



Universidade Federal da Bahia
Instituto de Matemática

Programa de Pós-Graduação em Ciência da Computação

**PREDIÇÃO DO DESFECHO CLÍNICO POR
LEPTOSPIROSE BASEADO NA ANÁLISE DE
EXPRESSÃO GÊNICA EM CASOS
HOSPITALIZADOS**

Nivison Ruy Rocha Nery Júnior

DISSERTAÇÃO DE MESTRADO

Salvador
9 de maio de 2017

NIVISON RUY ROCHA NERY JÚNIOR

**PREDIÇÃO DO DESFECHO CLÍNICO POR LEPTOSPIROSE
BASEADO NA ANÁLISE DE EXPRESSÃO GÊNICA EM CASOS
HOSPITALIZADOS**

Esta Dissertação de Mestrado foi apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal da Bahia, como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Orientadora: Profa. Dra. Daniela Barreiro Claro

Salvador
9 de maio de 2017

Sistema de Bibliotecas - UFBA

Nery Jr., Nivison Ruy Rocha.

Predição do Desfecho Clínico por Leptospirose baseado na Análise de Expressão Gênica em Casos Hospitalizados / Nivison Ruy Rocha Nery Júnior – Salvador, 2017.

81p.: il.

Orientadora: Prof. Dr. Profa. Dra. Daniela Barreiro Claro.

Dissertação (Mestrado) – Universidade Federal da Bahia, Instituto de Matemática, 2017.

1. Leptospirose. 2. Classificação. 3. Expressão gênica humana. I. Barreiro Claro, Daniela. II. Universidade Federal da Bahia. Instituto de Matemática. III Título.

CDD – XXX.XX

CDU – XXX.XX.XXX

TERMO DE APROVAÇÃO

NIVISON RUY ROCHA NERY JÚNIOR

PREDIÇÃO DO DESFECHO CLÍNICO POR LEPTOSPIROSE BASEADO NA ANÁLISE DE EXPRESSÃO GÊNICA EM CASOS HOSPITALIZADOS

Esta Dissertação de Mestrado foi julgada adequada à obtenção do título de Mestre em Ciência da Computação e aprovada em sua forma final pelo Programa de Pós-Graduação em Ciência da Computação da Universidade Federal da Bahia.

Salvador, 09 de MAIO de 2017

Profa. Dra. Fátima L. S. Nunes
Universidade de São Paulo - USP

Prof. Dra. Tatiane Nogueira Rios
Universidade Federal da Bahia - UFBA

Profa. Dra. Janet C Lindow
Montana State University

Prof. Dra. Daniela B. Claro
Universidade Federal da Bahia - UFBA

A minha mãe, Gisélia Gomes da Silva, por ser um exemplo de força e superação. A minha filha Beatriz, minha fonte de inspiração e motivação.

AGRADECIMENTOS

Agradeço a Deus por permitir a conclusão deste trabalho, me guiando e fortalecendo nos momentos difíceis. Agradeço a minha família: mãe, tia, irmãos, sobrinhos e em especial a minha esposa Elane e a minha filha Beatriz Sophia, por me incentivarem e principalmente por compreenderem o meu afastamento em muitos momentos, disponibilizando o seu apoio incondicional.

A todo o grupo de Leptospirose e Zika do Instituto Gonçalo Moniz, FIOCRUZ-BA, a minha família profissional, que tanto contribui para o meu crescimento e com muita dedicação não medem esforços em auxiliar a saúde pública em nosso país, principalmente na produção científica.

Agradeço aos Doutores professores Mitermayer Galvão dos Reis, Guilherme de Souza Ribeiro e Joice Neves Reis Pedreira por registrarem em cartas, recomendações que foram de grande relevância para o meu ingresso ao curso.

Agradeço aos Doutores Albert Ko, Federico Costa, José Hagan e Janet Lindow por todo apoio disponibilizado, as diversas reuniões e auxílios valorosos, fundamentais para a realização deste trabalho.

Agradeço em especial ao Doutor Luciano Kalabric, por me incentivar a fazer o Mestrado neste curso e, por direcionamento de Deus, apresentar-me a professora Daniela Barreiro Claro, a qual palavras não seriam suficiente para descrevê-la, principalmente pelos ensinamentos, orientação, paciência nos momentos de discordâncias e encontros motivadores. Agradeço aos membros do grupo FORMAS pelos encontros enriquecedores. Agradeço aos professores membros da banca de qualificação, Tatiane Nogueira e João Rocha Júnior, pelas críticas e sugestões que tanto contribuíram para a conclusão deste trabalho.

Finalmente, gostaria de agradecer ao Programa de Ciência da Computação do Instituto de Matemática, por abrirem as portas para que eu pudesse realizar este sonho que é a minha Dissertação de Mestrado.

"O homem se torna muitas vezes o que ele próprio acredita que é. Se insisto em repetir para mim mesmo que não posso fazer uma determinada coisa, é possível que acabe me tornando realmente incapaz de fazê-la. Ao contrário, se tenho a convicção de que posso fazê-la, certamente adquirirei a capacidade de realizá-la, mesmo que não a tenha no começo."

—MAHATMA GANDHI

RESUMO

A leptospirose é uma doença febril aguda, negligenciada, que atinge populações de diversas regiões tropicais do planeta. São aproximadamente 1 milhão de casos anuais de leptospirose no mundo, sendo que 5-15% destes casos podem desenvolver a forma grave, alcançando quase 60.000 óbito. As principais causas que levam a óbito os indivíduos doentes por leptospirose com as mesmas características clínicas dos sobreviventes ainda não foram identificadas. No entanto, estudos recentes indicam que a resposta imunológica difere entre indivíduos sobreviventes e mortos. A técnica de microarranjos em amostras de pacientes hospitalizados com desfechos de cura e óbito tem sido utilizada com a finalidade de identificar possíveis associações dos genes ao desfecho clínico. Isso demanda tempo por parte dos especialistas para analisarem a expressão gênica. Adicionalmente, os riscos de erro humano na análise empregada estimulam ainda mais a adoção de técnicas computadorizadas para o auxílio destas atividades. Neste sentido, o presente trabalho propôs o desenvolvimento de modelos de predição baseado em dados clínicos e epidemiológicos para auxiliar no diagnóstico da leptospirose. Uma metodologia para análise do conjunto completo de expressão gênica foi proposta com o intuito de predizer o desfecho clínico. Os resultados obtidos foram avaliados como relevantes por especialistas e podem contribuir para o desenvolvimento de novas abordagens terapêuticas para o tratamento de casos graves da leptospirose. Assim, os modelos resultantes deste trabalho podem auxiliar os profissionais de saúde na rotina diária do hospital, especialmente em áreas endêmicas de leptospirose, acelerando o tratamento e minimizando a exacerbação e mortalidade da doença.

Palavras-chave: Leptospirose, Classificação, Expressão gênica humana.

ABSTRACT

Leptospirosis is an acute, neglected, febrile illness that strikes populations of many tropical regions of the planet. There are approximately 1 million annual cases of leptospirosis worldwide, and 5-15% of these cases can develop in severe form, reaching almost 60,000 deaths. The main causes leading to death of leptospirosis patients with the same clinical characteristics of the survivors have not yet been identified. However, recent studies indicate that the immune response differs between surviving and dead individuals. The microarray technique in samples of hospitalized patients with cure and death outcomes has been used in order to identify possible associations of the genes to the clinical outcome. This takes time from experts to analyze gene expression. Additionally, the risks of human error in the analysis employed further stimulate the adoption of computerized techniques to aid these activities. In this sense, the present work proposed the development of prediction models based on clinical and epidemiological data to aid in the diagnosis of leptospirosis. A methodology for the analysis of the complete gene expression dataset was proposed in order to predict the clinical outcome. The results obtained were evaluated as relevant by specialists and may contribute to the development of new therapeutic approaches for the treatment of severe cases of leptospirosis. Thus, the models resulting from this work can help health professionals in the daily routine of the hospital, especially in endemic areas of leptospirosis, accelerating the treatment and minimizing the exacerbation and mortality of the disease.

Keywords: Leptospirosis, Classification, Human gene expression.

SUMÁRIO

Capítulo 1—Introdução	1
1.1 Proposta	2
1.2 Hipóteses	3
1.3 Publicações	3
1.4 Organização do Trabalho	4
Capítulo 2—Revisão Bibliográfica	7
2.1 Mineração de Dados	7
2.1.1 Agrupamento	8
2.1.1.1 Método de Particionamento	10
2.1.1.2 Método Hierárquico.	10
2.1.2 Classificação	11
2.1.3 Redução de Dimensionalidade	12
2.1.3.1 Agregação.	12
2.1.3.2 Seleção de Atributos.	13
2.1.4 Técnicas de Avaliação	13
2.2 Mineração de dados aplicada a Saúde	15
2.2.1 Leptospirose	17
2.2.2 Expressão de Genes aplicados a Leptospirose	18
2.2.2.1 Técnica de Análise de Expressão de Genes.	18
2.2.3 Agrupamento em dados de Microarranjos	19
2.3 Considerações Finais	20
Capítulo 3—Predição do Desfecho Clínico	21
3.1 Metodologia	22
3.1.1 Casos de leptospirose baseados em dados clínicos e epidemiológicos	22
3.1.2 Agrupamento de indivíduos doentes baseados nos desfechos clínicos	22
3.1.3 Predição do desfecho clínico na análise gênica	23
3.1.3.1 Método por Transcritos.	23
3.1.3.2 Método por Genes.	23
3.2 Considerações Finais	24

Capítulo 4—Análise dos dados para Leptospirose	25
4.1 Conjuntos de Dados	25
4.1.1 Base de dados de casos hospitalizados de Leptospirose	27
4.1.2 Análise dos atributos	27
4.1.3 Conjuntos de dados de expressão gênica	29
4.1.4 Conjunto com 389 transcritos	30
4.1.5 Conjunto completo com 50.575 transcritos	31
4.1.6 Conjunto completo com 32.080 genes	31
4.1.7 Redução da dimensionalidade	32
4.1.7.1 Análise de Componentes Principais - PCA	33
4.1.7.2 Volcano Plot.	36
4.1.7.3 Conjuntos de dados por transcritos e genes.	40
4.2 Parametrização dos Algoritmos	41
4.2.1 Algoritmos de Agrupamento	41
4.2.2 Algoritmos de Classificação	42
4.2.2.1 Avaliação das funções de <i>kernel</i> do SVM.	42
4.3 Considerações Finais	43
Capítulo 5—Experimentos	45
5.1 Experimento A	46
5.1.1 Testes preditivos	46
5.2 Experimento B	47
5.2.1 Teste 1	47
5.2.2 Teste 2	47
5.2.3 Teste 3	47
5.2.4 Teste 4	48
5.3 Experimento C	48
5.3.1 Classificação de dados de expressão gênica	48
5.4 Considerações Finais	48
Capítulo 6—Resultados	49
6.1 Resultado A	49
6.2 Resultado B	51
6.2.1 Resultado do Teste 1	51
6.2.2 Resultado do Teste 2	51
6.2.3 Resultado do Teste 3	54
6.2.4 Resultado do Teste 4	55
6.3 Resultado C	56
6.3.1 Resultado da Classificação dos dados por Transcritos	56
6.3.2 Resultado da Classificação dos dados por Genes	59
6.3.3 Genes diferentemente expressos	61
6.3.3.1 Por Transcritos	61

6.3.3.2	Por Genes	62
6.4	Discussão	64
6.5	Considerações Finais	65
Capítulo 7—Conclusões		67
7.1	Trabalhos futuros	69
Apêndice A—Genes e transcritos diferentemente expressos		71

LISTA DE FIGURAS

2.1	Etapas do Processo de KDD adaptada de (FAYYAD; PIATECKY-SHAPIRO; SMYTH, 1996b).	8
2.2	Métodos de Agrupamento adaptados de (HAN; KAMBER; PEI, 2011).	9
4.1	Resumo dos conjuntos de dados utilizados nesta dissertação.	26
4.2	Fluxo das etapas de seleção de características (a), Pré-processamento (b) e avaliação da relevância por especialista (c).	28
4.3	Conjuntos de dados utilizados no Experimento A.	29
4.4	Avaliação do perfil imunológico dos grupos de amostras obtida com a técnica de microarranjos (LINDOW et al., 2016).	30
4.5	Conjunto de dados utilizado no Experimento B.	31
4.6	Conjuntos de dados utilizados no Experimento C.	32
4.7	Screen Plot obtido com o PCA nas bases de dados por transcritos (A) e genes (B) e Diagrama de Veen comparando os atributos do PC1 e do PC2 das referidas base de dados.	36
4.8	Volcano Plot obtido na comparação das 13 amostras agudas de indivíduos sobreviventes e as 03 que foram a óbito no conjunto de dados por transcritos (A) e por genes (B).	37
4.9	Volcano Plot obtido na comparação das 13 amostras convalescentes de indivíduos sobreviventes e as 03 amostras agudas dos que foram a óbito no conjunto de dados por transcritos (A) e por genes (B).	38
4.10	Volcano Plot obtido na comparação das 13 amostras convalescentes de indivíduos sobreviventes e as 04 amostras de indivíduos saudáveis no conjunto de dados por transcritos (A) e por genes (B).	39
4.11	Diagrama de Venn obtido na comparação 01 e 02 do Volcano Plot. Parte A por transcritos e parte B por genes.	39
5.1	Resumo da metodologia da dissertação.	45
5.2	Fluxo dos experimentos e conjuntos de dados utilizados no experimento A.	46
6.1	Comparação da precisão de instâncias e do valor obtido com a estatística Kappa, por algoritmos e técnicas de fragmentação (NERY; CLARO; LINDOW, 2016).	49
6.2	Comparação de experimentos de sensibilidade e especificidade (JUNIOR; CLARO; LINDOW, 2017).	50
6.3	Resultado da aplicação do algoritmo Fuzzy.	53

6.4	Dendrograma apresentando três grandes grupos de amostras avaliadas na análise de expressão gênica.	54
6.5	Medida da Silhueta na avaliação do K-means com 3 clusters.	56
6.6	Comparação da acurácia dos conjuntos de dados por transcritos com os 04 grupos de amostras avaliados.	57
6.7	Comparação da acurácia dos conjuntos de dados por transcritos sem o grupo de amostras dos indivíduos saudáveis.	58
6.8	Comparação dos gráficos biplots obtidos com a base de 50.575 e com a base dos 51 transcritos.	58
6.9	Comparação da acurácia dos conjuntos de dados por genes com os 04 grupos de amostras avaliados.	59
6.10	Comparação da acurácia dos conjuntos de dados por genes sem o grupo de amostras dos indivíduos saudáveis.	60
6.11	Comparação dos gráficos biplots obtidos com a base de 32080 e com a base dos 30 genes.	60
6.12	Diagrama de Veen comparando os genes diferentemente expressos citados no trabalho de (LINDOW et al., 2016) com os transcritos do dataset 51T.	64
6.13	Diagrama de Veen comparando os genes diferentemente expressos entre os dataset 51T e dataset 30G.	65

LISTA DE TABELAS

2.1	Métodos de Agrupamento	9
2.2	Exemplo de uma Matriz de confusão	13
4.1	Grupos avaliados por técnicas convencionais e análise manual por profissionais especializados em expressão genética do IGM/FIOCRUZ-BA.	31
4.2	Aplicação do PCA no conjunto de dados dos 50.575 transcritos.	34
4.3	Aplicação do PCA no conjunto de dados dos 32.080 genes.	35
4.4	Conjuntos de dados por transcritos utilizados nos experimentos	40
4.5	Conjuntos de dados por genes utilizados nos experimentos	41
4.6	Comparação da acurácia do SVM em relação aos 4 tipos de kernel avaliados em um conjunto de dados com 33 amostras.	42
4.7	Comparação da acurácia do SVM em relação aos 4 tipos de kernel avaliados em um conjunto de dados com 29 amostras.	43
6.1	Comparação dos grupos obtidos com a aplicação do algoritmo K-means com k=4 em relação aos tipos de amostras avaliadas pela técnica de microarranjos	51
6.2	Grau de pertinência das amostras por grupo obtido com a aplicação do algoritmo Fuzzy.	52
6.3	Comparação dos grupos obtidos com a aplicação do algoritmo Fuzzy com k=4 em relação aos tipos de amostras avaliadas pela técnica de microarranjos	53
6.4	Comparação dos grupos obtidos com a aplicação do algoritmo K-means com k=3 em relação aos tipos de amostras avaliadas pela técnica de microarranjos	55
6.5	Comparação das métricas de avaliação dos Grupos	55
6.6	Matriz de Confusão obtida com a aplicação do SVM no dataset 51T. . .	57
6.7	Matriz de Confusão obtida com a aplicação do SVM no dataset 30G. . .	59
6.8	Principais transcritos diferentemente expressos pertencentes ao dataset 51T.	61
6.9	Genes diferentemente expressos pertencentes ao dataset 30G (Parte 1). .	62
6.10	Genes diferentemente expressos pertencentes ao dataset 30G (Parte 2). .	63
A.1	Transcritos diferentemente expressos pertencentes ao dataset 51T (Parte 1).	71
A.2	Transcritos diferentemente expressos pertencentes ao dataset 51T (Parte 2).	72
A.3	Transcritos diferentemente expressos pertencentes ao dataset 51T (Parte 3).	73
A.4	Transcritos diferentemente expressos pertencentes ao dataset 51T (Parte 4).	74
A.5	Genes diferentemente expressos pertencentes ao dataset 30G (Parte 3). .	75

Capítulo

1

Neste capítulo são descritas a contextualização, motivação e proposta deste trabalho.

INTRODUÇÃO

A leptospirose, uma zoonose causada por uma espiroqueta denominada *Leptospira*, é uma doença febril aguda, classificada como negligenciada, sendo um desafio para a saúde pública em comunidades urbanas. Atualmente, esta doença tem sido a principal causa zoonótica de morbidade e mortalidade em todo o mundo. Historicamente, a leptospirose era uma doença de base rural, de agricultores de subsistência e pastores (MYERS; VARELA-DIAZ, 1979). Porém, a expansão das áreas urbanas e o consequente aumento da população humana residindo em comunidades carentes, sem estrutura sanitária e coleta de lixo, favoreceram a transmissão urbana dessa doença. Nessas áreas, essa transmissão ocorre principalmente pelo contato com água ou solo contaminado com urina de roedores sinantrópicos infectados.

No mundo, estima-se que ocorram aproximadamente um milhão de casos de leptospirose por ano, dos quais 15% apresentam a forma grave (HAGAN et al., 2015), ocorrendo 60.000 óbitos por ano desta doença. No Brasil, 10.000 casos são reportados por ano pelo sistema de notificação, com taxa de letalidade maior que 10% dos casos notificados (KO et al., 1999a). A taxa de óbito aumenta para 50% quando o indivíduo desenvolve a síndrome de hemorragia pulmonar grave, levando à morte em aproximadamente 48 horas (GOUVEIA et al., 2008).

A causa do óbito em indivíduos com leptospirose grave com as mesmas características dos sobreviventes ainda não foi identificada. Uma das hipóteses existentes está relacionada à diferença da resposta imune em pacientes com desfecho de sobrevivência em relação aos que foram a óbito.

Com o objetivo de entender as diferenças entre indivíduos com a forma grave da doença que sobreviveram e os que foram a óbito, o grupo de Leptospirose do Instituto Gonçalo Moniz (IGM/FIOCRUZ-BA) em parceria com a Universidade americana de Yale realizou diversas pesquisas sobre a doença. Em uma destas pesquisas, foram avaliados o genoma humano de indivíduos com e sem a doença, com o objetivo de verificar se há ou não diferenças no sistema imunológico entre os grupos investigados. O genoma humano

possui informações genéticas sobre o desenvolvimento e funcionamento do organismo, possuindo cerca de 20-30 mil genes (VENTER et al., 2001).

Os genes produzem o RNA (*ribonucleic acid*) mensageiro (mRNA) que é traduzido em proteína. O nível de uma sequência particular de RNA mensageiro pode ser determinado utilizando uma sonda ou transcritos, o qual contém a sequência complementar exata do RNA mensageiro (DENOUE et al., 2008).

A quantificação de todos os níveis de mRNAs produzidos em um tempo específico por todos os genes no genoma ou por um subgrupo particular de célula, em sangue total ou em tecidos específicos, chama-se transcrição. A identificação de perfil específico de todos os genes em grupos de indivíduos que compartilham características comuns de uma doença pode ser um importante método para a avaliação de marcadores específicos, principalmente para doenças graves, como a leptospirose.

Dentre as diversas técnicas utilizadas para analisar a expressão gênica, a mais usada é a do real-time PCR (AHMED et al., 2009), pois, por meio dela é possível analisar a expressão de até 4 genes simultaneamente. Porém, essa técnica torna-se inviável quando se objetiva analisar o perfil de uma expressão gênica no nível de genoma, já que um genoma bacteriano pode conter cerca de 4000 genes, sendo necessária a realização de 1000 análises. Em se tratando de uma análise do genoma, como por exemplo o humano, a técnica indicada é a de microarranjos (BRAZMA et al., 2001), pois é uma técnica de hibridização de ácidos nucleicos em grande escala.

Enquanto a abordagem de microarranjos aumenta as chances de identificar biomarcadores, que são indicações de ocorrência de uma determinada função de um organismo (VOLP et al., 2008). A técnica de microarranjos também produz uma grande quantidade de dados, requerendo instrumentos de análise especializados, a maioria dos quais com grandes limitações, resultando em análises manuais por parte dos profissionais de expressão de genes. Assim, o presente trabalho visa auxiliar a análise de identificação de biomarcadores com diferença de expressão entre indivíduos doentes por leptospirose que foram a óbito em comparação aos que sobreviveram.

1.1 PROPOSTA

No estudo do IGM/FIOCRUZ-BA foram realizadas análises de expressão gênica de 20 indivíduos, sendo: 13 com amostras de sangue coletadas na fase aguda (dentro de 72 horas de internação) e na fase convalescente (32-90 dias pós-admissão) da doença dos indivíduos sobreviventes; 3 amostras coletadas na fase aguda, mas que os indivíduos foram a óbito; e amostras de 4 indivíduos saudáveis. O presente trabalho tem como principal objetivo utilizar algoritmos da Mineração de Dados (MD) (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996a) para identificar os genes que apresentam diferenças genéticas para sobrevivência ou morte de indivíduos com leptospirose.

Inicialmente, o presente trabalho avaliou a identificação de casos da doença baseado em dados clínicos e epidemiológicos. Em seguida, utilizou-se algoritmos de agrupamento para confirmar a suspeita de que há diferenças genéticas entre os grupos de amostras avaliadas e posteriormente o uso de algoritmos de classificação para identificar os genes que apresentam esta diferença. Além disso, características específicas da base de dados

demandaram a utilização de técnicas de redução da dimensionalidade e ajustes específicos nos algoritmos utilizados.

As avaliações permitiram validar os agrupamentos dos genes e corroborar com a identificação dos mais relevantes na determinação da morte por leptospirose.

Além das contribuições científicas obtidas com este trabalho (NERY; CLARO; LINDOW, 2016; JUNIOR; CLARO; LINDOW, 2017), observou-se contribuições tecnológicas, pois os algoritmos podem ser utilizados para o auxílio dos profissionais da saúde, tornando o processo de identificação de casos da doença mais célere, bem como a descoberta dos genes diferentemente expressos podem contribuir com o desenvolvimento de melhores antibióticos ou até vacinas. Outrossim, são as contribuições sociais, visto que estes resultados podem melhorar o tratamento, consequentemente minimizando o agravamento e por sua vez o desfecho fatal de pacientes hospitalizados pela doença.

1.2 HIPÓTESES

Este trabalho propôs validar três hipóteses:

1. O uso de algoritmos de classificação baseado em dados clínicos e epidemiológicos pode auxiliar profissionais de saúde na identificação de casos de leptospirose.
2. A caracterização do desfecho clínico em pacientes doentes por leptospirose pode ser baseada na similaridade da expressão gênica.
3. O uso de algoritmos de classificação de desfecho clínico baseado em dados de expressão gênica pode auxiliar no tratamento de casos hospitalizados por leptospirose.

1.3 PUBLICAÇÕES

Alguns dos resultados parciais desta dissertação já foram publicados em conferência e revista científica, como pode ser observado abaixo:

- Nery, Nivison Ruy R., Daniela Barreiro Claro, and Janet C. Lindow. “Classification model analysis for the prediction of leptospirosis cases.” Information Systems and Technologies (CISTI), 2016 11th Iberian Conference on. IEEE, 2016 (NERY; CLARO; LINDOW, 2016). Neste trabalho foi realizada uma análise dos algoritmos de classificação, dos métodos de Árvore de Decisão, Regras de Classificação e Classificação Bayesiana, com a finalidade de identificar o melhor modelo de predição de casos da doença baseado em dados clínicos e epidemiológicos. Dos algoritmos avaliados, o JRIP foi o modelo com melhor desempenho, obtendo 85% de sensibilidade e 81% de especificidade. Os algoritmos previram com sucesso a doença e podem representar uma nova ferramenta para auxiliar os profissionais de saúde na rotina diária do hospital, especialmente em áreas endêmicas de leptospirose, acelerando o tratamento direcionado e minimizando a exacerbação e mortalidade da doença.
- Junior, Nivison, Daniela Claro, and Janet Lindow. “Prediction of Leptospirosis Cases using Classification Algorithms.” IET Software Jornal (2017) (JUNIOR;

CLARO; LINDOW, 2017). Neste trabalho foi realizada uma análise de casos de leptospirose clinicamente e epidemiologicamente definidos para predizer a doença utilizando algoritmos de classificação. Foram conduzidos quatro conjuntos de experimentos para avaliar o desempenho dos algoritmos, usando diferentes conjuntos de dados de treinamento e teste. O algoritmo JRIP atingiu 84% de sensibilidade usando um conjunto de dados apenas de casos confirmados de leptospirose e uma especificidade de 99% usando um conjunto de dados apenas de casos confirmados de dengue. Portanto, esta abordagem previu, com sucesso, casos de leptospirose, diferenciando-os de doenças febris semelhantes, podendo representar uma nova ferramenta para auxiliar profissionais de saúde, particularmente em áreas endêmicas da doença.

Os últimos resultados obtidos nesta dissertação estão em fase de organização para serem publicados como segue abaixo:

- Nery, Nivison Ruy R., Daniela Barreiro Claro, and Janet C. Lindow. “Characterization of Leptospirosis Clinical Outcomes Based on Similarity of Gene Expression”. Neste trabalho descrevemos os experimentos e resultados obtidos com a hipótese 2 desta dissertação. Utilizamos um subconjunto de dados, contendo 389 transcritos, obtidos pela técnica de microarranjos de amostras de pacientes hospitalizados, com desfecho de cura e morte, para identificar possíveis associações de genes que diferenciam cada grupo. A identificação de grupos ocorreu de acordo com a similaridade de expressão genética, a fim de caracterizar as amostras avaliadas por uso de algoritmos de agrupamento. Os resultados obtidos com a implementação dos algoritmos ratificaram a distinção do sistema imunológico entre os indivíduos sobreviventes e aqueles que foram a óbito.
- Nery, Nivison Ruy R., Daniela Barreiro Claro, and Janet C. Lindow. “Prediction of Clinical Outcome by Leptospirosis Based on Gene Expression Analysis in Hospitalized Cases”. Neste trabalho avaliamos o uso de um conjunto de dados completo por transcritos para melhor predizer o desfecho clínico de casos hospitalizados por leptospirose. Adicionalmente foram avaliadas técnicas de redução da dimensionalidade com o propósito de auxiliar os profissionais de saúde na análise dos genes responsáveis pelo desfecho fatal proveniente da doença. Este trabalho será a última contribuição desta dissertação, estando em fase de conclusão e revisão por parte dos autores.

