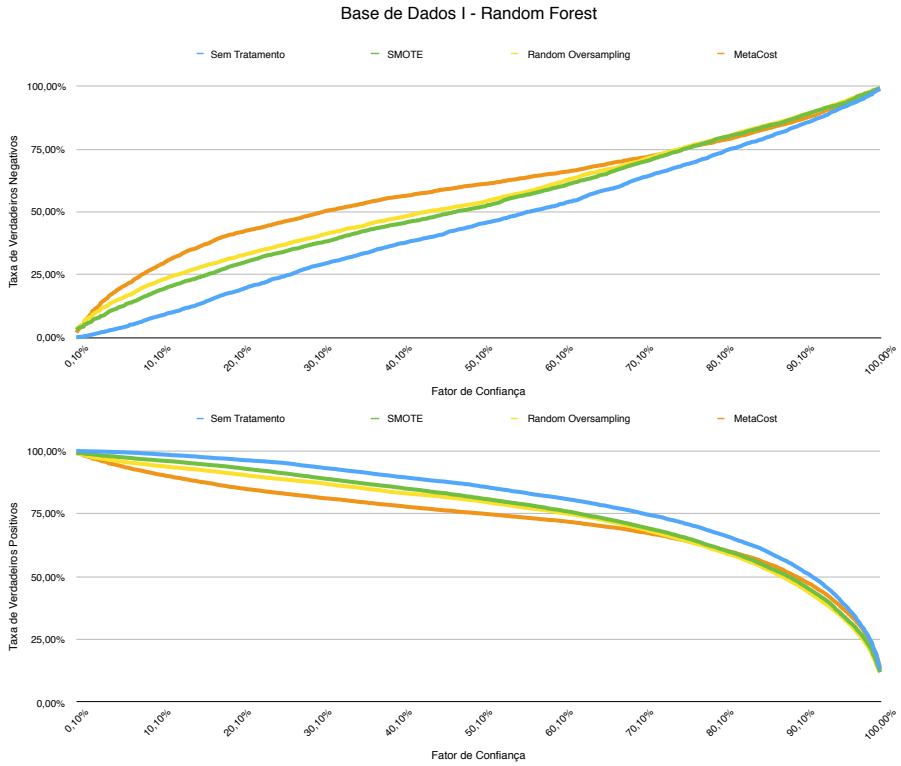

Figura 29 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados I com o algoritmo *Random Forest*.

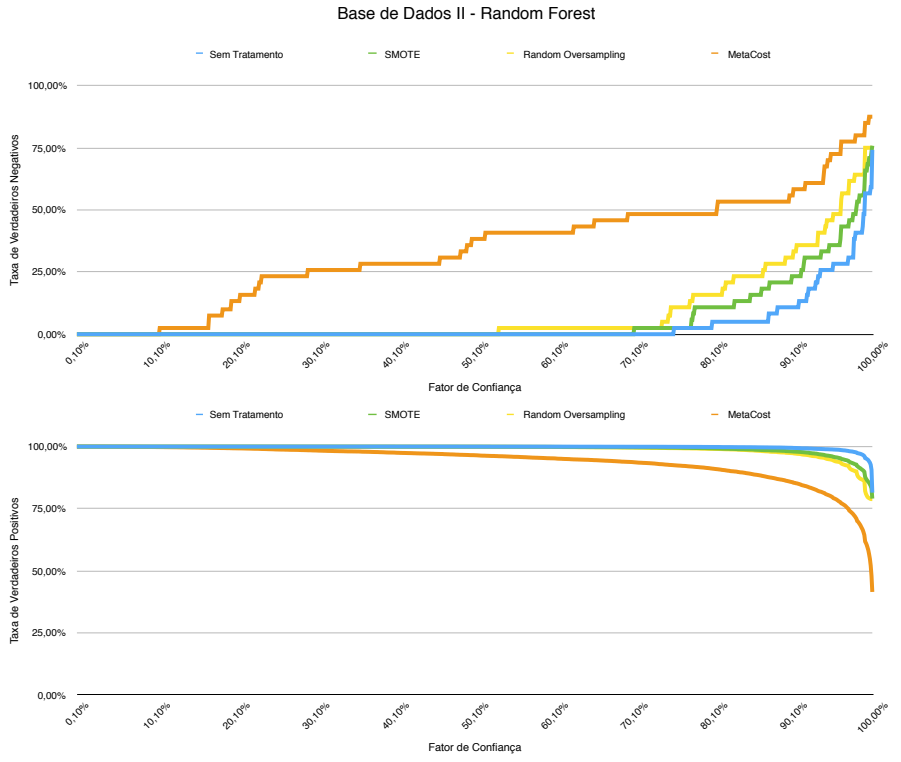

Figura 30 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados II com o algoritmo *Random Forest*.

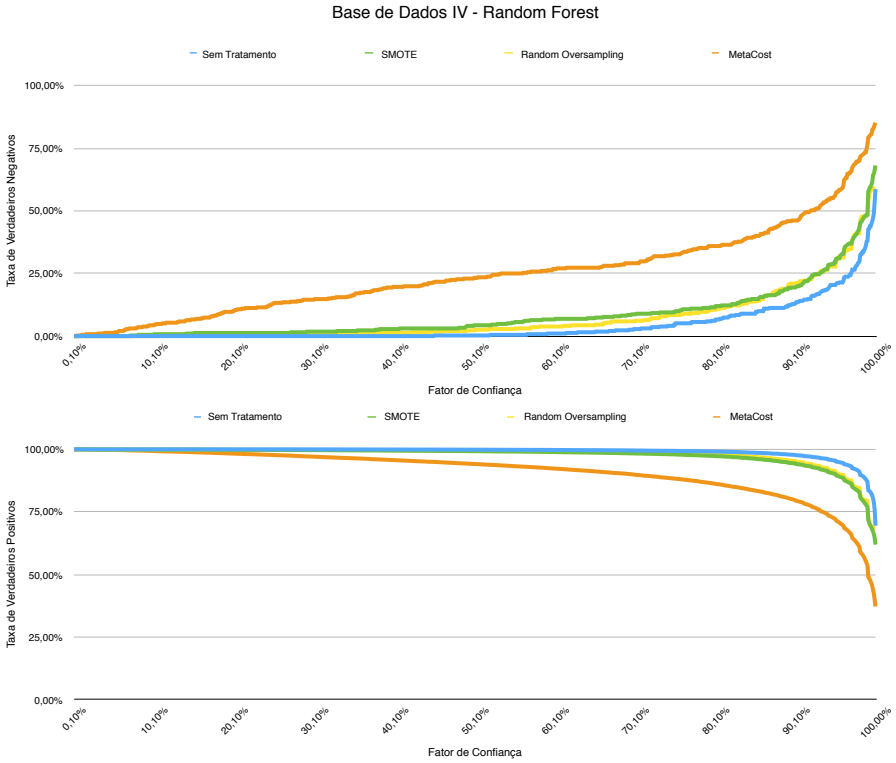

Figura 31 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados IV com o algoritmo *Random Forest*.

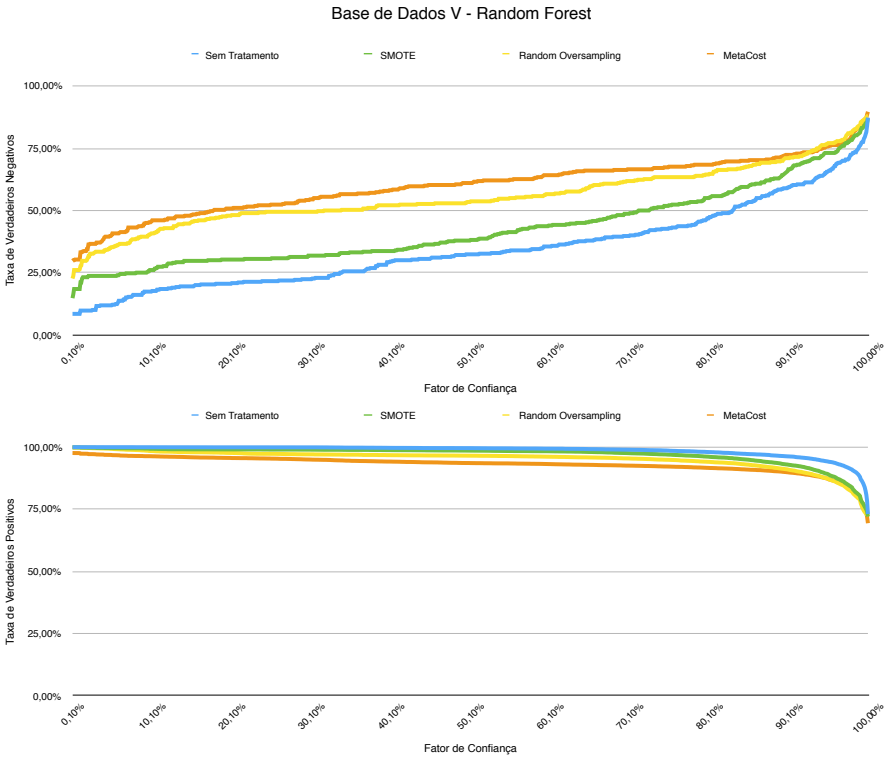

Figura 32 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados V com o algoritmo *Random Forest*.

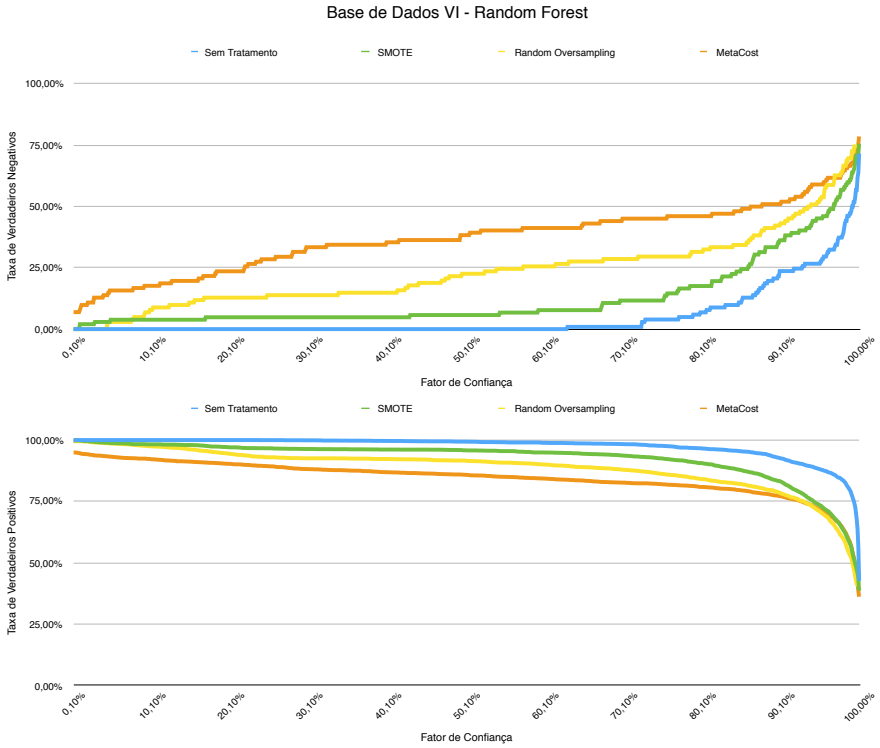

Figura 33 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados VI com o algoritmo *Random Forest*.

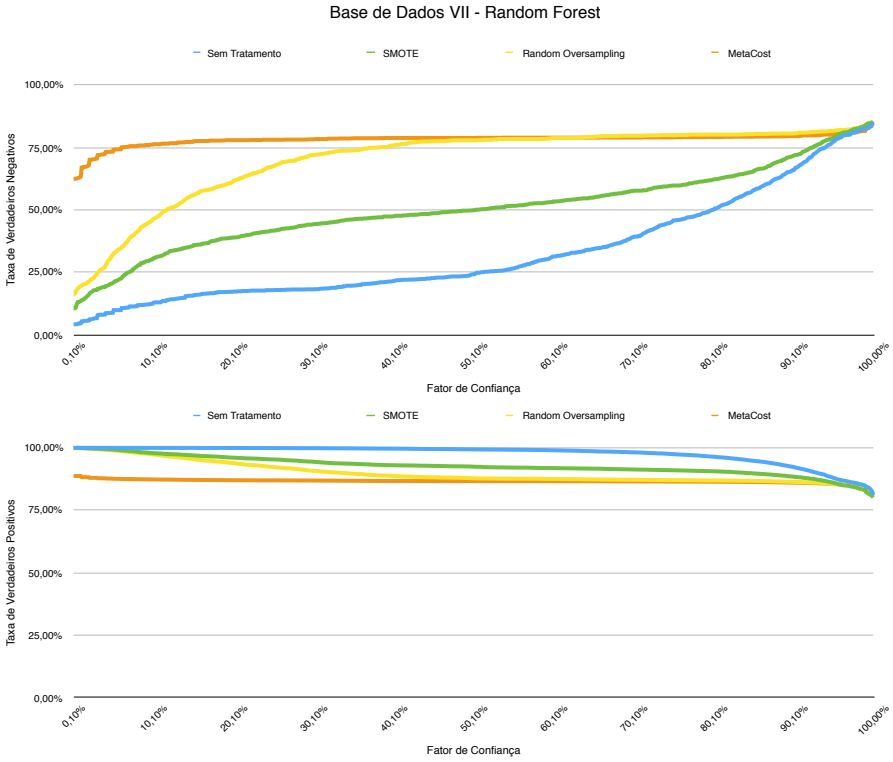

Figura 34 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados VII com o algoritmo *NaiveBayes*.

Tabela 11 – Fatores de confiança mínimos e máximos, por base de dados, para o classificador *Random Forest* - Sem Tratamento e SMOTE.

BD	Fator de Confiança	TVN	TVP	Diferença
		Sem Tratamento		
I	56,20% 98,70%	50,06% 96,53%	83,17% 25,23%	42,50%
II	99,10% 100,00%	56,67% 74,17%	95,37% 81,46%	0,90%
IV	99,90% 100,00%	53,85% 58,69%	75,68% 69,51%	0,10%
V	83,00% 100,00%	50,00% 87,37%	97,60% 73,10%	17,00%
VI	99,30% 100,00%	51,73% 71,55%	75,40% 42,67%	0,70%
VII	80,00% 100,00%	50,08% 84,84%	96,46% 81,22%	20,00%
		SMOTE		
I	47,30% 98,10%	50,09% 97,10%	82,58% 25,07%	50,80%
II	98,00% 100,00%	50,83% 75,83%	92,72% 79,07%	2,00%
IV	99,10% 100,00%	57,89% 68,13%	71,84% 61,96%	0,90%
V	71,20% 100,00%	50,00% 87,11%	97,47% 72,08%	28,80%
VI	96,80% 100,00%	50,82% 75,55%	68,48% 38,66%	3,20%
VII	50,40% 100,00%	50,08% 85,40%	92,40% 80,14%	49,60%

I, V e VII) enquanto nas bases de dados mais desbalanceadas (Bases de Dados II, VI e VI) houve um leve aumento da TVN e uma diminuição da TVP.

5.2.3.1.5 SVM

As Figuras 35, 36, 37, 38, 39 e 40 ilustram o comportamento da TVN e da TVP causado pela alteração do fator de confiança e a Tabela 13 exibe as faixas de valores para o fator de confiança nas quais os requisitos mínimos foram atendidos (as bases de dados que não apresentaram uma faixa de fatores de confiança dentro dos valores mínimos definidos não foram incluídas na tabela).

Diferentemente dos outros algoritmos, o classificador SVM sem a utilização de técnicas para o tratamento de desbalanceamento não produziu uma faixa de fatores de confiança dentre os valores mínimos estipulados para as métricas de TVN e TVP na Base

Tabela 12 – Fatores de confiança mínimos e máximos, por base de dados, para o classificador *Random Forest* - *Random Oversampling* e *MetaCost*.

BD	Fator de Confiança	TVN	TVP	Diferença
		Random Oversampling		
I	43,70%	50,04%	82,38%	54,10%
	97,80%	96,71%	25,26%	
II	96,10%	54,17%	93,21%	3,90%
	100,00%	75,00%	78,87%	
IV	99,10%	54,35%	72,93%	0,90%
	100,00%	58,94%	68,67%	
V	31,70%	50,00%	97,05%	68,30%
	100,00%	87,89%	72,83%	
VI	93,90%	50,82%	73,29%	6,10%
	100,00%	74,45%	40,89%	
VII	11,80%	50,08%	96,65%	88,20%
	100,00%	84,79%	81,25%	
		MetaCost		
I	30,80%	50,09%	81,26%	67,80%
	98,60%	96,83%	25,33%	
II	80,50%	50,83%	90,90%	19,50%
	100,00%	87,50%	41,52%	
IV	92,00%	50,24%	77,35%	8,00%
	100,00%	85,21%	37,26%	
V	17,80%	50,00%	95,79%	82,20%
	100,00%	89,74%	69,41%	
VI	87,70%	50,91%	78,24%	12,30%
	100,00%	78,45%	36,19%	
VII	0,10%	62,47%	88,73%	99,90%
	100,00%	83,79%	80,76%	

de Dados I. Os únicos resultados dentro dos valores mínimos foram para as Bases de Dados II e IV, porém as faixas encontradas são muito curtas para serem considerados bons resultados.

A utilização da técnica SMOTE melhorou os resultados para a Base de Dados VI, aumentando a TVP máxima e a faixa de fatores de confiança. Com exceção das Bases de Dados II e IV, nas quais não foi possível produzir uma faixa de fatores de confiança, as todas as bases de dados apresentaram valores semelhantes para as métricas discutidas mas as faixas de fatores de confiança nas Bases de Dados VI e VII foram mais curtas.

A utilização do *Random Oversampling* obteve resultados similares ao da técnica SMOTE, apresentando uma fraca melhora no valor máximo da TVN para as Bases de Dados I, V e VI, bem como um aumento da faixa de valores de fatores de confiança para a Base de Dados VI.

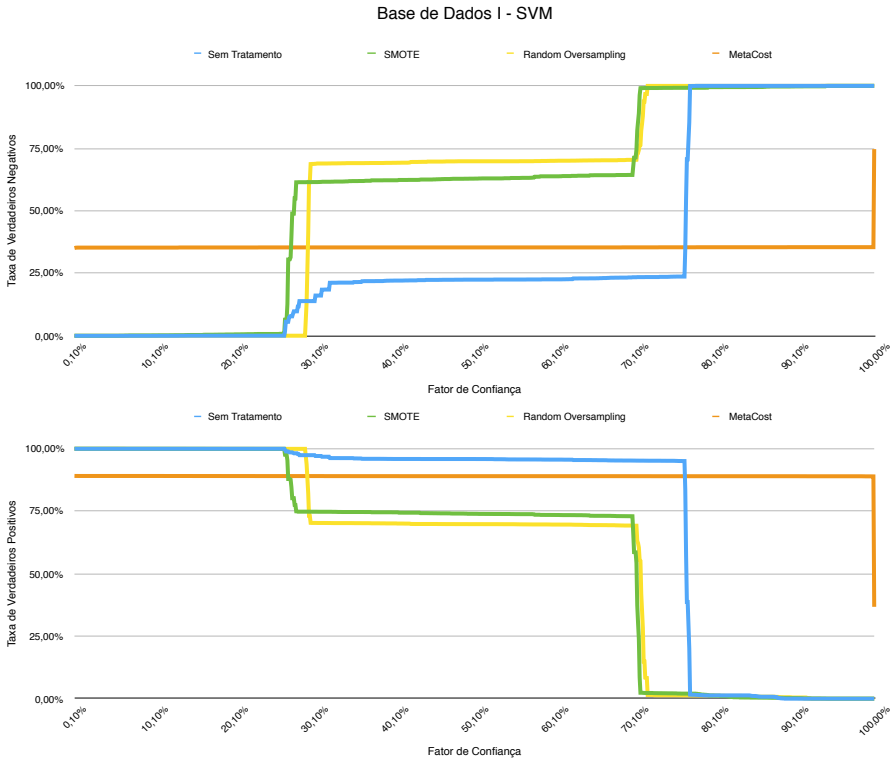

Figura 35 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados I com o algoritmo SVM.

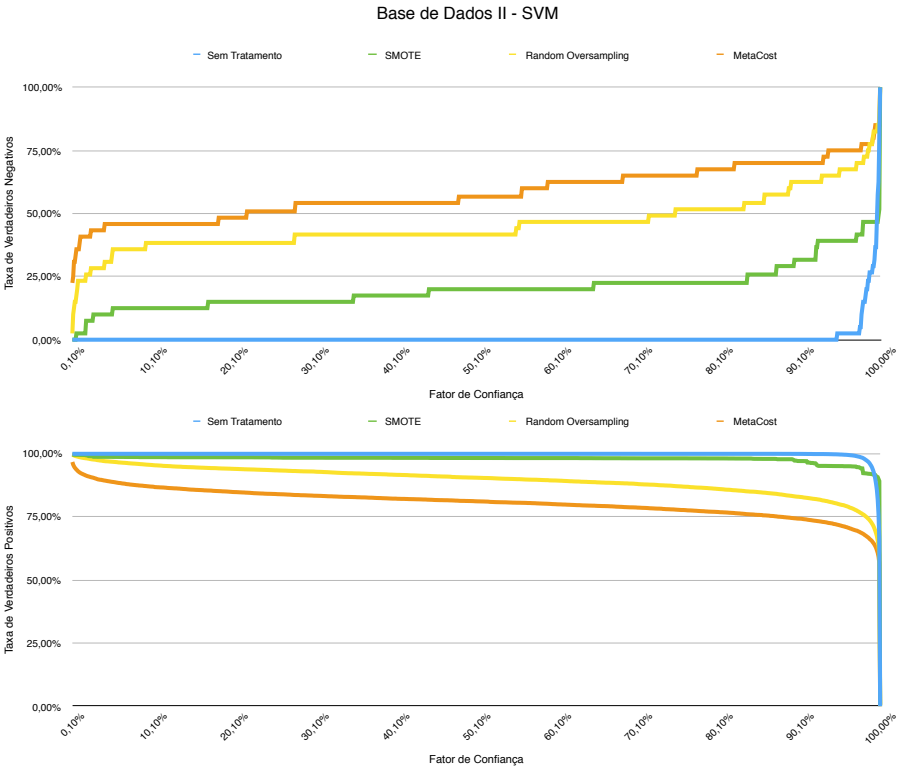

Figura 36 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados II com o algoritmo SVM.

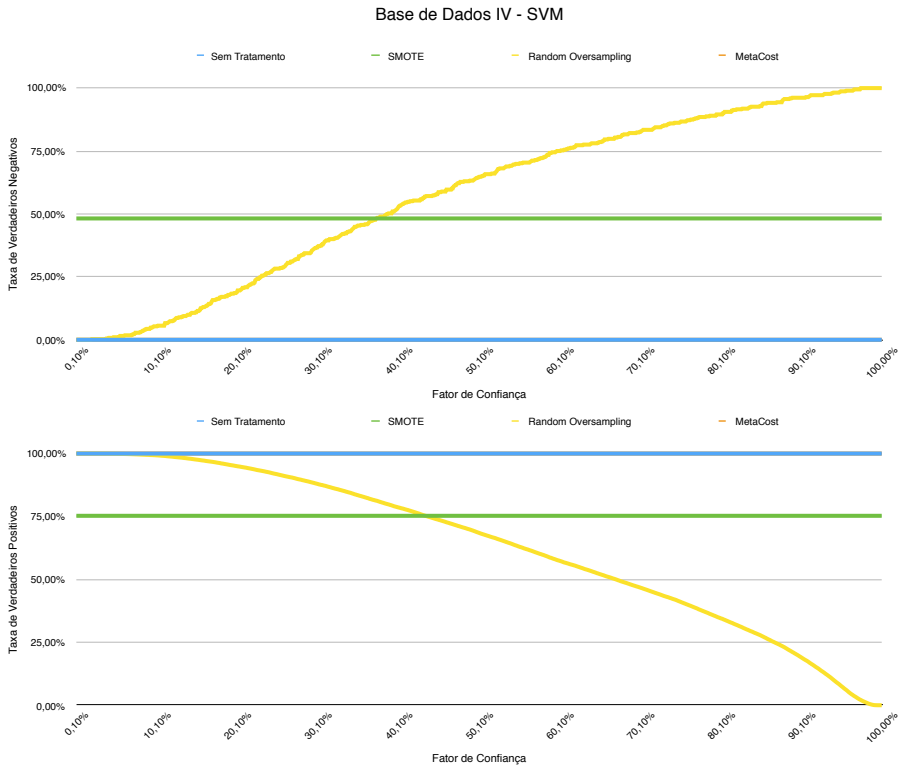

Figura 37 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados IV com o algoritmo SVM.

