UNIVERSIDADE FEDERAL FLUMINENSE

NADINE MELLONI NEUMANN

P-Valor<0,05 é Suficiente?
Um Estudo na Avaliação de Classificadores

NITERÓI

2018

UNIVERSIDADE FEDERAL FLUMINENSE

NADINE MELLONI NEUMANN

P-Valor<0,05 é Suficiente? Um Estudo na Avaliação de Classificadores

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Mestre em Computação. Área de concentração: Engenharia de Sistemas e Informação

Orientador:
ALEXANDRE PLASTINO DE CARVALHO

NITERÓI

2018

Ficha catalográfica automática - SDC/BEE

M527p Melloni Neumann, Nadine
P-valor<0,05 é Suficiente? Um Estudo na Avaliação de Classificadores / Nadine Melloni Neumann ; Alexandre Plastino de Carvalho, orientador. Niterói, 2018.
117 f. : il.

Dissertação (mestrado)-Universidade Federal Fluminense, Niterói, 2018.

DOI: http://dx.doi.org/10.22409/PGC.2018.m.14274607798

1. Teste de hipótese. 2. Aprendizado de máquina. 3. Mineração de dados (Computação). 4. Produção intelectual. I. Título II. Plastino de Carvalho,Alexandre, orientador. III. Universidade Federal Fluminense. Escola de Engenharia.

CDD -

Bibliotecária responsável: Fabiana Menezes Santos da Silva - CRB7/5274

NADINE MELLONI NEUMANN

P-Valor<0,05 é Suficiente?
Um Estudo na Avaliação de Classificadores

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Mestre em Computação. Área de concentração: Engenharia de Sistemas e Informação

Aprovada em Julho de 2018.

BANCA EXAMINADORA

Prof. Alexandre Plastino de Carvalho - Orientador, UFF

Prof. Jony Arrais Pinto Junior, UFF

Prof. Aline Marins Paes Carvalho, UFF

Prof. Gustavo Henrique Mitraud Assis Rocha, ENCE

Niterói

2018

Resumo

Uma ferramenta comumente utilizada no processo de comparação de classificadores é a análise da significância estatística, realizada através de teste de hipóteses. Entretanto, percebe-se que muitos pesquisadores estão buscando cegamente a significância estatística por meio da condição p-valor<0,05 e ignorando conceitos importantes como o tamanho do efeito e o poder do teste. Neste trabalho, são evidenciados possíveis problemas causados pelo mau uso dessa ferramenta e como o tamanho do efeito e o poder do teste acrescentam informações para uma melhor tomada de decisão. Para tanto, são realizados estudos empíricos com diferentes classificadores e 50 bases de dados, comparando-se os resultados por meio do teste t de Student e do teste de Wilcoxon. Além disso, dados sintéticos que simulam os resultados de classificadores são utilizados para ampliar as análises. Os resultados mostram que a análise isolada do p-valor pode levar a conclusões equivocadas e que o cálculo do tamanho do efeito e do poder do teste colaboram para que a tomada de decisão seja mais fundamentada e responsável.

Palavras-chave: comparação de classificadores, significância estatística, p-valor, tamanho do efeito, poder do teste, teste de hipóteses, teste t de Student, teste de Wilcoxon.

Abstract

A common tool used in the process of comparing classifiers is the statistical significance analysis, performed through the hypothesis test. However, there are many researchers attempting to obtain statistical significance through a blinding evaluating of the p-value<0.05 condition, ignoring important concepts such as the effect size and statistical power. This work highlight possible problems caused by the misuse of the hypothesis test and how the effect size and the statistical power can provide information for a better decision making. For this, empirical studies with different classifiers and 50 datasets are performed, comparing the results using the Student's t-test and the Wilcoxon test. In addition, synthetic data that simulate the results of classifiers are used to increase the analyzes. The results show that the isolated p-value analysis can lead to wrong conclusions and that the evaluation of the effect size and the statistical power contribute to a more informed and responsible decision-making.

Keywords: comparison of classifiers, statistical significance, p-value, effect size, statistical power, statistical hypothesis test, Student's t-test, Wilcoxon signed-ranks test.