1.4 ORGANIZAÇÃO DO TRABALHO

Esta dissertação está organizada em capítulos: o capítulo 2 apresenta a fundamentação teórica que contribuiu para o entendimento deste trabalho; o capítulo 3 descreve os objetivos e a metodologia utilizada; o capítulo 4 apresenta os conjuntos de dados utilizados para a análise de expressão gênica, bem como as parametrizações utilizadas nos algoritmos avaliados; o capítulo 5 apresenta os 3 conjuntos de experimentos realizados para avaliação deste trabalho; o capítulo 6 descreve os resultados, bem como discute-os; por fim, no

capítulo 7 são apresentadas as considerações finais do trabalho, bem como as contribuições obtidas.

Capítulo

2

Este capítulo apresenta a fundamentação teórica, que permite melhor compreender os conceitos apresentados neste trabalho e alguns dos principais trabalhos relacionados.

REVISÃO BIBLIOGRÁFICA

Neste capítulo serão apresentados os trabalhos encontrados nas bases de dados do IEEE (*Institute of Electrical and Electronics Engineers*), ACM (*Association for Computing Machinery*) e MEDLINE (*Medical Literature Analysis and Retrieval System Online*) relacionados a aprendizagem de máquina, mais especificamente nas técnicas de classificação e agrupamento de dados, com o intuito de auxiliar no diagnóstico e prognósticos acerca de doenças.

2.1 MINERAÇÃO DE DADOS

Extrair informações relevantes a partir de uma base de dados é uma tarefa complexa, executada há muitos anos, mas que a cada dia, empresas de diversas áreas buscam realizar, com a intenção de inovar e até mesmo se destacar perante seus concorrentes. O processo de descoberta de conhecimento em base de dados é conhecido na literatura como *Knowledge Discovery in Databases* (KDD). O KDD é um processo criado desde a década de 80, cuja finalidade é de transformar dados de baixo nível em conhecimento (GOEBEL; GRUENWALD, 1999).

Para realizar a descoberta de conhecimento, o KDD possui diversas etapas, como pode ser visto na Figura 2.1 (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996b). Destas etapas, a Mineração de dados (MD) é tida por muitos autores como a principal de todo o processo de descoberta de conhecimento (GOLDSCHMIDT; PASSOS, 2005), com métodos e ferramentas incorporadas de outros domínios como estatísticas, aprendizagem de máquina, reconhecimento de padrões, sistemas de banco de dados e de armazenamento, recuperação de informação, visualização, algoritmos, computação de alto desempenho, e muitos domínios de aplicação (HAN; KAMBER; PEI, 2011).



Figura 2.1: Etapas do Processo de KDD adaptada de (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996b).

Segundo (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996a), a Mineração de Dados é uma etapa no processo de descoberta de conhecimento, ao qual são realizadas análises e aplicações de algoritmos com o propósito de identificar padrões a partir de um conjunto de dados. Já (BERRY; LINOFF, 2004) definem MD como a exploração e a análise por meio automático ou semiautomático de grandes quantidades de dados a fim de descobrir padrões e regras relevantes, que auxiliam especialistas da área do conhecimento na avaliação das informações identificadas.

A mineração de dados possui duas atividades consideradas de alto nível conhecidas como: atividades preditivas e descritivas (CORTES; PORCARO; LIFSCHITZ, 2002). A atividade preditiva tem a finalidade de aprender com as instâncias dos atributos do banco de dados visando prever valores desconhecidos ou futuros de outros atributos de interesse. Já a atividade descritiva tem a função de encontrar padrões e tendências interpretáveis por humanos para descrever os dados. Enquanto que a atividade preditiva pode ser subdividida em tarefas de Classificação e Regressão; A atividade descritiva pode ser subdividida em tarefas de Associação, Agrupamento, Sumarização e entre outras que divergem na literatura. Neste trabalho foram utilizados algoritmos das tarefas de Agrupamento e Classificação.

2.1.1 Agrupamento

Um dos principais papéis do agrupamento é ajudar os especialistas de uma determinada área de atuação a analisar, descrever e utilizar informações valiosas identificadas dentro dos grupos.

Segundo (HAN; KAMBER; PEI, 2011), o agrupamento é o processo de particionar um conjunto de objetos em subconjuntos. Cada subconjunto é uma coleção de objetos que são similares a outro objeto dentro do grupo e dissimilares a objetos de outros grupos. A tarefa de agrupamento também é conhecida na literatura como *Cluster* ou *Clustering*.

Cluster é também chamado de segmentação de dados em algumas aplicações, porque particionam os grandes conjuntos de dados em grupos de acordo com a sua similaridade. *Clusters* também são utilizados para detecção de discrepâncias, que são valores fora de qualquer *cluster*, utilizados em detecção de fraudes de cartões de crédito e monitoramento de atividades criminosas em comércio eletrônico (HAN; KAMBER; PEI, 2011).

A tarefa de agrupamento é conhecida como aprendizagem não supervisionada, porque

a classe etiquetada não está presente no início do processo.

Com a finalidade de identificar o melhor algoritmo para a tarefa de agrupamento dos transcritos, foi realizado um estudo na literatura dos principais métodos de agrupamento, sendo identificados cinco tipos: Particionamento, Hierárquico, baseado em Densidade, em Grade e em Modelo. A Figura 2.2 apresenta os cinco métodos e seus respectivos tipos ou abordagens.

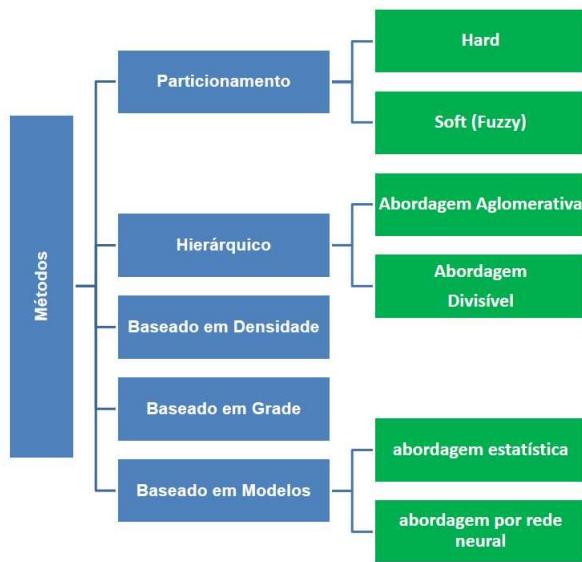


Figura 2.2: Métodos de Agrupamento adaptados de (HAN; KAMBER; PEI, 2011).

A Tabela 2.1 apresenta os métodos de agrupamento (HAN; KAMBER; PEI, 2011) e exemplos de algoritmos identificados na revisão da literatura.

Tabela 2.1: Métodos de Agrupamento

Método	Type	Exemplo de Algoritmos
Particionamento	Hard	PAM, CLARA, CLARANS, K-means
	Soft	Fuzzy C-Means
Hierárquico	Aglomerativa	AGNES, BIRCH, CURE, CHAMELEON e ROCK
	Divisível	DIANA
Baseado em Densidade		DBSCAN, OPTICS e DENCLUE
Baseado em Grade		STING, CLIQUE e WaveCluster
Baseado em Modelos	Estatísticos	COBWEB, CLASSIT e AutoClass
	Rede Neural	SOMs

2.1.1.1 Método de Particionamento Dado um conjunto de N objetos, um método de particionamento constrói K partições de dados, onde cada partição representa um *cluster* e $K \leq N$, sendo que N é a divisão de dados em K grupos, cada grupo precisa ter pelo menos um objeto. Em outras palavras, métodos de particionamento conduz um nível de divisão de conjunto de dados.

Geralmente o método de particionamento tipicamente adota separação exclusiva de *cluster*, que cada objeto precisa pertencer exatamente a um grupo. Este requisito pode ser flexível com a utilização da técnica de particionamento fuzzy.

K-means. É um algoritmo de particionamento, onde cada centro dos clusters é representado pelo valor médio da distância dos objetos no *cluster* (HAN; KAMBER; PEI, 2011). O algoritmo k-means define o centroide de um cluster com a média do valor dos objetos dentro do cluster. Primeiro, randomicamente seleciona K dos objetos em D , onde D é o conjunto de dados e K o número de *clusters*, onde inicialmente cada um representa uma média ou centro do *cluster* (KANUNGO et al., 2002). Para cada remanejamento de objetos, um objeto é avaliado, para que todos os objetos sejam similares no *cluster*, baseado na métrica de distância entre os objetos e a média do *cluster*. O algoritmo k-means provê iterativamente a variação dentro do *cluster*. Para cada *cluster*, ele computa uma nova média usando os objetos avaliados para o *cluster* em iterações prévias. Todos os objetos são então reatribuídos usando uma atualização média com o novo centro do *cluster*. As iterações continuam até se tornarem estáveis, que ocorre quando os *clusters* formados na rodada atual são os mesmos formados na rodada anterior, ou quando atingir algum critério de parada.

O algoritmo k-means é sensível a discrepâncias, quando objetos estão longe da maioria dos outros objetos, e assim, são atribuído para um *cluster* estes objetos podem distorcer dramaticamente o valor médio do *cluster*. Isto inadvertidamente afeta a atribuição de outros objetos para o *cluster*.

Fuzzy C-means. É um algoritmo de agrupamento flexível que permite um objeto pertencer dois ou mais grupos. O algoritmo foi desenvolvido por Dunn em 1973 e melhorado por Bezdek em 1981 e é frequentemente utilizado no reconhecimento de padrões. O algoritmo Fuzzy C-means (FCM) tenta partitionar um conjunto finito de pontos em uma coleção de C conjuntos difusos em relação a alguns critérios estabelecidos (BEZDEK; EHRLICH; FULL, 1984). Assim, os objetos na borda de um *cluster*, podem ter no *cluster* um menor grau do que os objetos do centro do grupo.

2.1.1.2 Método Hierárquico. Este método cria uma decomposição hierárquica dos conjuntos de objetos. Um método hierárquico pode ser classificado como aglomerativo ou divisível (HAN; KAMBER; PEI, 2011). A abordagem aglomerativa, também chamada de *bottom-up* (de baixo para cima), inicia com cada objeto formando um grupo separado. Sucessivamente objetos ou grupos fechados fazem *merge* (juntam) para outro, até todos os grupos estarem juntos, até um se tornar o maior nível hierárquico ou atendam algum critério de parada. A abordagem divisível, também conhecida como *top-down*, inicia com todos os objetos no mesmo *cluster*, em cada sucessível interação, um *cluster* é dividido

em pequenos *clusters*, até eventualmente cada objeto pertencer a um *cluster* ou atender uma condição de parada.

Clusters hierárquicos podem ser baseados em distância ou densidade e baseado em continuidade. Métodos hierárquicos param de fato quando uma etapa (*merge* ou *split*) (mesclagem ou divisão) é concluída (HAN; KAMBER; PEI, 2011).

Cluster Hierárquico. Neste estudo foi utilizada a abordagem aglomerativa, iniciando cada objeto como seu próprio conjunto e, em seguida, o algoritmo procede de forma iterativa, em cada fase junta-se os dois agrupamentos mais semelhantes, continuando até que haja apenas um único conjunto. Em cada fase, as distâncias entre os agrupamentos são recalculadas pela fórmula da atualização de dissimilaridade de acordo com o método de agrupamento utilizado (OKSANEN, 2010).

2.1.2 Classificação

Classificação é utilizada para criar modelos que descrevem classes de dados importantes. Um modelo classificador é construído para prever uma classe etiquetada com base em um conjunto de treinamento, no processo conhecido como aprendizagem supervisionada, devido à indicação da classe que as observações do conjunto de dados pertence (HAN; KAMBER; PEI, 2011). Após o aprendizado, o modelo precisa ser avaliado em um conjunto de teste, com o intuito de analisar a acurácia do classificador, ou seja, avaliar a taxa de acertos de classificação do modelo baseado na classe de referência.

Na literatura podem ser encontrados diversos algoritmos de classificação, todos com características distintas, que podem tornar alguns deles mais adequado para seus propósitos do que outros. Abaixo são citados alguns algoritmos agrupados por métodos:

- Métodos baseados em Distâncias, cuja hipótese é que os dados similares tendem a estar concentrados em uma mesma região, distantes dos dados não similares (FACELI et al., 2011). Exemplos de algoritmos deste método: 1 vizinho mais próximo (1-NN) e k-NN.
- Métodos Probabilísticos, utiliza algoritmos baseados no teorema de Bayes, como exemplo o Naive Bayes, considerado ingênuo por tratar os atributos como independentes (FACELI et al., 2011).
- Método baseado em Procura, utiliza uma função de avaliação de hipóteses para solução do problema de aprendizagem de máquina (FACELI et al., 2011). O J48, a uma implementação em java do algoritmo c4.5 (BHARGAVA et al., 2013; WITTEN; FRANK, 2005), é um exemplo de algoritmo do tipo de árvore de decisão que utiliza o método de procura, assim como o JRIP (WITTEN; FRANK, 2005) é um algoritmo do tipo de Regras de Decisão, o qual avalia a condição para resolução do problema, assumindo uma relação entre um atributo e os valores do domínio (FACELI et al., 2011).
- Métodos baseados em Otimização, cujo intuito é minimizar ou maximizar uma função objetivo. Como exemplo as Máquinas de Vetores de Suporte (SVM). No

processo de treinamento, o SVM envolve a solução de otimização quadrática, cuja finalidade é de maximizar a margem de separação entre os objetos de diferentes classes (FACELI et al., 2011).

Os algoritmos utilizados neste trabalho foram: Naive Bayes (NB), J48, JRIP, REP-Tree, OneR, PART, DecisionTable (DT) e SVM.

Algoritmo de Máquina de Vetores de Suporte. O algoritmo escolhido para a tarefa de classificação dos dados de expressão de genes foi o Máquina de Vetores de Suporte, do inglês, *Support Vector Machines* (SVM). A alta dimensionalidade do conjunto de dados de expressão gênica utilizado neste trabalho foi um fator decisivo para escolha do algoritmo SVM. O SVM além de apresentar uma robustez em conjuntos de dados com alta dimensionalidade, apresenta bom desempenho preditivo na tarefa de classificação (FACELI et al., 2011). Apesar do bom desempenho do SVM em dados com alta dimensionalidade, foram avaliadas técnicas de redução do conjunto de dados com a finalidade avaliar o desempenho do algoritmo e auxiliar na identificação dos genes relevantes.

2.1.3 Redução de Dimensionalidade

Um conjunto de dados de expressão gênica quando é trabalhado com as amostras como objetos e os genes como atributos, torna-se um conjunto com alta dimensionalidade. Segundo (FACELI et al., 2011), um conjunto de dados com muitos atributos pode prejudicar o desempenho dos algoritmos de aprendizagem de máquina. Exemplificando: cada atributo representa uma coordenada no espaço d-dimensional, em que d é o número de atributos, e, a consequente adição de novos atributos acarretaria em um aumento exponencial do volume de dados que esta dimensão representa. Segundo (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996a), um conjunto de dados de alta dimensionalidade cria problemas em termos de aumentar o tamanho do espaço de pesquisa para a indução do modelo, aumentando as chances de que um algoritmo encontre falsos padrões. A redução da dimensionalidade tem como objetivo melhorar o desempenho do algoritmo, reduzindo o custo computacional e tornando mais compreensíveis os resultados. Ainda segundo (FACELI et al., 2011), técnicas de diferentes áreas, como reconhecimento de padrões, teoria da informação e estatística podem ser utilizadas para a redução de um conjunto de dados. As técnicas de redução da dimensionalidade podem ser divididas em duas grandes abordagens: Agregação e Seleção de Atributos (FACELI et al., 2011).

2.1.3.1 Agregação. As técnicas de agregação substituem os atributos originais por combinações de atributos, através de funções lineares ou não lineares. Neste trabalho, a técnica de agregação utilizada foi a Análise de Componentes Principais (ACP ou em inglês PCA, de Principal Component Analysis) (LÊ; JOSSE; HUSSON, 2008) que correlacionam estatisticamente os exemplos com a finalidade de eliminar redundâncias do conjunto de dados.

O PCA é um método multivariado para a redução da dimensionalidade dos dados. Durante a sua execução, uma análise do conjunto de dados numéricos é realizada com a

finalidade de reduzir o número de atributos. No processo de redução, o PCA mantém a máxima variabilidade dos dados com o intuito de minimizar a perda de informação. Ao final do processo, ocorre o agrupamento das maiores variações do conjunto de dados nos primeiros componentes.

2.1.3.2 Seleção de Atributos. A seleção de atributos tem como objetivo remover atributos irrelevantes e redundantes. Neste trabalho foi utilizado o Volcano Plot (CUI; CHURCHILL, 2003) como técnica de seleção de atributos. O uso do Volcano Plot permite realizar teste de hipótese para avaliar a significância estatística e por sua vez, identificar quais atributos são relevantes na comparação dos grupos avaliados.

2.1.4 Técnicas de Avaliação

Nesta dissertação foram utilizadas métricas para avaliar o desempenho dos algoritmos tanto na predição quanto no agrupamento dos conjuntos de dados experimentados. No aspecto preditivo, foram avaliadas métricas para quantificar os acertos de classificação por parte do modelo. Para essa quantificação, foram avaliadas as métricas: Acurácia, Sensibilidade, Especificidade, Verdadeiro Positivo (VP), Verdadeiro Negativo (VN), Falso Positivo (FP) e Falso Negativo (FN). Todas as métricas avaliadas utilizaram como base a Matriz de Confusão.

Na avaliação do agrupamento, foram utilizadas medidas de validação relativa, cuja função é identificar o número de *clusters* que melhor descreva o conjunto de dados avaliado (FACELI et al., 2011). Neste trabalho 3 índices de validação relativa foram utilizadas: Conectividade, Índice de Dunn e Silhueta. Estas técnicas são descritas em maiores detalhes nas seções seguintes.

Matriz de Confusão A matriz de confusão é uma tabela de contingência que compara os resultados do modelo preditivo em relação ao atributo classe utilizado como referência pelo algoritmo de classificação.

Tabela 2.2: Exemplo de uma Matriz de confusão

		Classe de referência	
		+	-
Classe Preditiva	+	Verdadeiro Positivo	Falso Negativo
	-	Falso Positivo	Verdadeiro Negativo

A análise da matriz de confusão permite avaliar quem são os:

- Verdadeiros Positivos (VP): são os acertos do modelo em classificar os verdadeiros casos positivos.
- Verdadeiros Negativos (VN): são os acertos do modelo em classificar os verdadeiros casos negativos.

- Falsos Positivos (FP): são os casos positivos que foram classificados como negativo pelo modelo.
- Falsos Negativos (FN): são os casos negativos que foram classificados como positivos pelo modelo.

Acurácia. É a taxa de acertos total do modelo, obtida pela soma do valor de VP + VN dividida por todos os elementos da matriz de confusão (FACELI et al., 2011).

$$\frac{VP + VN}{VP + VN + FP + FN} \quad (2.1)$$

Sensibilidade. É a taxa de acerto do modelo na classificação dos casos positivos baseado na classe de referência (FACELI et al., 2011). Como pode ser observado na função abaixo:

$$\frac{VP}{VP + FN} \quad (2.2)$$

Especificidade. É a taxa de acerto do modelo na classificação dos casos negativos, quando eles realmente são negativos (FACELI et al., 2011). Como pode ser observado na função abaixo:

$$\frac{VN}{VN + FP} \quad (2.3)$$

Conectividade. Segundo (FACELI et al., 2011), a conectividade reflete o grau com que os objetos são agrupados no mesmo *cluster* de acordo com a matriz de similaridade. A conectividade é calculada na equação abaixo:

$$con(\pi) = \sum_{x_i \in X} \sum_{j=1}^v f(X_i, nn_{ij}) \quad (2.4)$$

Índice de Dunn. O índice de Dunn mede a razão da separação interna e externa dos clusters, através da equação abaixo (FACELI et al., 2011):

$$D(\pi) = \min_{a=1,\dots,k} \left\{ \min_{b=a+1,\dots,k} \left\{ \frac{d(C_a, C_b)}{\max_{l=1,\dots,k} d(C_l)} \right\} \right\} \quad (2.5)$$

Silhueta. A medida da silhueta avalia a similaridade dos objetos de um *cluster* e a dissimilaridade em relação aos objetos de outros *clusters* (FACELI et al., 2011; HAN; KAMBER; PEI, 2011). A medida pode ser obtida através da fórmula abaixo (FACELI et al., 2011):

$$sil(X_i) = \begin{cases} 1 - a(X_i, C_i)/b(X_i), & a(X_i, C_i) < b(X_i) \\ 0, & a(X_i, C_i) = b(X_i) \\ b(X_i)/a(X_i, C_i) - 1, & a(X_i, C_i) > b(X_i) \end{cases} \quad (2.6)$$

Nesta dissertação os algoritmos e métricas foram utilizados para auxiliar profissionais da saúde na identificação dos casos de leptospirose e dos genes responsáveis pelo agravamento da doença. Na próxima seção serão citados trabalhos que também utilizaram a MD aplicadas a saúde.

2.2 MINERAÇÃO DE DADOS APLICADA A SAÚDE

Existem vários trabalhos na literatura relacionados com a descoberta de conhecimento em bancos de dados (KDD) aplicada aos dados de saúde (GARCÍA-LAENCINA et al., 2015; YEH; WU; TSAO, 2011; YEH; CHENG; CHEN, 2011; FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996a, 1996b; BAKAR et al., 2011; BIER, 2013). Para realizar a descoberta de conhecimento, o KDD tem várias etapas, como mostrado na Figura 1 (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996b). As etapas Seleção, Processamento e Transformação de Atributos são essenciais para a preparação de dados para a etapa mineração de dados (CORTES; PORCARO; LIFSCHITZ, 2002), particularmente quando se trabalha com informações clínicas incompletas, como dados de pacientes hospitalizados (GARCÍA-LAENCINA et al., 2015).

O tema de análise dos modelos de predição tem sido usado por muitos pesquisadores para identificar uma variedade de doenças usando MD. Os trabalhos citados mostraram técnicas analisadas, algoritmos e o mérito da pesquisa. Os autores (GARCÍA-LAENCINA et al., 2015) compararam os algoritmos de predição KNN (k-vizinhos mais próximos), árvore de decisão, regressão logística e SVM (Support Vector Machine) com métodos de imputação de dados para identificar o melhor modelo de previsão de sobrevida de casos de cancro da mama ao longo de cinco anos de dados históricos. O melhor modelo de predição foi obtido com a aplicação do algoritmo KNN com 81% de precisão e 0,78 área sob a curva ROC (Receiver Operating Characteristic).

No trabalho de (YEH; WU; TSAO, 2011), aplicaram técnicas de MD a dados bioquímicos de pacientes em diálise para prever a probabilidade de hospitalização e descobriram que o tratamento imediato está associado a taxas mais baixas de hospitalização. Utilizando algoritmos de Árvore de Decisão e Regras de Associação, eles determinaram ainda que o índice de albumina era um fator importante na previsão de hospitalização de pacientes, com uma precisão que variava de 71% a 100% (YEH; WU; TSAO, 2011).

Em (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996a), construíram um modelo preditivo para melhorar o diagnóstico e prognóstico da doença cardiovascular. Os autores analisaram os algoritmos Árvore de Decisão, Rede Bayesiana e Redes Neurais com a

técnica de validação cruzada com 10 folds. O algoritmo que apresentou o melhor desempenho para os casos de classificação de doença foi a Árvore de Decisão com acurácia e sensibilidade de mais de 99% (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996a).

Como a leptospirose, que mata milhões de pessoas em todo o mundo, a malária é outra doença febril com altas taxas de mortalidade, especialmente na África. As principais razões são a escassez de profissionais de saúde e hospitais equipados para detecção e tratamento adequados, especialmente em áreas rurais. Em (OGUNTIMILEHIN; ADETUNMBI; ABIOLA, 2015), realizou uma revisão do trabalho na área de desenvolvimento de modelos preditivos para assistência no diagnóstico e tratamento médico da malária. Quinze estudos foram apresentados com diferentes metodologias de apoio ao diagnóstico precoce da doença, o que reduz a mortalidade. O autor enfatizou a importância de ter ferramentas computacionais que poderiam auxiliar no diagnóstico da doença, mas alertou para a necessidade de avaliar a exatidão e confiabilidade dessas ferramentas antes de colocá-las em uso clínico (OGUNTIMILEHIN; ADETUNMBI; ABIOLA, 2015). De particular relevância para o trabalho aqui apresentado, (UGWU; ONYEJEGBU; OBAG-BUWA, 2013) usaram a árvore de decisão como parte do diagnóstico e recomendação de tratamento para a malária. Este é um exemplo específico de um algoritmo desenvolvido que serve como uma importante ferramenta de saúde em uma região com escassez de profissionais especializados.

No trabalho de (SAHLE; MESHESHA, 2013), a redução das taxas de mortalidade por malária também foi a motivação para o uso de técnicas de mineração de dados. Neste trabalho, os algoritmos J48, JRIP e MLP foram utilizados para os métodos Árvore de Decisão, Regras de Classificação e Redes Neurais, respectivamente. Os experimentos mostraram resultados promissores para os modelos avaliados com uma taxa de acurácia de 97% de casos para MLP e 96% para os modelos obtidos com J48 e JRIP (SAHLE; MESHESHA, 2013). Assim, esta estratégia representa uma metodologia promissora para o diagnóstico precoce da doença, particularmente em regiões globais sem apoio médico adequado.

Em (BAKAR et al., 2011), foi desenvolvido um modelo de detecção de surtos de dengue usando vários classificadores baseados em regras, como Árvore de Decisão e Classificador Bayesiano. Eles descobriram que o uso de vários classificadores melhorou a precisão e a qualidade das regras em comparação com um único classificador. Por exemplo, os melhores resultados para análise de classificador único resultaram em uma curva ROC de 0,729, enquanto que as experiências usando classificadores múltiplos renderam uma curva ROC de 0,761.

Até onde conhecemos, apenas um estudo aplicou MD a dados de leptospirose (BIER, 2013). Os autores aplicaram o algoritmo de Árvore de Decisão aos dados epidemiológicos e sorológicos para avaliar os fatores de risco para a leptospirose canina e verificaram que, embora a especificidade do algoritmo J48 fosse 87 %, a sensibilidade do algoritmo foi de apenas 66 %. Ao contrário da leptospirose humana, em que o diagnóstico é dificultado por sintomas comuns a outras síndromes febris, a suspeita de leptospirose em cães é mais frequente por causa da maior exposição a reservatórios bacterianos. Assim, o presente trabalho é o primeiro a aplicar esses métodos a coortes humanas com e sem leptospirose confirmada laboratorialmente, incorporando dados epidemiológicos.