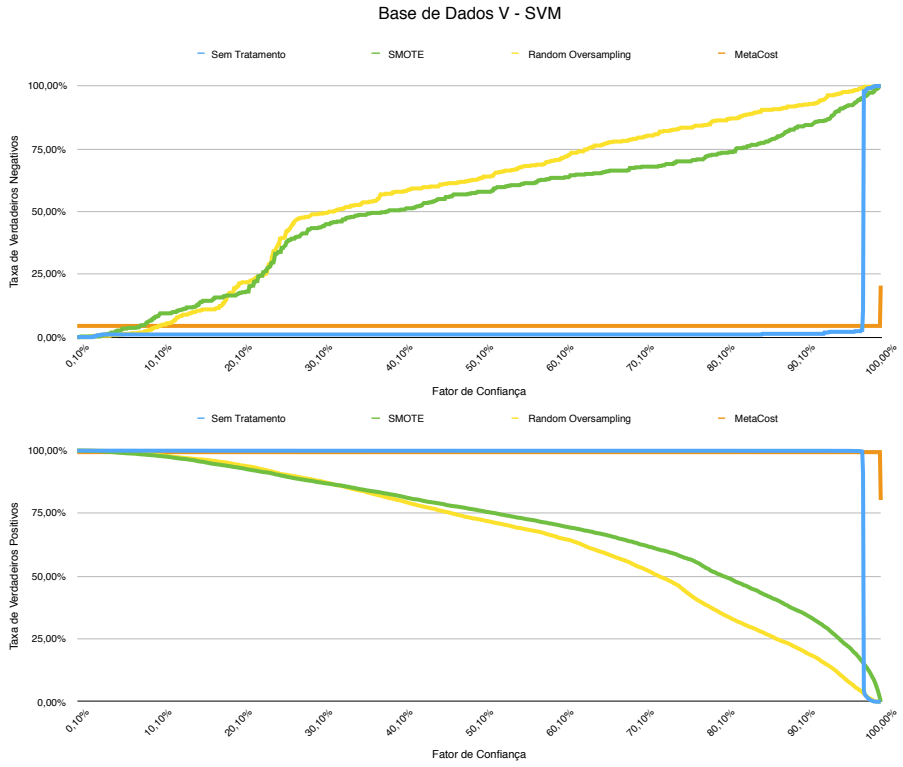

Figura 38 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados V com o algoritmo SVM.

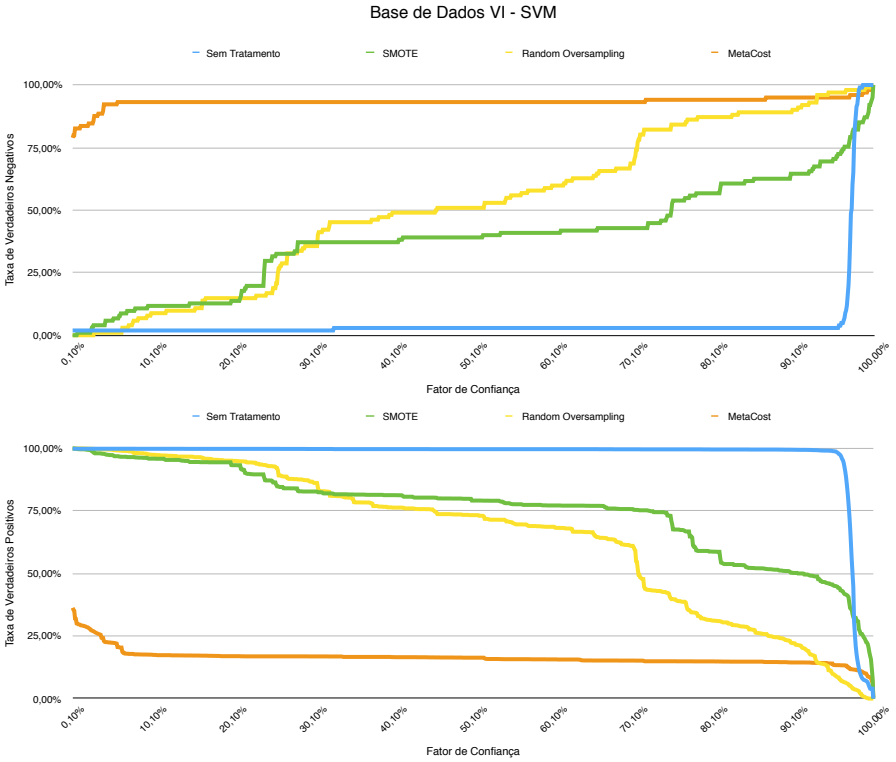

Figura 39 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados VI com o algoritmo SVM.

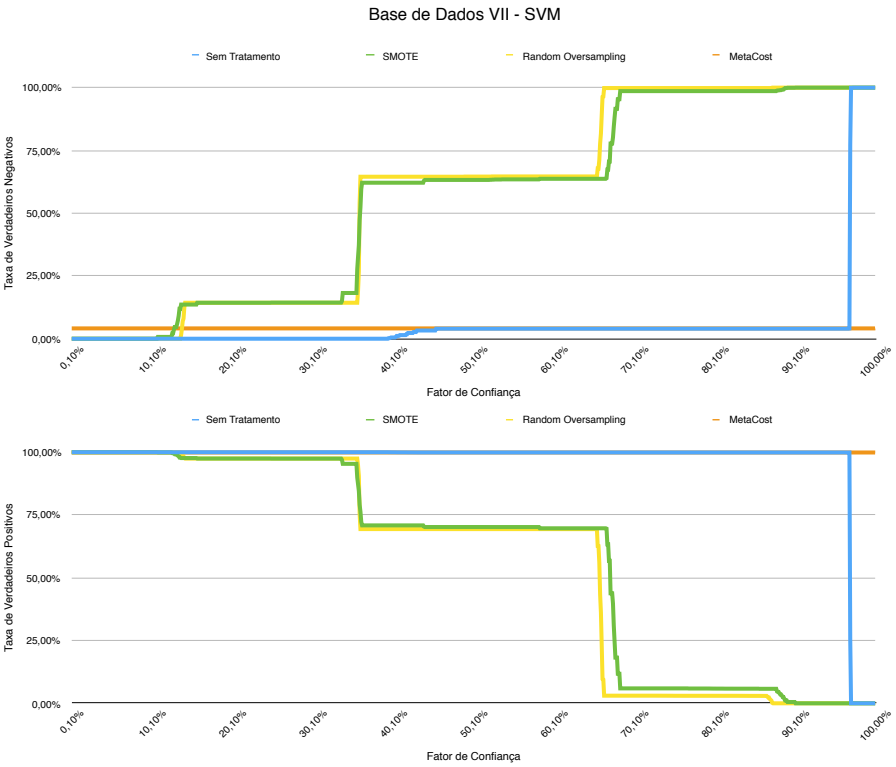

Figura 40 – Comportamento da TVN e da TVP causado pela alteração do fator de confiança na Base de Dados VII com o algoritmo SVM.

Tabela 13 – Fatores de confiança mínimos e máximos, por base de dados, para o classificador SVM.

BD	Fator de Confiança	TVN	TVP	Diferença
		Sem Tratamento		
II	99,70%	57,50%	81,85%	0,20%
	99,90%	80,00%	66,16%	
VI	97,30%	50,64%	55,69%	0,40%
	97,70%	83,55%	25,53%	
		SMOTE		
I	27,60%	54,99%	77,54%	42,90%
	70,50%	85,64%	30,34%	
V	38,50%	50,00%	82,94%	56,30%
	94,80%	90,00%	25,22%	
VI	74,90%	52,82%	69,70%	23,80%
	98,70%	85,09%	25,18%	
VII	36,00%	52,35%	76,29%	31,50%
	67,50%	84,50%	30,89%	
		Random Oversampling		
I	29,40%	61,60%	73,14%	41,70%
	71,10%	87,92%	28,34%	
II	74,70%	51,67%	87,19%	25,20%
	99,90%	87,50%	59,71%	
IV	38,80%	50,22%	79,81%	47,80%
	86,60%	94,13%	25,15%	
V	31,40%	50,00%	87,03%	55,60%
	87,00%	90,79%	25,09%	
VI	45,60%	50,91%	73,74%	41,30%
	86,90%	89,09%	25,02%	
VII	35,90%	54,22%	75,28%	30,00%
	65,90%	81,69%	36,14%	
		MetaCost		
II	21,70%	50,83%	84,61%	78,20%
	99,90%	85,00%	56,67%	
VI	0,10%	78,91%	36,28%	3,50%
	3,60%	88,55%	25,51%	

Assim como nos experimentos sem tratamento, a Base de Dados I também não produziu resultados com a técnica *MetaCost*, porém para a Base de Dados II houve um aumento significativo no tamanho da faixa de valores de fatores de confiança disponível.

5.2.3.1.6 Considerações

Como esperado, os melhores resultados foram obtidos ao se trabalhar com a base de dados menos desbalanceada, Base de Dados I, na qual, com exceção do algoritmo *Ripper*,

foi possível definir uma faixa de fatores de confiança grande o suficiente para permitir a configuração gradual entre autorizar mais ou menos solicitações automaticamente. Os algoritmos C4.5, *Naive Bayes* e *Random Forest* obtiveram valores semelhantes para as métricas de TVN e TVP tanto para os seus respectivos fatores de confiança mínimos (TVN ≈ 50% e TVP ≈ 80%) e máximos (TVN ≈ 90% e TVP ≈ 25%) tanto para os experimentos sem tratamento de desbalanceamento quanto para os experimentos com as técnicas de tratamento para o desbalanceamento.

Para as bases de dados mais desbalanceadas, Bases de Dados II e IV, os algoritmos C4.5 e *Ripper* não obtiveram resultados relevantes. A utilização dos algoritmos *Naive Bayes* e *Random Forest* em conjunto com as técnicas de *Random Oversampling* e *Meta Cost* permitiu uma melhora nos resultados dessa base, mas seguiram o mesmo padrão. Para a Base de Dados II, o algoritmo *Random Forest* combinado com a técnica *Meta Cost* atingiu valores bons para os fatores de confiança mínimo (TVN ≈ 50% e TVP ≈ 90%) e máximo (TVN ≈ 90% e TVP ≈ 40%), mas a diferença entre eles não foi muito grande (19,5%). Para a Base de Dados IV, o algoritmo *Naive Bayes* combinado com as técnica *Random Oversampling* e *Meta Cost* atingiu valores para as métricas de TVN e TVP semelhantes aos da Base de Dados I, tanto para o fator de confiança mínimo (TVN ≈ 50% e TVP ≈ 80%) quanto para o máximo (TVN ≈ 90% e TVP ≈ 25%) com uma boa diferença entre eles (71,70% para *Random Oversampling* e 34,90% para *Meta Cost*).

Apesar de apresentarem um alto grau de desbalanceamento as Bases de Dados V, VI e VII, quando experimentadas com os algoritmos *Naive Bayes* e *Random Forest*, obtiveram resultados semelhantes aos da Base de Dados I (TVP ≈ 50% e TVP ≈ 80% para o fator de confiança mínimo e TVP ≈ 90% e TVP ≈ 25% para o máximo). Porém, quando não submetidas a nenhuma técnicas de tratamento de desbalanceamento, não produziram faixas de fatores de confiança grandes o suficiente para permitir uma escolha gradativa entre autorizar automaticamente mais ou menos solicitações.

5.2.4 Avaliação do impacto

O resultado final é analisado segundo o impacto na redução da carga de trabalho e do desperdício assistencial no processo tradicional de regulação. Esse impacto pode ser avaliado por meio de duas métricas: valor preditivo positivo (5.1) e valor preditivo negativo (5.2). Essas duas métricas indicam o percentual de acerto nas classificações de solicitações adequadas e inadequadas, respectivamente.

$$\text{Valor Preditivo Positivo} = \frac{\text{Solicitações Adequadas Autorizadas Automaticamente}}{\text{Total de Solicitações Autorizadas Automaticamente}} \times 100 \tag{5.1}$$

$$\text{Valor Preditivo Negativo} = \frac{\text{Solicitações Inadequadas Encaminhadas para Regulação}}{\text{Total de Solicitações Encaminhadas para Regulação}} \times 100 \quad (5.2)$$

Para facilitar a discussão será tomado como exemplo um universo de 10.000 solicitações, balanceadas de acordo com as características iniciais de cada base de dados.

5.2.4.1 Base de Dados I

A Tabela 14 exibe os melhores fatores de confiança para a combinação entre a Base de Dados I, a técnica *Random Oversampling* e o algoritmo *Naive Bayes*.

Tabela 14 – Fatores de confiança para a combinação entre a Base de Dados I, a técnica *Random Oversampling* e o algoritmo *Naive Bayes*.

Fator de Confiança	VPN	VPP	TVN	TVP
10,10%	51,30%	81,10%	50,58%	81,46%
20,00%	46,46%	85,64%	69,87%	69,05%
30,00%	46,08%	86,12%	71,55%	67,82%
40,00%	45,33%	86,58%	73,36%	65,98%
50,00%	42,47%	88,31%	79,80%	58,42%
60,00%	40,25%	89,24%	83,59%	52,24%
70,00%	38,32%	90,06%	86,66%	46,34%
80,00%	36,70%	90,81%	89,25%	40,80%
90,00%	35,36%	93,53%	93,91%	33,99%
98,30%	33,59%	96,81%	97,75%	25,72%

A partir do fator de confiança de 10,10% (Figura 41) 81,46% das solicitações adequadas (cerca de 5.885 solicitações) seriam autorizadas automaticamente porém 49,42% das solicitações inadequadas (cerca de 1.372 solicitações) também o seriam, isso significa uma redução de cerca de 72% no trabalho manual da regulação, porém com um erro de 18,90% nas autorizações automáticas. No outro extremo, com um fator de confiança de 98,30% (Figura 42), 97,75% das solicitações inadequadas (cerca de 2.714 solicitações) seriam encaminhadas para a regulação juntamente com 74,28% das solicitações adequadas (cerca de 5.366 solicitações), isso significa uma redução de 19% no trabalho manual de regulação com um erro inferior a 5,00% nas autorizações automáticas.

5.2.4.2 Base de Dados II

A Tabela 15 exibe os melhores fatores de confiança para a combinação entre a Base de Dados II, a técnica *MetaCost* e o algoritmo *Random Forest*.

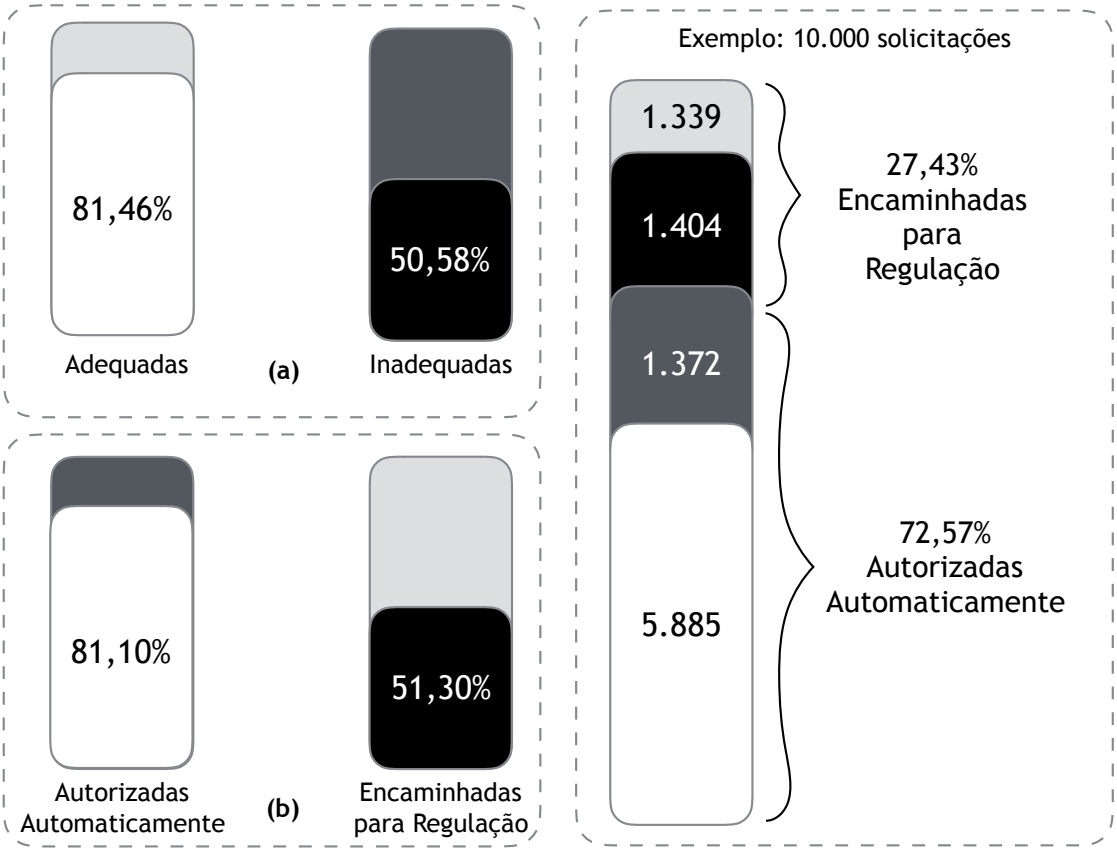

Figura 41 – Fator de Confiança de 10,10% para a combinação entre a Base de Dados I, a técnica *Random Oversampling* e o algoritmo *Naive Bayes*.

Tabela 15 – Fatores de confiança para a combinação entre a Base de Dados II, a técnica *MetaCost* e o algoritmo *Random Forest*.

Fator de Confiança	VPN	VPP	TVN	TVP
10,00%	0,00%	99,76%	0,00%	99,81%
20,00%	4,11%	99,79%	13,33%	99,30%
30,00%	3,70%	99,82%	25,83%	98,43%
40,00%	2,70%	99,83%	28,33%	97,57%
50,00%	2,65%	99,85%	38,33%	96,53%
60,00%	2,02%	99,85%	40,83%	95,25%
70,00%	1,82%	99,87%	48,33%	93,72%
80,00%	1,29%	99,87%	48,33%	91,10%
80,50%	1,33%	99,87%	50,83%	90,90%
90,00%	0,93%	99,88%	55,83%	85,65%
100,00%	0,35%	99,93%	87,50%	41,52%

A partir do fator de confiança de 80,50% (Figura 43) 90,90% das solicitações adequadas (cerca de 9.068 solicitações) seriam autorizadas automaticamente porém 49,17% das solicitações inadequadas (cerca de 12 solicitações) também o seriam, isso significa uma redução de cerca de 90,80% no trabalho manual da regulação, com um erro inferior a 1,00% nas autorizações automáticas. No outro extremo, com um fator de confiança de

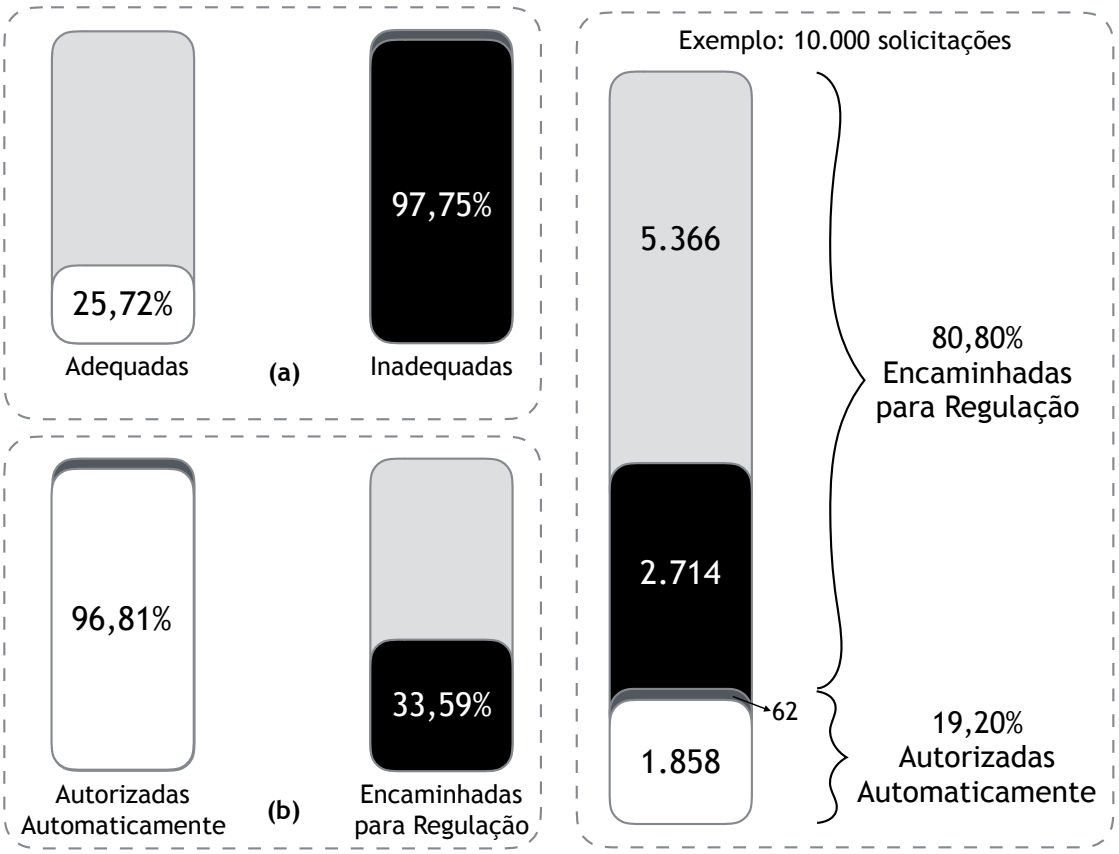

Figura 42 – Fator de Confiança de 98,30% para a combinação entre a Base de Dados I, a técnica *Random Oversampling* e o algoritmo *Naive Bayes*.

100% (Figura 44) 87,50% das solicitações inadequadas (cerca de 21 solicitações) seriam encaminhadas para a regulação juntamente com 58,48% das solicitações adequadas (cerca de 5.834 solicitações), isso significa uma redução de 41,45% no trabalho manual de regulação com um erro inferior a 1,00% nas autorizações automáticas.

5.2.4.3 Base de Dados IV

A Tabela 16 exibe os melhores fatores de confiança para a combinação entre a Base de Dados IV, a técnica *Random Oversampling* e o algoritmo *SVM*.