2.2.1 Leptospirose

A leptospirose é uma doença infecciosa febril, que afeta diversas regiões do planeta. Sua ocorrência está relacionada às precárias condições de infra-estrutura, saneamento básico e por sua vez, infestação de roedores infectados, mas também pode ser transmitida por animais domésticos e silvestres (SAUDE, 2009).

A doença mostra-se de maneira polimórfica, devido aos quadros clínicos apresentados por indivíduos infectados, podendo ter a forma leve, moderada e grave (SAUDE, 2009) até resultar em óbito. No quadro leve, a leptospirose é muitas vezes confundida com outras doenças febris, como a dengue, hepatite, gripe ou outra virose passageira, devido aos sintomas iniciais inespecíficos como febre alta, calafrios, cefaleia e mialgias (SAUDE, 2009). Havendo suspeita da doença, nem sempre é possível realizar os testes laboratoriais específicos para confirmação e quando é possível, os resultados levam alguns dias para serem disponibilizados, atrasando o início do tratamento, interferindo na evolução clínica do paciente e na conclusão do caso nos serviços públicos de notificação epidemiológica (SAUDE, 2009).

No Brasil os testes laboratoriais específicos, mais utilizados para o diagnóstico da doença, são o ELISA-IgM e MAT (Micro-aglutinação) (SAUDE, 2009). Caso o resultado destes testes, sejam negativos ou não reagentes na primeira amostra de sangue, coletada antes do 7º dia do início dos sintomas, não deve-se descartar a suspeita, sendo necessária uma segunda coleta após o 7º dia, devido ao período de incubação da bactéria (SAUDE, 2014).

Clinicamente, a doença apresenta-se de duas formas: Forma Anictérica, responsável por 90% a 95% dos casos da doença (SAUDE, 2009), podendo apresentar quadros leves, moderados ou graves, de início súbito com febre, cefaleia, dores musculares, anorexia, náuseas e vômitos, podendo ser curada em poucos dias, sem deixar sequelas; Forma Ictérica, com quadros moderados ou graves, apresentando disfunção renal, fenômenos hemorrágicos, alterações cardíacas, pulmonares e de consciência (SAUDE, 2009).

A leptospirose já foi considerada uma doença rural ou ocupacional (MYERS; VARELA-DIAZ, 1979), associada aos trabalhadores de lavouras ou em algumas profissões que implique exposição a ambientes ou a animais contaminados, como trabalhadores de minas, açougueiros e agricultores (NAJERA et al., 2005). Atualmente a leptospirose tem sido associada a alterações demográficas, ambientais e climáticas (KO et al., 1999b), principalmente em eventos extremos, como tsunamis e dilúvios, responsáveis por epidemias urbanas, como por exemplo nas Filipinas, em 2009 (AMILASAN et al., 2012) e na Austrália, em 2011 (SMITH et al., 2013).

Hoje em dia a leptospirose tem sido relatada predominantemente nos grandes centros urbanos, causada pela rápida urbanização, crescimento exacerbado das comunidades carentes em países em desenvolvimento (UN-HABITAT, 2010). Em 2001, o número total de moradores em comunidades carentes no mundo era de aproximadamente 924 milhões de pessoas, o que representa 32% da população urbana mundial (UN-HABITAT, 2004). Nos últimos anos, o número de moradores em comunidades carentes tem triplicado, alcançando 2,6 bilhões de pessoas (UN-HABITAT, 2010). Condições precárias de moradia fazem com que os moradores das comunidades carentes fiquem mais expostos aos agentes

causadores das doenças zoonóticas como, por exemplo, as leptospiras (REIS et al., 2008).

Só no Brasil, mais de 10.000 casos de leptospirose são notificados durante epidemias que ocorrem em períodos de alta precipitação e atingem as comunidades carentes urbanas (SAUDE, 2009). Esse alto número de notificações acontece por a leptospirose ser uma doença de notificação compulsória ao Ministério da Saúde, sendo obrigatória, a todos os profissionais de saúde, a notificação de casos suspeitos, isolados ou de surtos. A taxa de letalidade da leptospirose é acima de 10% dos casos notificados (KO et al., 1999b), mas a proporção de mortos pode aumentar para 50% se o indivíduo apresentar quadros de hemorragia pulmonar, o que pode levá-lo a morte em aproximadamente 48 horas (GOUVEIA et al., 2008).

2.2.2 Expressão de Genes aplicados a Leptospirose

No trabalho de (LINDOW et al., 2016) foi realizada uma análise multiparamétrica em dados de microarranjos, com a finalidade de melhor entender o processo da infecção de leptospira em humanos e identificar características do sistema imunológico associadas a morte ou sobrevivência de indivíduos hospitalizados. Para identificar fatores do hospedário que contribuem para a fatalidade, foram recrutados 16 pacientes internados com leptospirose aguda (13 sobreviventes e 3 casos fatais) e 4 voluntários saudáveis para caracterização aprofundada do curso clínico e respostas imunes. Em (LINDOW et al., 2016) foram realizadas comparações entre os grupos de amostras dos indivíduos experimentados, obtendo os seguintes achados:

- As respostas imunes inatas e adaptativas distinguem a doença nas fases agudas e de convalescência, identificando 1089 genes significantes. Consultando os termos da Ontologia de Genes (termos GO), os genes identificados são responsáveis pelas respostas às bactérias, de defesa, ligação de antígeno e genes da imunoglobulina.
- Na comparação do grupo de sobreviventes na fase aguda e amostras de indivíduos que foram a óbito, foram identificados 389 genes diferentemente expressos. Ao avaliar os termos GO, foram identificados funções de Atividade receptora de Interleucina 1 e Processos Biossintéticos com Composto de Enxofre. Adicionalmente foram identificados transcrições de vias pró-inflamatórias nos casos fatais.
- O transcrito com maior diferença em abundância entre os grupos de sobreviventes e óbito foi o gene da catelicidina (cAMP), apresentando uma diminuição da expressão, sendo o único péptido antimicrobiano significativamente diminuída em casos fatais.
- Não foram identificados genes significativamente diferentes entre o grupo de sobreviventes na fase de convalescência e grupos de voluntários saudáveis, o que indica que o estado imunológico tinha retornado à linha de base de 1-3 meses após a hospitalização.

2.2.2.1 Técnica de Análise de Expressão de Genes. A análise de expressões genéticas é um grande avanço científico, gerando bases de dados que podem ser utilizadas para a compreensão da complexidade dos processos biológicos.

zadas para diversas finalidades, inclusive no desenvolvimento de vacinas e no auxílio do diagnóstico precoce de doenças.

O corpo humano é composto por bilhões de células. Um núcleo de células contém 46 cromossomos individuais ou 23 pares. Cada cromossomo é composto por até milhares de genes. A análise de expressão gênica em humanos contém cerca de 30.000 genes, cuja função é de levar informações que determinam as características herdadas dos pais.

Segundo (SEALFON; CHU, 2011), os níveis de expressão dos genes de uma célula determinam o tipo de célula, estado de desenvolvimento, funções celulares e/ou estado patológico. Alterações na expressão de genes podem ser originadas de deleções ou inserções na codificação de proteínas ou sequências reguladoras de DNA.

Em se tratando de uma análise do genoma, como por exemplo o humano, a técnica indicada para a análise de expressões de genes é a microarranjos ou *microarray*, pois é uma técnica de hibridização de ácidos nucleicos em grande escala, permitindo a análise de milhares de genes simultaneamente. A técnica de *microarray* pode ser utilizada para medir os níveis de RNA ou DNA. A técnica de *microarray* de RNA é utilizado para identificar genes, vias ou redes de genes regulados em células e tecidos na comparação de condições biológicas (SEALFON; CHU, 2011), como realizada no trabalho de (LINDOW et al., 2016)

A técnica de *microarray* funciona aplicando uma solução contendo o cDNA corado com compostos fluorescentes em lâminas de vidro de microarray, realizando a hibridização por complementaridade dos ácidos nucléicos entre o cDNA (DNA complementar com um nucleótido fluorescente) e a sonda (transcritos), em seguida o *microarray* é colocado em um scanner especial que irá fazer a leitura da intensidade da fluorescência nos comprimentos de onda, digitalizando a imagem em alta resolução para posterior análise em software específico que converterá a cor e a intensidade da fluorescência em um conjunto de valores numéricos.

2.2.3 Agrupamento em dados de Microarranjos

Analisar expressão genética através da técnica de microarranjos é uma oportunidade para avaliar as funções de milhões de genes, podendo auxiliar no entendimento de doenças e do sistema biológico não só de humanos, mas também de diversas espécies. As técnicas de mineração de dados (MD) são importantes ferramentas de auxílio na interpretação dos dados e entendimento das funções biológicas. Agrupamento é uma das principais técnicas de MD utilizadas na análise de microarranjos. De acordo com (JIANG; TANG; ZHANG, 2004), agrupamento é uma forma de classificar genes de uma forma não supervisionada, não dependendo de classes predefinidas para classificar os dados de expressão.

Segundo (QUACKENBUSH, 2001), a identificação de grupos de genes tem a hipótese de que os genes pertencentes a um grupo compartilham alguma função ou elementos reguladores. O uso de agrupamento não visa oferecer respostas absolutas, mas apresentar relacionamentos nos dados para que sejam posteriormente explorados. Ainda segundo (QUACKENBUSH, 2001), o agrupamento hierárquico tornou-se uma das técnicas mais utilizadas na análise de dados de microarranjos, gerando uma relação hierárquica de similaridade de expressão de genes. Caso haja conhecimento do número de *clusters* que

devam ser representados, o K-means é uma alternativa ao uso do método hierárquico (QUACKENBUSH, 2001).

O uso de algoritmos de agrupamentos em dados de microarranjos é comumente encontrado na literatura, mas até a presente data não foi realizado na análise de expressão gênica em pacientes com leptospirose. No trabalho de (MAKRETSOV et al., 2004) utilizou-se algoritmos hierárquico para avaliar a melhoria no prognóstico em pacientes com câncer de mama invasivo. A análise hierárquica identificou grupos com diferenças significativas em relação ao desfecho clínico dos indivíduos com amostras avaliadas.

No trabalho de (WANG et al., 2003), foi utilizado o algoritmo Fuzzy C-means na análise de dados de microarranjos, com a finalidade de agrupar tumores baseado em dados classificados de leucemia, câncer do colo, tumores cerebrais e um conjunto de dados contendo uma variedade de tipos de câncer humanos. O algoritmo fuzzy foi utilizado para avaliar o grau de pertinência da expressão gênica do tumor em relação às classes preditoras, no intuito de conhecer a função biológica dos genes pertencentes a cada tumor.

O uso do algoritmo Fuzzy C-means na análise de dados de microarranjos também foi utilizado no trabalho de (TARI; BARAL; KIM, 2009), os autores combinaram o uso do algoritmo com anotações da ontologia de genes, que contempla uma gama de termos biológicos, com o objetivo de guiar o processo de agrupamento de genes funcionalmente expressos. Segundo os autores, o uso do Fuzzy permite que genes tenham associações em múltiplos clusters, haja vista que os produtos genéticos estão normalmente envolvidos em múltiplos papéis no fundamento da célula. O experimento foi realizado com dados de expressão gênica de levedura. O autor concluiu que a importância do conhecimento prévio ou classe de amostras, melhora a coerência dos clusters em relação ao domínio do conhecimento (TARI; BARAL; KIM, 2009).

2.3 CONSIDERAÇÕES FINAIS

O uso de técnicas de mineração de dados auxilia na análise e obtenção de conhecimentos a partir de um conjunto de dados. Esta dissertação aborda a análise de dados de expressão gênica resultante da técnica de microarranjos realizada em amostras de pacientes com leptospirose que tiveram desfecho fatal ou de sobrevivência obtidas do trabalho de (LINDOW et al., 2016). Os algoritmos K-means, Fuzzy C-means e Hierárquico foram descritos na seção de agrupamento por auxiliarem na avaliação da viabilidade de se trabalhar com expressão gênica. O algoritmo SVM também foi descrito, devido a sua aplicação na classificação dos dados de expressão gênica de acordo com o desfecho clínico das amostras avaliadas. Uma seção sobre técnicas de redução de dimensionalidade foi apresentada devido ao uso de dados provindos da técnica de microarranjos, divididas por técnica de agregação e seleção de atributos. As métricas de avaliação abordadas nesta dissertação também foram citadas neste capítulo, como a Sensibilidade e Especificidade, métricas comumente utilizadas na avaliação de testes de diagnósticos para avaliar o desempenho dos modelos em predizer os casos e não casos de leptospirose. Por fim, foram apresentados trabalhos relacionados ao uso de mineração de dados aplicado a saúde, que tiveram fundamental contribuição para o desenvolvimento deste trabalho.

Capítulo

3

Este capítulo descreve os objetivos e a metodologia do trabalho.

PREDIÇÃO DO DESFECHO CLÍNICO

O diagnóstico de casos de leptospirose hospitalizados precocemente não é uma tarefa simples, devido aos sinais e sintomas da doença serem similares aos de outras enfermidades, como a dengue. Além disso, identificar a causa de morte de indivíduos doentes por leptospirose com as mesmas características dos que sobrevivem é importante para a melhoria do tratamento.

Atualmente a análise dos níveis dos genes durante uma doença aguda, como leptospirose, para identificar biomarcadores, ainda é uma tarefa manual e que requer muito tempo. Somando a isso, os riscos de erro humano na análise adotada estimulam ainda mais a adoção de técnicas computadorizadas para o auxílio destas atividades.

O IGM/FIOCRUZ-BA possui profissionais especializados no estudo da leptospirose, realizando diversos estudos simultaneamente relacionados ao tema. Uma das hipóteses avaliadas por esse grupo de pesquisas refere-se à resposta imune distinta entre indivíduos sobreviventes e os que foram a óbito. Estudos preliminares indicam que esta diferença existe e que possivelmente ocorre devido a duas possibilidades:

- O sistema imune alterado causa dano aos tecidos dos pacientes que foram a óbito, prejudicando a produção de anticorpos.
- Amostras de fase aguda de pacientes que foram a óbito não tem anticorpos suficientes para combater a bactéria.

Assim, o presente trabalho propõe auxiliar na identificação precoce dos casos hospitalizados da doença, tornando mais célere o tratamento. Suplementarmente pretende-se identificar os genes associados à morte ou à sobrevivência de indivíduos doentes através do uso de algoritmos de classificação na avaliação da diferença de expressão gênica dos grupos de amostras analisadas.

Neste sentido, o presente trabalho pretende ajustar os algoritmos de classificação na análise dos fatores clínicos, epidemiológicos e de expressão gênica com o intuito de auxiliar a identificação de casos, bem como oferecer uma metodologia que permita conhecer os genes responsáveis pelo desfecho clínico da doença.

3.1 METODOLOGIA

Com a finalidade de melhor entender a doença e seus fatores de risco, tanto na concepção clínica quanto na epidemiológica, e consequentemente entender os fatores genéticos, este trabalho foi dividido em 3 etapas com o intuito de: avaliar algoritmos que permitissem classificar os casos hospitalizados como suspeito ou não pela doença; avaliar a viabilidade de agrupar dados de expressão gênica por desfecho clínico; e por fim, predizer o desfecho clínico de acordo com os dados de expressão gênica, identificando os genes com diferença de expressão entre indivíduos sobreviventes e os que foram a óbito. Estas três etapas corresponderam as respectivas hipóteses deste trabalho.

3.1.1 Casos de leptospirose baseados em dados clínicos e epidemiológicos

O uso de um modelo de predição baseado em dados clínicos somados a dados epidemiológicos pode sugerir que um indivíduo é suspeito para a doença, restando realizar exames laboratoriais específicos para a confirmação. Essa predição pode auxiliar no tratamento eficaz, favorecer o prognóstico da doença para casos graves e/ou até mesmo a redução dos óbitos (NERY; CLARO; LINDOW, 2016; JUNIOR; CLARO; LINDOW, 2017).

Assim, o objetivo desta etapa foi analisar os modelos de classificação que apresentem o melhor ajuste para a identificação dos casos da leptospirose aplicados a base de dados clínicos e epidemiológicos do IGM/FIOCRUZ-BA. A identificação precoce de casos de leptospirose auxiliará os profissionais de saúde no tratamento dos pacientes, minimizando o risco ou evolução para a forma grave da doença (NERY; CLARO; LINDOW, 2016; JUNIOR; CLARO; LINDOW, 2017).

Objetivos específicos

- Conhecer os fatores clínicos e epidemiológicos da leptospirose com o intuito de melhor gerir os dados a serem minerados.
- Aplicar técnicas de pré-processamento para a base de dados do estudo.
- Avaliar os algoritmos de classificação na identificação dos casos da doença.
- Avaliar a sensibilidade dos melhores modelos de predição em um conjunto de dados contendo apenas casos confirmados laboratorialmente por leptospirose.
- Avaliar a especificidade dos melhores modelos de predição em um conjunto de dados contendo apenas casos confirmados laboratorialmente por dengue.

3.1.2 Agrupamento de indivíduos doentes baseados nos desfechos clínicos

Nesta etapa do trabalho, foram realizados experimentos para avaliar a viabilidade de analisar o perfil de expressão gênica dos indivíduos doentes por meio de técnicas de mineração de dados em um conjunto de dados reduzido (389 transcritos). A tarefa de mineração de dados escolhida foi a de agrupamento (HAN; KAMBER; PEI, 2011; FACELI et al.,

2011), pois o intuito inicial era de avaliar de forma automática se o uso de algoritmos poderiam identificar os grupos de amostras avaliadas, pela técnica de microarranjos, de acordo com o valor de expressão e seus respectivos desfechos clínicos.

Objetivos específicos. O foco da pesquisa é a aplicação de algoritmos utilizando as técnicas de agrupamento na análise de genes da leptospirose a fim de auxiliar na identificação da similaridade entre perfis de amostras avaliadas. Deste modo, especificamente esta etapa pretendia:

- Compreender as técnicas de análise dos genes com o intuito de verificar quais os tipos de dados a serem minerados.
- Avaliar os métodos e algoritmos de agrupamento mais indicado para a identificação dos grupos de amostras avaliadas.
- Avaliar os resultados obtidos com o intuito de melhor ajustar os algoritmos utilizados.
- Avaliar a hipótese dos profissionais da saúde de que há diferenças no perfil de expressão entre indivíduos que sobrevivem e os que morrem.

3.1.3 Predição do desfecho clínico na análise gênica

O objetivo desta etapa foi predizer o desfecho clínico de pacientes hospitalizados por leptospirose, baseado em dados completos de expressão gênica resultante da técnica de microarranjos. Avaliar um conjunto de dados de expressão gênica é uma tarefa que requer metodologias e ferramentas que auxiliem o trabalho. Assim, nesta hipótese foram avaliados dois métodos de análise completa de expressão gênica: por transcritos e por genes.

3.1.3.1 Método por Transcritos. Neste método, foi utilizado o conjunto completo de transcritos obtido da análise de expressão gênica dos 20 indivíduos participantes dos estudos do IGM/FIOCRUZ-BA. Devido à alta dimensionalidade dos dados, técnicas de redução e identificação dos genes relevantes foram experimentadas com o uso de algoritmo de classificação, avaliando a acurácia em identificar os genes diferentemente expressos entre os grupos de sobreviventes e óbitos.

3.1.3.2 Método por Genes. Neste método foi utilizado o conjunto completo de genes, convertido da base de dados por transcritos. Neste método, foi avaliada a conversão do conjunto de dados, com o propósito de melhorar o desempenho das técnicas de redução da dimensionalidade, bem como na predição dos casos de óbito e sobreviventes.

Objetivos específicos. Nesta etapa do trabalho aplicou-se algoritmos utilizando as técnicas de classificação na análise de genes da leptospirose a fim de auxiliar nas previsões clínicas dos casos da doença. Especificamente:

- Identificar e avaliar técnicas de redução da dimensionalidade para a aplicação em dados genéticos.
- Aplicar o algoritmo SVM na análise dos dados de expressão gênica resultantes dos experimentos de redução da dimensionalidade, comparando a acurácia do experimento de predição.
- Avaliar os resultados obtidos com o intuito de melhor ajustar o algoritmo utilizado.
- Identificar os genes e sua respectiva função biológica contido no conjunto de dado com maior taxa de acurácia.
- Avaliar a importância clínica dos genes identificados.

3.2 CONSIDERAÇÕES FINAIS

Neste capítulo foi apresentado o problema da pesquisa acometido pela doença, bem como a proposta de identificar os genes associados à morte ou sobrevivência de indivíduos doentes por leptospirose através do uso de algoritmos de mineração de dados. Foi descrita a metodologia do trabalho com as 3 etapas detalhadas para alcançar o objetivo proposto.

Capítulo

4

Este capítulo descreve os conjuntos de dados utilizados neste trabalho.

ANÁLISE DOS DADOS PARA LEPTOSPIROSE

4.1 CONJUNTOS DE DADOS

Neste capítulo serão apresentados os conjuntos de dados utilizados nesta dissertação, organizados por conjuntos de experimentos:

- No Experimento A foram utilizados conjuntos de dados oriundos da base de dados da Vigilância Hospitalar de Leptospirose do IGM/FIOCRUZ-BA. Dados clínicos e epidemiológicos foram extraídos para serem experimentadas com o intuito de predizer casos da doença sem a utilização de informações laboratoriais específicas.
- Já nos experimentos B e C, foram utilizados dados de expressão gênica, com o objetivo de avaliar se há distinção na expressão entre amostras de indivíduos com diferentes desfechos clínicos.

A Figura 4.1 apresenta um resumo dos conjuntos de dados utilizados neste trabalho, divididos por experimentos realizados para validar as hipóteses desta dissertação.

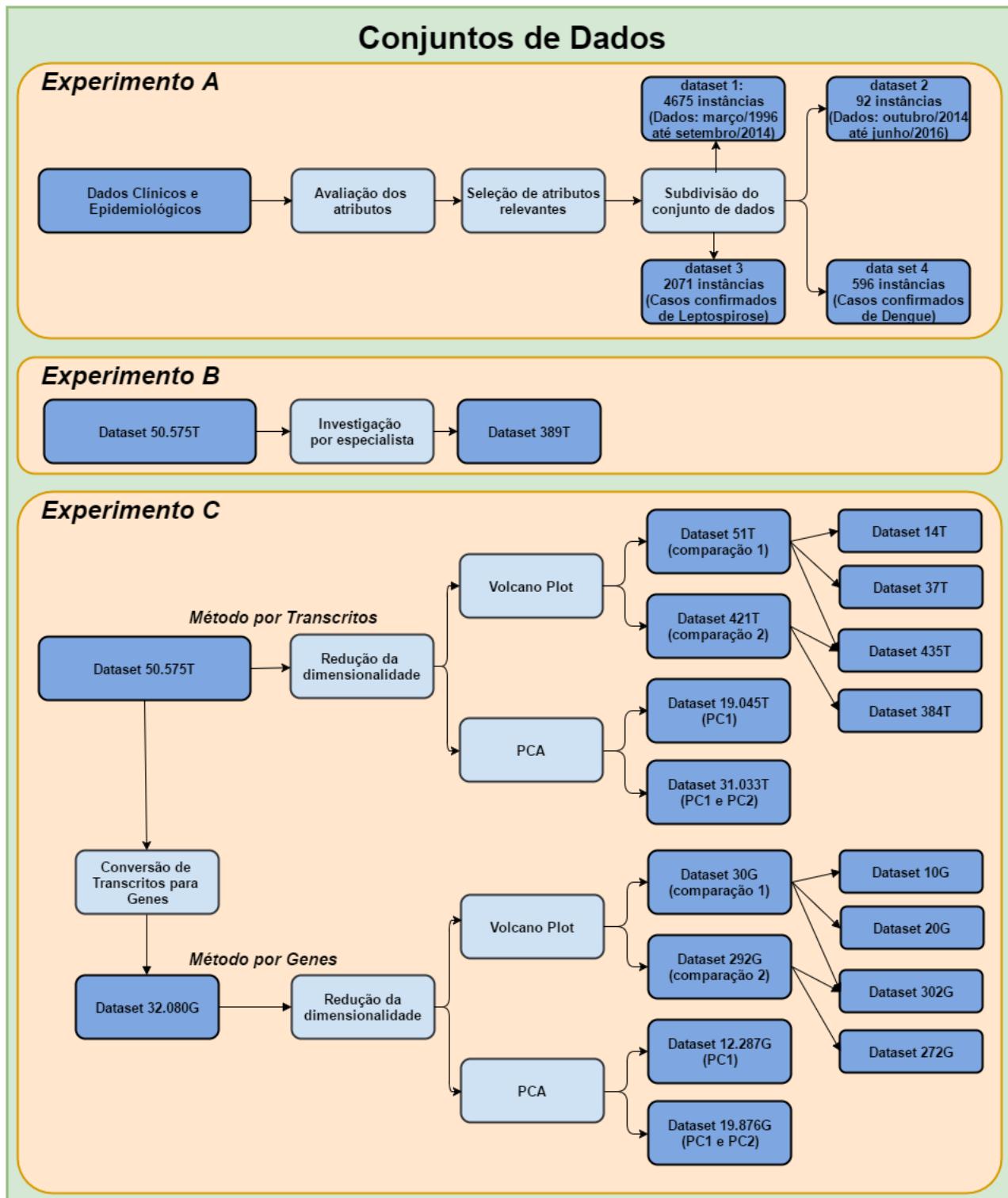


Figura 4.1: Resumo dos conjuntos de dados utilizados nesta dissertação.

4.1.1 Base de dados de casos hospitalizados de Leptospirose

Desde 1996, o grupo de Leptospirose do IGM/FIOCRUZ-BA tem realizado vigilância ativa para identificar casos suspeitos da doença no Hospital Couto Maia, em Salvador, no Brasil. Este estudo utilizou dados coletados durante a vigilância hospitalar, compostos de informações coletadas de entrevistas de pacientes, revisão de prontuários e resultados de testes laboratoriais. Os dados da vigilância hospitalar consistiram em:

- Dados demográficos, incluindo o local de residência, idade e sexo do participante.
- Dados epidemiológicos coletados no momento da internação do paciente para determinar se os pacientes tinham fatores de risco para leptospirose e resultados iniciais de exames não específicos, como hemograma completo, radiografia de tórax e sumário de urina. Após a alta hospitalar, a equipe coletou dados epidemiológicos sobre o desfecho clínico final e história relevante do prontuário médico.
- Dados relativos aos últimos 30 dias de atividades para identificar fatores de risco individuais, tais como possíveis exposições (contato com lama, esgoto e/ou resíduos), características socioeconômicas e/ou vivendo ou trabalhando em áreas de risco.
- Dados sobre os fatores de risco das casas dos participantes: informações como proximidade de esgoto, acúmulo de lixo, número de ratos vistos em/perto de casa, vegetação e animais.
- Georeferenciamento de dados na localização exata da casa do participante para análises espaciais posteriores.

A base de dados da vigilância hospitalar consistiu de 1.715 atributos, com 4.675 casos suspeitos coletados de março de 1996 a setembro de 2014.

4.1.2 Análise dos atributos

A principal limitação encontrada na base de dados da vigilância ativa do grupo de leptospirose do IGM/FIOCRUZ-BA é que a maioria dos atributos foram criados nos últimos cinco anos. Assim, muitas variáveis estavam faltando dados ao longo de 13 anos, resultando em dados incompletos.

Como o objetivo deste modelo era prever o diagnóstico de leptospirose com base em dados clínicos e epidemiológicos, foram removidos alguns atributos como os resultados laboratoriais que confirmam o diagnóstico, os atributos incluídos com menos de 2 anos, informações sobre coleta de material biológico, resultados clínicos e laboratoriais e informações diárias sobre internamento em UTI também foram removidas. Assim, 267 atributos mantidos foram inicialmente avaliados, como mostrado na Figura 4.2 a.

Em seguida, os valores em falta para cada atributo foram analisados, removendo cada variável que ultrapassou 80% dos dados em falta, resultando em 99 atributos, conforme ilustrado na Figura 4.2 b. O conjunto de dados resultante possuía 84% de atributos discretos e 16% nominal.

Foi realizada uma revisão manual, por um profissional de saúde, dos 99 atributos para remover dados de diagnóstico ou clínicos irrelevantes. Este passo excluiu variáveis que careciam de relevância clínica. Além disso, os dados de diagnóstico clínico foram removidos para evitar um enviesamento dos resultados. Como resultado desta revisão, a base de dados ficou com 76 atributos clínicos e epidemiológicos, incluindo 80% de atributos discretos e 20% nominal, como mostrado na Figura 4.2 c.

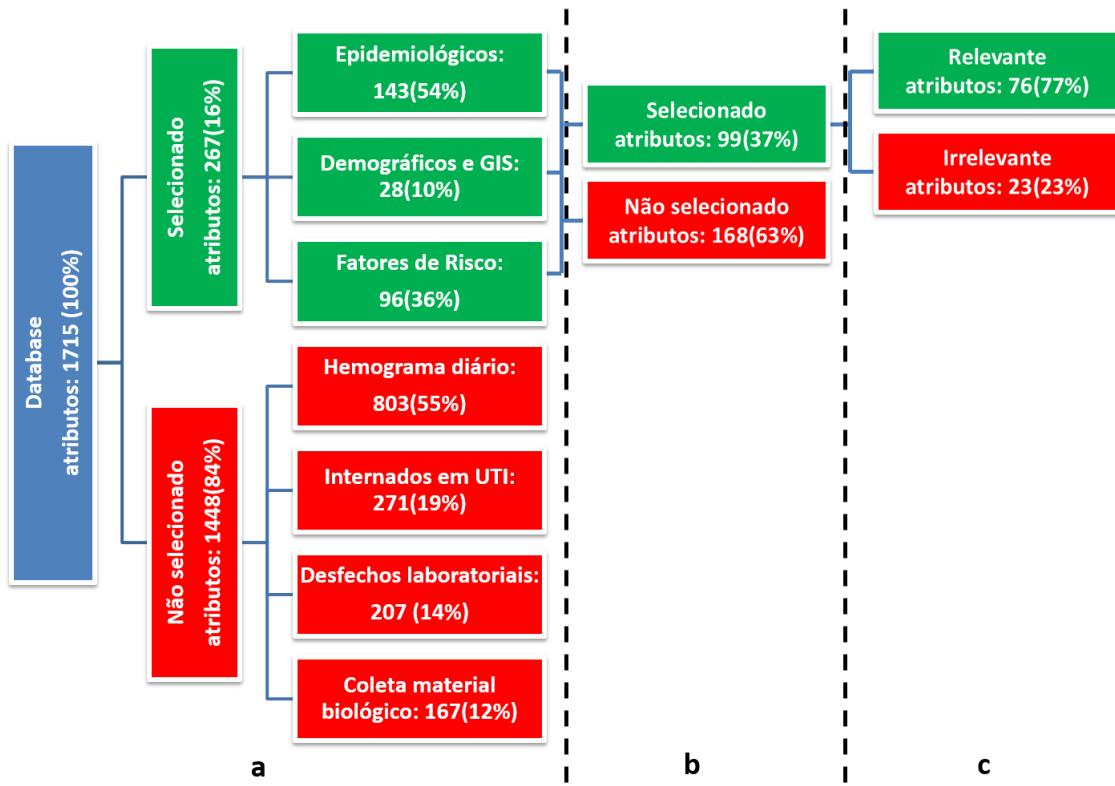


Figura 4.2: Fluxo das etapas de seleção de características (a), Pré-processamento (b) e avaliação da relevância por especialista (c).

Dos 4.675 casos suspeitos de leptospirose da base de dados do IGM/FIOCRUZ-BA: 2.046 foram confirmados laboratorialmente (44%) e 2.629 não confirmados (56%). Todos os 76 atributos do conjunto inicial com 4.675 instâncias foram divididos em dados de treinamento e teste através das técnicas de validação cruzada e divisão percentual. Dos 2.629 casos de leptospirose não confirmados, 596 casos foram confirmados laboratorialmente para dengue. Estes 596 foram utilizados para avaliar a especificidade dos algoritmos JRIP e J48. A confirmação laboratorial foi utilizada para validar a eficácia dos modelos. Finalmente, foi criado um atributo de predição baseado em testes de diagnóstico laboratorial denominados LEPTO cujas opções de resposta foram "Confirmado" ou "Não Confirmado".

A Figura 4.3 apresenta um resumo dos conjuntos de dados utilizado no experimento A.

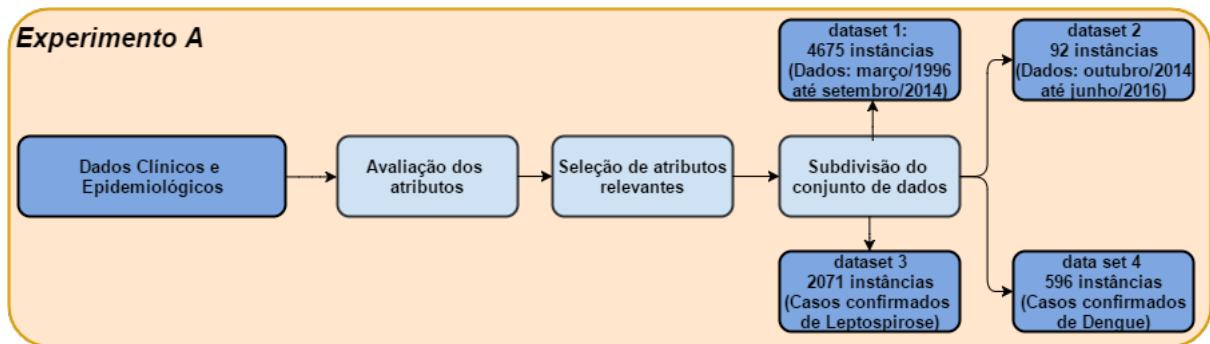


Figura 4.3: Conjuntos de dados utilizados no Experimento A.

4.1.3 Conjuntos de dados de expressão gênica

A base de dados utilizada neste trabalho foi resultante da técnica de microarranjos a partir da avaliação do material biológico coletados de 20 indivíduos pertencentes a quatro grupos de participantes dos estudos do IGM/FIOCRUZ-BA, destes: 13 são indivíduos com duas coletas de amostras, a primeira na fase aguda (S) e a segunda na fase convalescente (C) da doença, 3 indivíduos com amostras coletadas na fase aguda, mas que foram a óbito por leptospirose (D) e 4 indivíduos sadios (H), totalizando 33 amostras avaliadas (LINDOW et al., 2016).

A técnica de microarranjos permitiu analisar o perfil genético do material avaliado. A Figura 4.4 apresenta duas análises realizadas por profissionais do IGM/FIOCRUZ-BA, com a finalidade de avaliar as amostras do experimento. A parte A da figura apresenta a comparação entre os grupos de amostras na fase aguda de indivíduos que sobreviveram (S) e os que foram a óbito (D), além de comparar com amostras de indivíduos saudáveis (H), indicando pela escala de cores que existem 3 perfis de genes: o primeiro na cor azul escuro, indicando os genes correspondentes a este perfil são mais semelhantes entre o grupo de sobreviventes e saudáveis quando comparados com o grupo dos óbitos. Já no perfil azul claro (perfil 2), os genes de indivíduos que foram a óbito é mais semelhante aos sobreviventes. Na análise do perfil na cor vermelha (perfil 3), óbito e sobreviventes são mais semelhantes quando comparados com amostras de indivíduos saudáveis, indicando possíveis genes afetados com a doença.

Ao analisar a parte B da Figura 4.4, observa-se que amostras de indivíduos na fase convalescente são semelhantes as amostras de indivíduos saudáveis, indicando que ao chegar neste período, o paciente está praticamente curado. A Figura 4.4 é de suma importância para avaliar a hipótese de que há diferenças no sistema imune dos grupos avaliados, servindo como base para uma possível identificação dos genes responsáveis pelo agravamento da doença.

Após a realização da técnica de microarranjos, os dados da leitura da fluorescência são convertidos em valores numéricos, tendo como atributos a identificação dos transcritos pertencentes aos genes analisados, formando a base de dados utilizada neste trabalho.

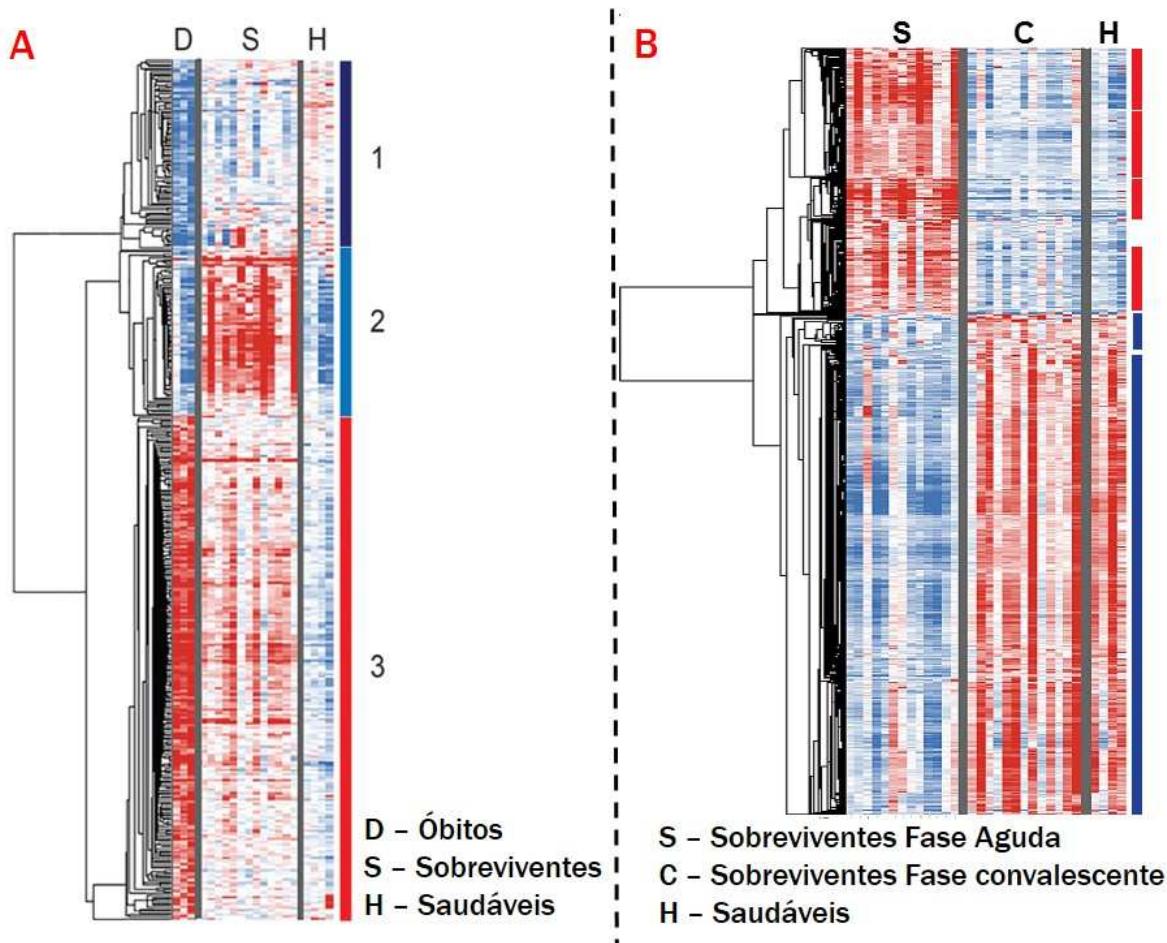


Figura 4.4: Avaliação do perfil imunológico dos grupos de amostras obtida com a técnica de microarranjos (LINDOW et al., 2016).

4.1.4 Conjunto com 389 transcritos

O primeiro conjunto de dados (dataset 389 transcritos) utilizado para a avaliação da viabilidade de se trabalhar com mineração de dados genéticos foi obtido através da investigação de especialistas do IGM/FIOCRUZ-BA em análise de microarranjos, que avaliaram os transcritos com diferença de expressão na comparação das amostras testadas (LINDOW et al., 2016).

A Tabela 4.1 apresenta o resultado obtido da investigação na comparação dos grupos de amostras. O conjunto de dados utilizado contém 389 transcritos resultantes da comparação das amostras do grupo de sobreviventes em relação ao grupo dos que foram a óbito. Nesta investigação, não foram encontrados transcritos com diferença de expressão na comparação entre indivíduos sobreviventes na fase convalescente em relação aos indivíduos saudáveis, o que adicionalmente pretendemos avaliar através do uso de técnicas de aprendizagem de máquina.

Tabela 4.1: Grupos avaliados por técnicas convencionais e análise manual por profissionais especializados em expressão genética do IGM/FIOCRUZ-BA.

Grupos avaliados		Transcritos
Sobreviventes Fase aguda	X	Sobreviventes Convalescentes (SvC) 1089
Sobreviventes	X	Óbitos (DvS) 389
Sobreviventes convalescentes	X	Saudáveis São semelhantes

A Figura 4.5 apresenta um resumo do conjunto de dados utilizado no experimento B.



Figura 4.5: Conjunto de dados utilizado no Experimento B.

4.1.5 Conjunto completo com 50.575 transcritos

Este conjunto de dados (dataset 50.575 transcritos) possui uma alta dimensionalidade contendo 50.575 transcritos como atributos contínuos com os valores de expressão obtidos durante a técnica de microarranjos e 1 atributo categórico, utilizado como classe preditora, contendo o tipo de cada amostra sequenciada: S – amostra na fase aguda dos sobreviventes, C amostra na fase convalescência dos sobreviventes, D – amostras na fase aguda dos óbito e H – amostra de indivíduos saudáveis. A base de dados possui 33 instâncias compostas pelos resultados da análise de expressão gênica das 33 amostras coletadas dos 20 indivíduos dos estudos do IGM/FIOCRUZ-BA.

4.1.6 Conjunto completo com 32.080 genes

Uma estratégia utilizada por alguns especialistas na análise de dados de microarranjos é a conversão do conjunto de dados de transcritos para genes, uma vez que um conjunto de dados por transcritos contém redundância de informações sobre genes, já que os transcritos são partes de um gene. Com a finalidade de avaliar os benefícios deste método, o conjunto de dados (dataset 50.575 transcritos) foi convertido em um conjunto de dados por gene (dataset 32.080 genes).

A conversão do conjunto de dados foi realizada no programa RStudio (RACINE, 2012), através do uso da função “avereps” obtida juntamente com o pacote Limma (SMYTH, 2005), específico para análise de microarranjos. A função avereps sumariza todos os transcritos do mesmo gene, atribuindo o valor médio das expressões contidas em cada transrito, resultando em um valor único de expressão por gene. Um exemplo do processo de conversão é o gene PDE4DIP (*Phosphodiesterase 4D Interacting Protein*, responsável pela codificação de proteínas), que possui 17 transcritos e na conversão teve seu valor

obtido da média destas 17 expressões. Ao final do processo de conversão dos dados de transcritos para genes, a base de dados possuía as seguintes características: 32.080 genes como atributos contínuos e o atributo categórico, utilizado como classe preditora.

A conversão do conjunto de dados de transcritos para genes permitiu reduzir 37% o número de atributos do conjunto de dados em relação à base de dados inicial (32.080/50.575), contudo, a alta dimensionalidade foi mantida, sendo necessário experimentar outras técnicas utilizadas para redução deste tipo de informação.

A Figura 4.6 apresenta um resumo dos conjuntos de dados utilizados no experimento C.

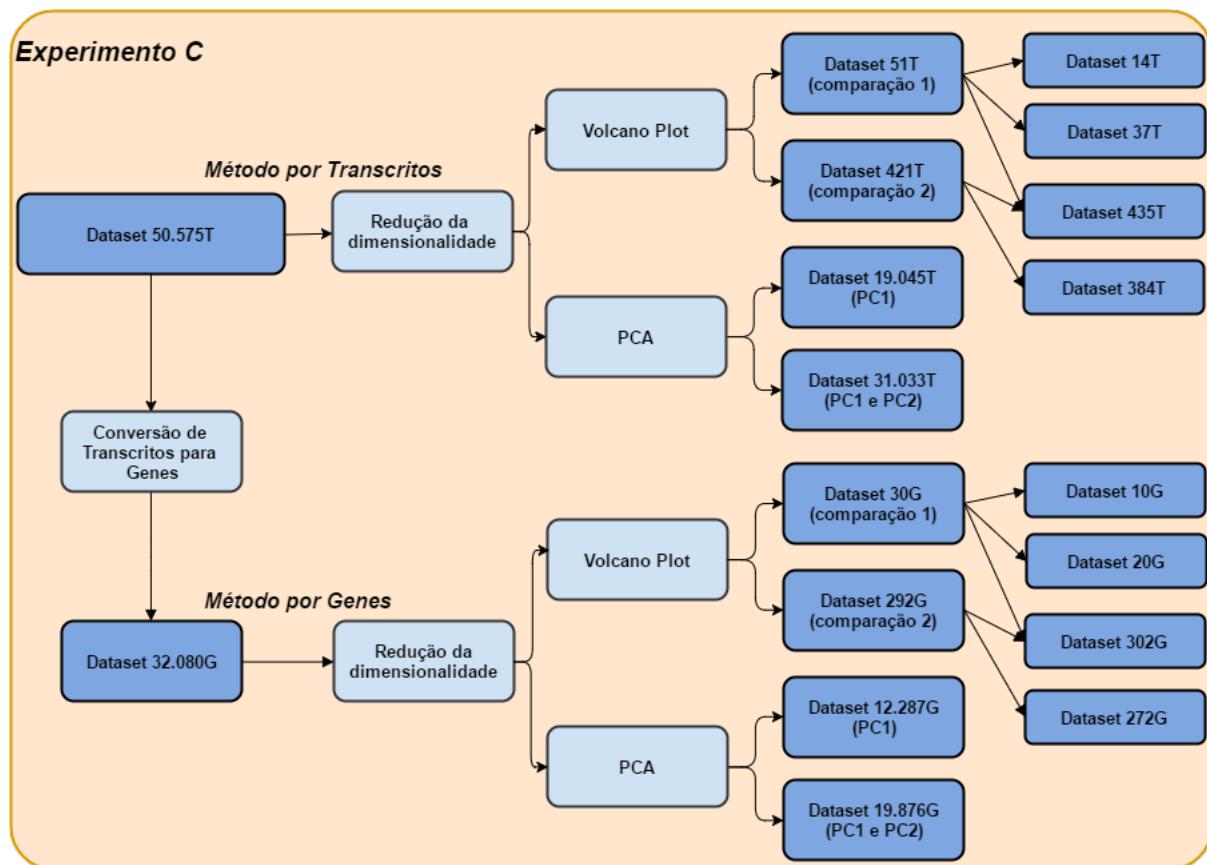


Figura 4.6: Conjuntos de dados utilizados no Experimento C.

4.1.7 Redução da dimensionalidade

Devido a alta dimensionalidade do conjunto de dados por transcritos (dataset 50.575 transcritos) e por genes (dataset 32.080 genes), tornou-se necessária a análise da redução da dimensionalidade. Assim, com o intuito de diminuir o custo computacional sem perda da qualidade dos dados e facilitar a interpretação dos resultados, duas técnicas foram avaliadas: Análise de Componentes Principais (PCA ou ACP) e o teste de hipótese com o uso do Volcano Plot.

Os experimentos realizados para reduzir a dimensionalidade dos dados de expressão

gênica neste trabalho, na avaliação dos métodos por transcritos e por genes, podem ser divididos em dois objetivos:

- Avaliar o impacto da redução da dimensionalidade através dos métodos: PCA e Volcano Plot;
- Identificar quais os genes responsáveis pela sobrevivência ou morte de indivíduos doentes por leptospirose.

Nas próximas seções são descritos os experimentos e resultados das duas técnicas de redução da dimensionalidade, PCA e Volcano Plot, aplicadas aos conjuntos de dados por transcritos e por genes.

4.1.7.1 Análise de Componentes Principais - PCA

Método por Transcritos. No experimento, realizou-se o PCA no conjunto de dados completo, contendo 50.575 transcritos. Na Tabela 4.2, observa-se a proporção da variância para cada componente, bem como a proporção acumulativa. Adicionalmente é apresentado o número de transcritos pertencentes aos cinco primeiros componentes, bem como o número de transcritos que podem ser adicionados ao conjunto de dados de acordo com a quantidade de componentes escolhidos. Caso seja definido que o primeiro componente atende os requisitos de variabilidade dos dados, teríamos um conjunto composto de 19.045 transcritos e uma proporção de variação de aproximadamente 15%, formando um dos conjuntos de dados (dataset 19.045T) utilizados nos experimentos.

O uso do método do PCA na seleção dos 2 primeiros componentes apresentou uma redução significativa (de 50.575 para 19.045 transcritos), entretanto a alta dimensionalidade foi mantida, mesmo tendo apenas 26% de variabilidade. Para termos uma proporção de variabilidade próxima a 50%, teríamos um conjunto de dados com 39.032 transcritos (soma dos transcritos exclusivos pertencentes aos PC1 até o PC5 descritos na Tabela 4.2), indicando que o conjunto de dados possivelmente possui variáveis pouco correlacionadas.

Método por Genes. Com a finalidade de reduzir a dimensionalidade da base de dados por genes, foi aplicado o PCA para avaliar a variabilidade e identificar os genes pertencentes a cada componente principal.

Na Tabela 4.3 são apresentadas as proporções de variâncias para cada componente e a proporção acumulativa. Observa-se que os resultados obtidos com a base de dados de genes são similares aos obtidos com o conjunto de dados dos transcritos. O PC1 apresenta aproximadamente 15% da variabilidade dos dados, tendo 12.287 genes agrupados neste componente. O PC2 é responsável por 11% de variabilidade, tendo 10.669 genes.

O método da Análise de Componentes Principais como técnica de redução de dimensionalidade não apresentou resultados satisfatórios no conjunto de dados por genes, assim como observado na aplicação da técnica no conjunto por transcritos. Na tabela 4.3 foi avaliada a quantidade de transcritos únicos pertencentes aos 5 primeiros componentes e juntos formariam um conjunto de dados com 24.905 genes únicos, o que corresponderia a 78% do tamanho da base de dados inicial e com menos de 50% variabilidade.

Tabela 4.2: Aplicação do PCA no conjunto de dados dos 50.575 transcritos.

Componentes	Variação	Proporção da Variação	Proporção Acumulativa	Transcritos por componentes	Número único de transcritos
PC1	7412.483	14.656	14.656	19045	19045
PC2	5666.406	11.204	25.86	16610	11988
PC3	4625.89	9.147	35.007	11624	2948
PC4	3515.747	6.952	41.959	9928	3049
PC5	2497.496	4.938	46.897	6419	2002
PC6	2054.508	4.062	50.959		
PC7	1698.928	3.359	54.318		
PC8	1580.22	3.125	57.443		
PC9	1412.483	2.793	60.236		
PC10	1330.103	2.63	62.866		
PC11	1209.192	2.391	65.256		
PC12	1188.951	2.351	67.607		
PC13	1070.999	2.118	69.725		
PC14	1061.368	2.099	71.824		
PC15	1011.877	2.001	73.824		
PC16	959.295	1.897	75.721		
PC17	939.888	1.858	77.58		
PC18	907.507	1.794	79.374		
PC19	889.199	1.758	81.132		
PC20	876.093	1.732	82.864		
PC21	841.833	1.665	84.529		
PC22	803.649	1.589	86.118		
PC23	784.984	1.552	87.67		
PC24	758.455	1.5	89.17		
PC25	745.727	1.474	90.644		
PC26	734.757	1.453	92.097		
PC27	720.966	1.426	93.522		
PC28	703.37	1.391	94.913		
PC29	678.305	1.341	96.254		
PC30	659.908	1.305	97.559		
PC31	645.643	1.277	98.836		
PC32	588.771	1.164	100		

A Figura 4.7 apresenta a comparação dos resultados da aplicação do PCA nas bases de dados por transcritos e por genes através dos gráficos da máxima variância entre os componentes, além de apresentar dois diagramas de Venn: na parte A, indicando os transcritos exclusivos do componente 1 (14.423), os transcritos que estão em ambos componentes (4.622) e os transcritos exclusivos do componente 2 (11.988), totalizando 31.033

Tabela 4.3: Aplicação do PCA no conjunto de dados dos 32.080 genes.

Componentes	Variação	Proporção da Variação	Proporção Acumulativa	Transcritos por componentes	Número único de genes
PC1	69.2466	0.1495	0.1495	12287	12287
PC2	60.0182	0.1123	0.2618	10669	7589
PC3	54.25035	0.09174	0.3535	7403	1856
PC4	47.54582	0.07047	0.42397	6389	1928
PC5	39.87156	0.04956	0.47353	4111	1245
PC6	36.06403	0.04054	0.51407		
PC7	32.7864	0.03351	0.54758		
PC8	31.5145	0.03096	0.57854		
PC9	29.6954	0.02749	0.60602		
PC10	28.96279	0.02615	0.63217		
PC11	27.65835	0.02385	0.65602		
PC12	27.30346	0.02324	0.67926		
PC13	26.09578	0.02123	0.70048		
PC14	25.72259	0.02063	0.72111		
PC15	25.25666	0.01988	0.74099		
PC16	24.66784	0.01897	0.75996		
PC17	24.33155	0.01845	0.77842		
PC18	24.00341	0.01796	0.79638		
PC19	23.64698	0.01743	0.81381		
PC20	23.52141	0.01725	0.83105		
PC21	22.95366	0.01642	0.84748		
PC22	22.47463	0.01575	0.86322		
PC23	22.09607	0.01522	0.87844		
PC24	21.76349	0.01476	0.89321		
PC25	21.59847	0.01454	0.90775		
PC26	21.46079	0.01436	0.92211		
PC27	21.16423	0.01396	0.93607		
PC28	21.07954	0.01385	0.94992		
PC29	20.58887	0.01321	0.96313		
PC30	20.40449	0.01298	0.97611		
PC31	20.03738	0.01252	0.98863		
PC32	19.1005	0.01137	1		
PC33	2.7E-13	0	1		

transcritos ao conjunto de dados; na parte B, indicando a quantidade de genes exclusivos de cada componente (PC1 9.207 e PC2 7.589) e os genes contidos nos dois componentes (3.080). Observa-se que os gráficos da máxima variância entre os componentes do PCA obtido com o dataset 50.575 transcritos é similar ao gráfico obtido com o dataset 32.080

genes, demonstrando que a variabilidade dos dados é significativamente reduzida a partir do componente 5, mantendo uma baixa variabilidade até o componente 32.