A partir do fator de confiança de 38,80% (Figura 45) 79,81% das solicitações adequadas (cerca de 7.912 solicitações) seriam autorizadas automaticamente porém 49,78% das solicitações inadequadas (cerca de 43 solicitações) também o seriam, isso significa uma redução de cerca de 79,55% no trabalho manual da regulação, com um erro inferior a 1,00% nas autorizações automáticas. No outro extremo, com um fator de confiança de 86,60% (Figura 46), 94,13% das solicitações inadequadas (cerca de 82 solicitações) seriam encaminhadas para a regulação juntamente com 74,85% das solicitações adequadas (cerca de 7.420 solicitações), isso significa uma redução de 24,98% no trabalho manual de regulação com um erro inferior a 1,00% nas autorizações automáticas.

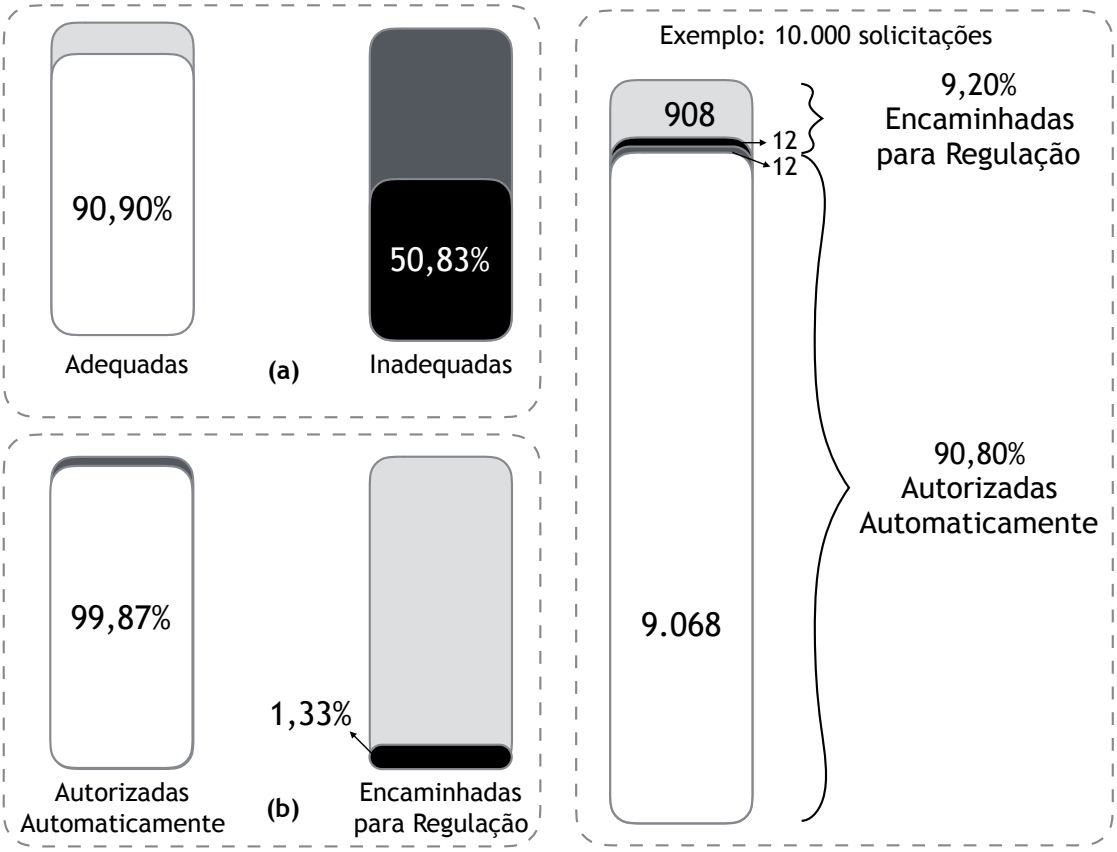

Figura 43 – Fator de Confiança de 80,50% para a combinação entre a Base de Dados II, a técnica *MetaCost* e o algoritmo *Random Forest*.

Tabela 16 – Fatores de confiança para a combinação entre a Base de Dados IV, a técnica *Random Oversampling* e o algoritmo *SVM*.

Fator de Confiança	VPN	VPP	TVN	TVP
10,00%	6,27%	99,17%	5,35%	99,29%
20,00%	3,19%	99,26%	18,84%	95,00%
30,00%	2,61%	99,37%	36,71%	87,94%
38,80%	2,13%	99,46%	50,22%	79,81%
40,00%	2,11%	99,47%	52,51%	78,63%
50,00%	1,77%	99,55%	64,52%	68,55%
60,00%	1,52%	99,62%	75,00%	57,43%
70,00%	1,35%	99,68%	82,65%	46,80%
80,00%	1,19%	99,74%	89,54%	34,58%
86,60%	1,09%	99,80%	94,13%	25,15%
90,00%	1,03%	99,82%	96,17%	19,05%
100,00%	0,87%	0,00%	100,00%	0,00%

5.2.4.4 Base de Dados V

A Tabela 17 exibe os melhores fatores de confiança para a combinação entre a Base de Dados V, a técnica *Random Oversampling* e o algoritmo *Naive Bayes*.

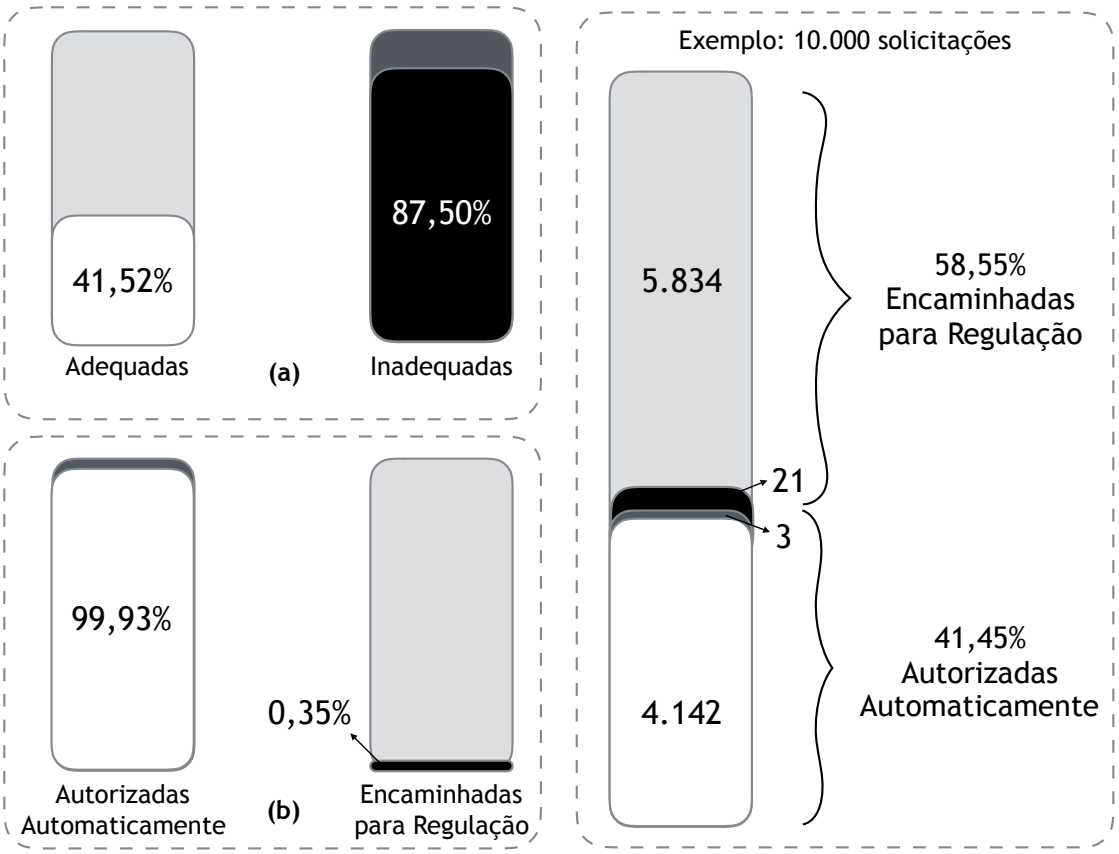

Figura 44 – Fator de Confiança de 100,00% para a combinação entre a Base de Dados II, a técnica *MetaCost* e o algoritmo *Random Forest*.

Tabela 17 – Fatores de confiança para a combinação entre a Base de Dados V, a técnica *Random Oversampling* e o algoritmo *Naive Bayes*.

Fator de Confiança	VPN	VPP	TVN	TVP
10,00%	9,56%	98,69%	48,16%	89,47%
11,00%	9,06%	98,72%	50,00%	88,41%
20,00%	6,59%	98,92%	62,11%	79,73%
30,00%	5,21%	99,03%	69,74%	70,73%
40,00%	4,46%	99,07%	74,21%	63,43%
50,00%	3,97%	99,07%	76,58%	57,45%
60,00%	3,64%	99,08%	78,68%	52,23%
70,00%	3,46%	99,10%	80,79%	48,28%
80,00%	3,35%	99,13%	82,37%	45,46%
90,00%	3,30%	99,19%	84,47%	43,22%
99,90%	2,86%	99,50%	93,68%	27,05%
100,00%	2,24%	0,00%	100,00%	0,00%

A partir do fator de confiança de 11,00% (Figura 47) 88,41% das solicitações adequadas (cerca de 8.643 solicitações) seriam autorizadas automaticamente porém 50,00% das solicitações inadequadas (cerca de 112 solicitações) também o seriam, isso significa uma redução de cerca de 86,25% no trabalho manual da regulação, com um erro inferior

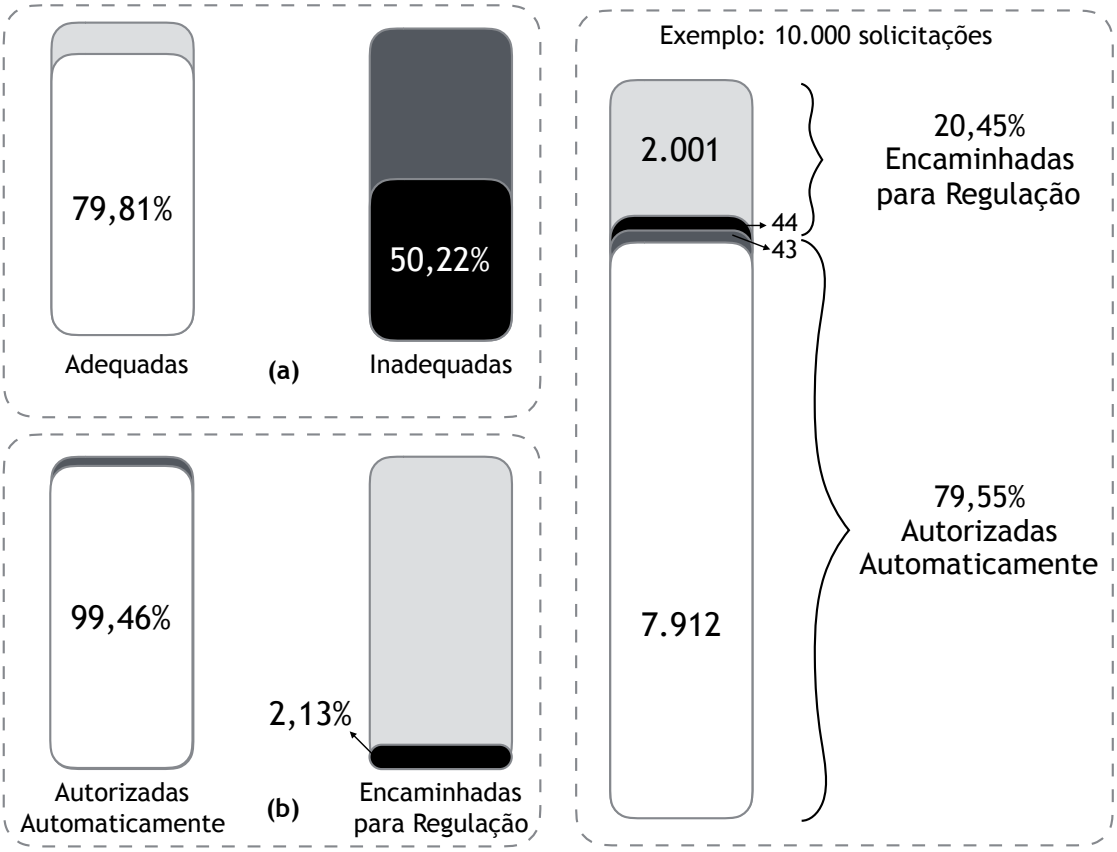

Figura 45 – Fator de Confiança de 38,80% para a combinação entre a Base de Dados IV, a técnica *Random Oversampling* e o algoritmo *SVM*.

a 2,00% nas autorizações automáticas. No outro extremo, com um fator de confiança de 99,90% (Figura 48), 93,68% das solicitações inadequadas (cerca de 163 solicitações) seriam encaminhadas para a regulação juntamente com 72,95% das solicitações adequadas (cerca de 7.132 solicitações), isso significa uma redução de 27,05% no trabalho manual de regulação com um erro inferior a 3,00% nas autorizações automáticas.

5.2.4.5 Base de Dados VI

A Tabela 18 exibe os melhores fatores de confiança para a combinação entre a Base de Dados VI, a técnica *Random Oversampling* e o algoritmo *Naive Bayes*.

A partir do fator de confiança de 17,90% (Figura 49) 81,42% das solicitações adequadas (cerca de 7.938 solicitações) seriam autorizadas automaticamente porém 50,00% das solicitações inadequadas (cerca de 125 solicitações) também o seriam, isso significa uma redução de cerca de 80,63% no trabalho manual da regulação, com um erro inferior a 2,00% nas autorizações automáticas. No outro extremo, com um fator de confiança de 99,10% (Figura 50) 92,09% das solicitações inadequadas (cerca de 230 solicitações) seriam encaminhadas para a regulação juntamente com 74,50% das solicitações adequadas (cerca de 7.264 solicitações), isso significa uma redução de 25,06% no trabalho manual de

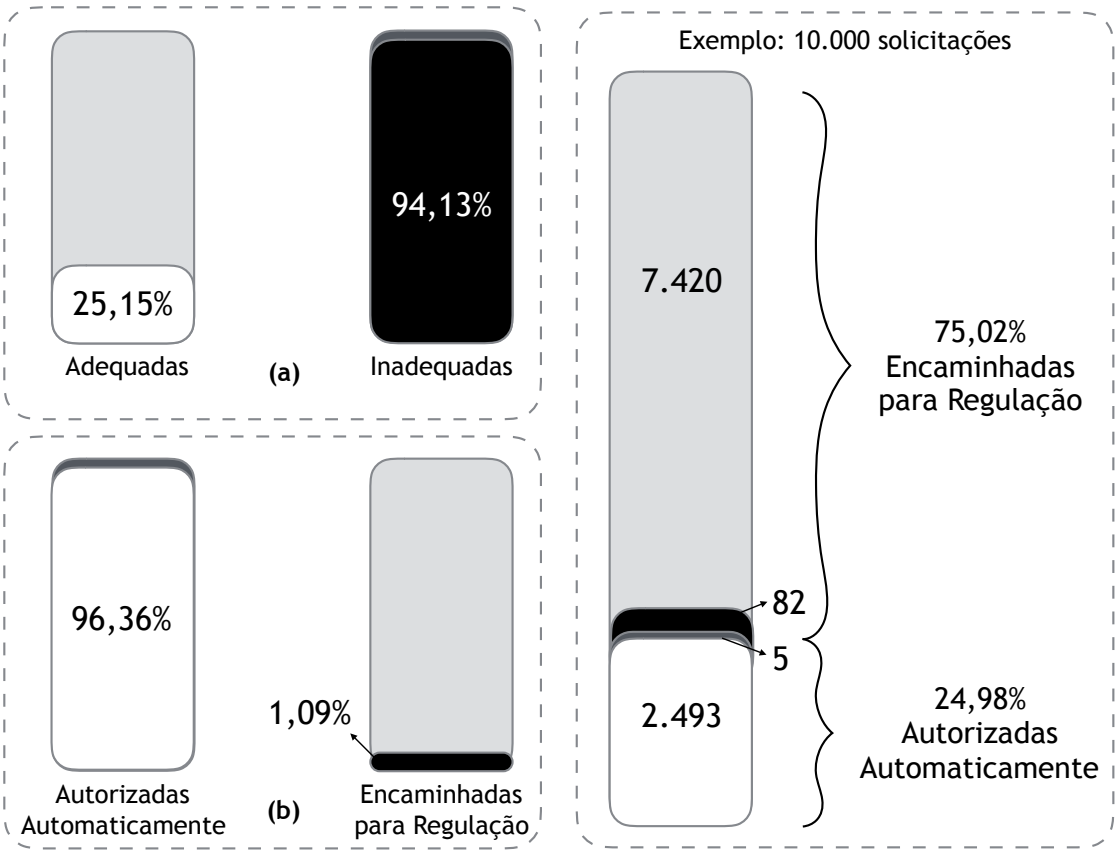

Figura 46 – Fator de Confiança de 86,60% para a combinação entre a Base de Dados IV, a técnica *Random Oversampling* e o algoritmo *SVM*.

Tabela 18 – Fatores de confiança para a combinação entre a Base de Dados VI, a técnica *Random Oversampling* e o algoritmo *Naive Bayes*.

Fator de Confiança	VPN	VPP	TVN	TVP
10,00%	6,65%	98,18%	37,45%	86,55%
17,90%	6,48%	98,46%	50,00%	81,42%
20,00%	6,52%	98,50%	52,00%	80,77%
30,00%	6,46%	98,60%	55,82%	79,22%
40,00%	6,42%	98,69%	59,64%	77,46%
50,00%	5,46%	98,72%	63,55%	71,56%
60,00%	4,42%	98,90%	74,27%	58,36%
70,00%	3,84%	99,07%	83,18%	46,26%
80,00%	3,44%	99,12%	87,09%	37,12%
90,00%	3,33%	99,17%	89,09%	33,61%
99,10%	3,07%	99,26%	92,09%	25,50%
100,00%	2,50%	0,00%	100,00%	0,00%

regulação com um erro inferior a 1,00% nas autorizações automáticas.

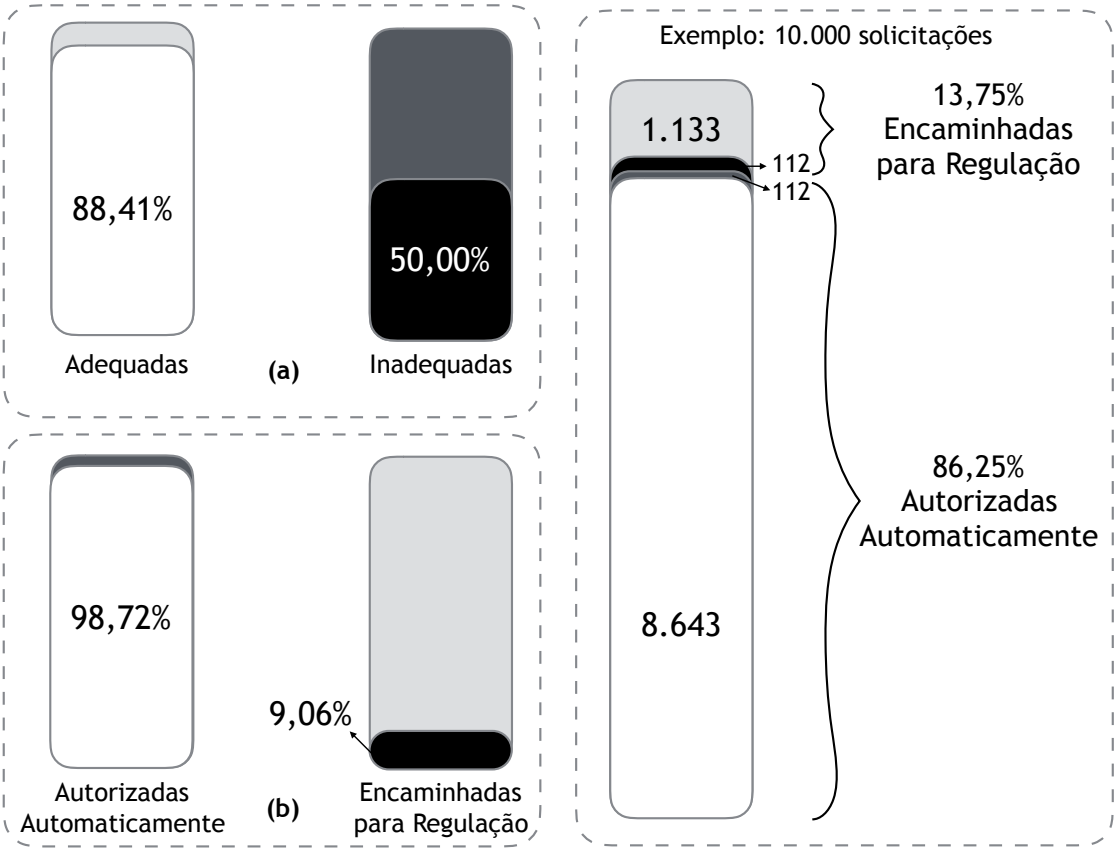

Figura 47 – Fator de Confiança de 11,00% para a combinação entre a Base de Dados V, a técnica *Random Oversampling* e o algoritmo *Naive Bayes*.

5.2.4.6 Base de Dados VII

A Tabela 19 exibe os melhores fatores de confiança para a combinação entre a Base de Dados VII, a técnica *Random Oversampling* e o algoritmo *Naive Bayes*.

Tabela 19 – Fatores de confiança para a combinação entre a Base de Dados VII, a técnica *Random Oversampling* e o algoritmo *Naive Bayes*.

Fator de Confiança	VPN	VPP	TVN	TVP
10,00%	20,28%	97,26%	17,60%	97,65%
20,00%	12,05%	97,56%	30,82%	92,43%
30,00%	9,17%	97,77%	40,81%	86,48%
40,00%	7,90%	97,93%	48,47%	81,13%
41,20%	7,83%	97,97%	50,03%	80,32%
50,00%	7,06%	98,21%	59,63%	73,80%
60,00%	5,94%	98,47%	70,68%	62,71%
70,00%	4,82%	98,66%	81,07%	46,59%
79,20%	3,92%	98,89%	91,56%	25,18%
80,00%	3,86%	98,93%	92,45%	23,23%
90,00%	3,42%	99,44%	98,78%	7,06%
100,00%	3,23%	0,00%	100,00%	0,00%

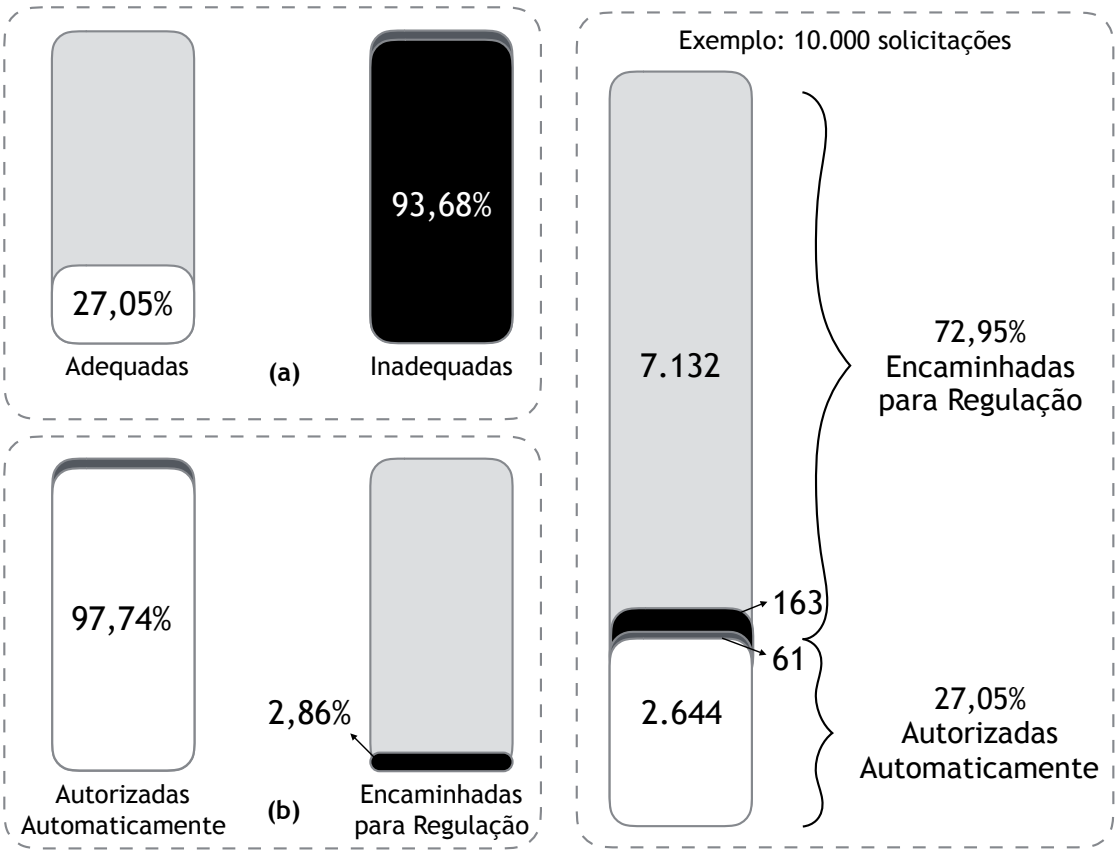

Figura 48 – Fator de Confiança de 99,90% para a combinação entre a Base de Dados V, a técnica *Random Oversampling* e o algoritmo *Naive Bayes*.

A partir do fator de confiança de 41,20% (Figura 51) 80,32% das solicitações adequadas (cerca de 7.773 solicitações) seriam autorizadas automaticamente porém 49,97% das solicitações inadequadas (cerca de 161 solicitações) também o seriam, isso significa uma redução de cerca de 79,34% no trabalho manual da regulação, com um erro inferior a 3,00% nas autorizações automáticas. No outro extremo, com um fator de confiança de 79,20% (Figura 52), 91,56% das solicitações inadequadas (cerca de 296 solicitações) seriam encaminhadas para a regulação juntamente com 74,82% das solicitações adequadas (cerca de 7.240 solicitações), isso significa uma redução de 24,64% no trabalho manual de regulação com um erro inferior a 2,00% nas autorizações automáticas.

5.3 Ameaças à validade

Uma questão fundamental a respeito dos resultados de um estudo experimental é o quão válido são os resultados obtidos. Portanto, é importante tratar os aspectos relacionados às ameaças à validade durante a fase de planejamento do experimento, para que seja possível antecipar e contornar problemas que tornem os resultados inválidos para a população em análise (WOHLIN et al., 2012).

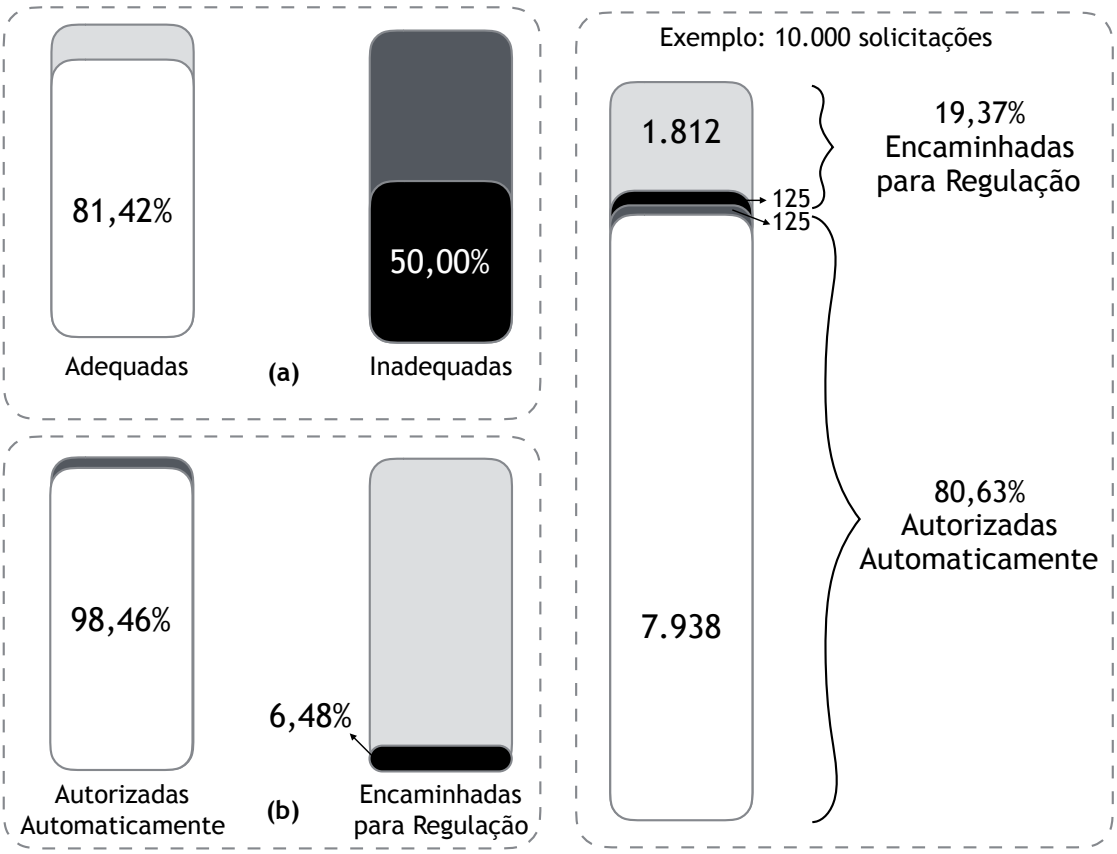

Figura 49 – Fator de Confiança de 17,90% para a combinação entre a Base de Dados VI, a técnica *Random Oversampling* e o algoritmo *Naive Bayes*.

5.3.1 Validade externa

Essa ameaça preocupa-se com a generalização dos resultados obtidos pelo estudo. Existem três riscos principais: utilizar participantes não adequados, realizar o experimento em ambiente não adequado ou com ferramentas inadequadas e executar o estudo em um período no qual a história possa afetar os resultados.

As bases de dados utilizadas nesta avaliação são de OPSs públicas/sem fins lucrativos. Em tais OPSs existe uma tendência de ser mais complacentes com solicitações do que em OPSs privadas, uma vez que o grau de contestação é baixo por causa do corporativismo do profissional de saúde. Os reguladores preferem evitar a indisposição com os médicos, e, pela falta de acompanhamento do trabalho de regulação, isso não é percebido pelos gestores. Isso foi notado por conta da da diferença de resultados existentes entre reguladores. Por conta disso, os resultados aqui obtidos podem ser diferentes dos resultados obtidos em bases eminentemente formada por operadoras privadas.

Apesar da metodologia ser aplicável em outros contextos, os resultados obtidos neste trabalho não podem ser extrapolados para outros locais, uma vez que são bastante dependentes das bases usadas. O nível de rigor e de consistência das regulações realizadas

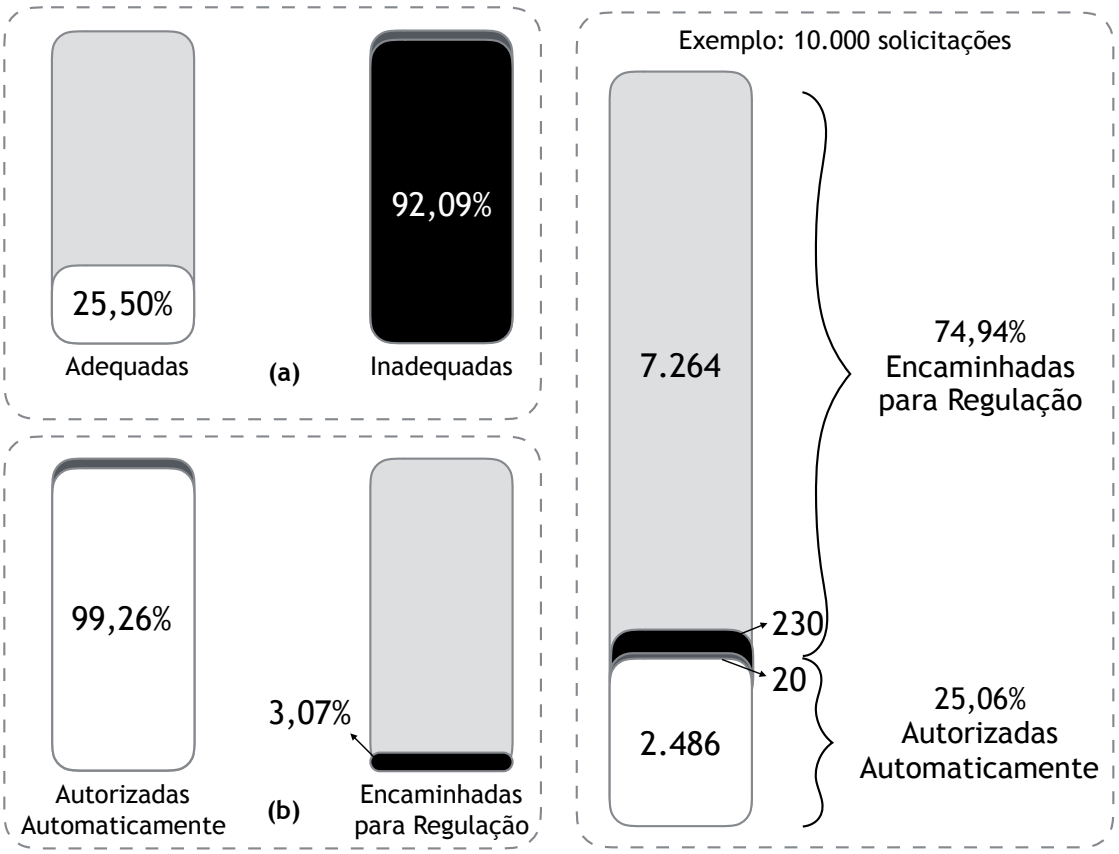

Figura 50 – Fator de Confiança de 99,10% para a combinação entre a Base de Dados VI, a técnica *Random Oversampling* e o algoritmo *Naive Bayes*.

definem o quão boa é a triagem.

5.3.2 Validade Interna

Se uma relação é observada entre o tratamento e os resultados, é necessário assegurar que essa relação é do tipo causa-efeito e que não é decorrente de um fator que não foi medido ou controlado. Em outras palavras, essa ameaça procura atestar que os tratamentos causam os resultados. Dentre os fatores que podem impactar na validade interna do experimento, destaca-se: a seleção e agrupamento dos participantes, o modo como os participantes são tratados ou o acontecimento de algum evento especial durante o estudo.

Uma ameaça à validade interna deste trabalho diz respeito à exploração de mais algoritmos de pré-processamento e mineração de dados. Um pequeno conjunto de algoritmos foi utilizados durante os experimentos devido ao fator limitante de tempo, assim, pode ser que as melhores técnicas não tenham sido empregadas.

Um outro ponto que remete a um trabalho futuro é o fato de não ter sido considerado, para o aprendizado, atributos que contêm texto desestruturados. Porém, tais atributos são comuns em sistemas médicos e normalmente contêm informações essenciais para se

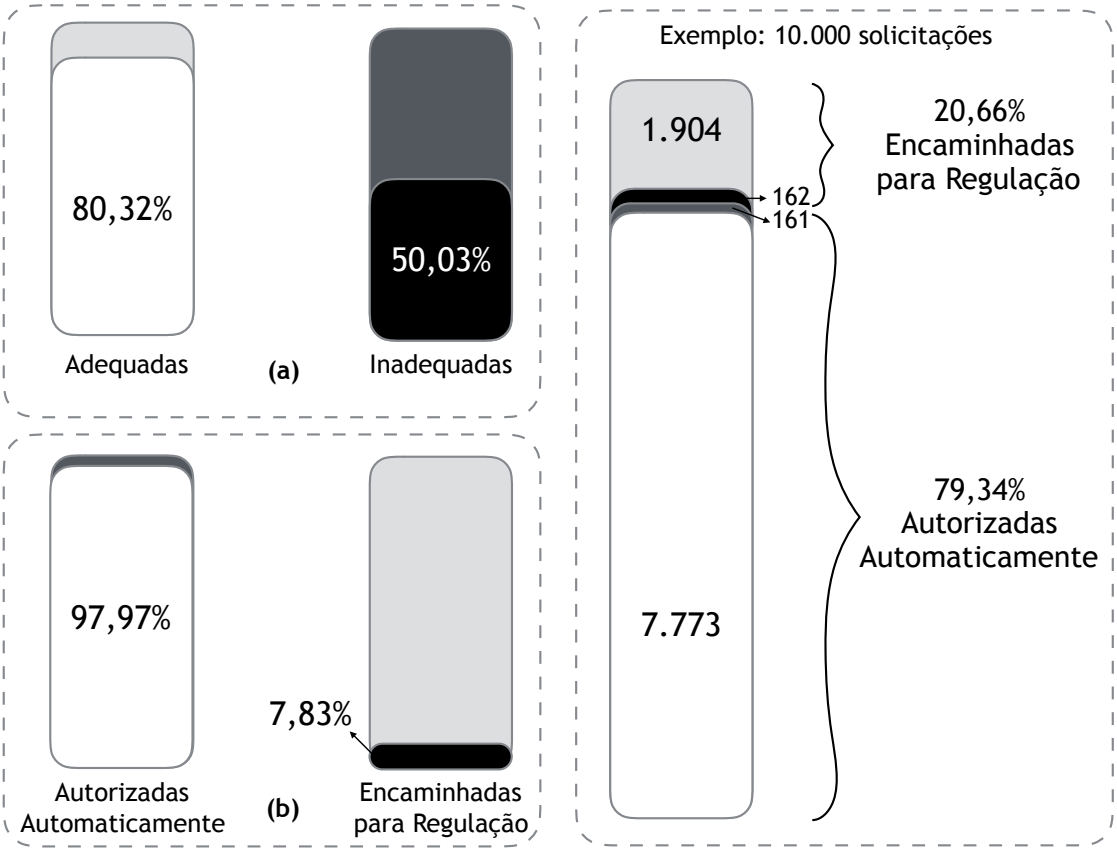

Figura 51 – Fator de Confiança de 41,20% para a combinação entre a Base de Dados VII, a técnica *Random Oversampling* e o algoritmo *Naive Bayes*.

usar em um mecanismo de apoio à decisão.

Além dos fatores relacionados, percebe-se que a falta de consistência nas regulações é um grande ponto limitador da triagem. Nas bases usadas foi percebido alguns casos dessa natureza. Outro ponto limitador das conclusões é o extremo desbalanceamento entre as classes. Parece irreal a existência de cenários com classes contendo mais de 98% de solicitações negadas. Esse valor não parece combinar com o que os especialistas da área entendem ser um comportamento comum nas solicitações de itens em saúde.

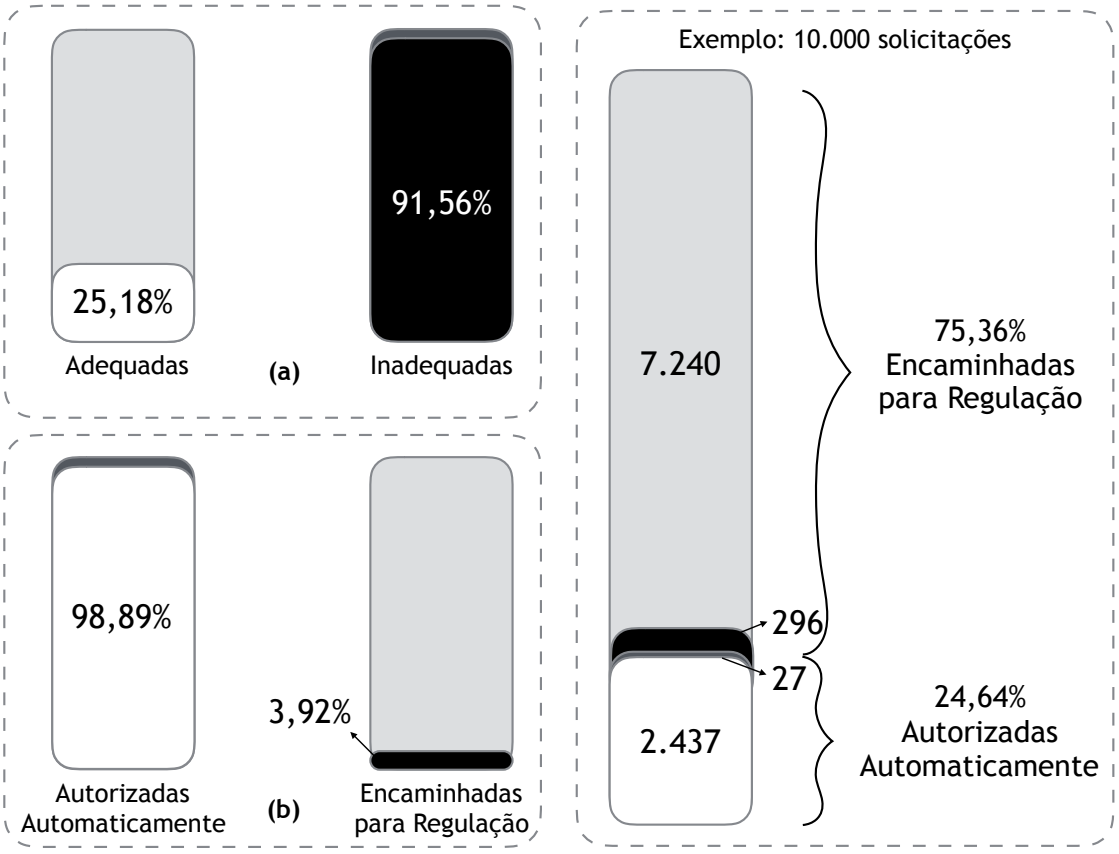

Figura 52 – Fator de Confiança de 79,20% para a combinação entre a Base de Dados VII, a técnica *Random Oversampling* e o algoritmo *Naive Bayes*.

6 Conclusões

Este trabalho propôs a triagem automática de solicitações de serviços assistenciais em saúde, com o intuito de reduzir a carga de trabalho no processo tradicional de regulação, porém, sem comprometer a qualidade do serviço. Somado a isso, foi estudada a identificação de solicitações adequadas e inadequadas em vários cenários distintos para estimar o impacto da triagem automática no nível de redução da carga de trabalho no processo tradicional de regulação. Por fim, foi avaliada a viabilidade da utilização da triagem automática para autorização imediata de solicitações adequadas.

Para julgar a triagem automática foram utilizadas setes bases de dados fornecidas por uma empresa parceira no trabalho, que oferece serviço de apoio à gestão de planos de saúde. Essas bases possuem características distintas como origem (pública ou privada), além de cobrirem procedimentos médicos e odontológicos, realizados nos âmbitos ambulatorial e hospitalar, e terem sido submetidas individualmente à metodologia descrita neste trabalho.

Na análise do fator de confiança foi identificado que os algoritmos *Naive Bayes* e *Random Forest* obtiveram os melhores resultados para todas as bases de dados, independentemente do grau de desbalanceamento e das técnicas utilizadas para balancea-las. Os algoritmos C4.5 e *Ripper* não obtiveram resultados tão animadores para bases de dados com alto grau de desbalanceamento, mas com a utilização das técnicas de *Random Oversampling* e SMOTE foi possível melhorar os resultados para estas bases. Por fim, com a utilização da técnica *MetaCost* foram encontrados os resultados mais balanceados em relação às Taxas de Verdadeiros Negativos e Positivos, porém, a identificação das solicitações inadequadas foi inferior as outras abordagens experimentadas.

A avaliação do impacto da triagem automática foi realizada em um cenário hipotético de 10.000 solicitações balanceadas, de acordo com a base de dados em questão. Nessa avaliação foram percebidos dois extremos: primeiro, com fatores de confiança baixos, uma OPS poderia diminuir a carga de trabalho na regulação tradicional em cerca de 80%, porém cerca de 50% das solicitações inadequadas seriam autorizadas automaticamente (desperdício assistencial); segundo, com fatores de confiança mais altos, é possível identificar cerca de 90% das solicitações inadequadas e encaminha-las para o processo tradicional de regulação, porém, a redução da carga de trabalho cairia de 80% para 25%. A partir da escolha de diferentes fatores de confiança dentro da faixa indicada, a OPS poderia decidir a priorização, de acordo com as circunstâncias, da diminuição da carga de trabalho ou diminuição dos desperdícios com assistência médica.

Conclui-se com este trabalho que a abordagem promoveu resultados promissores para a redução da carga de trabalho manual de regulação, além da apresentação de um

método configurável para autorização automática de solicitações de serviços assistenciais, por meio de um fator de confiança variável, de acordo com a situação e carga de trabalho manual na regulação. Além disso, foi descrito um roteiro para a aplicação da abordagem em um cenário real.

6.1 Limitações

Durante o desenvolvimento do corrente trabalho, vários elementos foram considerados como possíveis limitações. Dentre os elementos, destacam-se:

- A abordagem não contempla fatores sazonais, nem externos às bases de dados utilizadas. Dessa maneira, surtos endêmicos como o do vírus da Zika, que poderiam ser de extrema importância para avaliação de uma solicitação, não são considerados;
- As bases dados oriundas de OPSs públicas/sem fins lucrativos, apesar de possuírem graus de desbalanceamento semelhantes, não apresentaram resultados semelhantes, isso levanta a dúvida sobre a confiabilidade e consistência das informações pondo em cheque o aprendizado realizado nessas bases;
- Foi notada uma diferença de performance entre reguladores, exibida na disparidade entre a quantidade de procedimentos não autorizados de certos profissionais;
- Devido ao alto grau de desbalanceamento, algoritmos de detecção de anomalia podem apresentar melhores resultados.
- Por fim, a triagem não foi implantada em nenhuma OPS para avaliar o seu desempenho em uma situação real.

6.2 Continuidade da Pesquisa

De acordo com as limitações encontradas e ideias complementares para o trabalho, a seguir encontram-se os pontos para a continuidade da pesquisa e o cronograma de execução das tarefas.

Tarefa 1: Utilizar uma janela deslizante para capturar eventos sazonais e a criar, com o auxílio dos especialistas da OPS, atributos para representar fatores externos que podem ser considerados na avaliação das solicitações, como surtos endêmicos;

Tarefa 2: Investigar, em conjunto com especialistas em regulação, a consistência das informações contidas nas bases de dados das OPSs públicas/sem fins lucrativos;

Tarefa 3: Experimentar a construção de modelos preditivos separados por regulador para avaliar se a diferença de performance dos profissionais é refletida nos classificadores;

Tarefa 4: Utilização de algoritmos de detecção de anomalias (e.g. *one-class learners*) para verificar o seu impacto nos resultados;

Tarefa 5: Utilização de dois fatores de confiança, um para cada classe, para indicar aos reguladores aqueles mais prováveis de não autorização;

Tarefa 6: Utilizar a abordagem em um cenário real simultaneamente à regulação tradicional para confirmar os resultados encontrados.

Referências

AGGARWAL, C. C. *Data mining: the textbook*. [S.l.]: Springer, 2015. Citado na página 15.

ALPAYDIN, E. *Introduction to Machine Learning*. 2nd. ed. [S.l.]: MIT Press, 2010. ISBN 026201243X, 9780262012430. Citado 3 vezes nas páginas 17, 18 e 19.

ANS, A. N. de S. S. *Dados Consolidados da Saúde Suplementar (dezembro 15)*. 2015. Citado 4 vezes nas páginas 11, 1, 2 e 3.