Com a finalidade de aumentar a proporção de variação dos dados de 15% para 26%, foi criado o conjunto de dados (dataset 31.033T) contendo os transcritos dos componentes 1 e 2, como pode ser visto no diagrama de Venn apresentado na parte A da Figura 4.7. Da mesma forma, foi criado o conjunto de dados (dataset 19.876G) contendo os genes pertencentes ao PC1 e PC2, totalizando 19.876 genes únicos.

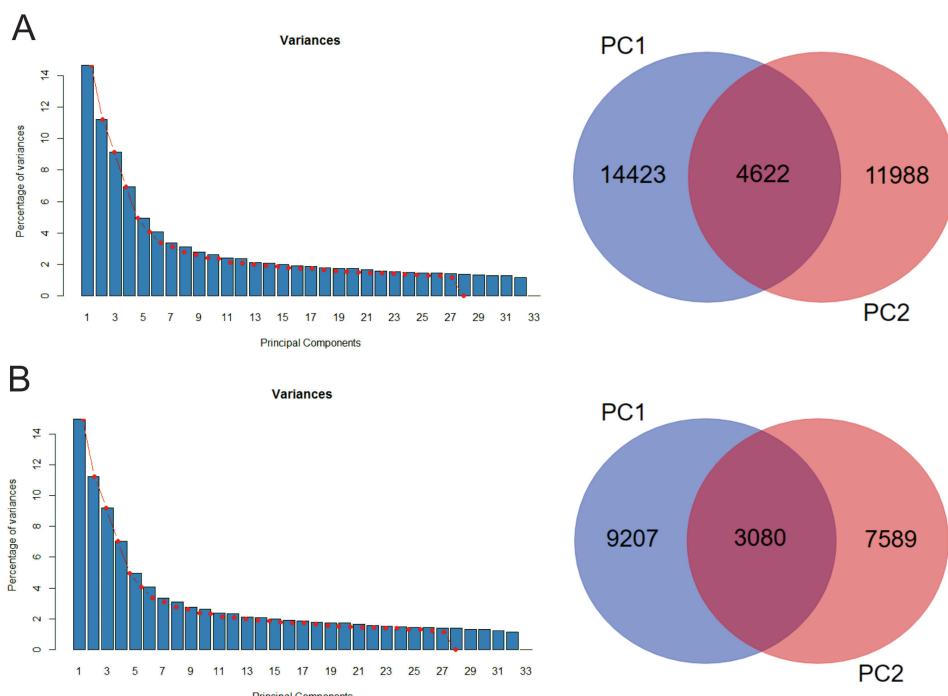


Figura 4.7: Screen Plot obtido com o PCA nas bases de dados por transcritos (A) e genes (B) e Diagrama de Veen comparando os atributos do PC1 e do PC2 das referidas base de dados.

A manutenção da alta dimensionalidade pode indicar que mesmo através do cálculo da média do valor de expressão na conversão por genes, o conjunto de dados manteve a baixa correlação observada no conjunto por transcritos. Assim, foi analisado o Volcano Plot com a finalidade de avaliar se o uso de teste de hipótese poderia apresentar melhores resultados na seleção de atributos relevantes.

4.1.7.2 Volcano Plot. Para identificar quais os transcritos ou genes são diferencialmente expressos, durante a execução do VolcanoPlot, foi avaliada a significância estatística, através de teste de hipótese obtido pelo teste t, avaliando se o valor de p era menor que 0,05 e se o fold change era 1.5 vezes maior ou menor. O fold change é uma técnica que avalia o logaritmo da média entre as duas condições comparadas pelo teste de significância estatística, identificando quais os transcritos diferem mais do que um

valor de referência. Neste experimento, o fold change foi de 1.5 vezes maior ou menor, ou seja, o transcrito ou gene é diferentemente expresso uma vez e meia negativamente ou positivamente em relação ao mesmo transcrito ou gene pertencente ao outro grupo comparado. A execução do VolcanoPlot foi realizada no programa R (RACINE, 2012), através do pacote ggplot2 (WICKHAM, 2009).

Foram realizadas 3 comparações com o uso de testes de hipótese:

Comparação 1 - Verificar se há diferença entre as amostras da fase aguda nos indivíduos sobreviventes e nos que foram a óbito.

A Figura 4.8 apresenta o Volcano plot criado na comparação dos grupos de amostras da fase aguda de indivíduos sobreviventes em relação aos indivíduos que foram a óbito. Os pontos mais diferentemente expressos são os que estão na cor verde (destacados na figura com um retângulo), tanto os do lado esquerdo como os que estão no lado direito.

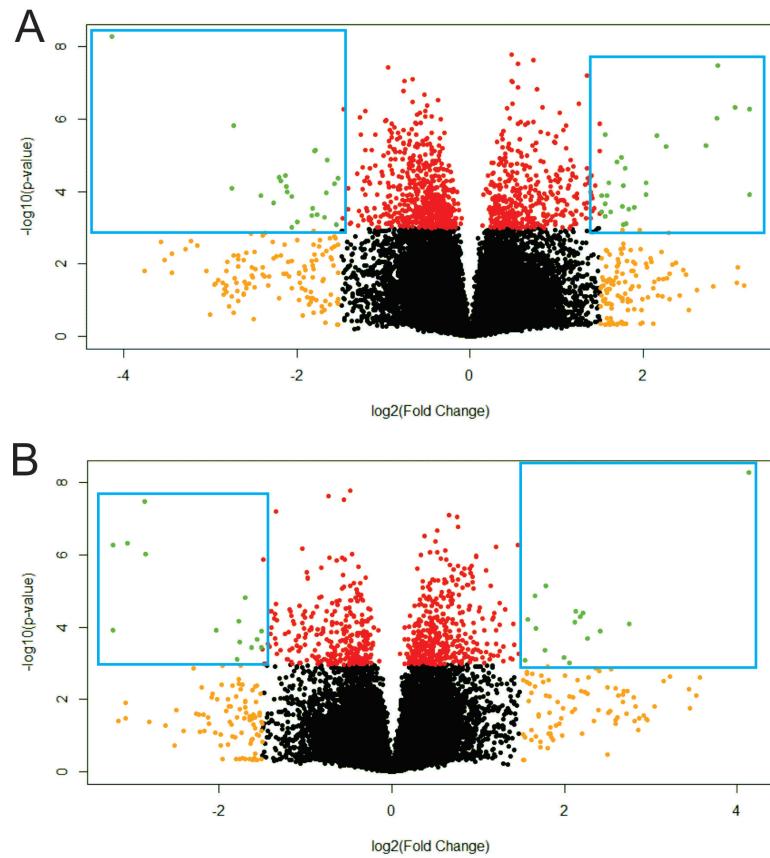


Figura 4.8: Volcano Plot obtido na comparação das 13 amostras agudas de indivíduos sobreviventes e as 03 que foram a óbito no conjunto de dados por transcritos (A) e por genes (B).

Na parte A da Figura 4.8 são apresentados os resultados obtidos com a base de dados por transcritos. Com a utilização do volcano plot, foram identificados 51 transcritos

diferentemente expressos entre amostras dos sobreviventes e dos óbitos, sendo posteriormente extraídos de forma automática, formando um dos conjuntos de dados utilizados nos experimentos deste trabalho (dataset 51T).

Na parte B da Figura 4.8 são apresentados os resultados obtidos com a base de dados por genes. Com a utilização do volcano plot, foram identificados 30 genes diferentemente expressos entre amostras dos sobreviventes e dos óbitos, formando o conjunto de dados (dataset 30G) utilizado nos experimentos deste trabalho.

Comparação 2 - Esta comparação objetivou verificar se há diferença na expressão entre as amostras da fase convalescente dos indivíduos sobreviventes e em comparação com as amostras da fase aguda dos que foram a óbito. O resultado desta comparação pode ser observado na Figura 4.9, que apresenta um número maior de pontos verdes (destacados na figura com um retângulo). Como resultado desta comparação, na base de dados por transcritos foram identificados 421 diferentemente expressos, formando um dos conjuntos de dados utilizados neste trabalho (dataset 421T). Já na comparação com a base de dados por genes, foram identificados 292 diferentemente expressos, formando o conjunto de dados (dataset 292G) utilizado nos experimentos.

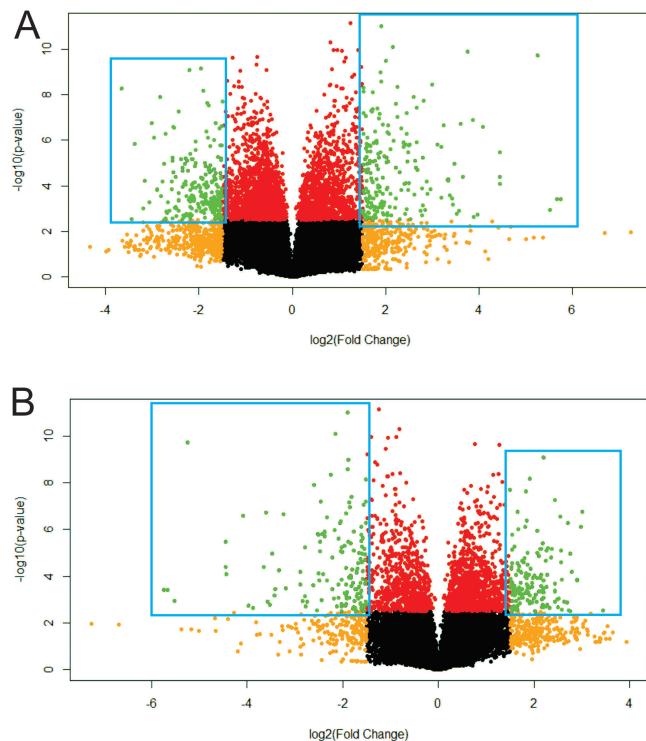


Figura 4.9: Volcano Plot obtido na comparação das 13 amostras convalescentes de indivíduos sobreviventes e as 03 amostras agudas dos que foram a óbito no conjunto de dados por transcritos (A) e por genes (B).

Comparação 3 - Nesta comparação verificou-se há diferença na expressão entre as amostras da fase convalescente dos indivíduos sobreviventes em comparação com as amos-

tras de indivíduos saudáveis utilizadas como um grupo de controle do experimento. Como pode ser observado na Figura 4.10, tanto na avaliação do conjunto de dados por transcritos quanto na dos genes, não foram identificados diferença de expressão significante, indicando que indivíduos doentes e hospitalizados, ao chegar na fase de convalescência estão praticamente recuperados, apresentando resposta imune semelhantes aos indivíduos saudáveis, validando os achados dos profissionais da saúde do IGM/FIOCRUZ-BA apresentados na Figura 4.4 B e na Tabela 4.1.

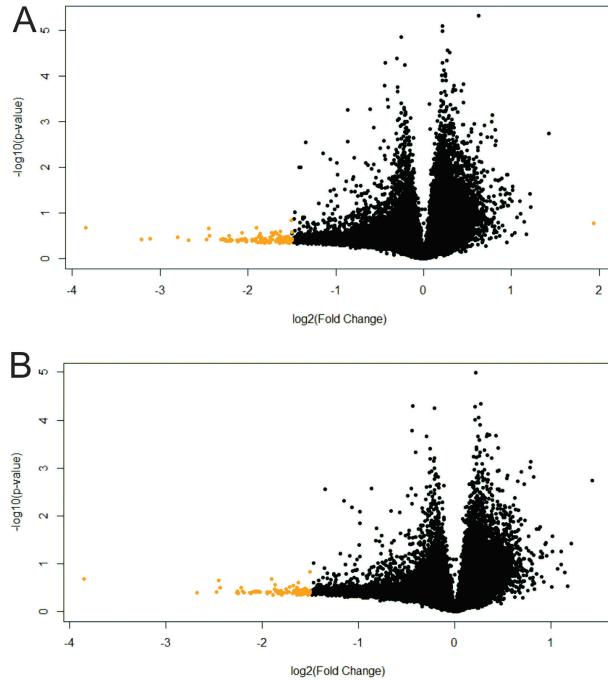


Figura 4.10: Volcano Plot obtido na comparação das 13 amostras convalescentes de indivíduos sobreviventes e as 04 amostras de indivíduos saudáveis no conjunto de dados por transcritos (A) e por genes (B).

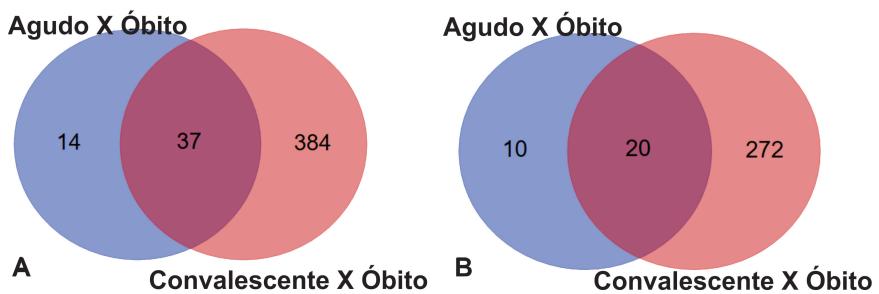


Figura 4.11: Diagrama de Venn obtido na comparação 01 e 02 do Volcano Plot. Parte A por transcritos e parte B por genes.

A Figura 4.11 apresenta os diagramas de Venn: na parte A, indicando o número de transcritos identificados na comparação 1 (51) e comparação 2 (421), bem como indi-

cando o número de transcritos exclusivos de cada comparação (comparação 1 com 14 e a comparação 2 com 384) e os transcritos identificados nas duas comparações (37); na parte B por genes, indicando o número de genes identificados na comparação 1 (30) e comparação 2 (292), bem como indicando o número de genes exclusivos de cada comparação (comparação 1 com 10 e a comparação 2 com 272 genes exclusivos) e os genes identificados nas duas comparações (20 genes).

Ao final das 03 comparações, foram identificados 435 transcritos exclusivos que apresentaram diferença de expressão no uso das métricas estatísticas utilizadas pelo Volcano Plot, formando o dataset 435T. Nas análises com a base de dados por genes foram identificados 302 genes exclusivos, formando o dataset 302G.

4.1.7.3 Conjuntos de dados por transcritos e genes. Devido à dimensionalidade dos conjuntos de dados, sejam eles por transcritos ou genes, estratégias de redução, PCA e Volcano Plot, foram avaliadas para criação de subconjuntos que permitam reduzir a dimensão sem perda de informações relevantes em relação à classe preditora.

As Tabelas 4.4 e 4.5 apresentam os conjuntos e subconjuntos de transcritos e de genes criados para serem experimentados através da aplicação de algoritmo de classificação cuja finalidade foi de avaliar a acurácia na predição das classes que representam os desfechos clínicos. Todos os conjuntos de dados possuem as informações das 33 amostras avaliadas e um atributo categórico contendo a classe preditora.

Tabela 4.4: Conjuntos de dados por transcritos utilizados nos experimentos

Conjunto de dados	Número de atributos	Descrição
dataset 50.575T	50.575	Base de dados completa por transcritos
dataset 19.045T	19.045	Transcritos identificados no PC1 pelo uso do PCA
dataset 31.033T	31.033	Transcritos identificados no PC1 e PC2 pelo uso do PCA
dataset 51T	51	Transcritos com diferenças de expressão entre o grupo de amostras agudas dos indivíduos sobreviventes e dos óbito
dataset 421T	421	Transcritos com diferenças de expressão entre o grupo de amostras convalescentes dos indivíduos sobreviventes e das amostras agudas dos indivíduos que foram a óbito.
dataset 435T	435	Junção de dois conjuntos de dados (dataset 51T e dataset 421T). .
dataset 14T	14	Transcritos identificados na parte A da Figura 4.11 como exclusivos do grupo de Acute X Death
dataset 37T	37	Transcritos identificados na parte A da Figura 4.11 com diferença de expressão nas duas comparações (Acute X Death e Convalescent X Death)
dataset 384T	384	Transcritos identificados na parte A da Figura 4.11 como exclusivos do grupo de Convalescent X Death.

Tabela 4.5: Conjuntos de dados por genes utilizados nos experimentos

Conjunto de Dados	Número de atributos	Descrição
dataset 32.080G	32.080	Base de dados completa por genes
dataset 12.287G	12.287	Genes identificados no PC1 pelo uso do PCA
dataset 19.876G	19.876	Genes identificados no PC1 e PC2 pelo uso do PCA
dataset 30G	30	Genes com diferenças de expressão entre o grupo de amostras agudas dos indivíduos sobreviventes e dos óbito
dataset 292G	292	Genes com diferenças de expressão entre o grupo de amostras convalescentes dos indivíduos sobreviventes e das amostras agudas dos indivíduos que foram a óbito.
dataset 302G	302	Junção de dois conjuntos de dados (dataset 30G e dataset 292G)
dataset 10G	10	Genes identificados na parte B da Figura 4.11 como exclusivos do grupo de Acute X Death
dataset 20G	20	Genes identificados na parte B da Figura 4.11 com diferença de expressão nas duas comparações (Acute X Death e Convalescent X Death)
dataset 272G	272	Genes identificados na parte B da Figura 4.11 como exclusivos do grupo de Convalescent X Death.

4.2 PARAMETRIZAÇÃO DOS ALGORITMOS

Nesta seção serão apresentados os algoritmos utilizados, bem como os ajustes realizados durante a avaliação dos dados de expressão gênica.

4.2.1 Algoritmos de Agrupamento

Com base nos principais algoritmos dos métodos de agrupamentos identificados na literatura, foram realizados experimentos através do programa RStudio, versão 0.98.1062 (RACINE, 2012).

Neste trabalho os algoritmos de agrupamento foram utilizados em dados de microarranjos com a finalidade de caracterizar o desfecho clínico de pacientes hospitalizados. Os algoritmos utilizados foram o K-means (HAN; KAMBER; PEI, 2011), Fuzzy C-means (BEZDEK; EHRLICH; FULL, 1984) e o algoritmo hierárquico (HAN; KAMBER; PEI, 2011). O algoritmo hierárquico foi avaliado para identificar o número de *clusters* ideal para os experimentos da análise de expressão gênica.

Na aplicação do K-means, foram ajustados o valor de K=4 no experimento 1 e o valor de K=3 no experimento 4. No experimento 2, realizado com o Fuzzy C-means, foi utilizado o agrupamento em 4 *clusters*. Com a finalidade de avaliar o número máximo de iterações realizadas pelo algoritmo, foi ajustado o valor para 100, finalizando a execução na iteração 35. Já no experimento 3, foi utilizado o método hierárquico do tipo divisível,

pois inicialmente todas as amostras seriam agrupadas no mesmo cluster e iterativamente seriam identificadas baseadas nas características do sistema imunológico das amostras avaliadas. Neste experimento, foi utilizada a função HClust (OKSANEN, 2010) com o método *Average* como estratégia de distância entre as amostras dos *clusters*. A medida utilizada para avaliar a dissimilaridade das amostras, baseado em seus dados de expressão, foi à distância euclidiana (HAN; KAMBER; PEI, 2011). Embora esta distância seja mais aplicada a clusters esféricos, com tamanho e densidade semelhantes, os resultados apresentados foram satisfatórios, agrupando as amostras avaliadas por desfecho clínico.

4.2.2 Algoritmos de Classificação

Na análise preditiva do desfecho clínico baseada em dados de expressão gênica, foi utilizado o algoritmo SVM (FACELI et al., 2011). O uso do SVM como classificador requer a escolha de uma função de kernel para calcular a margem de separação entre os objetos. Foram avaliadas nesta dissertação, as 4 principais funções de kernel: linear, função de base radial, polinomial e sigmoidal (FACELI et al., 2011). Segundo Faceli et al. (FACELI et al., 2011), a escolha da função de kernel é importante na avaliação do desempenho do classificador. Para avaliar o desempenho do SVM foi utilizada a métrica de acurácia para quantificar os acertos de predição.

4.2.2.1 Avaliação das funções de *kernel* do SVM. Quatro funções de *kernel* foram avaliadas com a finalidade de obter o melhor desempenho por parte do algoritmo SVM. As Tabelas 4.6 e 4.7 apresentam a acurácia por *fold*, bem como a média obtida nos 10 *folds* utilizados na validação cruzada, organizados pelas funções de *kernel* avaliadas.

A Tabela 4.6 foi obtida através do experimento do SVM no conjunto de dados com os 50.575 transcritos.

Tabela 4.6: Comparação da acurácia do SVM em relação aos 4 tipos de kernel avaliados em um conjunto de dados com 33 amostras.

fold	Kernel			
	Linear	Radial	Polinomial	Sigmoidal
1	100	33	33	100
2	67	100	67	100
3	100	100	67	100
4	75	100	50	75
5	100	100	33	33
6	33	100	67	100
7	50	75	25	75
8	67	100	33	67
9	67	67	67	33
10	75	100	75	100
Média	73	88	52	78

A Tabela 4.7 foi obtida através do experimento do SVM em um conjunto de dados contendo 29 amostras avaliadas pela técnica de microarranjos: 13 amostras da fase aguda e 13 amostras da fase convalescente de indivíduos que sobreviveram e 3 amostras da fase aguda de indivíduos que foram a óbito, diferenciando do conjunto de dados resultante da Tabela 4.6 por não possuir o grupo de amostras de indivíduos saudáveis.

Tabela 4.7: Comparação da acurácia do SVM em relação aos 4 tipos de kernel avaliados em um conjunto de dados com 29 amostras.

fold	VC			
	Linear	Radial	Polinomial	Sigmoidal
1	100	100	50	100
2	100	100	100	100
3	100	100	67	100
4	100	100	67	100
5	100	100	67	67
6	100	100	33	100
7	100	100	67	100
8	100	100	100	67
9	100	100	67	100
10	100	100	100	67
Média	100	100	72	90

Observou-se que o uso da função de kernel='radial', apresentou as melhores taxas de acurácia, tanto na avaliação do conjunto de dados contendo 33 amostras, quanto no conjunto com 29 amostras, apresentando taxas de 88% e 100% respectivamente.

4.3 CONSIDERAÇÕES FINAIS

Neste capítulo foram descritos os conjuntos de dados utilizados na avaliação da viabilidade e predição do desfecho clínico baseado em dados de expressão gênica. Nos experimentos de avaliação da viabilidade, foi utilizado o conjunto de dados contendo 389 transcritos identificados por (LINDOW et al., 2016) na comparação do grupo de óbito em relação as amostras agudas dos sobreviventes. Para a predição do desfecho clínico, foram utilizados dois métodos: um contendo o conjunto de dados por transcritos (50.575 instâncias) e outro por genes (30.080 instâncias). Seções para descrever os parâmetros ajustados nos algoritmos de agrupamento e classificação foram apresentadas para auxiliar na reproduzibilidade do trabalho.

Capítulo

5

Neste capítulo os experimentos são apresentados de acordo a metodologia proposta.

EXPERIMENTOS

Neste capítulo os métodos e experimentos foram divididos de acordo com a metodologia proposta. Assim, este trabalho pode ser dividido em três conjuntos de experimentos, conforme a Figura 5.1.

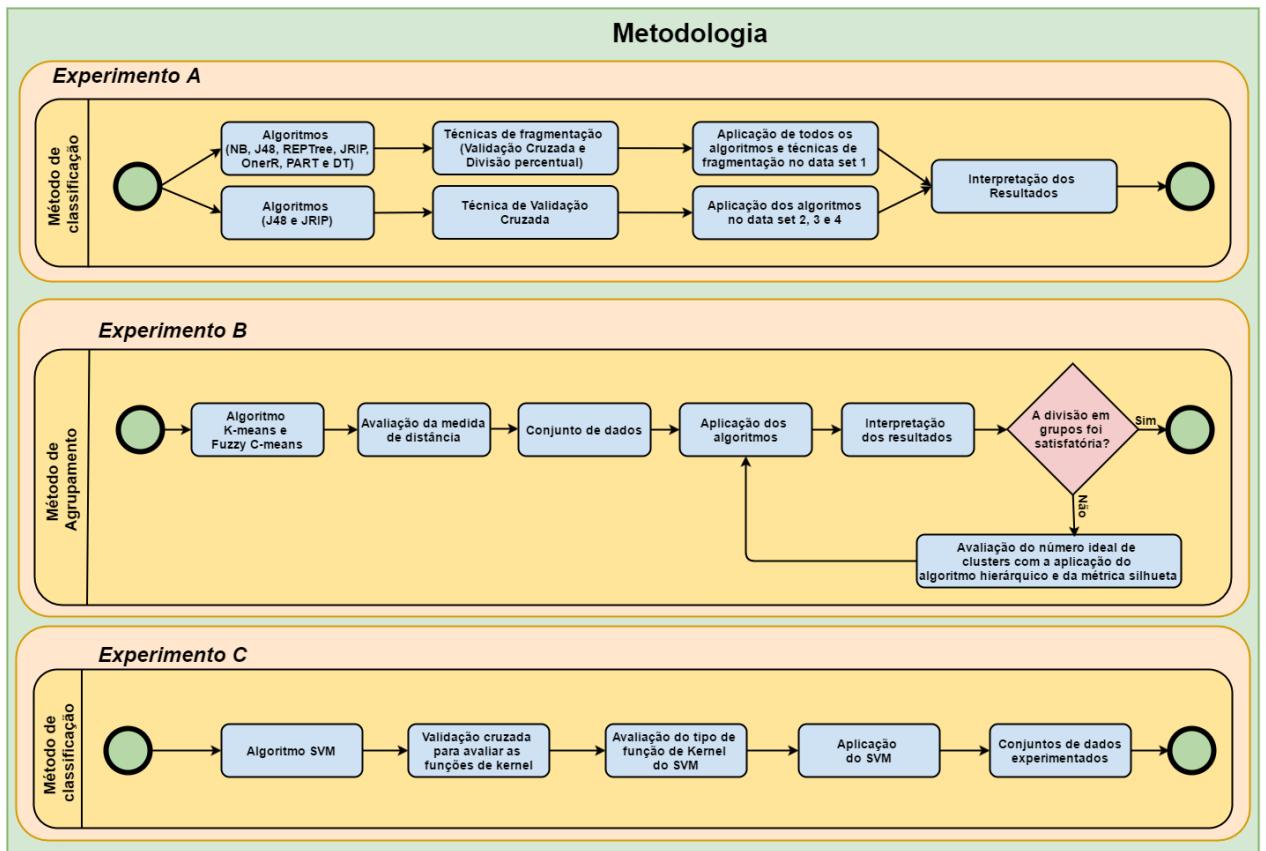


Figura 5.1: Resumo da metodologia da dissertação.

- A - No experimento A, foram avaliados algoritmos de classificação com a finalidade de identificar o melhor modelo de predição de casos de leptospirose baseado em dados clínicos e epidemiológicos.
- B - No experimento B, um subconjunto de dados de microarranjos foi utilizado para avaliar a viabilidade de caracterizar o desfecho clínico de amostras de pacientes hospitalizados baseado na similaridade de expressão gênica.
- C - No experimento C, (por transcritos e por genes) classificou-se os dados de expressão de acordo com o desfecho clínico e posterior identificação dos genes responsáveis pelo agravamento do estado de saúde do indivíduo.