ANS, A. N. de S. S. *Histórico*. 2015. Accessed: 2015-12-15. Disponível em: <http://www.ans.gov.br/aans/quem-somos/historico>. Citado na página 1.

ARAúJO, F. H.; SANTANA, A. M.; SANTOS NETO, P. de A. Using machine learning to support healthcare professionals in making preauthorisation decisions. *International Journal of Medical Informatics*, v. 94, p. 1 – 7, 2016. ISSN 1386-5056. Citado 2 vezes nas páginas 12 e 25.

BOUHADDOU, O. et al. Implementation of practice guidelines in a clinical setting using a computerized knowledge base (iliad). *Proceedings / the . Annual Symposium on Computer Application [sic] in Medical Care. Symposium on Computer Applications in Medical Care*, p. 258–262, 1993. ISSN 01954210. Citado 2 vezes nas páginas 9 e 10.

BRASIL, C. do. *Constituição da República Federativa do Brasil (Artigos 196 a 200)*. 1988. Accessed: 2015-12-15. Disponível em: <http://conselho.saude.gov.br/14cns/docs/constituicaofederal.pdf>. Citado na página 1.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Citado na página 20.

CARROLL, N. d. et al. Evaluation of an automated system for prior authorization: A cox-2 inhibitor example. *American Journal of Managed Care*, v. 12, n. 9, p. 501–508, 2006. ISSN 10880224. Citado na página 10.

CHAWLA, N. V. et al. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, v. 16, p. 321–357, 2002. Citado na página 22.

COHEN, W. W. Fast effective rule induction. In: *Proceedings of the twelfth international conference on machine learning*. [S.l.: s.n.], 1995. p. 115–123. Citado na página 20.

DOMINGOS, P. Metacost: A general method for making classifiers cost-sensitive. In: ACM. *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.], 1999. p. 155–164. Citado na página 22.

ECLIPSE. 2016. Accessed: 2016-08-07. Disponível em: <https://eclipse.org>. Citado na página 32.

FARMAN, A.; FARAG, A.; YEAP, P. Expediting prior approval and containing third-party costs for dental care. *Annals of the New York Academy of Sciences*, v. 670, p. 269–276, 1992. ISSN 00778923. Citado na página 9.

FAWCETT, T. An introduction to roc analysis. *Pattern recognition letters*, Elsevier, v. 27, n. 8, p. 861–874, 2006. Citado na página 23.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI magazine*, v. 17, n. 3, p. 37, 1996. Citado 2 vezes nas páginas 15 e 16.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. *The elements of statistical learning*. [S.l.]: Springer series in statistics Springer, Berlin, 2001. v. 1. Citado na página 21.

HILLERMAN, T. P.; CARVALHO, R. N.; REIS, A. C. B. Analyzing suspicious medical visit claims from individual healthcare service providers using k-means clustering. In: SPRINGER. *International Conference on Electronic Government and the Information Systems Perspective*. [S.l.], 2015. p. 191–205. Citado na página 13.

JAMIESON, D. Ecosystem health: some preventive medicine. *Environmental Values*, White Horse Press, v. 4, n. 4, p. 333–344, 1995. Citado na página 1.

JAVA. 2016. Accessed: 2016-08-07. Disponível em: <https://java.com/en/>. Citado na página 32.

KITCHENHAM, B. Procedures for performing systematic reviews. *Keele, UK, Keele University*, v. 33, n. 2004, p. 1–26, 2004. Citado na página 9.

KITCHENHAM, B.; CHARTERS, S. *Guidelines for performing Systematic Literature Reviews in Software Engineering*. 2007. Citado na página 9.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995. (IJCAI'95), p. 1137–1143. ISBN 1-55860-363-8. Disponível em: <http://dl.acm.org/citation.cfm?id=1643031.1643047>. Citado na página 23.

KOSE, I.; GOKTURK, M.; KILIC, K. An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance. *Applied Soft Computing*, Elsevier, v. 36, p. 283–299, 2015. Citado na página 13.

LUNDEEN, K. et al. Evaluation of a novel web-based prior approval application for palivizumab prophylaxis of respiratory syncytial virus in a state medicaid program. *JOURNAL OF MANAGED CARE PHARMACY*, v. 19, p. 115–124, 2013. Citado na página 10.

MAFRA, S. N.; TRAVASSOS, G. H. *Estudos Primários e Secundários apoiando a busca por Evidência em Engenharia de Software*. [S.l.], 2006. v. 687, n. 06. Citado na página 9.

MAIMON, O.; ROKACH, L. Introduction to knowledge discovery in databases. In: *Data Mining and Knowledge Discovery Handbook*. [S.l.: s.n.], 2005. p. 1–17. Citado 3 vezes nas páginas 11, 16 e 17.

MITCHELL, T. M. Machine learning. *New York*, 1997. Citado na página 20.

MOHRI, M.; ROSTAMIZADEH, A.; TALWALKAR, A. *Foundations of Machine Learning*. [S.l.]: The MIT Press, 2012. ISBN 026201825X, 9780262018258. Citado 3 vezes nas páginas 18, 19 e 23.

MURTHY, S. K.; KASIF, S.; SALZBERG, S. A system for induction of oblique decision trees. *Journal of artificial intelligence research*, 1994. Citado na página 20.

ORTEGA, P. A.; FIGUEROA, C. J.; RUZ, G. A. A medical claim fraud/abuse detection system based on data mining: A case study in chile. *DMIN*, Citeseer, v. 6, p. 26–29, 2006. Citado na página 12.

QUINLAN, J. R. Simplifying decision trees. *International journal of man-machine studies*, Elsevier, v. 27, n. 3, p. 221–234, 1987. Citado na página 20.

SHAVLIK, J. W.; DIETTERICH, T. G. *Readings in machine learning*. [S.l.]: Morgan Kaufmann, 1990. Citado na página 18.

SIMON, H. A. Why should machines learn? In: *Machine learning*. [S.l.]: Springer, 1983. p. 25–37. Citado na página 18.

SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, v. 45, n. 4, p. 427 – 437, 2009. ISSN 0306-4573. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0306457309000259>. Citado na página 21.

TERRY, K. Automated precertification lags behind, but new solutions emerging. *Medical economics*, v. 92, n. 2, p. 25–7, 30, 2015. ISSN 00257206. Citado na página 9.

TERRY, K. Electronic prior authorization. the solution to physicians' headaches? *Medical economics*, v. 92, n. 1, p. 26–7, 31–2, 2015. ISSN 00257206. Citado na página 9.

WEKA. 2016. Accessed: 2016-08-07. Disponível em: <https://weka.wikispaces.com>. Citado 2 vezes nas páginas 23 e 32.

WHO, W. H. O. *Programmes: Health and development*. 2010. Accessed: 2016-09-13. Disponível em: <http://www.who.int/hdp/en/>. Citado na página 1.

WHO, W. H. O. *The World Health Report [2010]: Health Systems Financing ; the Path to Universal Coverage*. [S.l.]: BWHO, 2010. Citado na página 2.

WOHLIN, C. et al. *Experimentation in software engineering*. [S.l.]: Springer Science & Business Media, 2012. Citado na página 70.

Apêndices

APÊNDICE A – Atributos Utilizados por Base de Dados

Base de Dados I: idade, sexo, valor, permiteMaterialComplementar, nivel, codigo, capitulo, subgrupo, mes, diaSemana, especialidade, procedimentosPorCapituloNivelMes, procedimentosPorGrupoNivelMes, procedimentosPorSubgrupoNivelMes, procedimentosPorCapituloNivelSemestre, procedimentosPorSubgrupoNivelSemestre, procedimentosPorGrupoNivelAno, procedimentosPorCapituloNivel, procedimentosPorSubgrupoNivel, outrosCapitulos, outrosSubgrupos, classe;

Base de Dados II: idade, sexo, valor, permiteMaterialComplementar, nivel, codigo, capitulo, grupo, subgrupo, mes, diaSemana, tempoPermanencia, limiteMensal, limiteSemestral, procedimentosPorCapituloNivelMes, procedimentosPorSubgrupoNivelMes, procedimentosPorGrupoNivelSemestre, procedimentosPorSubgrupoNivelSemestre, procedimentosPorGrupoNivelAno, procedimentosPorCapituloNivel, procedimentosPorSubgrupoNivel, outrosCapitulos, outrosSubgrupos, classe;

Base de Dados III: excluída por falta de procedimentos negativos;

Base de Dados IV: idade, sexo, valor, codigo, capitulo, grupo, subgrupo, mes, diaSemana, procedimentosPorCapituloNivelMes, procedimentosPorGrupoNivelMes, procedimentosPorSubgrupoNivelMes, procedimentosPorCapituloNivelSemestre, procedimentosPorSubgrupoNivelSemestre, procedimentosPorGrupoNivelAno, procedimentosPorCapituloNivel, procedimentosPorSubgrupoNivel, outrosCapitulos, outrosSubgrupos, classe;

Base de Dados V: idade, sexo, valor, permitematerialcomplementar, nivel, codigo, capitulo, grupo, subgrupo, mes, diaSemana, tempoPermanencia, procedimentosPorCapituloNivelMes, procedimentosPorGrupoNivelMes, procedimentosPorCapituloNivelSemestre, procedimentosPorSubgrupoNivelSemestre, procedimentosPorGrupoNivelAno, procedimentosPorCapituloNivel, procedimentosPorSubgrupoNivel, classe;

Base de Dados VI: sexo, valor, permitematerialcomplementar, quantidade, capitulo, grupo, mes, diaSemana, tempoPermanencia, procedimentosPorCapituloNivelMes, procedimentosPorGrupoNivelMes, procedimentosPorSubgrupoNivelMes, procedimentosPorCapituloNivelSemestre, procedimentosPorSubgrupoNivelSemestre, procedimentosPorGrupoNivelAno, procedimentosPorCapituloNivel, procedimentosPorSubgrupoNivel, outrosCapitulos, outrosSubgrupos, classe;

Base de Dados VII: idade, sexo, valor, permitematerialcomplementar, quantidade, codigo, grupo, subgrupo, diaSemana, procedimentosPorCapituloNivelMes, proce-

dimentosPorSubgrupoNivelMes, procedimentosPorCapituloNivelSemestre, procedimentosPorGrupoNivelSemestre, procedimentosPorCapituloNivelAno, procedimentosPorSubgrupoNivelAno, procedimentosPorGrupoNivel, quantidadeTotalExames, outrosGrupos, classe;

APÊNDICE B – Resultados

Banco de Dados I - C4.5

Sem Tratamento

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	75,82%	72,96%	4,08%	99,49%
20,00%	74,41%	73,30%	5,94%	99,22%
30,00%	68,12%	73,79%	9,06%	98,31%
40,00%	65,51%	77,03%	26,61%	94,56%
50,00%	62,53%	77,72%	30,77%	92,77%
60,00%	56,84%	78,67%	37,09%	88,98%
65,60%	51,62%	81,05%	50,11%	81,52%
70,00%	48,55%	82,58%	58,16%	76,21%
80,00%	40,34%	86,61%	77,15%	56,13%
85,20%	33,16%	93,51%	95,02%	25,65%
90,00%	32,12%	96,50%	98,07%	20,19%
100,00%	28,65%	93,04%	98,94%	5,26%

SMOTE

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	54,09%	73,92%	11,71%	96,19%
20,00%	57,09%	78,12%	34,32%	90,08%
30,00%	52,49%	80,99%	49,42%	82,80%
30,60%	52,17%	81,11%	50,06%	82,36%
40,00%	50,83%	82,11%	54,86%	79,60%
50,00%	49,89%	82,31%	56,23%	78,28%
60,00%	48,57%	82,93%	59,35%	75,85%
70,00%	46,20%	84,46%	66,33%	70,27%
80,00%	40,95%	86,29%	76,06%	57,81%
90,00%	35,63%	89,24%	87,69%	39,09%
97,60%	32,68%	88,90%	90,84%	28,00%
100,00%	29,34%	82,32%	91,41%	15,39%

Random Oversampling

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	50,46%	73,75%	11,26%	95,78%
20,00%	55,20%	78,01%	34,41%	89,33%
27,10%	51,04%	81,02%	50,38%	81,34%
30,00%	50,07%	82,90%	58,26%	77,68%
40,00%	48,12%	84,07%	63,68%	73,60%
50,00%	47,49%	84,53%	65,66%	72,07%
60,00%	47,00%	85,12%	67,91%	70,53%
70,00%	44,81%	85,73%	71,30%	66,23%
80,00%	39,66%	86,51%	77,84%	54,46%
90,00%	36,08%	86,87%	82,72%	43,70%
99,80%	31,45%	83,23%	84,80%	28,96%
100,00%	30,01%	80,49%	84,80%	23,98%

MetaCost

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,0%	53,8%	79,4%	41,9%	86,2%
17,9%	52,5%	81,2%	50,3%	82,5%
20,0%	52,2%	81,8%	52,7%	81,5%
30,0%	51,2%	83,1%	58,4%	78,6%
40,0%	50,8%	83,4%	59,6%	77,8%
50,0%	50,6%	83,4%	59,9%	77,5%
60,0%	50,5%	83,5%	60,4%	77,2%
70,0%	50,0%	83,9%	62,1%	76,1%
80,0%	49,0%	84,3%	63,9%	74,3%
90,0%	47,3%	85,0%	67,2%	71,2%
100,0%	42,6%	85,7%	73,0%	62,0%

Banco de Dados I - Naive Bayes

Sem Tratamento

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	69,95%	75,67%	18,94%	96,87%
20,00%	57,31%	78,13%	34,41%	90,07%
30,00%	48,03%	84,15%	64,03%	73,37%
40,00%	46,34%	85,71%	70,17%	68,76%
50,00%	45,93%	86,04%	71,40%	67,68%
60,00%	45,61%	86,22%	72,15%	66,92%
70,00%	44,14%	87,05%	75,44%	63,29%
80,00%	39,88%	89,60%	84,60%	50,96%
90,00%	36,77%	90,95%	89,43%	40,87%
98,90%	33,45%	96,45%	97,57%	25,38%

SMOTE

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	52,77%	80,46%	46,99%	83,84%
20,00%	48,34%	84,00%	63,31%	73,99%
30,00%	46,73%	85,31%	68,68%	69,88%
40,00%	44,58%	86,59%	73,91%	64,68%
50,00%	42,02%	88,00%	79,48%	57,83%
60,00%	40,06%	88,82%	82,89%	52,29%
70,00%	38,69%	89,81%	85,96%	47,60%
80,00%	37,34%	90,75%	88,64%	42,80%
90,00%	35,65%	93,07%	93,14%	35,35%
98,80%	33,43%	96,39%	97,52%	25,35%

Random Oversampling

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	51,57%	80,98%	49,91%	81,90%
20,00%	46,46%	85,64%	69,87%	69,05%
30,00%	46,08%	86,12%	71,55%	67,82%
40,00%	45,33%	86,58%	73,36%	65,98%
50,00%	42,47%	88,31%	79,80%	58,42%
60,00%	40,25%	89,24%	83,59%	52,24%
70,00%	38,32%	90,06%	86,66%	46,34%
80,00%	36,70%	90,81%	89,25%	40,80%
90,00%	35,36%	93,53%	93,91%	33,99%
98,30%	33,59%	96,81%	97,75%	25,72%

MetaCost

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	42,87%	87,21%	76,53%	60,49%
20,00%	41,66%	88,19%	79,87%	56,73%
30,00%	41,40%	88,39%	80,52%	55,90%
40,00%	41,28%	88,52%	80,89%	55,51%
50,00%	41,07%	88,65%	81,46%	54,88%
60,00%	40,83%	88,94%	82,42%	54,00%
70,00%	40,66%	89,22%	83,24%	53,27%
80,00%	40,51%	89,27%	83,46%	52,86%
90,00%	40,22%	89,29%	83,73%	52,11%
99,90%	37,49%	90,10%	87,42%	43,97%

Banco de Dados I - Random Forest

Sem Tratamento

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	71,87%	73,68%	8,20%	98,76%
20,00%	67,77%	75,50%	18,42%	96,63%
30,00%	63,27%	77,34%	28,62%	93,61%
40,00%	58,24%	78,80%	37,14%	89,79%
50,00%	55,59%	80,32%	45,01%	86,19%
56,20%	53,33%	81,26%	50,06%	83,17%
60,00%	52,18%	81,77%	52,76%	81,42%
70,00%	49,85%	84,23%	63,16%	75,56%
80,00%	46,13%	86,75%	73,33%	67,07%
90,00%	40,96%	89,96%	84,55%	53,14%
98,70%	33,17%	94,99%	96,53%	25,23%
100,00%	30,39%	97,08%	98,98%	12,87%

SMOTE

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	65,82%	75,41%	18,22%	96,36%
20,00%	62,68%	77,38%	28,94%	93,41%
30,00%	57,62%	78,83%	37,48%	89,43%
40,00%	54,50%	80,21%	45,06%	85,55%
47,30%	52,47%	81,16%	50,09%	82,58%
50,00%	51,63%	81,44%	51,70%	81,40%
60,00%	49,62%	83,26%	59,92%	76,59%
70,00%	47,22%	85,64%	69,40%	70,14%
80,00%	43,99%	88,46%	79,20%	61,20%
90,00%	39,24%	91,20%	88,02%	47,60%
98,10%	33,25%	95,75%	97,10%	25,07%
100,00%	30,23%	97,68%	99,26%	11,92%

Random Oversampling

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	59,32%	75,84%	21,91%	94,22%
20,00%	57,23%	77,67%	32,04%	90,82%
30,00%	55,13%	79,19%	40,18%	87,43%
40,00%	52,56%	80,57%	47,59%	83,49%
43,70%	52,19%	81,10%	50,04%	82,38%
50,00%	50,97%	81,78%	53,48%	80,21%
60,00%	49,29%	83,68%	61,62%	75,60%
70,00%	46,78%	85,77%	70,04%	69,35%
80,00%	43,36%	88,29%	79,20%	60,22%
90,00%	38,64%	91,07%	88,19%	46,16%
97,80%	33,22%	95,24%	96,71%	25,26%
100,00%	30,23%	97,39%	99,18%	12,00%

MetaCost

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	54,39%	76,67%	27,95%	90,97%
20,00%	52,42%	79,20%	41,59%	85,45%
30,00%	50,80%	80,75%	49,37%	81,59%
30,80%	50,72%	80,91%	50,09%	81,26%
40,00%	49,75%	82,24%	56,00%	78,24%
50,00%	48,61%	83,35%	60,88%	75,23%
60,00%	47,65%	84,52%	65,54%	72,30%
70,00%	46,08%	86,02%	71,28%	67,90%
80,00%	43,61%	87,99%	78,31%	61,07%
90,00%	39,68%	90,57%	86,61%	49,39%
98,60%	33,26%	95,43%	96,83%	25,33%
100,00%	30,58%	97,01%	98,89%	13,72%

Banco de Dados I - Ripper

Sem Tratamento

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	26,00%	72,28%	0,22%	99,98%
20,00%	47,98%	72,34%	0,59%	99,94%
30,00%	65,09%	73,15%	5,45%	98,93%
40,00%	66,55%	75,84%	20,43%	95,97%
50,00%	62,12%	77,08%	27,58%	93,57%
60,00%	62,12%	77,08%	27,58%	93,57%
70,00%	62,12%	77,08%	27,58%	93,57%
77,40%	48,69%	46,45%	57,48%	56,01%
77,70%	45,33%	38,75%	64,73%	46,68%
80,00%	27,76%	0,00%	100,00%	0,00%
90,00%	27,76%	0,00%	100,00%	0,00%
100,00%	27,76%	0,00%	100,00%	0,00%

SMOTE

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	54,81%	72,38%	0,82%	99,91%
20,00%	65,11%	75,53%	19,01%	96,05%
30,00%	54,20%	78,24%	35,82%	88,06%
33,30%	51,36%	81,12%	50,11%	81,16%
40,00%	48,51%	83,86%	62,74%	74,24%
50,00%	47,43%	84,08%	64,27%	72,40%
60,00%	47,43%	84,08%	64,27%	72,40%
70,00%	47,43%	84,08%	64,27%	72,40%
74,20%	36,45%	33,59%	85,02%	30,17%
80,00%	27,76%	0,00%	100,00%	0,00%
90,00%	27,76%	0,00%	100,00%	0,00%
100,00%	27,76%	0,00%	100,00%	0,00%

Random Oversampling

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	68,96%	72,51%	1,44%	99,90%
20,00%	55,61%	74,74%	13,94%	97,61%
30,00%	57,83%	80,38%	43,39%	85,57%
32,10%	51,27%	82,72%	55,96%	78,93%
40,00%	48,77%	85,74%	68,63%	72,00%
50,00%	48,07%	86,37%	71,01%	70,36%
60,00%	47,70%	86,73%	72,17%	69,26%
70,00%	46,70%	87,26%	73,91%	66,45%
73,10%	35,72%	44,30%	89,53%	29,20%
80,00%	28,82%	18,83%	99,16%	5,04%
90,00%	28,52%	19,48%	99,73%	3,51%
100,00%	27,91%	20,00%	100,00%	0,73%

MetaCost

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	57,76%	80,31%	44,14%	87,55%
20,00%	57,45%	80,60%	45,53%	87,00%
30,00%	57,28%	80,69%	45,95%	86,78%
40,00%	57,24%	80,71%	46,08%	86,73%
50,00%	57,24%	80,71%	46,08%	86,73%
60,00%	57,24%	80,71%	46,08%	86,73%
70,00%	57,24%	80,71%	46,08%	86,73%
80,00%	57,24%	80,71%	46,08%	86,73%
90,00%	57,24%	80,71%	46,08%	86,73%
99,10%	54,05%	72,57%	51,27%	77,97%
99,50%	42,35%	40,27%	72,75%	43,38%
100,00%	27,76%	0,00%	100,00%	0,00%

Base de Dados I - SVM

Sem Tratamento

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	0,00%	72,24%	0,00%	100,00%
20,00%	10,00%	72,24%	0,02%	100,00%
30,00%	47,29%	74,70%	13,87%	97,43%
40,00%	67,54%	76,21%	22,08%	95,91%
50,00%	67,82%	76,30%	22,46%	95,89%
60,00%	66,62%	76,28%	22,58%	95,65%
70,00%	65,45%	76,40%	23,42%	95,26%
76,50%	50,29%	83,97%	55,05%	57,55%
76,80%	39,73%	87,66%	78,19%	29,44%
80,00%	28,03%	90,00%	100,00%	1,31%
90,00%	27,77%	10,00%	100,00%	0,01%
100,00%	27,76%	0,00%	100,00%	0,00%

SMOTE

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	40,00%	72,26%	0,15%	99,99%
20,00%	87,33%	72,33%	0,52%	99,95%
27,60%	53,62%	82,26%	54,99%	77,54%
30,00%	48,43%	83,49%	61,50%	74,80%
40,00%	48,49%	83,71%	62,22%	74,56%
50,00%	48,16%	83,83%	62,86%	73,97%
60,00%	48,09%	84,09%	63,80%	73,51%
70,00%	43,58%	86,24%	71,29%	58,63%
70,50%	35,97%	87,30%	85,64%	30,34%
80,00%	27,91%	86,72%	99,48%	1,23%
90,00%	27,76%	78,95%	99,75%	0,25%
100,00%	27,76%	0,00%	100,00%	0,00%

Random Oversampling

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	10,00%	72,23%	0,02%	99,96%
20,00%	10,00%	72,23%	0,02%	99,96%
29,40%	42,18%	83,96%	61,60%	73,14%
30,00%	47,12%	85,41%	68,75%	70,32%
40,00%	47,11%	85,56%	69,20%	70,12%
50,00%	47,09%	85,76%	69,82%	69,84%
60,00%	47,02%	85,80%	69,99%	69,66%
70,00%	46,78%	85,88%	70,39%	69,21%
71,10%	35,25%	93,76%	87,92%	28,34%
80,00%	28,01%	95,00%	99,98%	1,24%
90,00%	27,85%	65,00%	99,98%	0,47%
100,00%	27,76%	0,00%	100,00%	0,00%