Estes experimentos serão detalhados nas seções seguintes:

5.1 EXPERIMENTO A

Um modelo de predição de casos de leptospirose baseado em dados clínicos e epidemiológicos pode servir como auxílio aos profissionais de saúde na avaliação dos pacientes, principalmente em regiões rurais ou que não tenham hospitais preparados para o diagnóstico. Neste sentido, foram realizados experimentos com os principais métodos de classificação com o objetivo de avaliar os melhores modelos de predição de casos da doença.

5.1.1 Testes preditivos

Foram realizados quatro experimentos utilizando conjuntos de dados específicos criados para treinamento, testes, avaliação da sensibilidade e especificidade obtidos da base de dados da vigilância ativa de leptospirose do IGM/FIOCRUZ-BA, como pode ser visto na Figura 5.2.

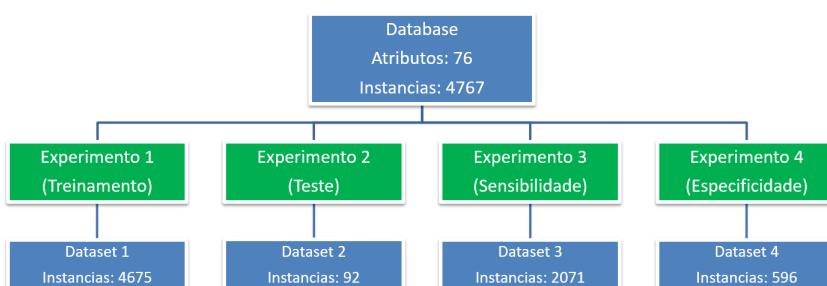


Figura 5.2: Fluxo dos experimentos e conjuntos de dados utilizados no experimento A.

O primeiro experimento teve como objetivo prever casos utilizando a base de dados de leptospirose do período de março de 1996 até setembro de 2014, que foi dividida em conjuntos de dados de treinamento e testes. Os algoritmos foram executados três vezes com técnicas de fragmentação diferente: na primeira execução, utilizou-se a técnica *Percentage Split* (PS), dividindo o conjunto de dados em 66% para treinamento e os 34%

restantes para teste. Na segunda execução, aplicou-se a técnica PS, utilizando 80% para treinamento e 20% para o teste. Na terceira execução, utilizou-se a técnica de Validação Cruzada (CV) com 10 *folds* (NERY; CLARO; LINDOW, 2016). Neste experimento foram avaliados o desempenho dos algoritmos: Naive Bayes (NB), J48, JRIP, REPTree, OneR, PART e DecisionTable (DT).

Os algoritmos JRIP e J48 apresentaram os melhores resultados no primeiro experimento e foram selecionados para experimentos adicionais. No segundo experimento, adicionou-se registros de pacientes recrutados no período de outubro de 2014 a junho de 2016 para avaliar o desempenho dos modelos.

O terceiro experimento teve como objetivo avaliar exclusivamente a sensibilidade dos modelos obtidos com os algoritmos JRIP e J48 no conjunto de treinamento. Foram testados dois modelos em um conjunto de dados contendo apenas casos de leptospirose clinicamente confirmados retirados da vigilância ativa do IGM/FIOCRUZ-BA. Este conjunto de dados contém 76 variáveis e 2071 casos de pacientes recrutados de março de 1996 a junho de 2016.

O objetivo do quarto experimento foi avaliar a especificidade dos algoritmos JRIP e J48 utilizando uma base de dados contendo apenas casos de dengue clinicamente confirmados, uma doença com sintomas iniciais semelhantes à leptospirose. Este conjunto de dados é oriundo da base de dados da vigilância ativa do IGM/FIOCRUZ-BA, contendo indivíduos que foram identificados como casos suspeitos de ter leptospirose, mas que foram confirmados laboratorialmente como dengue. Este conjunto de dados continha 76 variáveis e 596 casos de pacientes recrutados a partir de março 1996 a junho de 2016.

5.2 EXPERIMENTO B

Este experimento teve como principal objetivo avaliar se havia similaridade nos dados de expressão gênica em relação ao desfecho clínico das amostras dos pacientes analisados. Este conjunto de experimento foi dividido em 4 testes:

5.2.1 Teste 1

Aplicação do K-means com K=4 para avaliar o agrupamento das amostras avaliadas pela técnica de microarranjos em suas 4 categorias: sobreviventes fase aguda; sobreviventes fase convalescente; óbito fase aguda; indivíduos saudáveis.

5.2.2 Teste 2

Nesta etapa foi utilizado o algoritmo Fuzzy C-means para avaliar o maior grau de pertinência das amostras em relação aos grupos dos óbito ou sobreviventes na fase aguda ou convalescente.

5.2.3 Teste 3

Nesta etapa foi utilizado o método hierárquico para avaliar, de forma não supervisionada, o numero ideal de *clusters* que melhor agrupasse as amostras.

5.2.4 Teste 4

Devido os resultados obtidos no teste 3, bem como pela avaliação da métrica de silhueta, foi realizado outro experimento com o algoritmo K-means definindo o valor de K=3.

5.3 EXPERIMENTO C

Neste experimento foram utilizados métodos com o uso de conjuntos de dados por transcritos e por genes, com a finalidade de avaliar qual estratégia de análise apresenta melhores resultados. O conjunto de dados por transcritos contém 50.575 atributos (dataset 50.575T) com dados de expressão, já o conjunto de dados por genes possui 32.080 atributos (dataset 32.080G) contendo a média de expressão dos transcritos pertencentes ao mesmo gene. Subconjuntos resultantes do uso das técnicas de redução da dimensionalidade foram experimentados para avaliar o desempenho do algoritmo de classificação em conjuntos de dados menores.

5.3.1 Classificação de dados de expressão gênica

Para avaliar o impacto da redução da dimensionalidade dos métodos por transcritos e por genes, foram realizados experimentos nos conjuntos de dados descritos nas Tabelas 4.4 e 4.5, aplicando o algoritmo de SVM, com a finalidade de classificar as amostras avaliadas de acordo com a classe de referência representada pelo seu desfecho clínico. Os modelos de predição, obtidos pelo algoritmo SVM. A técnica de fragmentação utilizada para dividir o conjunto de dados em conjuntos de treinamento e teste foi à validação cruzada (CV) com 10 *folds*, que segundo (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996a), esta técnica soluciona o problema de *overfitting*, principalmente ao se trabalhar com conjuntos de dados com poucas observações.

5.4 CONSIDERAÇÕES FINAIS

Neste capítulo foram descritos os 3 conjuntos de experimentos que nortearam esta dissertação: o experimento A mensurou algoritmos de classificação na predição de casos de leptospirose; o experimento B analisou a viabilidade de se trabalhar com expressão gênica com a finalidade de caracterizar o desfecho clínico das amostras avaliadas por similaridade de expressão; o experimento C considerou o uso de técnicas de redução de dimensionalidade para auxiliar na predição do desfecho clínico e identificação dos genes responsáveis pelo agravamento de indivíduos doentes, mensurando o uso de conjuntos de dados completo por transcritos e por genes, além de avaliar subconjuntos resultantes das técnicas de redução da dimensionalidade.

Capítulo

6

Este capítulo apresenta os resultados obtidos através dos experimentos apresentados.

RESULTADOS

Os resultados deste trabalho foram divididos em 3 conjuntos de acordo com os experimentos:

Resultado A: Predição de casos de leptospirose

Resultado B: Agrupamento por similaridade de expressão gênica

Resultado C: Classificação de dados de expressão gênica por desfecho clínico

6.1 RESULTADO A

A Figura 6.1 apresenta a proporção de classificações corretas e seus respectivos valores obtidos através das estatísticas Kappa. Observou-se nesta figura que ao avaliar o desempenho dos algoritmos Naive Bayes (NB), J48, JRIP, REPTree, OneR, PART e DecisionTable (DT), os melhores resultados foram com os algoritmos JRIP e J48. Enquanto o modelo obtido com o J48 produziu resultados promissores, o modelo do algoritmo JRIP apresentou os melhores resultados para a precisão, especialmente na técnica de fragmentação de percentagem de divisão 80.

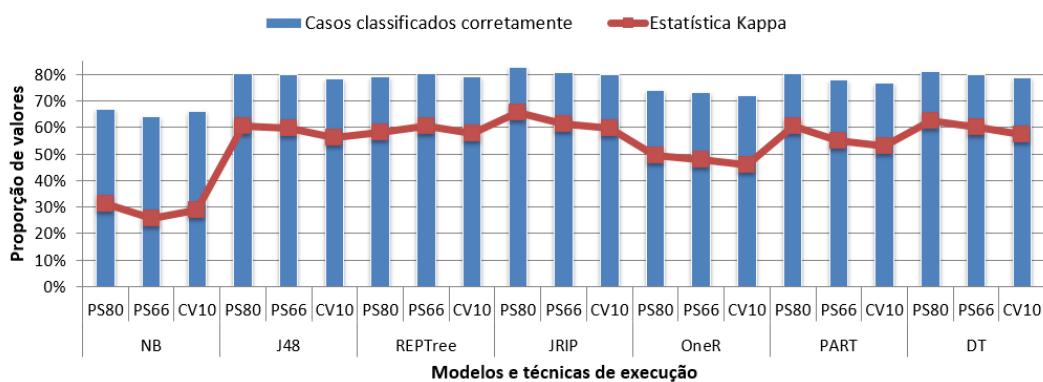


Figura 6.1: Comparação da precisão de instâncias e do valor obtido com a estatística Kappa, por algoritmos e técnicas de fragmentação (NERY; CLARO; LINDOW, 2016).

Em experimentos de teste (dataset 2), o modelo obtido com o algoritmo JRIP manteve a sensibilidade de 84%, mas a especificidade diminuiu para 52%. A especificidade do modelo obtido com J48 neste conjunto de dados também foi de 52%, a sensibilidade aumentou de 75% para 82% em comparação com o Experimento 1 (dataset 1) devido ao uso de um conjunto de dados com atributos relevantes.

Na avaliação da sensibilidade (dataset 3) dos algoritmos JRIP e J48 com o uso de um conjunto de dados contendo apenas casos confirmados por leptospirose, o JRIP reduziu a sensibilidade de 84% para 82%, enquanto J48 aumentou a sensibilidade de 75% para 85%.

Na avaliação da especificidade (dataset 4) dos algoritmos JRIP e J48 com o uso de um conjunto de dados contendo apenas casos confirmados por dengue, o JRIP e J48 apresentaram especificidades de 99% e 98%, respectivamente, indicando que os modelos têm alta especificidade na discriminação entre pacientes sem leptospirose.

A Figura 6.2 ilustra uma comparação entre as taxas de sensibilidade e de especificidade obtidos no Experimento A. Na comparação da sensibilidade só foi possível avaliar os experimentos 1, 2 e 3, pois no conjunto de dados do experimento 4 não tinham casos de leptospirose. Da mesma forma, o experimento 3 não relata dados de especificidade do experimento 3, uma vez que o conjunto de dados utilizado possuía apenas casos de leptospirose.

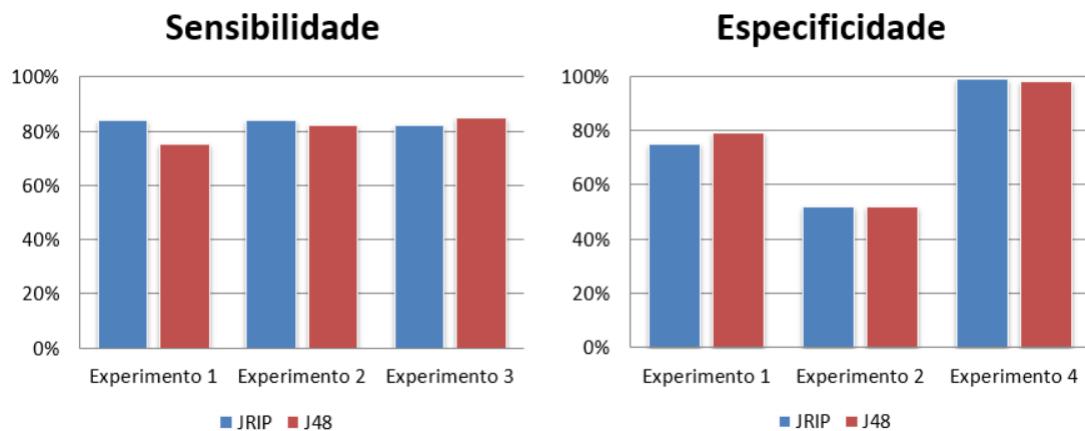


Figura 6.2: Comparação de experimentos de sensibilidade e especificidade (JUNIOR; CLARO; LINDOW, 2017).

6.2 RESULTADO B

Conforme descrito no Experimento B, no capítulo anterior, este resultado também foi dividido em 4 testes:

6.2.1 Resultado do Teste 1

No primeiro experimento com o uso do algoritmo k-means, o valor de K definido foi 4 por corresponder aos quatro grupos de amostras avaliadas: das fase Agudas (S) e convalescentes (C) dos sobreviventes, amostras da fase aguda dos óbito (D) e das amostras de indivíduos saudáveis (H). De acordo com a Tabela 6.1, ao determinar o número 4 como o K, foi observado que o algoritmo agrupou as amostras convalescentes e saudáveis no grupo 1, no grupo 2 todas as amostras agudas de óbito, mas subdividiu as amostras agudas dos sobrevivente em dois grupos: o grupo 3 com 4 amostras e o grupo 4 com 09 amostras.

Tabela 6.1: Comparação dos grupos obtidos com a aplicação do algoritmo K-means com k=4 em relação aos tipos de amostras avaliadas pela técnica de microarranjos

Grupos	Amostras			
	S	C	D	H
1	0	13	0	4
2	0	0	3	0
3	4	0	0	0
4	9	0	0	0

6.2.2 Resultado do Teste 2

A Tabela 6.2 apresenta o resultado da aplicação do experimento com o Fuzzy C-means. Nesta tabela observa-se o grau de pertinência de cada amostra avaliada para os 4 grupos identificados na execução do algoritmo. Como pode ser observado na Tabela 6.2, o maior grau de pertinência de cada amostra está destacado na cor verde, com a finalidade de facilitar as análises. Todas as amostras avaliadas possuem uma identificação de registro e uma letra que representa o seu desfecho clínico referente ao estudo do grupo de leptospirose do IGM/FIOCRUZ-BA. As letra S e C são amostras na fase aguda e convalescente respectivamente dos indivíduos doentes que sobreviveram, a letra D são amostras agudas de indivíduos que foram a óbito e por fim a letra H são amostras de indivíduos saudáveis. Os resultados apresentados na Tabela 6.2 podem ser resumidos analisando a matriz de confusão apresentada na Tabela 6.3.

Tabela 6.2: Grau de pertinência das amostras por grupo obtido com a aplicação do algoritmo Fuzzy.

Amostras	Grupos			
	1	2	3	4
1-S	0,04351055	0,17676424	0,70145433	0,07827088
2-S	0,04563856	0,66257987	0,21261233	0,07916924
3-S	0,04961012	0,2128893	0,62490962	0,11259096
4-S	0,09164181	0,29529195	0,5082042	0,10486205
5-S	0,1918241	0,21123257	0,48981271	0,10713062
6-S	0,03836718	0,6014605	0,23023789	0,12993442
7-S	0,04886232	0,33720973	0,49054591	0,12338205
8-S	0,11553767	0,27080768	0,51924279	0,09441186
9-S	0,05791931	0,61980063	0,19934785	0,12293221
10-S	0,05183463	0,46773976	0,40365527	0,07677034
11-S	0,07470781	0,15167388	0,65965159	0,11396672
12-S	0,0654981	0,12990061	0,70917631	0,09542498
13-S	0,02829976	0,76276399	0,15176722	0,05716903
14-D	0,75127292	0,06844219	0,12256995	0,05771494
15-D	0,81474834	0,04387127	0,08471906	0,05666133
16-D	0,79506919	0,05492055	0,08866914	0,06134111
1-C	0,06086937	0,08704366	0,16840474	0,68368223
2-C	0,04522272	0,21182822	0,20503699	0,53791207
3-C	0,06243109	0,09604344	0,1311042	0,71042127
4-C	0,04436249	0,11661377	0,18787638	0,65114736
5-C	0,03002282	0,08510836	0,1094435	0,77542532
6-C	0,05501254	0,138806	0,24127835	0,5649031
7-C	0,05084056	0,08554065	0,12456005	0,73905873
8-C	0,04066961	0,17022328	0,21397515	0,57513195
9-C	0,08492704	0,13316738	0,30891279	0,47299279
10-C	0,02808886	0,05628515	0,07901137	0,83661462
11-C	0,05801731	0,09243841	0,17582304	0,67372124
12-C	0,03025083	0,06812929	0,09782151	0,80379837
13-C	0,05110059	0,24363056	0,20919673	0,49607211
17-H	0,02391468	0,05696652	0,0898774	0,8292414
18-H	0,02935284	0,08522907	0,12539557	0,76002252
19-H	0,04733887	0,07238281	0,11171715	0,76856118
20-H	0,0891391	0,15372549	0,19209067	0,56504474

Tabela 6.3: Comparação dos grupos obtidos com a aplicação do algoritmo Fuzzy com $k=4$ em relação aos tipos de amostras avaliadas pela técnica de microarranjos

Grupos	Amostras			
	S	C	D	H
1	0	0	3	0
2	5	0	0	0
3	8	0	0	0
4	0	13	0	4

O algoritmo Fuzzy apresentou resultados similares com os obtidos pela aplicação do k-means com o número 4 como k, repetindo o agrupamento de todas as amostras dos óbito apresentada no 1º grupo, juntando as amostras da fase convalescente e saudáveis no 4º grupo e também subdividindo as amostras da fase aguda dos sobreviventes em dois grupos, diferenciando apenas a quantidade de objetos, tendo no 2º grupo 5 amostras e no 3º grupo 8 amostras, como pode ser observado na Figura 6.3. Esta subdivisão de amostras na fase aguda pode indicar um estado mais crítico na saúde dos indivíduos pertencentes ao grupo 2 ou 3, o que pode ser verificado com uma análise dos dados clínicos e laboratoriais.

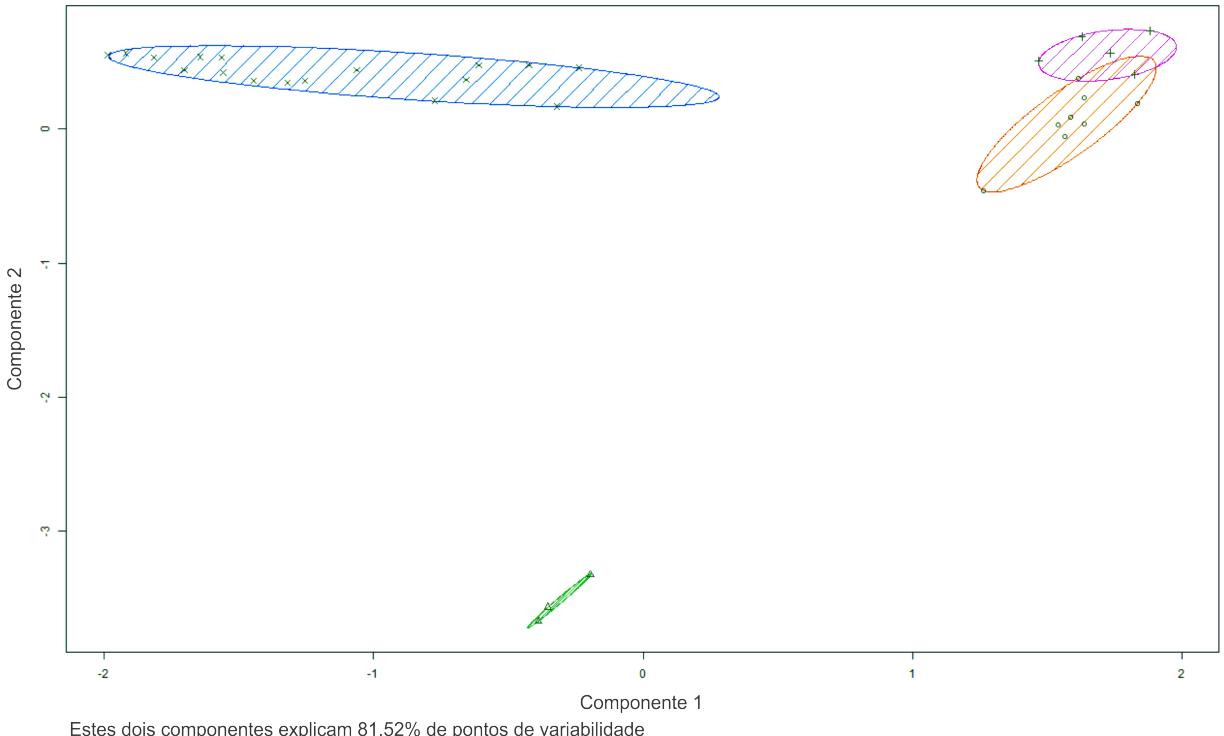


Figura 6.3: Resultado da aplicação do algoritmo Fuzzy.

6.2.3 Resultado do Teste 3

Na Figura 6.4 é apresentado o dendrograma com o agrupamento das amostras avaliadas. O algoritmo hierárquico implementado agrupou as amostras em 3 grandes grupos: um grupo contendo todas as 3 amostras dos indivíduos que foram a óbito (14-D, 15-D e 16-D), outro contendo todas as amostras dos indivíduos na fase aguda da doença e um grupo contendo as 13 amostras dos indivíduos sobreviventes na fase convalescente e as amostras dos 04 indivíduos saudáveis.

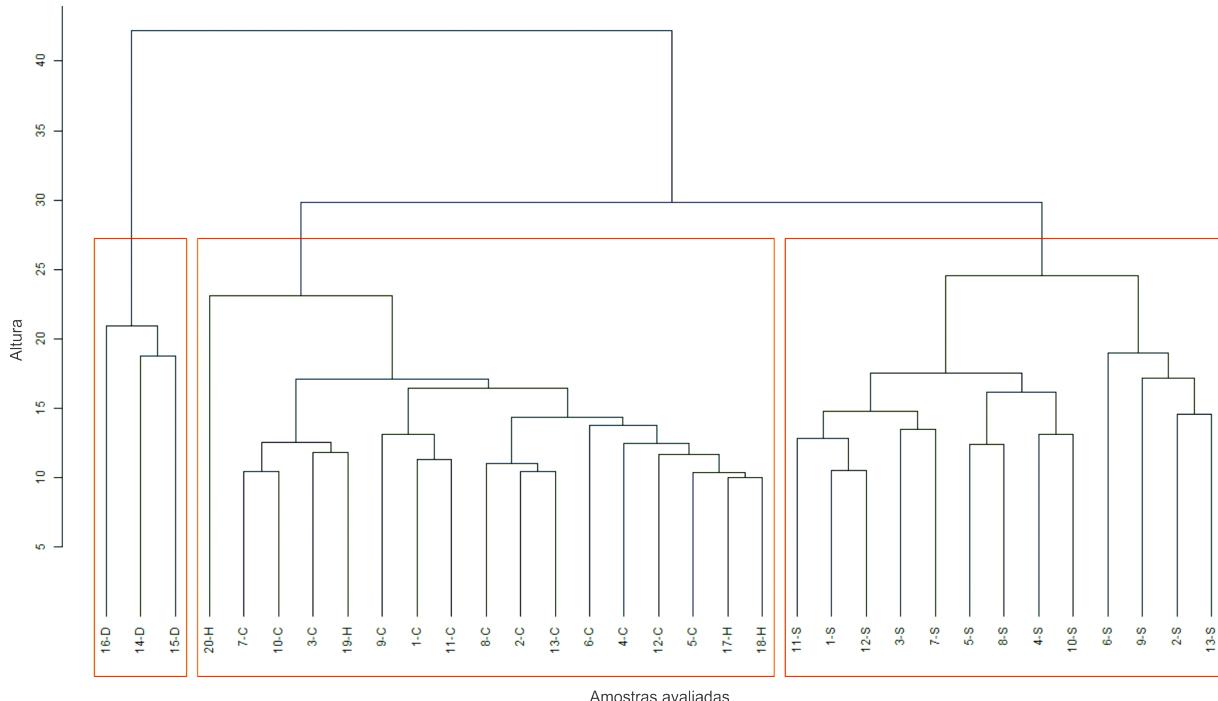


Figura 6.4: Dendrograma apresentando três grandes grupos de amostras avaliadas na análise de expressão gênica.

Com o resultado obtido com o dendrograma, o número de agrupamento foi alterado, como poder ser visto na próxima seção.

6.2.4 Resultado do Teste 4

Neste experimento, repetiu-se o algoritmo K-means ajustando o valor de $k=3$, onde as amostras da fase convalescente dos sobreviventes foram agrupadas com as amostras dos indivíduos saudáveis, bem como pelo resultado obtido com o algoritmo hierárquico. Na Tabela 6.4 é apresentado o resultado da execução deste experimento, no qual pode-se observar as mesmas características do experimento 1 com 3 grupos.

Tabela 6.4: Comparação dos grupos obtidos com a aplicação do algoritmo K-means com $k=3$ em relação aos tipos de amostras avaliadas pela técnica de microarranjos

Grupos	Amostras			
	S	C	D	H
1	13	0	0	0
2	0	0	3	0
3	0	13	0	4

Validação do agrupamento Foram utilizadas as métricas de validação interna: conectividade (FACELI et al., 2011), silhueta (FACELI et al., 2011) e índice de Dunn (FACELI et al., 2011), através do pacote c1Valid (BROCK et al., 2011), com a finalidade de quantificar o número ideal de clusters, bem como avaliar a qualidade do agrupamento. Na Tabela 6.5 são apresentados os resultados de cada métrica em relação a aplicação do K-means com 2, 3 e 4 clusters.

Tabela 6.5: Comparação das métricas de avaliação dos Grupos

Medidas de validação	Grupos		
	2	3	4
Conectividade	11.8639	6.5226	12.1679
Dunn	0.277	0.4939	0.5547
Silhueta	0.3247	0.3949	0.3681

Os melhores resultados apresentados na Tabela 6.5 estão destacados, indicando que 2 das 3 métricas avaliadas apresentaram melhores resultados para o uso do K-means com 3 clusters. Diferente das métricas Silhueta e Índice de Dunn, a métrica Conectividade considera como melhor partição a que apresentar o menor valor de índice. A Figura 6.5 apresenta o resultado da avaliação da métrica Silhueta, indicando no eixo vertical as amostras avaliadas e no eixo horizontal o valor da silhueta.

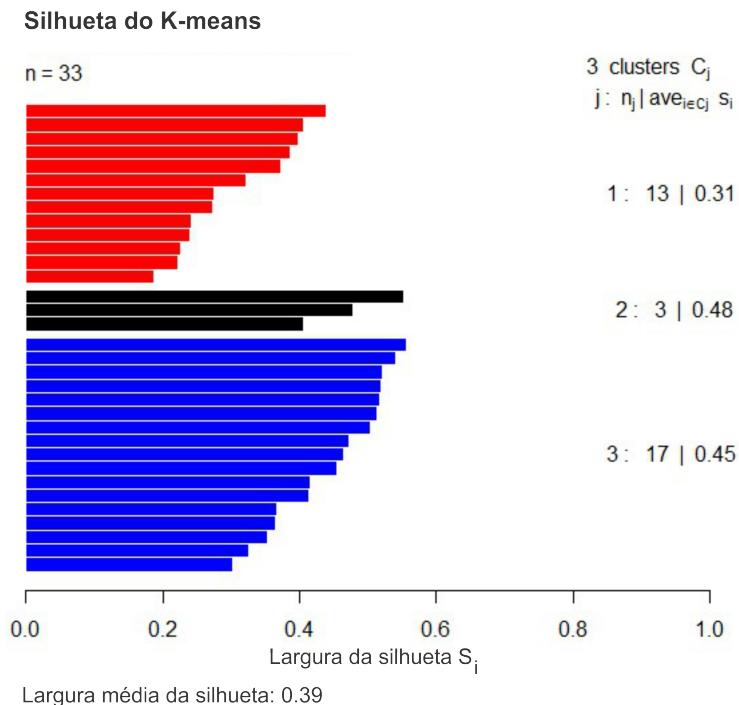


Figura 6.5: Medida da Silhueta na avaliação do K-means com 3 clusters.

Embora os valores da silhueta tenham sido pouco expressivos, o uso da métrica contribuiu para a escolha do número ideal de *clusters* que permitissem descrever os dados analisados.

Estes experimentos apresentaram resultados relevantes do ponto de vista da análise de expressão gênica, indicando que realmente há diferenças no sistema imunológico entre indivíduos que foram a óbito em relação aos indivíduos sobreviventes, principalmente na fase aguda da doença. Outro fator importante identificado neste experimento foi à similaridade entre os indivíduos na fase convalescente e indivíduos saudáveis, corroborando com os resultados obtidos em (LINDOW et al., 2016).

6.3 RESULTADO C

6.3.1 Resultado da Classificação dos dados por Transcritos

A Figura 6.6 apresenta a comparação da acurácia resultante da aplicação do SVM nos conjuntos de dados por transcritos descritos na Tabela 4.4. Observou-se que ao utilizar a base de dados completa (50575 transcritos, conjunto A), a acurácia foi de 76%.

Ao avaliar os conjuntos de dados obtidos com o método PCA, observou-se que o dataset 19.045T obtido com o primeiro componente (19.045 transcritos) e com o dataset 31.033T contendo o primeiro e segundo componente (31.033 transcritos) apresentaram a proporção de acurácia de 76%, a mesma obtida com o uso do dataset 50.575T. Indicando que mesmo com uma variabilidade dos dados de 25.86%, os primeiros componentes do PCA mantiveram as características no processo de classificação do conjunto de dados completo.

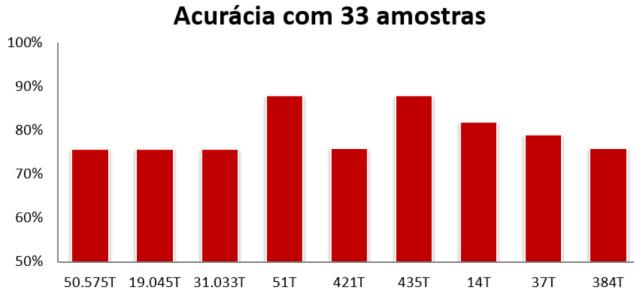


Figura 6.6: Comparação da acurácia dos conjuntos de dados por transcritos com os 04 grupos de amostras avaliados.

Ao avaliar os conjuntos de dados obtidos com os métodos estatísticos através do uso do Volcano Plot, constatou-se que a acurácia dos conjuntos de dados foram iguais ou superiores aos obtidos com o uso do PCA. O dataset 51T, juntamente com o dataset 435T, apresentaram as melhores proporções de acurácia, ambas com 88%. Os datasets 14T e 37T, subdivisões do dataset 51T, apresentaram proporções de acertos de 82% e 79% respectivamente, apresentando melhores resultados que ao utilizar a base de dados completa. Os conjuntos dataset 421T e dataset 435T apresentaram a mesma taxa de acerto obtida com o dataset 50.575T (76%).

Ao avaliar as matrizes de confusão obtidas nos experimentos, foi percebido que as 4 amostras pertencentes ao grupo de indivíduos saudáveis foram o que mais influenciou nos erros de predição de cada modelo, como pode ser visto na Tabela 6.6, que apresenta a matriz de confusão obtida pelo modelo ao utilizar o conjunto D.

Tabela 6.6: Matriz de Confusão obtida com a aplicação do SVM no dataset 51T.

Amostras	C	D	H	S
C	13	0	0	0
D	0	3	0	0
H	4	0	0	0
A	0	0	0	13

Como observado na Figura 4.10, não foram encontrado transcritos com diferença de expressão através do uso do Volcano Plot, indicando que o sistema imune de indivíduos doentes por leptospirose ao chegar na fase de convalescência assemelha-se ao sistema imune de indivíduos saudáveis. Por esta razão, em todos os experimentos realizados nos nove conjuntos de dados, o algoritmo SVM errou em classificar as amostras de indivíduos saudáveis, classificando as amostras saudáveis como amostras da fase de convalescência. Sendo assim, foram repetidos os experimentos nos nove conjuntos de dados removendo as 04 instâncias das amostras pertencentes aos indivíduos saudáveis, com a finalidade de avaliar a acurácia do SVM em cada conjunto de dados.

A Figura 6.7 apresenta a proporção de acertos de classificação nos nove conjuntos de dados sem as 4 amostras de indivíduos saudáveis. Constatou-se que o desempenho do

SVM melhorou em todos os conjuntos de dados experimentados, subindo de 76% para 86% nos conjuntos dataset 50.575, dataset 19.045, dataset 31.033 e dataset 384T. No dataset 421T, a taxa subiu de 76% para 90%, no dataset 14T, subiu de 82% para 90%, no dataset 37T, subiu de 79% para 93%. Os melhores resultados mantiveram para os experimentos com o dataset 51T e o dataset 435T, subindo a taxa de acertos de 88% para 100%.

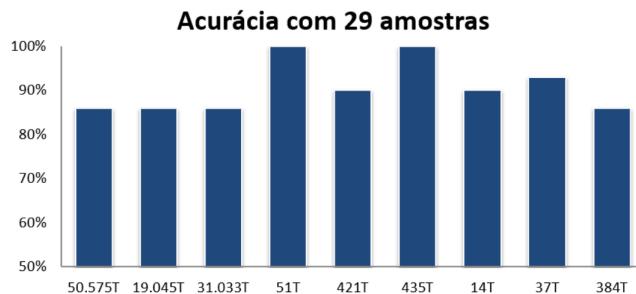


Figura 6.7: Comparação da acurácia dos conjuntos de dados por transcritos sem o grupo de amostras dos indivíduos saudáveis.

De acordo com os experimentos, o uso de métodos estatísticos como testes de hipóteses e fold change, realizados através do Volcano Plot, apresentaram os melhores resultados quando utilizados como técnica de redução da dimensionalidade, apresentando as melhores taxas de acurácia na classificação de amostras avaliadas com a aplicação do algoritmo SVM.

A Figura 6.8 apresenta dois biplots: um com o dataset 50.575T e o outro com o dataset 51T. No primeiro biplot, devido a variabilidade dos dados, não é possível separar as instâncias do conjunto de dados por amostras avaliadas, o que pode ser observado no segundo biplot, que agrupou as amostras dos 3 óbito, as 13 amostras de fase aguda dos sobreviventes e devido as características, agrupou as 13 amostras dos sobreviventes da fase convalescente com as 4 amostras dos indivíduos saudáveis.

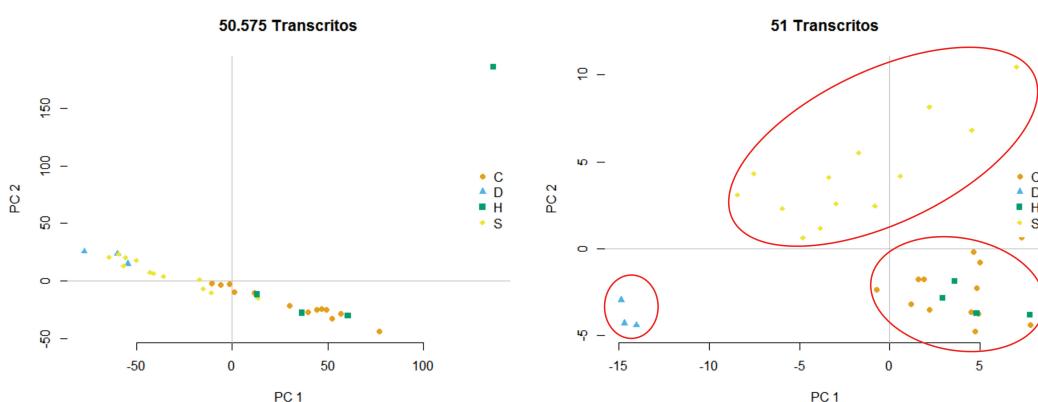


Figura 6.8: Comparação dos gráficos biplots obtidos com a base de 50.575 e com a base dos 51 transcritos.

6.3.2 Resultado da Classificação dos dados por Genes

Após a realização dos experimentos de redução da dimensionalidade e criação dos subconjuntos apresentados na Tabela 4.5, foram realizados experimentos de classificação dos grupos de amostras e avaliação da taxa de acurácia. A Figura 6.9 apresenta a comparação da taxa de acurácia obtida após a execução do SVM nos conjuntos de dados por genes. No dataset 32.080G, a acurácia foi de 76%, a mesma apresentada no experimento da base completa por transcritos.

Ao avaliar os conjuntos de dados obtidos com o método PCA, o dataset 12.287G, obtido com o primeiro componente (12.287 genes), e com o dataset 19.876G, contendo os dois primeiros componentes (19.876 genes), apresentaram a proporção de acurácia de 76%, a mesma obtida com o uso do dataset 32.080G. Indicando que mesmo com uma variabilidade dos dados de 26.18%, os primeiros componentes do PCA mantiveram as características no processo de classificação do conjunto de dados completo.

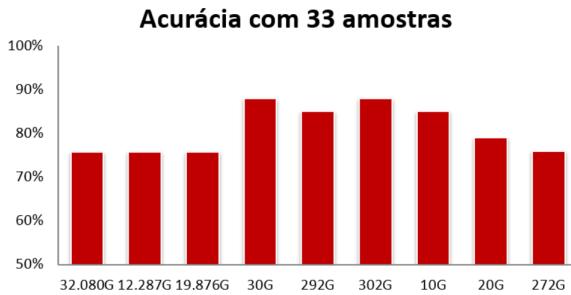


Figura 6.9: Comparação da acurácia dos conjuntos de dados por genes com os 04 grupos de amostras avaliados.

Avaliando os conjuntos de dados obtidos com os métodos estatísticos através do uso do Volcano Plot, a acurácia dos conjuntos de dados foi igual ou superior aos resultados obtidos com o uso do PCA ou do conjunto de dados inicial. O dataset 30G e o dataset 302G apresentaram as melhores proporções de acurácia, ambas com 88%. Os conjuntos dataset 10G, dataset 20G e dataset 272G são subdivisões do dataset 302G e apresentaram proporções de acertos de 85%, 79% e 76% respectivamente, apresentando resultados melhores ou semelhantes em relação a utilização da base de dados completa.

Tabela 6.7: Matriz de Confusão obtida com a aplicação do SVM no dataset 30G.

Amostras	C	D	H	S
C	13	0	0	0
D	0	3	0	0
H	4	0	0	0
A	0	0	0	13

Os experimentos de classificação dos nove conjuntos de dados foram repetidos, removendo as 4 instâncias das amostras pertencentes aos indivíduos saudáveis, com o intuito

de reavaliar a taxa acurácia do SVM. A Figura 6.10 apresenta a proporção de acertos de classificação da repetição dos conjuntos de dados sem as 4 amostras de indivíduos saudáveis. O desempenho do SVM melhorou em todos os conjuntos de dados experimentados, subindo de 76% para 86% nos conjuntos dataset 32.080G, dataset 12.287G, dataset 19.876G e dataset 272G. No dataset 20G, a taxa subiu de 79% para 90%, nos conjuntos: dataset 292G e dataset 10G, a taxa aumentou de 85% para 97%. Os melhores resultados mantiveram para os experimentos com os dataset 30G e dataset 302G, subindo a taxa de acertos de 88% para 100%.

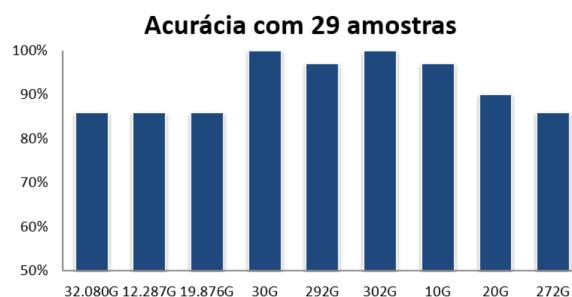


Figura 6.10: Comparação da acurácia dos conjuntos de dados por genes sem o grupo de amostras dos indivíduos saudáveis.

A Figura 6.11 apresenta dois biplots: um com o dataset 32.080G e o outro com o dataset 30G. Assim como observado no biplot da Figura 6.8, neste primeiro biplot, devido a variabilidade dos dados, não é possível separar as instâncias do conjunto de dados por amostras avaliadas, o que pode ser observado no segundo biplot, que agrupou as amostras dos 3 óbito, as 13 amostras de fase aguda dos sobreviventes e devido as características, agrupou as 13 amostras dos sobreviventes da fase convalescente com as 4 amostras dos indivíduos saudáveis.

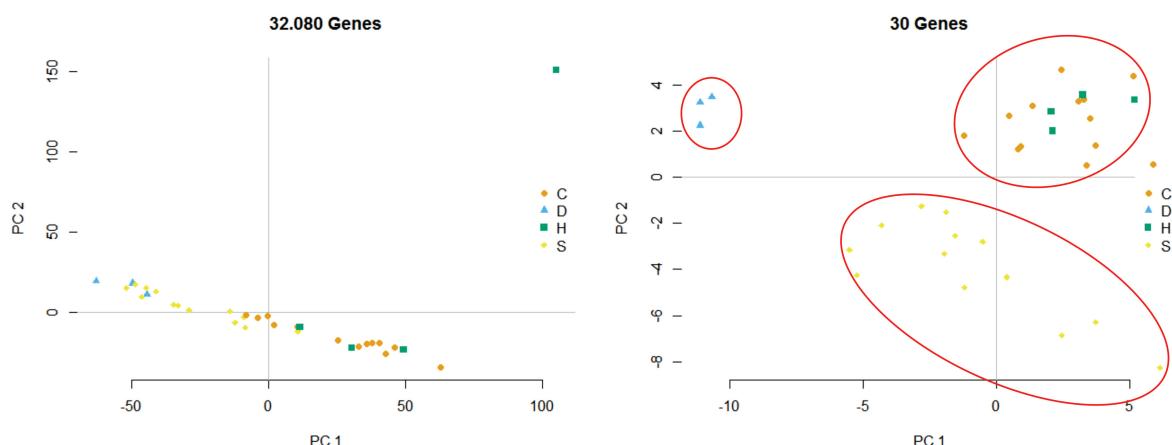


Figura 6.11: Comparação dos gráficos biplots obtidos com a base de 32080 e com a base dos 30 genes.

6.3.3 Genes diferentemente expressos

Nesta seção serão apresentados os transcritos/genes diferentemente expressos nos experimentos de redução da dimensionalidade pertencentes aos conjuntos de dados que apresentaram os melhores desempenhos na classificação dos grupos de amostras avaliadas com a aplicação do algoritmo SVM.

6.3.3.1 Por Transcritos O experimento com os testes de significância estatística através do Volcano Plot, na comparação das amostras de indivíduos que foram a óbito em relação as agudas dos sobreviventes, resultaram em conjunto de dados (dataset 51T) contendo 51 transcritos diferentemente expressos. Com a finalidade de auxiliar o entendimento da função destes transcritos e seus respectivos genes, foi utilizada a *Gene Ontology* (CONSORTIUM et al., 2004), uma ontologia criada para definir termos dos domínios de componente celular, função molecular e processo biológico. A Tabela 6.8, apresenta os transcritos com maiores diferenças de expressão agrupados por função ou tipo, nome do gene e uma breve descrição. A relação completa dos outros transcritos pode ser observada no Apêndice A desta dissertação. O dataset 51T possui 47 genes, pois alguns transcritos fazem parte do mesmo gene, como os A_24_P208567, A_33_P3211666 e A_33_P3251876 que são transcritos do gene IL18R1, assim como os transcritos A_23_P209625 e A_33_P3290343 que são transcritos do gene CYP1B1.

Tabela 6.8: Principais transcritos diferentemente expressos pertencentes ao dataset 51T.

Transcritos	Gene	Descrição do gene
<i>Resposta imune inata: Peptído antimicrobiano</i>		
A_23_P253791	CAMP	Peptídeo antimicrobiano de catelicidina
<i>Resposta imune: Vias/receptores de citocinas ou quimocinas ou Quimiotaquia</i>		
A_24_P63019	IL1R2	Interleucina 1. Tipo II
A_24_P208567	IL18R1	Receptor 1 de interleucina 18
A_33_P3211666	IL18R1	Receptor 1 de interleucina 18
A_33_P3251876	IL18R1	Receptor 1 de interleucina 18
A_23_P28334	IL18RAP	Proteína acessória do receptor da interleucina 18: subunidade acessória do receptor heterodimérico da interleucina 18 (IL18), uma citocina pró-inflamatória envolvida na indução de imunidade mediada por células
A_23_P162300	IRAK3	Quinase 3 associada ao receptor de interleucina-1
A_33_P3328254	IL5RA	Interleucina 5. Alfa: Receptor de IL-5
A_23_P152838	CCL5	Quimioquina (motivo C-C) ligando 5
A_23_P163567	SMPD3	Esfingomielina fosfodiesterase 3. membrana neutra (esfingomielinase II neutra)

6.3.3.2 Por Genes Assim como observado nos resultados realizados no método por transcritos, os experimentos com os testes de significância estatística através do Volcano Plot apresentam os melhores resultados. Na comparação das amostras de indivíduos que foram a óbito em relação as agudas dos sobreviventes, o uso de teste de hipótese resultou em conjunto de dados (dataset 30G) contendo 30 genes diferentemente expressos. As Tabelas 6.9 e 6.10, apresentam os genes agrupados por função ou tipo extraído da *Gene Ontology* (CONSORTIUM et al., 2004), e uma breve descrição.

Tabela 6.9: Genes diferentemente expressos pertencentes ao dataset 30G (Parte 1).

Gene	Descrição do gene
Resposta imune: Regulação de TLR	
TLR8	Pode detectar RNA de "single-stranded" bacteriano
Função do citoesqueleto	
MPP3	Interagem com o citoesqueleto e regulam a proliferação celular, vias de sinalização e junções intercelulares.
ARPC2	Regula a polimerização de actina. A actina é uma proteína do citoesqueleto e está envolvida na forma celular, mobilidade e adesão.
Transporte mediado pela vesícula.	
VPS26B	Componente do complexo de cargas (CSC). Atua redundantemente com VSP26A na reciclagem endocítica mediada por SNX-27 de SLC2A1 / GLUT1
KIF26B	Ativação plaquetária, sinalização e agregação e transporte mediado pela vesícula.
Vias de sinalização	
SYNJ2BP	Regulamenta a endocitose das receptorinas quinases da activina tipo 2 através da via dependente de Ral / RALBP1 e pode estar envolvido na supressão da transdução de sinal induzida por activina.
RGS19	RGS (reguladores de G-proteína de sinalização) e especificamente interage com a proteína G, GAI3. Esta proteína é uma guanosina trifosfatase-ativando proteína que funciona para down-regular Galphai / Galphaq-linked sinalização.
PMEPA1	Funciona como um regulador negativo da sinalização TGF-beta e, assim, provavelmente desempenha um papel na proliferação celular, diferenciação, apoptose, motilidade, produção de matriz extracelular e imunossupressão.

Tabela 6.10: Genes diferentemente expressos pertencentes ao dataset 30G (Parte 2).

Gene	Descrição do gene
Estrutura do tecido	
VWA1	Estrutura e função da cartilagem.
Regulamentação da transcrição / tradução	
EIF2C2	Necessário para o silenciamento de genes mediado por RNA (RNAi) pelo complexo de silenciamento induzido por RNA (RISC).
EIF1AD	Actividade do factor de iniciação da ligação de RNA e da tradução.
C20orf4 (AAR2)	Homólogo da proteína de repressão de A1-alfa 2 de levedura que está envolvida em splicing de mRNA.
C12orf26 (METTL25)	Metiltransferase putativa.
Regulação do metabolismo de lípidos e glicose	
PRKACG	A proteína quinase A (PKA, também conhecida como proteína cinase dependente de cAMP).
Degradação de proteínas	
BEAN1	Uma das várias proteínas que interagem com NEDD4, um membro de uma família de ubiquitina-proteína ligases; Visando o substrato para destruição pelo proteassoma.
DNAJC6	Regular atividade molecular chaperone; Recruta HSPA8 / HSC70 para vesículas revestidas com clatrina e promove o revestimento de vesículas revestidas com clatrina.

Ao avaliar a Tabela 6.8 e as tabelas do Apêndice A, observa-se que muitos transcritos identificados possuem função biológica relacionadas ao sistema de resposta imune, ou seja, ao sistema de defesa do organismo. Estes achados sugere que indivíduos que foram a óbito não conseguiam produzir anticorpos suficientes para combater a infecção.

Ao avaliar a função biológica dos genes pertencentes ao dataset 30G, a função da resposta imune foi identificada, mas com apenas 1 gene (TLR8), diferente do observado ao utilizar o método por transcritos, que identificaram 21 genes para a função citada. Foram identificadas funções importantes como: degradação de proteínas, vias de sinalização e regulação do metabolismo de lipídios e glicose. Neste método foram identificados 3 genes de função desconhecida e um pseudo gene também sem função descrita na *Gene Ontology* (CONSORTIUM et al., 2004), além de identificar 10 RNAs não codificantes, que podem ser vistos no apêndice, Tabela A.5.

6.4 DISCUSSÃO

Ao avaliar os genes identificados como diferentemente expressos no dataset 51T, apresentados nas Tabelas 6.8, A.2, A.3, e A.4, constatou-se que 43% (22/51) são relacionados a resposta imune, fundamental para a produção de anti-corpos. O gene que apresentou a maior diferença de expressão na comparação dos grupos de sobreviventes e óbito foi o CAMP, como apresentado na Tabela 6.8, ele é um peptídeo antimicrobiano de catelicidina, que faz parte do sistema de resposta imune inata. No trabalho de (LINDOW et al., 2016) foi realizada a primeira análise de transcriptoma humano abrangente de sangue periférico durante a leptospirose aguda, identificando a catelicidina também como o transcrito que apresentou a maior diferença de expressão, sendo o único péptido antimicrobiano com expressão significativamente diminuída em casos fatais. Nos experimentos laboratoriais com hamster, Lindow constatou que níveis reduzidos de catelicidina circulante poderiam contribuir para a elevação da carga bacteriana.

No trabalho de (LINDOW et al., 2016) foram utilizados parâmetros estatísticos para identificar os genes diferentemente expressos como os utilizados neste trabalho, entretanto a metodologia aplicada, bem como os experimentos realizados não permitem comparar todos os genes identificados no trabalho da autora e citado na Tabela 4.1, haja vista que as dimensões dos conjuntos de dados comparados são fatores que influenciam na avaliação da significância estatística. A Figura 6.12 apresenta os transcritos diferentemente expressos da Tabela 4.1 com os 51 identificados na comparação 1 do Volcano Plot, realizado no experimento C. Foi identificado 12 transcritos exclusivos do dataset 51T, não fazendo parte dos conjuntos DvS nem do conjunto SvC, isso pode ter ocorrido devido ao uso do valor de *fold change* 2 ou -2, utilizado por Lindow. Neste trabalho foi utilizado um valor de *fold change* de 1.5 ou -1.5. Ao avaliar a função dos 12 transcritos exclusivos, identificou-se que eles fazem parte do sistema de sinalização celular (A_23_P112260 e A_24_P235266), regulação imunológica (A_23_P68601), metabolismo e resposta ao estresse (A_33_P3364864, A_23_P29422, A_33_P3304983 e A_33_P3290343), regulação de RNA ou síntese de DNA (A_24_P7121) e resposta imune: Vias/receptores de citocinas ou quimocinas ou quimiotaxia (A_24_P208567, A_33_P3251876 e A_23_P28334); Vias de complemento ou coagulação (A_33_P3265030).

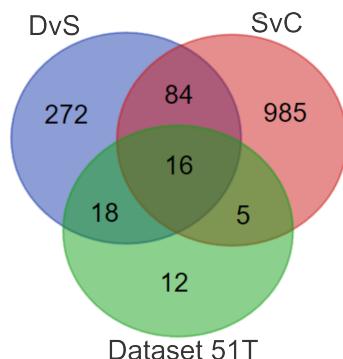


Figura 6.12: Diagrama de Venn comparando os genes diferentemente expressos citados no trabalho de (LINDOW et al., 2016) com os transcritos do dataset 51T.

A Figura 6.13 apresenta o diagrama de Venn comparando os genes pertencentes aos conjuntos dataset 51T e dataset 30G. Os genes diferentemente expressos são distintos, enfatizando a importância da decisão em analisar por transcritos ou por genes. A resposta imune na regulação de TLR é a função que está em comum nos dois conjuntos de dados. O gene TLR8 do dataset 30G é o único gene em comum com os citados no trabalho de (LINDOW et al., 2016). Como limitação da abordagem utilizada resultante do dataset 30G, 47% (14/30) dos genes identificados são de função desconhecida ou que o RNA não é codificante, além de identificar um pseudogene sem função.

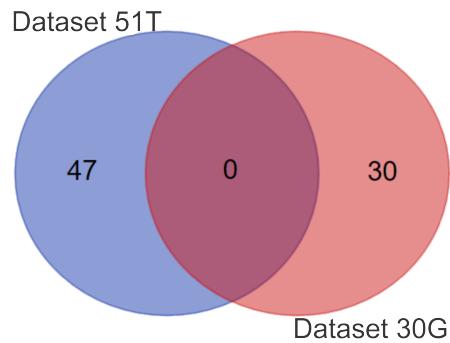


Figura 6.13: Diagrama de Veen comparando os genes diferentemente expressos entre os dataset 51T e dataset 30G.