MetaCost

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	55,56%	78,18%	35,28%	89,10%
20,00%	55,54%	78,19%	35,33%	89,08%
30,00%	55,53%	78,20%	35,36%	89,07%
40,00%	55,47%	78,19%	35,36%	89,04%
50,00%	55,45%	78,19%	35,36%	89,03%
60,00%	55,45%	78,19%	35,36%	89,03%
70,00%	55,44%	78,19%	35,38%	89,02%
80,00%	55,46%	78,21%	35,43%	89,01%
90,00%	55,47%	78,21%	35,45%	89,01%
100,00%	39,51%	50,81%	74,68%	36,76%

Banco de Dados II - C4.5

Sem Tratamento

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	0,00%	99,76%	0,00%	100,00%
20,00%	0,00%	99,76%	0,00%	100,00%
30,00%	0,00%	99,76%	0,00%	100,00%
40,00%	0,00%	99,76%	0,00%	100,00%
50,00%	0,00%	99,76%	0,00%	100,00%
60,00%	0,00%	99,76%	0,00%	100,00%
70,00%	0,00%	99,76%	0,00%	100,00%
80,00%	0,00%	99,76%	0,00%	100,00%
90,00%	0,00%	99,76%	0,00%	100,00%
99,70%	0,00%	99,76%	0,00%	100,00%
99,80%	0,24%	0,00%	100,00%	0,00%
100,00%	0,24%	0,00%	100,00%	0,00%

SMOTE

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	1,67%	99,77%	2,50%	99,76%
20,00%	1,67%	99,77%	2,50%	99,76%
30,00%	1,67%	99,77%	2,50%	99,69%
40,00%	1,67%	99,77%	2,50%	99,63%
50,00%	1,67%	99,77%	2,50%	99,61%
60,00%	1,67%	99,77%	2,50%	99,58%
70,00%	1,43%	99,77%	2,50%	99,55%
80,00%	1,43%	99,77%	2,50%	99,53%
90,00%	5,32%	99,79%	10,00%	99,44%
99,90%	0,43%	99,86%	56,67%	70,98%
100,00%	0,29%	99,82%	59,17%	52,20%
100,00%	0,29%	99,82%	59,17%	52,20%

Random Oversampling

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	2,11%	99,78%	7,50%	99,11%
20,00%	2,11%	99,78%	7,50%	99,09%
30,00%	1,85%	99,78%	7,50%	98,93%
40,00%	1,80%	99,78%	7,50%	98,82%
50,00%	1,80%	99,78%	7,50%	98,82%
60,00%	1,80%	99,78%	7,50%	98,82%
70,00%	1,80%	99,78%	7,50%	98,80%
80,00%	1,80%	99,78%	7,50%	98,80%
90,00%	1,80%	99,78%	7,50%	98,80%
100,00%	1,80%	99,78%	7,50%	98,80%

MetaCost

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	0,00%	99,76%	0,00%	100,00%
20,00%	0,00%	99,76%	0,00%	100,00%
30,00%	0,00%	99,76%	0,00%	100,00%
40,00%	0,00%	99,76%	0,00%	100,00%
50,00%	0,00%	99,76%	0,00%	100,00%
60,00%	0,00%	99,76%	0,00%	100,00%
70,00%	0,00%	99,76%	0,00%	100,00%
80,00%	0,00%	99,76%	0,00%	100,00%
90,00%	0,00%	99,76%	0,00%	100,00%
100,00%	0,00%	99,76%	0,00%	100,00%

Banco de Dados II - Naive Bayes

Sem Tratamento

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	0,00%	99,76%	0,00%	99,89%
20,00%	0,00%	99,76%	0,00%	99,78%
30,00%	0,00%	99,76%	0,00%	99,72%
40,00%	0,00%	99,76%	0,00%	99,63%
50,00%	0,00%	99,76%	0,00%	99,56%
60,00%	0,00%	99,76%	0,00%	99,42%
70,00%	1,27%	99,78%	5,83%	99,26%
80,00%	0,87%	99,77%	5,83%	98,95%
90,00%	0,60%	99,77%	5,83%	98,32%
99,60%	1,25%	99,88%	54,17%	89,32%
99,90%	0,90%	99,90%	66,67%	82,57%
100,00%	0,24%	0,00%	100,00%	0,00%

SMOTE

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	7,26%	99,78%	7,50%	99,72%
20,00%	4,11%	99,78%	7,50%	99,52%
30,00%	2,46%	99,78%	7,50%	99,32%
40,00%	3,38%	99,79%	12,50%	99,19%
50,00%	4,18%	99,81%	17,50%	99,09%
60,00%	3,66%	99,81%	17,50%	98,98%
70,00%	4,62%	99,82%	25,00%	98,91%
80,00%	4,20%	99,82%	25,00%	98,82%
90,00%	4,21%	99,83%	27,50%	98,64%
100,00%	0,30%	99,92%	90,00%	28,84%

Random Oversampling

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	1,32%	99,80%	15,00%	97,66%
20,00%	1,37%	99,81%	20,00%	96,98%
30,00%	1,88%	99,83%	30,83%	96,28%
40,00%	1,72%	99,84%	33,33%	95,64%
50,00%	1,56%	99,84%	35,83%	94,91%
60,00%	1,56%	99,85%	41,67%	94,17%
70,00%	1,48%	99,86%	44,17%	93,51%
80,00%	1,36%	99,86%	46,67%	92,52%
90,00%	1,13%	99,86%	46,67%	90,99%
99,70%	0,76%	99,87%	51,67%	84,54%
99,90%	0,73%	99,88%	60,00%	81,41%
100,00%	0,24%	100,00%	100,00%	1,93%

MetaCost

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	0,90%	99,83%	36,67%	87,90%
20,00%	0,92%	99,84%	44,17%	86,32%
24,80%	1,02%	99,86%	51,67%	85,79%
30,00%	0,95%	99,86%	51,67%	85,37%
40,00%	0,89%	99,87%	54,17%	84,66%
50,00%	0,82%	99,87%	54,17%	83,97%
60,00%	0,77%	99,87%	54,17%	83,30%
70,00%	0,73%	99,87%	54,17%	82,67%
80,00%	0,68%	99,87%	54,17%	81,77%
90,00%	0,70%	99,88%	59,17%	80,61%
99,90%	0,58%	99,89%	66,67%	72,98%
100,00%	0,24%	0,00%	100,00%	0,00%

Banco de Dados II - Random Forest

Sem Tratamento

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	0,00%	99,76%	0,00%	100,00%
20,00%	0,00%	99,76%	0,00%	100,00%
30,00%	0,00%	99,76%	0,00%	100,00%
40,00%	0,00%	99,76%	0,00%	100,00%
50,00%	0,00%	99,76%	0,00%	99,98%
60,00%	0,00%	99,76%	0,00%	99,96%
70,00%	0,00%	99,76%	0,00%	99,93%
80,00%	8,33%	99,78%	5,00%	99,84%
90,00%	7,24%	99,79%	10,83%	99,48%
99,10%	2,80%	99,89%	56,67%	95,37%
100,00%	0,93%	99,93%	74,17%	81,46%

SMOTE

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	0,00%	99,76%	0,00%	99,98%
20,00%	0,00%	99,76%	0,00%	99,98%
30,00%	0,00%	99,76%	0,00%	99,97%
40,00%	0,00%	99,76%	0,00%	99,96%
50,00%	0,00%	99,76%	0,00%	99,95%
60,00%	0,00%	99,76%	0,00%	99,89%
70,00%	0,00%	99,76%	0,00%	99,70%
80,00%	3,69%	99,79%	10,83%	99,29%
90,00%	2,73%	99,82%	23,33%	98,04%
98,00%	1,62%	99,88%	50,83%	92,72%
100,00%	0,85%	99,93%	75,83%	79,07%

Random Oversampling

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	0,00%	99,76%	0,00%	99,99%
20,00%	0,00%	99,76%	0,00%	99,99%
30,00%	0,00%	99,76%	0,00%	99,96%
40,00%	0,00%	99,76%	0,00%	99,95%
50,00%	0,00%	99,76%	0,00%	99,93%
60,00%	3,33%	99,77%	2,50%	99,84%
70,00%	2,00%	99,77%	2,50%	99,61%
80,00%	3,85%	99,80%	15,83%	99,06%
90,00%	2,61%	99,83%	30,83%	97,35%
96,10%	1,84%	99,88%	54,17%	93,21%
100,00%	0,82%	99,92%	75,00%	78,87%

MetaCost

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	0,00%	99,76%	0,00%	99,81%
20,00%	4,11%	99,79%	13,33%	99,30%
30,00%	3,70%	99,82%	25,83%	98,43%
40,00%	2,70%	99,83%	28,33%	97,57%
50,00%	2,65%	99,85%	38,33%	96,53%
60,00%	2,02%	99,85%	40,83%	95,25%
70,00%	1,82%	99,87%	48,33%	93,72%
80,00%	1,29%	99,87%	48,33%	91,10%
80,50%	1,33%	99,87%	50,83%	90,90%
90,00%	0,93%	99,88%	55,83%	85,65%
100,00%	0,35%	99,93%	87,50%	41,52%

Banco de Dados II - Ripper

Sem Tratamento

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	0,00%	99,76%	0,00%	99,99%
20,00%	0,00%	99,76%	0,00%	99,99%
30,00%	0,00%	99,76%	0,00%	99,99%
40,00%	0,00%	99,76%	0,00%	99,99%
50,00%	0,00%	99,76%	0,00%	99,99%
60,00%	0,00%	99,76%	0,00%	99,99%
70,00%	0,00%	99,76%	0,00%	99,99%
80,00%	0,00%	99,76%	0,00%	99,99%
90,00%	0,00%	99,76%	0,00%	99,99%
99,70%	0,00%	99,76%	0,00%	99,99%
99,80%	0,24%	0,00%	100,00%	0,00%
100,00%	0,24%	0,00%	100,00%	0,00%

SMOTE

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	0,00%	99,76%	0,00%	99,80%
20,00%	0,00%	99,76%	0,00%	99,80%
30,00%	0,00%	99,76%	0,00%	99,80%
40,00%	0,00%	99,76%	0,00%	99,80%
50,00%	0,00%	99,76%	0,00%	99,80%
60,00%	0,00%	99,76%	0,00%	99,80%
70,00%	0,00%	99,76%	0,00%	99,80%
80,00%	0,00%	99,76%	0,00%	99,80%
90,00%	0,00%	99,76%	0,00%	99,80%
99,70%	0,06%	99,77%	3,33%	98,77%
99,80%	0,27%	19,97%	86,67%	18,50%
100,00%	0,25%	9,99%	96,67%	5,94%

Random Oversampling

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	3,80%	99,79%	10,83%	99,50%
20,00%	3,80%	99,79%	10,83%	99,50%
30,00%	3,80%	99,79%	10,83%	99,50%
40,00%	3,80%	99,79%	10,83%	99,50%
50,00%	3,80%	99,79%	10,83%	99,50%
60,00%	3,80%	99,79%	10,83%	99,50%
70,00%	3,80%	99,79%	10,83%	99,50%
80,00%	3,80%	99,79%	10,83%	99,50%
90,00%	3,80%	99,79%	10,83%	99,50%
100,00%	3,80%	99,79%	10,83%	99,50%

MetaCost

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	4,76%	99,78%	5,00%	99,76%
20,00%	4,58%	99,78%	5,00%	99,73%
30,00%	2,11%	99,78%	5,00%	99,64%
40,00%	2,11%	99,78%	5,00%	99,63%
50,00%	2,11%	99,78%	5,00%	99,59%
60,00%	2,11%	99,78%	5,00%	99,59%
70,00%	2,11%	99,78%	5,00%	99,59%
80,00%	2,11%	99,78%	5,00%	99,59%
90,00%	2,11%	99,78%	5,00%	99,59%
99,80%	2,11%	99,78%	5,00%	99,59%
99,90%	1,30%	19,96%	82,50%	19,93%
100,00%	0,24%	0,00%	100,00%	0,00%

Base de Dados II - SVM

Sem Tratamento

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	0,00%	99,76%	0,00%	100,00%
20,00%	0,00%	99,76%	0,00%	100,00%
30,00%	0,00%	99,76%	0,00%	100,00%
40,00%	0,00%	99,76%	0,00%	100,00%
50,00%	0,00%	99,76%	0,00%	100,00%
60,00%	0,00%	99,76%	0,00%	100,00%
70,00%	0,00%	99,76%	0,00%	100,00%
80,00%	0,00%	99,76%	0,00%	99,99%
90,00%	0,00%	99,76%	0,00%	99,97%
99,70%	0,73%	99,87%	57,50%	81,85%
99,90%	0,55%	99,93%	80,00%	66,16%
100,00%	0,24%	0,00%	100,00%	0,00%

SMOTE

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	1,88%	99,79%	12,50%	98,68%
20,00%	2,41%	99,80%	15,00%	98,63%
30,00%	2,12%	99,80%	15,00%	98,45%
40,00%	2,53%	99,80%	17,50%	98,40%
50,00%	2,70%	99,81%	20,00%	98,34%
60,00%	2,61%	99,81%	20,00%	98,31%
70,00%	2,86%	99,82%	22,50%	98,22%
80,00%	2,79%	99,82%	22,50%	98,18%
90,00%	2,57%	99,83%	31,67%	97,14%
99,90%	1,05%	99,87%	51,67%	89,20%
100,00%	0,24%	20,00%	100,00%	0,62%

Random Oversampling

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	1,94%	99,85%	38,33%	95,55%
20,00%	1,47%	99,85%	38,33%	94,05%
30,00%	1,29%	99,85%	41,67%	92,97%
40,00%	1,10%	99,85%	41,67%	91,70%
50,00%	0,96%	99,85%	41,67%	90,54%
60,00%	0,97%	99,86%	46,67%	89,41%
70,00%	0,85%	99,86%	46,67%	88,05%
74,70%	0,89%	99,87%	51,67%	87,19%
80,00%	0,82%	99,87%	51,67%	86,07%
90,00%	0,82%	99,89%	62,50%	82,95%
99,90%	0,50%	99,95%	87,50%	59,71%
100,00%	0,24%	0,00%	100,00%	0,00%

MetaCost

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	0,83%	99,85%	45,83%	87,00%
20,00%	0,75%	99,86%	48,33%	84,86%
21,70%	0,77%	99,87%	50,83%	84,61%
30,00%	0,75%	99,87%	54,17%	83,40%
40,00%	0,70%	99,87%	54,17%	82,25%
50,00%	0,69%	99,87%	56,67%	81,19%
60,00%	0,71%	99,89%	62,50%	80,00%
70,00%	0,70%	99,89%	65,00%	78,68%
80,00%	0,67%	99,90%	67,50%	76,93%
90,00%	0,63%	99,90%	70,00%	74,31%
99,90%	0,46%	99,94%	85,00%	56,67%
100,00%	0,24%	60,00%	100,00%	0,58%

Banco de Dados IV - C4.5

Sem Tratamento

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	0,00%	99,13%	0,00%	100,00%
20,00%	0,00%	99,13%	0,00%	100,00%
30,00%	0,00%	99,13%	0,00%	100,00%
40,00%	0,00%	99,13%	0,00%	100,00%
50,00%	0,00%	99,13%	0,00%	100,00%
60,00%	0,00%	99,13%	0,00%	100,00%
70,00%	0,00%	99,13%	0,00%	100,00%
80,00%	0,00%	99,13%	0,00%	100,00%
90,00%	0,00%	99,13%	0,00%	100,00%
99,10%	0,00%	99,13%	0,00%	100,00%
99,20%	0,87%	0,00%	100,00%	0,00%
100,00%	0,87%	0,00%	100,00%	0,00%

SMOTE

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	3,39%	99,17%	5,61%	98,63%
20,00%	3,24%	99,17%	5,61%	98,57%
30,00%	3,20%	99,17%	5,61%	98,55%
40,00%	3,19%	99,17%	5,61%	98,55%
50,00%	3,18%	99,17%	5,61%	98,55%
60,00%	3,22%	99,17%	5,87%	98,50%
70,00%	3,22%	99,17%	6,12%	98,42%
80,00%	3,24%	99,17%	6,38%	98,38%
90,00%	3,29%	99,18%	8,43%	97,86%
99,50%	1,45%	99,39%	51,03%	69,64%
100,00%	0,87%	99,12%	66,31%	33,56%

Random Oversampling

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	2,64%	99,18%	7,92%	97,42%
20,00%	2,60%	99,18%	7,92%	97,39%
30,00%	2,60%	99,18%	7,92%	97,38%
40,00%	2,60%	99,18%	7,92%	97,38%
50,00%	2,59%	99,18%	7,92%	97,37%
60,00%	2,57%	99,18%	7,92%	97,34%
70,00%	2,57%	99,18%	7,92%	97,34%
80,00%	2,57%	99,18%	7,92%	97,34%
90,00%	2,57%	99,18%	7,92%	97,34%
100,00%	2,57%	99,18%	7,92%	97,34%

MetaCost

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	0,00%	99,13%	0,00%	100,00%
20,00%	0,00%	99,13%	0,00%	100,00%
30,00%	0,00%	99,13%	0,00%	100,00%
40,00%	0,00%	99,13%	0,00%	100,00%
50,00%	0,00%	99,13%	0,00%	100,00%
60,00%	0,00%	99,13%	0,00%	100,00%
70,00%	0,00%	99,13%	0,00%	100,00%
80,00%	0,00%	99,13%	0,00%	100,00%
90,00%	0,00%	99,13%	0,00%	100,00%
100,00%	0,00%	99,13%	0,00%	100,00%

Banco de Dados IV - Naive Bayes

Sem Tratamento

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	0,00%	99,13%	0,00%	99,93%
20,00%	0,00%	99,13%	0,00%	99,90%
30,00%	0,00%	99,13%	0,00%	99,86%
40,00%	0,00%	99,13%	0,00%	99,82%
50,00%	0,00%	99,13%	0,00%	99,76%
60,00%	1,02%	99,13%	0,51%	99,63%
70,00%	1,56%	99,13%	1,01%	99,43%
80,00%	3,29%	99,15%	4,07%	98,82%
90,00%	3,81%	99,21%	11,72%	97,34%
98,70%	1,85%	99,46%	53,57%	75,15%
99,80%	1,18%	99,81%	92,88%	31,81%
100,00%	0,87%	0,00%	100,00%	0,00%

SMOTE

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	0,00%	99,13%	0,00%	99,99%
20,00%	0,00%	99,13%	0,00%	99,96%
30,00%	1,25%	99,13%	0,26%	99,93%
40,00%	3,33%	99,13%	0,77%	99,87%
50,00%	5,83%	99,14%	1,28%	99,80%
60,00%	4,19%	99,14%	1,28%	99,71%
70,00%	5,37%	99,15%	2,31%	99,59%
80,00%	3,62%	99,15%	2,81%	99,34%
90,00%	4,02%	99,17%	6,12%	98,75%
99,90%	1,88%	99,38%	43,35%	80,11%
100,00%	0,87%	100,00%	100,00%	0,49%

Random Oversampling

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	2,91%	99,30%	25,22%	92,58%
20,00%	2,45%	99,40%	41,04%	85,71%
27,90%	2,25%	99,46%	50,24%	80,79%
30,00%	2,22%	99,48%	53,05%	79,50%
40,00%	1,98%	99,53%	60,69%	73,62%
50,00%	1,80%	99,58%	67,34%	67,66%
60,00%	1,60%	99,59%	70,91%	61,69%
70,00%	1,47%	99,61%	75,26%	55,60%
80,00%	1,36%	99,64%	79,58%	49,22%
90,00%	1,26%	99,67%	83,92%	42,11%
99,60%	1,05%	99,66%	89,78%	25,98%
100,00%	0,87%	0,00%	100,00%	0,00%

MetaCost

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	1,94%	99,27%	26,29%	87,72%
20,00%	1,77%	99,28%	30,09%	84,78%
30,00%	1,68%	99,30%	33,91%	82,13%
40,00%	1,64%	99,32%	37,99%	79,62%
50,00%	1,63%	99,36%	43,10%	77,02%
60,00%	1,59%	99,39%	47,70%	74,12%
65,00%	1,59%	99,40%	50,22%	72,59%
70,00%	1,58%	99,42%	52,77%	71,08%
80,00%	1,48%	99,45%	58,39%	66,02%
90,00%	1,46%	99,55%	68,86%	59,39%
99,90%	1,22%	99,72%	88,03%	37,19%
100,00%	0,87%	0,00%	100,00%	0,00%

Banco de Dados IV - Random Forest

Sem Tratamento

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	0,00%	99,13%	0,00%	100,00%
20,00%	0,00%	99,13%	0,00%	100,00%
30,00%	0,00%	99,13%	0,00%	99,98%
40,00%	0,00%	99,13%	0,00%	99,93%
50,00%	1,00%	99,13%	0,26%	99,86%
60,00%	3,89%	99,14%	1,03%	99,75%
70,00%	5,36%	99,15%	2,81%	99,55%
80,00%	6,01%	99,18%	6,13%	99,13%
90,00%	4,79%	99,22%	13,03%	97,74%
99,90%	1,91%	99,47%	53,85%	75,68%
100,00%	1,66%	99,48%	58,69%	69,51%

SMOTE

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	3,10%	99,13%	0,51%	99,88%
20,00%	5,52%	99,14%	1,28%	99,79%
30,00%	4,84%	99,14%	1,79%	99,68%
40,00%	4,80%	99,15%	2,81%	99,49%
50,00%	5,00%	99,16%	4,34%	99,26%
60,00%	5,45%	99,18%	6,64%	98,92%
70,00%	4,44%	99,19%	8,69%	98,36%
80,00%	3,63%	99,21%	11,50%	97,34%
90,00%	2,92%	99,26%	19,40%	94,36%
99,10%	1,77%	99,49%	57,89%	71,84%
100,00%	1,55%	99,55%	68,13%	61,96%

Random Oversampling

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	0,00%	99,13%	0,00%	99,90%
20,00%	2,43%	99,13%	0,51%	99,84%
30,00%	2,14%	99,13%	0,51%	99,77%
40,00%	3,67%	99,14%	1,28%	99,68%
50,00%	4,38%	99,15%	2,31%	99,55%
60,00%	4,83%	99,16%	3,83%	99,34%
70,00%	4,94%	99,17%	5,88%	98,95%
80,00%	4,39%	99,20%	10,48%	97,99%
90,00%	3,84%	99,28%	20,94%	95,40%
99,10%	1,73%	99,45%	54,35%	72,93%
100,00%	1,62%	99,48%	58,94%	68,67%

MetaCost

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	5,46%	99,16%	4,33%	99,35%
20,00%	4,86%	99,20%	9,95%	98,30%
30,00%	4,21%	99,23%	14,53%	97,07%
40,00%	3,88%	99,27%	19,65%	95,70%
50,00%	3,38%	99,29%	23,22%	94,15%
60,00%	2,97%	99,31%	26,79%	92,33%
70,00%	2,47%	99,31%	29,10%	89,89%
80,00%	2,25%	99,35%	35,96%	86,28%
90,00%	1,96%	99,41%	46,15%	79,70%
92,00%	1,91%	99,44%	50,24%	77,35%
100,00%	1,18%	99,65%	85,21%	37,26%

Banco de Dados IV - Ripper

Sem Tratamento

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	0,00%	99,13%	0,00%	100,00%
20,00%	0,00%	99,13%	0,00%	100,00%
30,00%	0,00%	99,13%	0,00%	100,00%
40,00%	0,00%	99,13%	0,00%	100,00%
50,00%	0,00%	99,13%	0,00%	99,99%
60,00%	0,00%	99,13%	0,00%	99,99%
70,00%	0,00%	99,13%	0,00%	99,99%
80,00%	0,00%	99,13%	0,00%	99,99%
90,00%	0,00%	99,13%	0,00%	99,99%
99,10%	0,00%	99,13%	0,00%	99,99%
99,20%	0,87%	0,00%	100,00%	0,00%
100,00%	0,87%	0,00%	100,00%	0,00%

SMOTE

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	2,92%	99,14%	1,78%	99,45%
20,00%	2,98%	99,14%	2,28%	99,34%
30,00%	2,87%	99,14%	2,28%	99,32%
40,00%	2,87%	99,14%	2,28%	99,32%
50,00%	2,87%	99,14%	2,28%	99,32%
60,00%	2,87%	99,14%	2,28%	99,32%
70,00%	2,87%	99,14%	2,28%	99,32%
80,00%	2,87%	99,14%	2,28%	99,32%
90,00%	2,87%	99,14%	2,28%	99,32%
99,00%	2,61%	69,40%	31,78%	69,55%
99,10%	0,87%	0,00%	100,00%	0,00%
100,00%	0,87%	0,00%	100,00%	0,00%

Random Oversampling

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	4,56%	99,20%	9,98%	98,23%
20,00%	4,56%	99,20%	9,98%	98,23%
30,00%	4,56%	99,20%	9,98%	98,23%
40,00%	4,56%	99,20%	9,98%	98,23%
50,00%	4,56%	99,20%	9,98%	98,23%
60,00%	4,56%	99,20%	9,98%	98,23%
70,00%	4,56%	99,20%	9,98%	98,23%
80,00%	4,56%	99,20%	9,98%	98,23%
90,00%	4,56%	99,20%	9,98%	98,23%
100,00%	4,56%	99,20%	9,98%	98,23%

MetaCost

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	5,22%	99,15%	2,31%	99,71%
20,00%	4,63%	99,15%	2,31%	99,69%
30,00%	4,50%	99,15%	2,31%	99,66%
40,00%	4,12%	99,15%	2,31%	99,61%
50,00%	4,12%	99,15%	2,31%	99,61%
60,00%	4,12%	99,15%	2,31%	99,61%
70,00%	4,12%	99,15%	2,31%	99,61%
80,00%	4,12%	99,15%	2,31%	99,61%
90,00%	4,12%	99,15%	2,31%	99,61%
99,90%	1,15%	39,65%	60,26%	39,85%
100,00%	0,87%	0,00%	100,00%	0,00%

Base de Dados IV - SVM

Sem Tratamento

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	0,00%	99,13%	0,00%	100,00%
20,00%	0,00%	99,13%	0,00%	100,00%
30,00%	0,00%	99,13%	0,00%	100,00%
40,00%	0,00%	99,13%	0,00%	100,00%
50,00%	0,00%	99,13%	0,00%	100,00%
60,00%	0,00%	99,13%	0,00%	100,00%
70,00%	0,00%	99,13%	0,00%	100,00%
80,00%	0,00%	99,13%	0,00%	100,00%
90,00%	0,00%	99,13%	0,00%	100,00%
100,00%	0,00%	99,13%	0,00%	100,00%

SMOTE

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	1,69%	99,40%	48,21%	75,27%
20,00%	1,69%	99,40%	48,21%	75,27%
30,00%	1,69%	99,40%	48,21%	75,27%
40,00%	1,69%	99,40%	48,21%	75,27%
50,00%	1,69%	99,40%	48,21%	75,27%
60,00%	1,69%	99,40%	48,21%	75,27%
70,00%	1,69%	99,40%	48,21%	75,27%
80,00%	1,69%	99,40%	48,21%	75,27%
90,00%	1,69%	99,40%	48,21%	75,27%
100,00%	1,69%	99,40%	48,21%	75,27%

Random Oversampling

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	6,27%	99,17%	5,35%	99,29%
20,00%	3,19%	99,26%	18,84%	95,00%
30,00%	2,61%	99,37%	36,71%	87,94%
38,80%	2,13%	99,46%	50,22%	79,81%
40,00%	2,11%	99,47%	52,51%	78,63%
50,00%	1,77%	99,55%	64,52%	68,55%
60,00%	1,52%	99,62%	75,00%	57,43%
70,00%	1,35%	99,68%	82,65%	46,80%
80,00%	1,19%	99,74%	89,54%	34,58%
86,60%	1,09%	99,80%	94,13%	25,15%
90,00%	1,03%	99,82%	96,17%	19,05%
100,00%	0,87%	0,00%	100,00%	0,00%

MetaCost

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	0,00%	99,13%	0,00%	100,00%
20,00%	0,00%	99,13%	0,00%	100,00%
30,00%	0,00%	99,13%	0,00%	100,00%
40,00%	0,00%	99,13%	0,00%	100,00%
50,00%	0,00%	99,13%	0,00%	100,00%
60,00%	0,00%	99,13%	0,00%	100,00%
70,00%	0,00%	99,13%	0,00%	100,00%
80,00%	0,00%	99,13%	0,00%	100,00%
90,00%	0,00%	99,13%	0,00%	100,00%
100,00%	0,00%	99,13%	0,00%	100,00%

Banco de Dados V - C4.5

Sem Tratamento

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	0,00%	97,76%	0,00%	100,00%
20,00%	0,00%	97,76%	0,00%	100,00%
30,00%	0,00%	97,76%	0,00%	100,00%
40,00%	0,00%	97,76%	0,00%	100,00%
50,00%	0,00%	97,76%	0,00%	100,00%
60,00%	0,00%	97,76%	0,00%	100,00%
70,00%	0,00%	97,76%	0,00%	100,00%
80,00%	0,00%	97,76%	0,00%	100,00%
90,00%	0,00%	97,76%	0,00%	100,00%
97,70%	0,00%	97,76%	0,00%	100,00%
97,80%	2,24%	0,00%	100,00%	0,00%
100,00%	2,24%	0,00%	100,00%	0,00%

SMOTE

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	32,79%	98,47%	33,42%	98,36%
20,00%	30,19%	98,47%	33,42%	98,18%
30,00%	29,11%	98,47%	33,68%	98,07%
40,00%	29,31%	98,49%	34,47%	98,04%
50,00%	29,24%	98,49%	34,47%	98,04%
60,00%	24,64%	98,55%	37,37%	97,34%
70,00%	23,93%	98,55%	37,63%	97,22%
80,00%	23,44%	98,58%	38,95%	97,04%
90,00%	19,82%	98,74%	46,84%	95,61%
92,40%	17,97%	98,80%	50,00%	94,66%
100,00%	3,67%	99,14%	80,53%	51,40%

Random Oversampling

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	27,90%	98,76%	46,84%	97,20%
11,60%	25,93%	98,84%	50,53%	96,68%
20,00%	22,82%	98,93%	55,00%	95,71%
30,00%	21,64%	98,96%	56,32%	95,28%
40,00%	21,60%	98,97%	56,84%	95,24%
50,00%	20,29%	99,00%	58,16%	94,75%
60,00%	20,31%	99,03%	59,47%	94,60%
70,00%	20,31%	99,03%	59,47%	94,60%
80,00%	20,31%	99,03%	59,47%	94,60%
90,00%	20,27%	99,03%	59,47%	94,59%
100,00%	20,27%	99,03%	59,47%	94,59%

MetaCost

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	23,23%	98,04%	13,95%	98,48%
20,00%	23,23%	98,04%	13,95%	98,48%
30,00%	23,23%	98,04%	13,95%	98,48%
40,00%	23,23%	98,04%	13,95%	98,48%
50,00%	23,23%	98,04%	13,95%	98,48%
60,00%	23,23%	98,04%	13,95%	98,48%
70,00%	23,16%	98,04%	13,95%	98,48%
80,00%	23,16%	98,04%	13,95%	98,48%
90,00%	23,16%	98,04%	13,95%	98,48%
100,00%	17,65%	97,89%	34,74%	73,07%

Banco de Dados V - Naive Bayes

Sem Tratamento

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	10,86%	97,79%	1,84%	99,66%
20,00%	8,29%	97,79%	2,11%	99,37%
30,00%	8,63%	97,82%	3,68%	99,11%
40,00%	7,53%	97,82%	4,21%	98,75%
50,00%	8,24%	97,85%	6,05%	98,28%
60,00%	7,85%	97,87%	7,37%	97,81%
70,00%	7,08%	97,92%	10,26%	96,84%
80,00%	8,34%	98,11%	20,00%	95,03%
90,00%	7,92%	98,42%	37,11%	90,03%
93,50%	6,82%	98,66%	50,26%	84,20%
99,90%	3,01%	99,50%	93,16%	31,11%
100,00%	2,24%	0,00%	100,00%	0,00%

SMOTE

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	11,64%	98,42%	33,95%	94,11%
20,00%	8,89%	98,55%	42,11%	90,10%
30,00%	7,56%	98,60%	46,32%	86,97%
40,00%	6,59%	98,62%	48,95%	84,02%
40,60%	6,67%	98,65%	50,00%	83,87%
50,00%	6,04%	98,68%	52,89%	81,05%
60,00%	5,55%	98,72%	55,79%	78,09%
70,00%	5,08%	98,75%	58,95%	74,64%
80,00%	4,70%	98,83%	63,68%	70,34%
90,00%	4,14%	98,87%	68,16%	63,80%
99,90%	3,14%	99,15%	84,74%	40,15%
100,00%	2,24%	0,00%	100,00%	0,00%

Random Oversampling

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	9,56%	98,69%	48,16%	89,47%
11,00%	9,06%	98,72%	50,00%	88,41%
20,00%	6,59%	98,92%	62,11%	79,73%
30,00%	5,21%	99,03%	69,74%	70,73%
40,00%	4,46%	99,07%	74,21%	63,43%
50,00%	3,97%	99,07%	76,58%	57,45%
60,00%	3,64%	99,08%	78,68%	52,23%
70,00%	3,46%	99,10%	80,79%	48,28%
80,00%	3,35%	99,13%	82,37%	45,46%
90,00%	3,30%	99,19%	84,47%	43,22%
99,90%	2,86%	99,50%	93,68%	27,05%
100,00%	2,24%	0,00%	100,00%	0,00%

MetaCost

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	4,94%	98,54%	49,21%	78,28%
10,40%	4,98%	98,56%	50,00%	78,07%
20,00%	4,76%	98,77%	60,53%	72,10%
30,00%	4,83%	99,00%	69,74%	68,30%
40,00%	4,74%	99,10%	73,95%	65,41%
50,00%	4,71%	99,22%	78,16%	63,13%
60,00%	4,61%	99,34%	82,63%	59,80%
70,00%	4,27%	99,44%	86,84%	54,39%
80,00%	3,68%	99,47%	90,00%	44,98%
90,00%	3,20%	99,50%	92,11%	36,09%
99,90%	3,06%	99,52%	93,16%	32,26%
100,00%	2,24%	0,00%	100,00%	0,00%

Banco de Dados V - Random Forest

Sem Tratamento

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	95,32%	98,14%	17,37%	99,98%
20,00%	89,89%	98,21%	20,53%	99,93%
30,00%	88,37%	98,25%	22,37%	99,92%
40,00%	72,67%	98,40%	29,47%	99,72%
50,00%	66,62%	98,47%	32,37%	99,61%
60,00%	58,82%	98,53%	35,53%	99,40%
70,00%	47,48%	98,62%	39,74%	98,97%
80,00%	35,69%	98,77%	46,84%	98,03%
83,00%	32,49%	98,84%	50,00%	97,60%
90,00%	26,88%	99,04%	59,21%	96,29%
100,00%	6,94%	99,61%	87,37%	73,10%

SMOTE

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	49,43%	98,32%	26,05%	99,34%
20,00%	44,56%	98,41%	30,26%	99,11%
30,00%	42,60%	98,45%	31,84%	99,00%
40,00%	39,58%	98,48%	33,68%	98,80%
50,00%	38,81%	98,58%	37,89%	98,60%
60,00%	38,31%	98,71%	43,95%	98,35%
70,00%	31,95%	98,81%	48,68%	97,58%
71,20%	31,56%	98,84%	50,00%	97,47%
80,00%	24,97%	98,94%	55,00%	96,16%
90,00%	17,80%	99,18%	66,58%	92,94%
100,00%	6,68%	99,59%	87,11%	72,08%

Random Oversampling

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	38,02%	98,64%	40,79%	98,44%
20,00%	32,25%	98,79%	47,89%	97,68%
30,00%	28,34%	98,82%	49,47%	97,11%
31,70%	28,06%	98,83%	50,00%	97,05%
40,00%	27,10%	98,88%	52,11%	96,79%
50,00%	26,40%	98,91%	53,42%	96,56%
60,00%	25,34%	98,97%	56,58%	96,15%
70,00%	23,83%	99,09%	61,84%	95,45%
80,00%	20,08%	99,14%	64,47%	94,10%
90,00%	15,30%	99,28%	71,05%	90,96%
100,00%	6,91%	99,62%	87,89%	72,83%

MetaCost

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	22,22%	98,72%	45,53%	96,35%
17,80%	21,44%	98,82%	50,00%	95,79%
20,00%	21,19%	98,83%	50,79%	95,65%
30,00%	20,08%	98,91%	54,47%	95,01%
40,00%	18,79%	98,99%	58,16%	94,20%
50,00%	18,02%	99,05%	61,05%	93,61%
60,00%	17,81%	99,13%	64,21%	93,19%
70,00%	17,09%	99,18%	66,58%	92,57%
80,00%	15,87%	99,22%	68,42%	91,65%
90,00%	14,10%	99,30%	72,37%	89,86%
100,00%	6,31%	99,66%	89,74%	69,41%

Banco de Dados V - Ripper

Sem Tratamento

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	97,08%	97,96%	9,21%	99,99%
20,00%	97,08%	97,98%	10,00%	99,99%
30,00%	92,08%	97,98%	10,00%	99,97%
40,00%	90,89%	97,98%	10,00%	99,96%
50,00%	90,89%	97,99%	10,53%	99,96%
60,00%	90,89%	97,99%	10,53%	99,96%
70,00%	90,89%	97,99%	10,53%	99,96%
80,00%	90,89%	97,99%	10,53%	99,96%
90,00%	90,89%	97,99%	10,53%	99,96%
98,00%	53,04%	58,77%	45,26%	59,97%
98,10%	9,16%	9,80%	91,32%	9,99%
100,00%	2,24%	0,00%	100,00%	0,00%

SMOTE

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	30,34%	98,45%	32,63%	98,24%
20,00%	28,83%	98,47%	33,42%	98,07%
30,00%	28,50%	98,48%	34,21%	98,00%
40,00%	27,99%	98,48%	34,21%	97,95%
50,00%	27,99%	98,48%	34,21%	97,95%
60,00%	27,99%	98,48%	34,21%	97,95%
70,00%	27,99%	98,48%	34,21%	97,95%
80,00%	27,99%	98,48%	34,21%	97,95%
90,00%	27,99%	98,48%	34,21%	97,95%
98,60%	13,18%	39,41%	74,47%	39,20%
100,00%	2,24%	0,00%	100,00%	0,00%

Random Oversampling

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
9,00%	28,21%	98,88%	52,11%	96,23%
10,00%	27,52%	98,90%	52,89%	96,09%
20,00%	23,85%	99,04%	59,47%	95,19%
30,00%	22,16%	99,05%	60,26%	94,77%
40,00%	21,87%	99,08%	61,32%	94,65%
50,00%	21,85%	99,08%	61,32%	94,64%
60,00%	21,85%	99,08%	61,32%	94,64%
70,00%	21,85%	99,08%	61,32%	94,64%
80,00%	21,85%	99,08%	61,32%	94,64%
90,00%	21,85%	99,08%	61,32%	94,64%
100,00%	21,85%	99,08%	61,32%	94,64%

MetaCost

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	46,98%	98,22%	21,32%	99,39%
20,00%	45,17%	98,24%	22,63%	99,31%
30,00%	42,36%	98,25%	23,16%	99,25%
40,00%	42,40%	98,27%	23,68%	99,24%
50,00%	42,13%	98,27%	23,68%	99,23%
60,00%	42,13%	98,27%	23,68%	99,23%
70,00%	42,13%	98,27%	23,68%	99,23%
80,00%	42,13%	98,27%	23,68%	99,23%
90,00%	42,13%	98,27%	23,68%	99,23%
100,00%	2,24%	0,00%	100,00%	0,00%

Base de Dados V - SVM

Sem Tratamento

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	40,00%	97,78%	1,05%	99,98%
20,00%	40,00%	97,78%	1,05%	99,96%
30,00%	35,00%	97,78%	1,05%	99,96%
40,00%	35,00%	97,78%	1,05%	99,95%
50,00%	35,00%	97,78%	1,05%	99,95%
60,00%	35,00%	97,78%	1,05%	99,95%
70,00%	35,00%	97,78%	1,05%	99,95%
80,00%	35,00%	97,78%	1,05%	99,95%
90,00%	35,00%	97,78%	1,32%	99,93%
97,80%	12,10%	97,80%	10,79%	90,77%
97,90%	2,30%	98,88%	97,89%	4,61%
100,00%	2,24%	0,00%	100,00%	0,00%

SMOTE

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	8,90%	97,89%	8,16%	97,99%
20,00%	5,63%	98,00%	17,11%	93,33%
30,00%	7,41%	98,54%	43,42%	87,48%
38,50%	6,31%	98,64%	50,00%	82,94%
40,00%	6,08%	98,64%	50,53%	82,07%
50,00%	5,28%	98,75%	57,89%	76,22%
60,00%	4,67%	98,82%	63,42%	70,34%
70,00%	4,01%	98,83%	67,63%	62,78%
80,00%	3,29%	98,80%	73,16%	50,63%
90,00%	2,92%	98,98%	83,95%	35,85%
94,80%	2,69%	99,09%	90,00%	25,22%
100,00%	2,24%	0,00%	100,00%	0,00%

Random Oversampling

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	4,58%	97,81%	3,95%	98,28%
20,00%	7,67%	98,09%	20,00%	94,48%
30,00%	8,58%	98,69%	48,95%	88,05%
31,40%	8,11%	98,70%	50,00%	87,03%
40,00%	6,32%	98,81%	57,89%	80,35%
50,00%	5,03%	98,85%	63,16%	72,66%
60,00%	4,48%	98,98%	70,79%	65,31%
70,00%	3,79%	99,13%	79,47%	53,68%
80,00%	2,98%	99,13%	86,32%	35,50%
87,00%	2,71%	99,17%	90,79%	25,09%
90,00%	2,61%	99,17%	92,37%	20,84%
100,00%	2,24%	0,00%	100,00%	0,00%

MetaCost

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	13,47%	97,84%	4,47%	99,40%
20,00%	13,38%	97,84%	4,47%	99,40%
30,00%	13,30%	97,84%	4,47%	99,39%
40,00%	13,30%	97,84%	4,47%	99,39%
50,00%	13,30%	97,84%	4,47%	99,39%
60,00%	13,30%	97,84%	4,47%	99,39%
70,00%	13,30%	97,84%	4,47%	99,39%
80,00%	13,30%	97,84%	4,47%	99,39%
90,00%	13,30%	97,84%	4,47%	99,39%
100,00%	7,53%	97,32%	20,53%	80,33%

Banco de Dados VI - C4.5

Sem Tratamento

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	0,00%	97,50%	0,00%	100,00%
20,00%	0,00%	97,50%	0,00%	100,00%
30,00%	0,00%	97,50%	0,00%	100,00%
40,00%	0,00%	97,50%	0,00%	100,00%
50,00%	0,00%	97,50%	0,00%	100,00%
60,00%	0,00%	97,50%	0,00%	100,00%
70,00%	0,00%	97,50%	0,00%	100,00%
80,00%	0,00%	97,50%	0,00%	100,00%
90,00%	0,00%	97,50%	0,00%	100,00%
97,40%	0,00%	97,50%	0,00%	100,00%
97,50%	1,96%	19,46%	80,00%	20,00%
100,00%	2,50%	0,00%	100,00%	0,00%

SMOTE

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	8,78%	97,64%	7,82%	97,54%
20,00%	4,95%	97,61%	8,82%	95,48%
30,00%	4,69%	97,61%	8,82%	95,28%
40,00%	4,66%	97,61%	8,82%	95,26%
50,00%	4,99%	97,63%	9,82%	95,13%
60,00%	4,48%	97,61%	9,82%	94,50%
70,00%	4,39%	97,63%	10,82%	93,90%
80,00%	4,71%	97,65%	11,82%	93,70%
90,00%	5,46%	97,77%	17,82%	92,19%
98,80%	3,36%	98,06%	52,18%	61,62%
100,00%	2,66%	97,78%	66,91%	37,80%

Random Oversampling

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	8,05%	97,77%	14,55%	95,86%
20,00%	6,49%	97,89%	22,55%	91,79%
30,00%	5,80%	97,91%	25,45%	89,48%
40,00%	5,69%	97,91%	25,45%	89,28%
50,00%	5,69%	97,91%	25,45%	89,28%
60,00%	5,69%	97,91%	25,45%	89,26%
70,00%	5,37%	97,89%	25,45%	88,40%
80,00%	5,37%	97,89%	25,45%	88,40%
90,00%	5,37%	97,89%	25,45%	88,40%
100,00%	5,37%	97,89%	25,45%	88,40%

MetaCost

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,0%	2,5%	0,0%	100,0%	0,0%
20,0%	2,5%	0,0%	100,0%	0,0%
30,0%	2,5%	0,0%	100,0%	0,0%
40,0%	2,5%	0,0%	100,0%	0,0%
50,0%	2,5%	0,0%	100,0%	0,0%
60,0%	2,5%	0,0%	100,0%	0,0%
70,0%	2,5%	0,0%	100,0%	0,0%
80,0%	2,5%	0,0%	100,0%	0,0%
90,0%	2,5%	0,0%	100,0%	0,0%
100,0%	2,5%	0,0%	100,0%	0,0%

Banco de Dados VI - Naive Bayes

Sem Tratamento

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	0,00%	97,50%	0,00%	99,85%
20,00%	0,00%	97,50%	0,00%	99,80%
30,00%	3,33%	97,51%	1,00%	99,47%
40,00%	7,50%	97,53%	1,91%	99,20%
50,00%	5,83%	97,55%	2,91%	98,85%
60,00%	8,53%	97,58%	4,91%	98,19%
70,00%	7,52%	97,59%	5,91%	97,54%
80,00%	7,77%	97,64%	8,82%	96,54%
90,00%	5,93%	97,81%	19,73%	91,89%
95,50%	5,82%	98,43%	50,82%	78,69%
99,80%	3,06%	99,14%	90,00%	27,13%
100,00%	2,50%	0,00%	100,00%	0,00%

SMOTE

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	6,62%	97,90%	23,36%	91,26%
20,00%	6,99%	98,15%	35,18%	87,90%
30,00%	5,85%	98,15%	38,09%	83,96%
40,00%	4,79%	98,13%	41,09%	79,09%
50,00%	4,34%	98,22%	47,91%	73,04%
58,70%	4,03%	98,22%	50,82%	69,00%
60,00%	3,96%	98,21%	50,82%	68,42%
70,00%	3,82%	98,28%	56,64%	63,15%
80,00%	3,96%	98,55%	66,64%	58,28%
90,00%	3,84%	98,74%	73,45%	52,73%
99,90%	2,90%	98,73%	85,09%	27,26%
100,00%	2,50%	0,00%	100,00%	0,00%

Random Oversampling

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	6,65%	98,18%	37,45%	86,55%
17,90%	6,48%	98,46%	50,00%	81,42%
20,00%	6,52%	98,50%	52,00%	80,77%
30,00%	6,46%	98,60%	55,82%	79,22%
40,00%	6,42%	98,69%	59,64%	77,46%
50,00%	5,46%	98,72%	63,55%	71,56%
60,00%	4,42%	98,90%	74,27%	58,36%
70,00%	3,84%	99,07%	83,18%	46,26%
80,00%	3,44%	99,12%	87,09%	37,12%
90,00%	3,33%	99,17%	89,09%	33,61%
99,10%	3,07%	99,26%	92,09%	25,50%
100,00%	2,50%	0,00%	100,00%	0,00%