6.5 CONSIDERAÇÕES FINAIS

Neste capítulo foram apresentados os resultados dos conjuntos de experimentos descritos no capítulo anterior, bem como uma discussão dos achados. No experimento de predição de casos baseados em dados clínicos e epidemiológicos, os modelos obtidos com os algoritmos JRIP e J48 apresentam as melhores taxas de sensibilidade e especificidade. No experimento de agrupamento dos dados de expressão gênica, o uso da métrica silhueta e do algoritmo hierárquico permitiu identificar que 3 clusters e a quantidade ideal de grupos para este experimento. Neste experimento, foi confirmado que indivíduos na fase de convalescência apresentam características de indivíduos saudáveis e que há diferenças de expressão entre indivíduos que foram a óbito em relação aos que sobreviveram. No experimento de classificação de dados de expressão gênica por desfecho clínico, observou-se que o uso de teste de hipótese através do gráfico Volcano Plot apresentou os melhores desempenhos como técnica de redução da dimensionalidade, tanto no método por transcritos quanto por genes, evidenciando as melhores taxas de acurácia obtida com a aplicação do SVM.

Capítulo

7

Neste capítulo são apresentadas as considerações finais e os trabalhos futuros.

CONCLUSÕES

A leptospirose é uma doença negligenciada de ocorrência global que surge principalmente em áreas rurais ou urbanas sem sistemas de saneamento de qualidade (COSTA et al., 2015). O diagnóstico rápido e o tratamento adequado da leptospirose aumenta a perspectiva de sobrevivência, desta doença potencialmente fatal.

Neste sentido, o presente trabalho ajustou algoritmos de classificação na análise dos fatores clínicos, epidemiológicos e de expressão gênica com o intuito de auxiliar na identificação precoce dos casos hospitalizados e na identificação dos genes associados à morte ou à sobrevivência de indivíduos doentes, tornando mais célere e eficaz o tratamento. Além disso, os modelos podem ser aplicados a bases de dados que contenham dados históricos, facilitando a estimativa do número médio de casos futuros, auxiliando em estratégias de intervenção e controle para a leptospirose em regiões de recursos limitados.

Com a finalidade de melhor entender a doença e seus fatores de risco, tanto na concepção clínica quanto na epidemiológica, e consequentemente entender os fatores genéticos, este trabalho realizou 3 conjuntos de experimentos: no experimento A, foram avaliados algoritmos com o propósito de classificar casos de leptospirose baseados em dados clínicos e epidemiológicos; no experimento B, avaliou-se o uso de dados de expressão gênica para agrupar indivíduos doentes baseados nos desfechos clínicos; por fim, no experimento C, foram avaliados conjuntos de dados por transcritos e por genes com o intuito de predizer o desfecho clínico na análise de expressão gênica.

Os resultados obtidos na avaliação dos algoritmos no experimento A, determinou-se que o JRIP é o melhor modelo de predição para identificação de casos de leptospirose, com 84,1% de classificação precisa de casos da doença, seguido por J48 (82%). Os resultados fornecem fortes evidências de que indivíduos com leptospirose podem ser identificados utilizando o modelo baseado em dados clínicos e epidemiológicos, em regiões com outras doenças febris apresentando sintomatologia semelhante (NERY; CLARO; LINDOW, 2016; JUNIOR; CLARO; LINDOW, 2017).

O uso de algoritmos de agrupamento, no experimento B, permitiu avaliar a hipótese de que há diferenças na expressão dos genes nos indivíduos experimentados. O resultado

obtido com o algoritmo K-means com o valor de k=3 identificou a similaridade dos indivíduos baseado em seu desfecho clínico, agrupando todas as amostras da fase aguda no primeiro *cluster*, todas as amostras convalescente e de indivíduos saudáveis no segundo *cluster* e por fim, no terceiro *cluster* todas as amostras dos indivíduos que foram a óbito. Resultados similares foram obtidos ao utilizar o método hierárquico, como pode ser observado na Figura 6.4. Os bons resultados obtidos com o Experimento B serviram de estímulo para a realização do Experimento C.

O experimento C avaliou dois métodos: por transcritos e por genes, utilizando o algoritmo SVM para classificar, haja vista que todas as amostras avaliadas já eram etiquetadas baseadas no desfecho clínico (óbito, sobreviventes na fase aguda, sobreviventes na fase convalescente e amostras de indivíduos saudáveis). Sendo assim, foi criado um atributo classe em cada conjunto de dados avaliados, com o objetivo de servir como classe de referência para o algoritmo de predição. O desafio deste experimento foi de avaliar metodologias que permitissem analisar um conjunto de dados completos de expressão gênica, de forma a facilitar as análises de profissionais da saúde, simplificando o trabalho em identificar genes significantes.

No método por transcritos foi utilizado como conjunto base, informações de 50.575 transcritos, já no método por genes, o conjunto de transcritos foi convertido em um conjunto por genes, contendo a média das expressões dos transcritos pertencentes ao mesmo gene, resultando em um conjunto de dados com 32.080 genes. Ambos os métodos iniciaram com conjuntos de dados como característica principal a alta dimensionalidade. Técnicas de redução da dimensionalidade foram avaliadas, criando os conjuntos de dados experimentados neste trabalho, com a finalidade de reduzir o tempo computacional mantendo a qualidade da informação, bem como auxiliar os profissionais de saúde na análise de um conjunto reduzido e específico, que permita identificar os genes responsáveis pelo desfecho clínico.

Os resultados obtidos no método por transcritos, no aspecto de redução da dimensionalidade, apresentaram os melhores desempenhos com os conjuntos de dados com 51 e com 435 transcritos significantes, ambos com 88% na avaliação da acurácia nos experimentos incluindo o grupo das amostras de indivíduos saudáveis. Ao avaliar as matrizes de confusão resultantes da aplicação do algoritmo SVM, foi confirmada a constatação de (LINDOW et al., 2016) sobre a similaridade de expressão gênica entre indivíduos sobreviventes com amostras da fase convalescente em relação a indivíduos saudáveis, o que resultou em erro de classificação por parte do algoritmo. Consequentemente, os experimentos foram repetidos removendo os indivíduos saudáveis e as melhores taxas de acurácia subiram para 100%, como pode ser observado na Figura 6.7. O conjunto de dados com 435 transcritos apresentou resultados similares, devido aos transcritos do dataset 51T estarem incluso neste conjunto. Resultados equivalentes foram obtidos com os dataset 30G e dataset 302G utilizados para testar o método por genes, apresentando taxas de acurácia de 88% para os experimentos com o grupo de saudáveis e 100% sem este grupo.

O uso do método Volcano Plot como seleção de atributos, bem como a aplicação do algoritmo SVM ao conjunto completo de transcritos permitiu reduzir a base de dados de 50.575 para 51 transcritos, apresentando 100% de acurácia ao avaliar o conjunto de da-

dos contendo as 29 amostras dos 16 indivíduos hospitalizados. A metodologia, bem como o modelo de classificação resultante poderá contribuir para o conhecimento mais aprofundado dos genes relevantes, bem como predizer o desfecho de pacientes hospitalizados baseado em marcadores genéticos.

No aspecto computacional, os métodos por transcritos e por genes apresentaram bons desempenhos, tanto no processo de redução da dimensionalidade, quanto na avaliação da acurácia por parte do algoritmo de classificação. Entretanto, ao avaliar o aspecto biológico, os genes pertencentes ao dataset 30G apresentaram como limitação o fato de 47% destes genes serem de função desconhecida ou que o RNA não é codificante. Já o método por transcritos apresentou genes com funções biológicas compatíveis com doenças infecciosas, sendo que 43% dos 51 transcritos são relacionados a respostas imune, fundamental para a produção de anticorpos.

Profissionais de um domínio, devido ao seu conhecimento, tendem a realizar análises subjetivas durante um processo investigativo. Sendo assim, neste trabalho foram avaliados métodos equânimis com a finalidade de simplificar o processo de análise e redução dos genes para um posterior estudo por parte de especialistas. Estes métodos computacionais foram validados com a leptospirose, mas que podem ser aplicados ao estudo de outras enfermidades, como a dengue.

Assim, as principais contribuições deste trabalho foram:

- Identificação de casos de leptospirose baseado em dados clínicos e epidemiológicos.
- Redução do tempo de análise e identificação dos genes diferentemente expressos, baseado na metodologia automatizada para auxílio de especialistas no processo investigativo.
- Identificação de potenciais genes responsáveis pela sobrevivência ou óbito de indivíduos doentes.
- Confirmação da catelicidina como um gene relevante para determinar o agravamento e potencial óbito de indivíduos hospitalizados.
- Modelo de predição de desfecho clínico de novas amostras baseados nos níveis de expressão de genes-chaves para a leptospirose.

A metodologia utilizada, bem como os modelos preditores obtidos, podem ser utilizados para vários tipos de decisões no aspecto clínico, como: avaliação de risco, testes de diagnóstico, estratificação prognóstica e seleção de tratamentos mais adequados, resultando em uma minimização do risco de óbito de pacientes hospitalizados.

7.1 TRABALHOS FUTUROS

Pretende-se investigar, com especialistas do domínio, os 12 transcritos identificados no dataset 51T e que não foram identificados no trabalho de (LINDOW et al., 2016).

Pretende-se avaliar o modelo de predição de desfecho clínico em uma base de dados com expressão gênica de amostras de outros indivíduos avaliados pela técnica de microarranjos.

Além disso, pretende-se combinar a metodologia empregada neste trabalho com o uso de ontologias que simplifique o trabalho dos profissionais de saúde na automatização do processo de identificação dos genes relevantes e conhecimento das suas respectivas funções biológicas.

Por fim, pretende-se desenvolver uma ferramenta para ser utilizada na rotina hospitalar, na avaliação clínica e na anamnese, com a finalidade de auxiliar os profissionais da saúde na suspeita clínica de leptospirose e posteriormente para outras doenças endêmicas.

Apêndice

A

GENES E TRANSCRITOS DIFERENTEMENTE EXPRESSOS

Tabela A.1: Transcritos differentemente expressos pertencentes ao dataset 51T (Parte 1).

Transcritos	Gene	Descrição do gene
<i>Resposta imune: Vias de complemento ou coagulação</i>		
A_33_P3265030	GP1BB	Glicoproteína Ib (plaquetas).
<i>Sinalização celular: Proteína G</i>		
A_23_P111701	GNG11	Guanina (proteína G). Gama 11
A_23_P112260	GNG10	Guanina (proteína G). Gama 10
<i>Resposta imune: Regulação de TLR</i>		
A_23_P85903	TLR5	Toll-like receptor 5: detecta a proteína flagelina bacteriana
A_23_P162300	IRAK3	Quinase 3 associada ao receptor de interleucina-1
A_23_P2601	HSP90B1	Proteína de choque térmico 90kDa beta (Grp94). Membro 1: trifosfato de adenosina (ATP)
<i>Resposta imune adaptativa: Ativação e regulação de células B</i>		
A_23_P167328	CD38	molécula CD38
A_23_P48088	CD27	molécula CD27

Tabela A.2: Transcritos diferentemente expressos pertencentes ao dataset 51T (Parte 2).

Transcritos	Gene	Descrição do gene
<i>Processo biossintético composto por enxofre</i>		
A_23_P50638	LRG1	Alucina-rica 2-glicoproteína 1
<i>Resposta imune adaptativa: Imunoglobulinas ->producao dos anticorpos</i>		
A_24_P161764	IGHV1-3	Imunoglobulina pesada variável 1-3
A_24_P318990	IGLV1-50	Imunoglobulina lambda variável 1-50 (não funcional)
A_33_P3247639	IGHV3OR16-10	Imunoglobulina pesada variável 3 / OR16-10 (não funcional)
A_23_P361654	IGKV1D-16	Immunoglobulin Kappa V 1D-16
A_33_P3331178	IGHV3-16	Imunoglobulina pesada variável 3-16 (não funcional)
A_33_P3399985	IGKV2D-24	Imunoglobulina kappa variável 2D-24 (não funcional)
A_24_P370172	LILRA5	Receptor tipo leucócito semelhante a imunoglobulina. Subfamília A (com domínio TM). Membro 5
<i>Possível galectina, envolvida na inflamação</i>		
A_23_P101683	CLC	Proteína de cristal de Charcot-Leyden (em eosinófilos, geralmente com alergias ou infecções parasitárias, às vezes em asma brônquica)
<i>Possível regulação imunológica</i>		
A_23_P68601	CST7	Cistatina F (leucocistatina): inibidor da cisteína protease glicosilada com um papel suposto na regulação imune através da inibição de um alvo único no sistema hematopoietico.
A_24_P38081	FKBP5	FK506 binding protein 5: família de proteínas de imunofilina, que desempenham um papel na imunorregulação e processos celulares básicos envolvendo dobramento e tráfego de proteínas. Interage com HSP90a
<i>Sinalização celular</i>		
A_24_P235266	GRB10	Receptor de fator de crescimento de ligação a proteínas-10: proteína adaptadora que modula o acoplamento de um número de quinases de receptores de superfície de células com vias de sinalização específicos.

Tabela A.3: Transcritos differentemente expressos pertencentes ao dataset 51T (Parte 3).

Transcritos	Gene	Descrição do gene
<i>Citoesqueleto e crescimento celular</i>		
A_24_P52004	PDS5A	PDS5. Regulador da manutenção da coesão.
<i>Síntese proteína</i>		
A_23_P25121	FKBP11	FK506 11. 19 kDa: catalisa a dobragem de polipeptídos contendo prolina
A_23_P208009	SEC11C	SEC11 homólogo C (<i>S. cerevisiae</i>): peptidase de sinal
<i>Metabolismo e resposta ao estresse</i>		
A_33_P3364864	NAMPT	Nicotinamida fosforibosiltransferase: citoquina (PBEF) que promove a maturação das células B e inibe a apoptose dos neutrófilos
A_23_P29422	GYG1	Glicogenina 1: biossíntese de glicogénio
A_33_P3304983	PRKAR2B	Proteína cinase. CAMP-dependente. Regulação. Tipo II. beta
A_23_P209625	CYP1B1	Citocromo P450. Família 1. subfamília B. polipeptído 1: As proteínas do citocromo P450 são monooxygenases que catalisam muitas reacções envolvidas no metabolismo de fármacos e síntese de colesterol, esteróides e outros lípidos
A_33_P3290343	CYP1B1	Citocromo P450. Família 1. subfamília B. polipeptído 1: As proteínas do citocromo P450 são monooxygenases que catalisam muitas reacções envolvidas no metabolismo de fármacos e síntese de colesterol, esteróides e outros lípidos
A_23_P165840	ODC1	Ornitina descarboxilase 1: A reacção de descarboxilação da ornitina catalisada pela ornitina descarboxilase é o primeiro e cometido passo na síntese de poliaminas, particularmente putrescina, espermidina e espermina.
A_24_P81900	SLC2A3	A família transportadora de soluto 2 (transportador de glicose facilitado).
A_23_P105012	HRASLS2	HRAS-como supressor 2: fosfolipase e atividades aciltransferase e atua como um supressor de tumor.
A_23_P56356	PLB1	Fosfolipase B1: fosfolipase associada à membrana.

Tabela A.4: Transcritos diferentemente expressos pertencentes ao dataset 51T (Parte 4).

Transcritos	Gene	Descrição do gene
<i>Regulação de RNA ou síntese de DNA</i>		
A_21_P0010806	XLOC_12_- 001496	LincRNA
A_21_P0014944	LOC100653206	LOC100653206 não caracterizado (RNA pequeno variado, função desconhecida)
A_23_P147805	UPP1	uridina fosforilase 1: catalisa a clivagem phosphorylytic reversível de uridina e desoxiuridina para uracilo e ribose- ou desoxirribose-1-fosfato (PubMed: 7488099).
A_24_P7121	NSUN7	Família de domínio NOP2 / Sun. Membro 7
A_23_P99397	ZDHHC20	Dedo de zinco. Tipo DHHC contendo 20 - função desconhecida
<i>Membrana plasma</i>		
A_23_P72668	SDPR	Resposta de privação de soro: proteína de ligação a fosfolípido independente de cálcio cuja expressão aumenta em células privadas de soro.
<i>Codificação de proteínas</i>		
A_23_P330561	C19orf59	Este gene codifica uma proteína transmembranar de passagem única.
A_24_P388528	ST6GAL1	Este gene codifica um membro da família de glicosiltransferase 29
A_23_P116264	NRGN	Neurogranina (substrato da proteína quinase C. RC3): substrato de proteína quinase pós-sináptica que se liga à calmodulina na ausência de cálcio.
<i>Desconhecido</i>		
A_33_P3245126		Gene desconhecido

Tabela A.5: Genes diferentemente expressos pertencentes ao dataset 30G (Parte 3).

Gene	Descrição do gene
<i>Função desconhecida</i>	
LDOC1L	
C15orf17	
(FAM219B)	
C9orf139	
<i>RNA não codificante.</i>	
XLOC_007562	
XLOC_012441	
LOC572558	
XLOC_012462	
LOC100129413	
XLOC_003548	
XLOC_12_-	
010863	
XLOC_001225	
XLOC_000326	
XLOC_013770	
<i>Pseudogene (sem função)</i>	
RPSAP52	Pseudogene (Ribosomal Protein SA Pseudogene 52).

REFERÊNCIAS BIBLIOGRÁFICAS

- AHMED, A. et al. Development and validation of a real-time pcr for detection of pathogenic leptospira species in clinical materials. *PLoS One*, v. 4, n. 9, p. e7093, 2009.
- AMILASAN, A.-s. T. et al. Outbreak of leptospirosis after flood, the philippines, 2009. *Emerging infectious diseases*, Centers for Disease Control and Prevention, v. 18, n. 1, p. 91–94, 2012.
- BAKAR, A. A. et al. Predictive models for dengue outbreak using multiple rulebase classifiers. In: IEEE. *Electrical Engineering and Informatics (ICEEI), 2011 International Conference on*. [S.l.], 2011. p. 1–6.
- BERRY, M. J.; LINOFF, G. S. *Data mining techniques: for marketing, sales, and customer relationship management*. [S.l.]: John Wiley & Sons, 2004.
- BEZDEK, J. C.; EHRLICH, R.; FULL, W. Fcm: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, Elsevier, v. 10, n. 2, p. 191–203, 1984.
- BHARGAVA, N. et al. Decision tree analysis on j48 algorithm for data mining. *Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering*, v. 3, n. 6, 2013.
- BIER, D. Distribuicao espacial e fatores de risco para leptospirose canina na vila pantanal, curitiba, parana, brasil. 2013.
- BRAZMA, A. et al. Minimum information about a microarray experiment (miame)—toward standards for microarray data. *Nature genetics*, Nature Publishing Group, v. 29, n. 4, p. 365–371, 2001.
- BROCK, G. et al. clvalid, an r package for cluster validation. *Journal of Statistical Software (Brock et al., March 2008)*, 2011.
- CONSORTIUM, G. O. et al. The gene ontology (go) database and informatics resource. *Nucleic acids research*, Oxford Univ Press, v. 32, n. suppl 1, p. D258–D261, 2004.
- CORTES, S. da C.; PORCARO, R. M.; LIFSCHITZ, S. *Mineracao de dados-funcionalidades, tecnicas e abordagens*. [S.l.]: PUC, 2002.
- COSTA, F. et al. Global morbidity and mortality of leptospirosis: a systematic review. *PLoS Negl Trop Dis*, Public Library of Science, v. 9, n. 9, p. e0003898, 2015.

- CUI, X.; CHURCHILL, G. A. Statistical tests for differential expression in cdna microarray experiments. *Genome biology*, BioMed Central, v. 4, n. 4, p. 210, 2003.
- DENOUEUD, F. et al. Annotating genomes with massive-scale rna sequencing. *Genome biology*, BioMed Central, v. 9, n. 12, p. R175, 2008.
- FACELI, K. et al. Inteligência artificial: Uma abordagem de aprendizado de máquina. *Rio de Janeiro: LTC*, v. 2, p. 192, 2011.
- FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI magazine*, v. 17, n. 3, p. 37, 1996a.
- FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, ACM, v. 39, n. 11, p. 27–34, 1996b.
- GARCÍA-LAENCINA, P. J. et al. Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values. *Computers in biology and medicine*, Elsevier, v. 59, p. 125–133, 2015.
- GOEBEL, M.; GRUENWALD, L. A survey of data mining and knowledge discovery software tools. *ACM SIGKDD Explorations Newsletter*, ACM, v. 1, n. 1, p. 20–33, 1999.
- GOLDSCHMIDT, R.; PASSOS, E. *Data Mining: um guia prático*. [S.l.]: Campus, 2005.
- GOUVEIA, E. L. et al. Leptospirosis-associated severe pulmonary hemorrhagic syndrome, salvador, brazil. *Emerging infectious diseases*, Centers for Disease Control, v. 14, n. 3, p. 505, 2008.
- HAGAN, J. E. et al. Global burden of disease due to leptospirosis: Systematic review of diseasespecific. 2015.
- HAN, J.; KAMBER, M.; PEI, J. *Data mining: concepts and techniques*. [S.l.]: Elsevier, 2011.
- JIANG, D.; TANG, C.; ZHANG, A. Cluster analysis for gene expression data: A survey. *IEEE Transactions on knowledge and data engineering*, IEEE, v. 16, n. 11, p. 1370–1386, 2004.
- JUNIOR, N.; CLARO, D.; LINDOW, J. Prediction of leptospirosis cases using classification algorithms. *IET Software*, IET, 2017.
- KANUNGO, T. et al. An efficient k-means clustering algorithm: Analysis and implementation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, IEEE, v. 24, n. 7, p. 881–892, 2002.
- KO, A. I. et al. Urban epidemic of severe leptospirosis in brazil. *The Lancet*, Elsevier, v. 354, n. 9181, p. 820–825, 1999.

- KO, A. I. et al. Urban epidemic of severe leptospirosis in brazil. *The Lancet*, Elsevier, v. 354, n. 9181, p. 820–825, 1999.
- LÊ, S.; JOSSE, J.; HUSSON, F. FactoMineR: A package for multivariate analysis. *Journal of Statistical Software*, v. 25, n. 1, p. 1–18, 2008.
- LINDOW, J. C. et al. Cathelicidin insufficiency in patients with fatal leptospirosis. *PLoS Pathog*, Public Library of Science, v. 12, n. 11, p. e1005943, 2016.
- MAKRETSOV, N. A. et al. Hierarchical clustering analysis of tissue microarray immunostaining data identifies prognostically significant groups of breast carcinoma. *Clinical cancer research*, AACR, v. 10, n. 18, p. 6143–6151, 2004.
- MYERS, D.; VARELA-DIAZ, V. The occurrence of leptospiral antibodies in rural inhabitants of argentina. *Tropical and geographical medicine*, v. 31, n. 2, p. 269–274, 1979.
- NAJERA, S. et al. Leptospirosis ocupacional en una region del caribe colombiano. *Salud publica de Mexico*, Instituto Nacional de Salud Publica, v. 47, n. 3, p. 240–244, 2005.
- NERY, N. R. R.; CLARO, D. B.; LINDOW, J. C. Classification model analysis for the prediction of leptospirosis cases. In: IEEE. *Information Systems and Technologies (CISTI), 2016 11th Iberian Conference on*. [S.l.], 2016. p. 1–6.
- OGUNTIMILEHIN, A.; ADETUNMBI, A.; ABIOLA, O. A review of predictive models on diagnosis and treatment of malaria fever. *International Journal of Computer Science and Mobile Computing*, Researchgate, v. 4, n. 5, p. 1087–1093, 2015.
- OKSANEN, J. Cluster analysis: tutorial with r. *University of Oulu, Oulu*, 2010.
- QUACKENBUSH, J. Computational analysis of microarray data. *Nature reviews genetics*, Nature Publishing Group, v. 2, n. 6, p. 418–427, 2001.
- RACINE, J. S. Rstudio: A platform-independent ide for r and sweave. *Journal of Applied Econometrics*, Wiley Online Library, v. 27, n. 1, p. 167–172, 2012.
- REIS, R. B. et al. Impact of environment and social gradient on leptospira infection in urban slums. *PLoS neglected tropical diseases*, Public Library of Science, v. 2, n. 4, p. e228, 2008.
- SAHLE, G.; MESHESHA, M. Uncovering knowledge that supports malaria prevention and control intervention program in ethiopia. *electronic Journal of Health Informatics*, v. 8, n. 1, p. e7, 2013.
- SAUDE, B. M. da. Guia de vigilancia epidemiologica. *Guia de vigilancia epidemiologica / Ministerio da Saude, Secretaria de Vigilancia em Saude, Departamento de Vigilancia Epidemiologica. 7. ed. Brasilia : Ministerio da Saude, 816 p. (Serie A. Normas e Manuais Tecnicos)*, 2009.

- SAUDE, B. M. da. Guia de leptospirose: Diagnostico e manejo clinico. *Leptospirose: diagnostico e manejo clinico / Ministerio da Saude, Secretaria de Vigilancia em Saude. Departamento de Vigilancia das Doencas Transmissiveis. Brasilia : Ministerio da Saude*, 2014.
- SEALFON, S. C.; CHU, T. T. Rna and dna microarrays. *Biological Microarrays: Methods and Protocols*, Springer, p. 3–34, 2011.
- SMITH, J. K. et al. Leptospirosis following a major flood in central queensland, australia. *Epidemiology and infection*, Cambridge Univ Press, v. 141, n. 03, p. 585–590, 2013.
- SMYTH, G. K. Limma: linear models for microarray data. In: *Bioinformatics and computational biology solutions using R and Bioconductor*. [S.l.]: Springer, 2005. p. 397–420.
- TARI, L.; BARAL, C.; KIM, S. Fuzzy c-means clustering with prior biological knowledge. *Journal of biomedical informatics*, Elsevier, v. 42, n. 1, p. 74–81, 2009.
- UGWU, C.; ONYEJEGBU, N.; OBAGBUWA, I. The application of machine learning technique for malaria diagnosis. In: *International and Interdisciplinary Studies in Green Computing*. [S.l.]: IGI Global, 2013. p. 263–272.
- UN-HABITAT. The challenge of slums: global report on human settlements 2003. *Management of Environmental Quality: An International Journal*, Emerald Group Publishing Limited, v. 15, n. 3, p. 337–338, 2004.
- UN-HABITAT. *State of the world's cities 2010/2011: bridging the urban divide*. [S.l.]: Earthscan, 2010.
- VENTER, J. C. et al. The sequence of the human genome. *science*, American Association for the Advancement of Science, v. 291, n. 5507, p. 1304–1351, 2001.
- VOLP, A. C. P. et al. Capacidade dos biomarcadores inflamatórios em predizer a síndrome metabólica: Inflammation biomarkers capacity in predicting the metabolic syndrome. *Arquivos Brasileiros de Endocrinologia & Metabologia*, SciELO Brasil, v. 52, n. 3, p. 537–549, 2008.
- WANG, J. et al. Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data. *BMC bioinformatics*, BioMed Central, v. 4, n. 1, p. 60, 2003.
- WICKHAM, H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. ISBN 978-0-387-98140-6. Disponível em: <http://ggplot2.org>.
- WITTEN, I. H.; FRANK, E. *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd. ed. San Francisco: Morgan Kaufmann, 2005.
- YEH, D.-Y.; CHENG, C.-H.; CHEN, Y.-W. A predictive model for cerebrovascular disease using data mining. *Expert Systems with Applications*, Elsevier, v. 38, n. 7, p. 8970–8977, 2011.

- YEH, J.-Y.; WU, T.-H.; TSAO, C.-W. Using data mining techniques to predict hospitalization of hemodialysis patients. *Decision Support Systems*, Elsevier, v. 50, n. 2, p. 439–448, 2011.