MetaCost

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	6,06%	98,70%	60,45%	75,95%
20,00%	5,99%	98,69%	60,45%	75,63%
30,00%	5,94%	98,69%	60,45%	75,43%
40,00%	5,90%	98,69%	60,45%	75,28%
50,00%	5,81%	98,68%	60,45%	74,90%
60,00%	5,76%	98,68%	60,45%	74,67%
70,00%	5,71%	98,67%	60,45%	74,42%
80,00%	5,68%	98,67%	60,45%	74,22%
90,00%	5,71%	98,70%	61,45%	73,95%
99,90%	4,75%	98,61%	62,45%	67,72%
100,00%	2,50%	0,00%	100,00%	0,00%

Banco de Dados IV - Random Forest

Sem Tratamento

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	0,00%	97,50%	0,00%	100,00%
20,00%	0,00%	97,50%	0,00%	100,00%
30,00%	0,00%	97,50%	0,00%	99,90%
40,00%	0,00%	97,50%	0,00%	99,65%
50,00%	0,00%	97,49%	0,00%	99,35%
60,00%	0,00%	97,48%	0,00%	98,87%
70,00%	1,67%	97,49%	0,91%	98,37%
80,00%	4,50%	97,59%	6,82%	96,59%
90,00%	7,62%	97,93%	23,64%	92,54%
99,30%	5,24%	98,39%	51,73%	75,40%
100,00%	3,10%	98,32%	71,55%	42,67%

SMOTE

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	7,50%	97,56%	3,82%	98,27%
20,00%	4,17%	97,55%	4,82%	97,04%
30,00%	3,29%	97,54%	4,82%	96,36%
40,00%	2,99%	97,53%	4,82%	96,13%
50,00%	3,42%	97,55%	5,73%	95,81%
60,00%	3,79%	97,58%	7,73%	94,93%
70,00%	4,66%	97,65%	11,64%	93,68%
80,00%	4,67%	97,73%	17,55%	90,61%
90,00%	5,35%	98,07%	36,18%	83,16%
96,80%	4,00%	98,20%	50,82%	68,48%
100,00%	3,06%	98,39%	75,55%	38,66%

Random Oversampling

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	6,94%	97,64%	7,82%	97,44%
20,00%	5,42%	97,69%	12,82%	94,40%
30,00%	4,50%	97,67%	13,82%	92,57%
40,00%	4,72%	97,69%	14,82%	92,27%
50,00%	6,50%	97,88%	22,55%	91,62%
60,00%	6,19%	97,93%	25,55%	90,03%
70,00%	5,75%	97,96%	28,55%	87,90%
80,00%	4,88%	97,95%	31,45%	84,11%
90,00%	4,96%	98,17%	43,09%	78,41%
93,90%	4,72%	98,31%	50,82%	73,29%
100,00%	3,13%	98,42%	74,45%	40,89%

MetaCost

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	5,81%	97,76%	17,64%	92,19%
20,00%	5,99%	97,88%	23,55%	90,31%
30,00%	6,82%	98,10%	33,36%	88,15%
40,00%	6,55%	98,13%	35,36%	86,90%
50,00%	6,65%	98,19%	38,27%	85,84%
60,00%	6,50%	98,24%	41,18%	84,31%
70,00%	6,39%	98,32%	45,00%	82,66%
80,00%	5,96%	98,32%	46,00%	80,90%
87,70%	5,74%	98,42%	50,91%	78,24%
90,00%	5,50%	98,40%	50,91%	77,16%
100,00%	3,05%	98,50%	78,45%	36,19%

Banco de Dados VI - Ripper

Sem Tratamento

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	0,00%	97,50%	0,00%	100,00%
20,00%	0,00%	97,50%	0,00%	100,00%
30,00%	0,00%	97,50%	0,00%	100,00%
40,00%	0,00%	97,50%	0,00%	100,00%
50,00%	0,00%	97,50%	0,00%	100,00%
60,00%	0,00%	97,50%	0,00%	100,00%
70,00%	0,00%	97,50%	0,00%	100,00%
80,00%	0,00%	97,50%	0,00%	100,00%
90,00%	0,00%	97,50%	0,00%	100,00%
97,40%	0,00%	97,50%	0,00%	100,00%
97,50%	1,96%	19,46%	80,00%	20,00%
100,00%	2,50%	0,00%	100,00%	0,00%

SMOTE

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	7,81%	97,62%	6,82%	97,77%
20,00%	4,90%	97,61%	7,82%	96,18%
30,00%	4,72%	97,60%	7,82%	96,03%
40,00%	4,72%	97,60%	7,82%	96,03%
50,00%	4,72%	97,60%	7,82%	96,03%
60,00%	4,72%	97,60%	7,82%	96,03%
70,00%	4,72%	97,60%	7,82%	96,03%
80,00%	4,72%	97,60%	7,82%	96,03%
90,00%	4,72%	97,60%	7,82%	96,03%
98,10%	3,70%	68,29%	34,82%	67,26%
98,20%	2,25%	9,74%	90,00%	9,52%
100,00%	2,50%	0,00%	100,00%	0,00%

Random Oversampling

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	4,80%	97,65%	7,64%	98,04%
20,00%	5,63%	98,03%	30,64%	87,48%
30,00%	5,55%	98,11%	36,64%	83,59%
40,00%	5,55%	98,11%	36,64%	83,59%
50,00%	5,55%	98,11%	36,64%	83,59%
60,00%	5,55%	98,11%	36,64%	83,59%
70,00%	5,55%	98,11%	36,64%	83,59%
80,00%	5,55%	98,11%	36,64%	83,59%
90,00%	5,49%	98,11%	36,64%	83,21%
100,00%	5,49%	98,11%	36,64%	83,21%

MetaCost

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	2,50%	0,00%	100,00%	0,00%
20,00%	2,50%	0,00%	100,00%	0,00%
30,00%	2,50%	0,00%	100,00%	0,00%
40,00%	2,50%	0,00%	100,00%	0,00%
50,00%	2,50%	0,00%	100,00%	0,00%
60,00%	2,50%	0,00%	100,00%	0,00%
70,00%	2,50%	0,00%	100,00%	0,00%
80,00%	2,50%	0,00%	100,00%	0,00%
90,00%	2,50%	0,00%	100,00%	0,00%
100,00%	2,50%	0,00%	100,00%	0,00%

Base de Dados VI - SVM

Sem Tratamento

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	20,00%	97,55%	1,91%	99,82%
20,00%	15,00%	97,55%	1,91%	99,80%
30,00%	15,00%	97,55%	1,91%	99,75%
40,00%	18,33%	97,57%	2,91%	99,70%
50,00%	18,33%	97,57%	2,91%	99,67%
60,00%	18,33%	97,57%	2,91%	99,67%
70,00%	18,33%	97,57%	2,91%	99,62%
80,00%	11,67%	97,56%	2,91%	99,55%
90,00%	11,67%	97,56%	2,91%	99,45%
97,30%	5,76%	97,83%	50,64%	55,69%
97,70%	2,82%	98,17%	83,55%	25,53%
100,00%	2,50%	0,00%	100,00%	0,00%

SMOTE

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	6,96%	97,70%	11,73%	95,88%
20,00%	5,50%	97,69%	13,73%	93,27%
30,00%	5,09%	98,11%	37,18%	82,76%
40,00%	4,69%	98,08%	37,18%	81,30%
50,00%	4,51%	98,08%	39,09%	79,14%
60,00%	4,30%	98,10%	40,91%	77,18%
70,00%	4,29%	98,12%	42,82%	75,70%
74,90%	4,30%	98,31%	52,82%	69,70%
80,00%	3,40%	98,20%	56,73%	58,71%
90,00%	3,24%	98,23%	64,55%	50,20%
98,70%	2,85%	98,64%	85,09%	25,18%
100,00%	2,50%	0,00%	100,00%	0,00%

Random Oversampling

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	6,32%	97,66%	8,82%	97,41%
20,00%	6,50%	97,76%	14,82%	95,08%
30,00%	7,45%	98,13%	35,64%	86,70%
40,00%	5,04%	98,33%	49,00%	76,33%
45,60%	4,76%	98,33%	50,91%	73,74%
50,00%	4,69%	98,32%	50,91%	73,32%
60,00%	4,70%	98,53%	59,82%	68,65%
70,00%	4,38%	98,73%	68,64%	61,02%
80,00%	3,17%	99,01%	87,18%	31,17%
86,90%	2,96%	98,97%	89,09%	25,02%
90,00%	2,88%	98,95%	90,09%	22,16%
100,00%	2,50%	0,00%	100,00%	0,00%

MetaCost

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
0,10%	3,10%	98,66%	78,91%	36,28%
3,60%	2,99%	98,89%	88,55%	25,51%
10,00%	2,84%	99,17%	93,18%	17,55%
20,00%	2,81%	99,16%	93,18%	16,95%
30,00%	2,81%	99,15%	93,18%	16,85%
40,00%	2,80%	99,14%	93,18%	16,57%
50,00%	2,79%	99,13%	93,18%	16,35%
60,00%	2,76%	99,12%	93,18%	15,67%
70,00%	2,75%	99,08%	93,18%	15,19%
80,00%	2,76%	99,17%	94,09%	14,87%
90,00%	2,78%	99,28%	95,00%	14,44%
100,00%	2,51%	10,00%	100,00%	0,55%

Banco de Dados VII - C4.5

Sem Tratamento

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	0,00%	96,77%	0,00%	100,00%
20,00%	0,00%	96,77%	0,00%	100,00%
30,00%	0,00%	96,77%	0,00%	100,00%
40,00%	0,00%	96,77%	0,00%	100,00%
50,00%	0,00%	96,77%	0,00%	100,00%
60,00%	0,00%	96,77%	0,00%	100,00%
70,00%	0,00%	96,77%	0,00%	100,00%
80,00%	0,00%	96,77%	0,00%	100,00%
90,00%	0,00%	96,77%	0,00%	100,00%
96,70%	0,00%	96,77%	0,00%	100,00%
96,80%	3,23%	0,00%	100,00%	0,00%
100,00%	3,23%	0,00%	100,00%	0,00%

SMOTE

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	27,01%	97,80%	34,81%	96,85%
20,00%	20,10%	97,99%	41,92%	94,43%
30,00%	17,72%	98,09%	45,86%	92,90%
40,00%	16,36%	98,12%	47,14%	91,96%
50,00%	16,24%	98,13%	47,41%	91,84%
60,00%	16,25%	98,14%	47,75%	91,79%
70,00%	16,08%	98,15%	48,30%	91,58%
80,00%	15,24%	98,18%	49,52%	90,80%
81,00%	15,29%	98,20%	50,08%	90,73%
90,00%	14,58%	98,28%	53,08%	89,59%
100,00%	4,94%	99,56%	94,84%	39,01%

Random Oversampling

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	32,99%	98,14%	45,03%	96,93%
11,60%	30,28%	98,30%	50,08%	96,15%
20,00%	23,28%	98,70%	63,24%	93,04%
30,00%	19,26%	99,02%	73,35%	89,73%
40,00%	17,42%	99,15%	77,40%	87,75%
50,00%	17,07%	99,20%	78,96%	87,19%
60,00%	17,05%	99,20%	79,01%	87,16%
70,00%	16,96%	99,21%	79,12%	87,06%
80,00%	16,66%	99,22%	79,46%	86,73%
90,00%	16,29%	99,23%	79,90%	86,30%
100,00%	14,95%	99,23%	80,23%	84,76%

MetaCost

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	19,74%	97,43%	23,99%	96,05%
20,00%	19,89%	97,45%	24,38%	96,04%
30,00%	19,90%	97,45%	24,44%	96,03%
40,00%	19,93%	97,45%	24,49%	96,03%
50,00%	19,89%	97,45%	24,49%	96,01%
60,00%	20,02%	97,45%	24,60%	96,01%
70,00%	20,02%	97,45%	24,60%	96,01%
80,00%	20,02%	97,45%	24,60%	96,01%
90,00%	19,96%	97,45%	24,60%	95,99%
100,00%	11,36%	97,48%	30,49%	89,45%

Banco de Dados VII - Naive Bayes

Sem Tratamento

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	0,00%	96,77%	0,00%	100,00%
20,00%	10,00%	96,77%	0,06%	100,00%
30,00%	33,33%	96,78%	0,22%	99,98%
40,00%	36,67%	96,78%	0,39%	99,97%
50,00%	51,69%	96,82%	1,61%	99,94%
60,00%	51,84%	96,91%	4,55%	99,85%
70,00%	31,20%	96,99%	7,44%	99,44%
80,00%	18,67%	97,31%	19,66%	97,11%
90,00%	10,82%	97,60%	33,04%	90,86%
95,70%	7,37%	97,94%	50,14%	78,97%
99,00%	3,91%	98,60%	88,01%	27,81%
100,00%	3,23%	0,00%	100,00%	0,00%

SMOTE

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	24,70%	97,19%	14,77%	98,46%
20,00%	13,90%	97,43%	24,93%	94,81%
30,00%	10,52%	97,64%	34,59%	90,15%
40,00%	8,94%	97,81%	42,76%	85,44%
47,30%	8,14%	97,99%	50,14%	81,09%
50,00%	7,79%	98,04%	52,58%	79,21%
60,00%	6,65%	98,22%	61,13%	71,36%
70,00%	5,51%	98,41%	71,24%	59,28%
80,00%	4,43%	98,60%	82,84%	40,35%
87,30%	3,87%	98,75%	90,45%	25,11%
90,00%	3,70%	98,74%	92,45%	19,78%
100,00%	3,23%	0,00%	100,00%	0,00%

Random Oversampling

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	20,28%	97,26%	17,60%	97,65%
20,00%	12,05%	97,56%	30,82%	92,43%
30,00%	9,17%	97,77%	40,81%	86,48%
40,00%	7,90%	97,93%	48,47%	81,13%
41,20%	7,83%	97,97%	50,03%	80,32%
50,00%	7,06%	98,21%	59,63%	73,80%
60,00%	5,94%	98,47%	70,68%	62,71%
70,00%	4,82%	98,66%	81,07%	46,59%
79,20%	3,92%	98,89%	91,56%	25,18%
80,00%	3,86%	98,93%	92,45%	23,23%
90,00%	3,42%	99,44%	98,78%	7,06%
100,00%	3,23%	0,00%	100,00%	0,00%

MetaCost

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	8,16%	97,64%	37,92%	85,75%
20,00%	8,11%	97,64%	37,98%	85,63%
30,00%	8,03%	97,65%	38,48%	85,26%
40,00%	7,80%	97,69%	40,37%	84,05%
50,00%	7,48%	97,72%	42,31%	82,53%
60,00%	7,30%	97,88%	48,31%	79,49%
70,00%	7,15%	97,90%	49,69%	78,42%
72,20%	7,13%	97,92%	50,14%	78,19%
80,00%	6,82%	98,12%	57,64%	73,62%
90,00%	5,19%	98,21%	67,96%	58,59%
99,90%	4,12%	98,18%	77,90%	39,60%
100,00%	3,23%	0,00%	100,00%	0,00%

Banco de Dados VII - Random Forest

Sem Tratamento

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	90,60%	97,16%	12,49%	99,96%
20,00%	89,35%	97,31%	17,27%	99,93%
30,00%	81,17%	97,34%	18,21%	99,86%
40,00%	68,62%	97,44%	21,49%	99,67%
50,00%	56,59%	97,52%	24,26%	99,37%
60,00%	51,14%	97,74%	31,32%	99,00%
70,00%	41,82%	97,97%	39,14%	98,18%
80,00%	32,13%	98,30%	50,08%	96,46%
80,00%	32,13%	98,30%	50,08%	96,46%
90,00%	22,37%	98,78%	65,85%	92,36%
100,00%	13,10%	99,38%	84,84%	81,22%

SMOTE

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	32,48%	97,68%	30,26%	97,89%
20,00%	24,78%	97,92%	38,92%	96,05%
30,00%	20,94%	98,07%	44,25%	94,42%
40,00%	18,49%	98,15%	47,41%	93,03%
50,00%	18,04%	98,22%	49,86%	92,44%
50,40%	18,04%	98,23%	50,08%	92,40%
60,00%	17,99%	98,34%	53,36%	91,88%
70,00%	18,17%	98,47%	57,58%	91,33%
80,00%	18,11%	98,63%	62,13%	90,61%
90,00%	17,16%	98,94%	71,57%	88,45%
100,00%	12,56%	99,40%	85,40%	80,14%

Random Oversampling

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	36,94%	98,19%	46,25%	97,36%
11,80%	33,38%	98,31%	50,08%	96,65%
20,00%	25,17%	98,65%	61,41%	93,89%
30,00%	20,66%	98,98%	71,90%	90,78%
40,00%	18,35%	99,10%	75,79%	88,75%
50,00%	17,58%	99,17%	77,85%	87,81%
60,00%	17,39%	99,20%	78,73%	87,52%
70,00%	17,21%	99,23%	79,68%	87,21%
80,00%	16,99%	99,24%	80,12%	86,94%
90,00%	16,50%	99,26%	80,57%	86,38%
100,00%	13,11%	99,38%	84,79%	81,25%

MetaCost

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
0,10%	15,61%	98,61%	62,47%	88,73%
10,00%	16,73%	99,10%	76,13%	87,35%
20,00%	16,71%	99,16%	77,90%	87,04%
30,00%	16,62%	99,17%	78,23%	86,90%
40,00%	16,53%	99,19%	78,79%	86,72%
50,00%	16,48%	99,19%	78,79%	86,67%
60,00%	16,47%	99,19%	78,90%	86,64%
70,00%	16,38%	99,20%	78,96%	86,54%
80,00%	16,29%	99,20%	79,18%	86,42%
90,00%	16,02%	99,21%	79,51%	86,08%
100,00%	12,69%	99,34%	83,79%	80,76%

Banco de Dados VII - Ripper

Sem Tratamento

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	10,00%	96,77%	0,06%	100,00%
20,00%	10,00%	96,77%	0,06%	100,00%
30,00%	30,00%	96,78%	0,17%	99,99%
40,00%	39,52%	96,78%	0,44%	99,98%
50,00%	34,52%	96,79%	0,61%	99,97%
60,00%	34,52%	96,79%	0,61%	99,97%
70,00%	34,52%	96,79%	0,61%	99,97%
80,00%	34,52%	96,79%	0,61%	99,97%
90,00%	34,52%	96,79%	0,61%	99,97%
96,80%	35,17%	77,44%	20,61%	79,98%
96,90%	3,23%	0,00%	100,00%	0,00%
100,00%	3,23%	0,00%	100,00%	0,00%

SMOTE

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	10,00%	96,77%	0,06%	100,00%
20,00%	10,00%	96,77%	0,06%	100,00%
30,00%	30,00%	96,78%	0,17%	99,99%
40,00%	39,52%	96,78%	0,44%	99,98%
50,00%	34,52%	96,79%	0,61%	99,97%
60,00%	34,52%	96,79%	0,61%	99,97%
70,00%	34,52%	96,79%	0,61%	99,97%
80,00%	34,52%	96,79%	0,61%	99,97%
90,00%	34,52%	96,79%	0,61%	99,97%
96,80%	35,17%	77,44%	20,61%	79,98%
96,90%	3,23%	0,00%	100,00%	0,00%
100,00%	3,23%	0,00%	100,00%	0,00%

Random Oversampling

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	6,76%	97,05%	9,20%	99,40%
20,00%	4,61%	97,16%	12,85%	98,56%
25,10%	7,60%	98,37%	51,52%	84,76%
30,00%	10,40%	99,57%	90,35%	70,25%
40,00%	10,02%	99,60%	91,18%	69,81%
50,00%	9,91%	99,61%	91,34%	69,67%
60,00%	9,91%	99,61%	91,34%	69,67%
70,00%	9,91%	99,61%	91,34%	69,67%
80,00%	9,91%	99,61%	91,34%	69,67%
90,00%	9,91%	99,61%	91,34%	69,67%
100,00%	7,33%	79,75%	95,06%	52,49%

MetaCost

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	38,02%	97,06%	9,77%	99,47%
20,00%	38,01%	97,08%	10,27%	99,44%
30,00%	37,47%	97,08%	10,38%	99,43%
40,00%	37,15%	97,08%	10,38%	99,42%
50,00%	37,09%	97,08%	10,49%	99,41%
60,00%	37,09%	97,08%	10,49%	99,41%
70,00%	37,09%	97,08%	10,49%	99,41%
80,00%	37,09%	97,08%	10,49%	99,41%
90,00%	37,09%	97,08%	10,49%	99,41%
99,90%	37,09%	97,08%	10,49%	99,41%
100,00%	3,23%	0,00%	100,00%	0,00%

Base de Dados VII - SVM

Sem Tratamento

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	0,00%	96,77%	0,00%	100,00%
20,00%	0,00%	96,77%	0,00%	100,00%
30,00%	0,00%	96,77%	0,00%	100,00%
40,00%	7,80%	96,79%	0,50%	99,97%
50,00%	56,84%	96,89%	4,00%	99,90%
60,00%	56,84%	96,89%	4,00%	99,90%
70,00%	56,84%	96,89%	4,00%	99,90%
80,00%	56,38%	96,89%	4,00%	99,90%
90,00%	56,38%	96,89%	4,00%	99,90%
96,90%	3,57%	57,81%	77,26%	25,53%
100,00%	3,23%	0,00%	100,00%	0,00%

SMOTE

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	0,00%	96,77%	0,00%	100,00%
20,00%	16,20%	97,15%	14,33%	97,48%
30,00%	16,11%	97,15%	14,38%	97,45%
36,00%	9,64%	98,01%	52,35%	76,29%
40,00%	6,65%	98,25%	62,13%	70,86%
50,00%	6,61%	98,28%	63,24%	70,20%
60,00%	6,57%	98,29%	63,69%	69,70%
67,50%	4,56%	98,92%	84,50%	30,89%
70,00%	3,38%	99,25%	98,61%	5,94%
80,00%	3,38%	99,24%	98,61%	5,86%
90,00%	3,24%	9,96%	99,94%	0,52%
100,00%	3,23%	0,00%	100,00%	0,00%

Random Oversampling

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	0,00%	96,77%	0,00%	100,00%
20,00%	16,17%	97,15%	14,33%	97,48%
30,00%	16,17%	97,15%	14,33%	97,48%
35,90%	8,42%	98,08%	54,22%	75,28%
40,00%	6,57%	98,32%	64,52%	69,39%
50,00%	6,58%	98,32%	64,58%	69,38%
60,00%	6,58%	98,33%	64,63%	69,38%
65,90%	4,87%	99,13%	81,69%	36,14%
70,00%	3,32%	99,83%	99,83%	3,03%
80,00%	3,32%	99,83%	99,83%	2,99%
90,00%	3,23%	0,00%	100,00%	0,00%
100,00%	3,23%	0,00%	100,00%	0,00%

MetaCost

Fator de Confiança	Valor Preditivo Negativo	Valor Preditivo Positivo	Taxa de Verdadeiros Negativos	Taxa de Verdadeiros Positivos
10,00%	54,02%	96,90%	4,11%	99,89%
20,00%	54,02%	96,90%	4,11%	99,89%
30,00%	54,02%	96,90%	4,11%	99,89%
40,00%	54,02%	96,90%	4,11%	99,89%
50,00%	54,02%	96,90%	4,11%	99,89%
60,00%	54,02%	96,90%	4,11%	99,89%
70,00%	54,02%	96,90%	4,11%	99,89%
80,00%	54,02%	96,90%	4,11%	99,89%
90,00%	54,02%	96,90%	4,11%	99,89%
100,00%	51,98%	96,90%	4,17%	99,88%