Lista de Figuras

2.1 Esquema de estimação com a variável X 8

2.2 Probabilidades do erro tipo I (α) e do erro tipo II (β), para diferentes definições da região crítica, considerando que a diferença entre as acurácias tem distribuição normal com média 0 e variância 16 quando H_0 é verdadeira e a média dessa distribuição é 5 quando H_0 é falsa 13

2.3 Função poder do teste para n=10 e n=30 21

3.1 Curvas do poder do teste com diversos valores de n 30

4.1 Representação dos resultados dos p-valores dos testes t aplicados com tamanhos de amostras 10, 20 e 30 e suas respectivas medidas de tamanho do efeito obtidas 47

4.2 Porcentual de bases sem significância estatística no teste t e com tamanho do efeito médio, dentre as bases analisadas em cada um dos 15 pares de classificadores, por tamanho da amostra 48

4.3 Porcentual de bases com significância estatística no teste t e com tamanho do efeito pequeno, dentre as bases analisadas em cada um dos 15 pares de classificadores, por tamanho da amostra 49

4.4 Boxplot dos valores obtidos na função poder do teste t por grupos e tamanhos das amostras 52

4.5 Boxplot das diferenças médias amostrais observadas por grupos do teste t e tamanhos das amostras 54

4.6 Boxplot dos desvios padrões das diferenças amostrais observadas por grupos do teste t e tamanhos das amostras 54

4.7 Representação dos resultados dos p-valores dos testes de Wilcoxon aplicados com tamanhos de amostras 10, 20 e 30 e suas respectivas medidas de tamanho do efeito obtidas 60

4.8 Porcentual de bases sem significância estatística no teste de Wilcoxon e com tamanho do efeito médio, dentre as bases analisadas em cada um dos 15 pares de classificadores, por tamanho da amostra 62

4.9 Porcentual de bases com significância estatística no teste de Wilcoxon e com tamanho do efeito pequeno, dentre as bases analisadas em cada um dos 15 pares de classificadores, por tamanho da amostra 63

4.10 Boxplot dos valores obtidos na função poder do teste de Wilcoxon por grupos e tamanhos das amostras 65

4.11 Boxplot das diferenças médias amostrais observadas por grupos do teste de Wilcoxon e tamanhos das amostras 67

4.12 Boxplot dos desvios padrões das diferenças amostrais observadas por grupos do teste de Wilcoxon e tamanhos das amostras 67

5.1 Representação dos resultados dos p-valores dos testes t aplicados aos dados simulados e suas respectivas medidas de tamanho do efeito obtidas por tamanho de amostra 71

5.2 Representação dos resultados dos p-valores dos testes de Wilcoxon aplicados aos dados simulados e suas respectivas medidas de tamanho do efeito obtidas por tamanho de amostra 72

5.3 Boxplot dos valores obtidos na função poder do teste t por grupos e tamanhos das amostras simuladas 73

5.4 Boxplot dos valores obtidos na função poder do teste de Wilcoxon por grupos e tamanhos das amostras simuladas 75

5.5 Boxplot das diferenças médias amostrais por grupos do teste t e tamanhos das amostras simuladas 76

5.6 Boxplot das diferenças médias amostrais por grupos do teste de Wilcoxon e tamanhos das amostras simuladas 77

5.7 Boxplot dos desvios padrões das diferenças amostrais por grupos do teste t e tamanhos das amostras simuladas 78

5.8 Boxplot dos desvios padrões das diferenças amostrais por grupos do teste de Wilcoxon e tamanhos das amostras simuladas 78

Lista de Tabelas

2.1 Decisões e erros associados em testes estatísticos 12

2.2 Amostras das acurácias obtidas através dos classificadores A e B em 10 partições de uma dada base, com postos para as diferenças 19

2.3 Possíveis sinais para os postos 20

2.4 Valores para a interpretação da medida *d de Cohen* de tamanho do efeito . 24

2.5 Valores para a interpretação da medida r de tamanho do efeito 25

3.1 Acurácias obtidas por partição para cada classificador 27

3.2 Acurácias observadas e o respectivo posto de cada par para o exemplo do Teste de Wilcoxon 32

4.1 Bases utilizadas, quantidade de instâncias, atributos e classes em cada base 38

4.1 Bases utilizadas, quantidade de instâncias, atributos e classes em cada base 39

4.2 Indicadores de aplicação do teste t para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 10, 20 e 30 41

4.3 Quantidade porcentual de testes aplicados nas quais o p-valor reduziu com o aumento do tamanho da amostra (de 10 para 20, de 10 para 30 e de 20 para 30) na aplicação do teste t em cada par de classificadores 43

4.4 Porcentuais de resultados alterados com o aumento do tamanho da amostra no teste t 45

4.5 Quantidade porcentual de testes de Wilcoxon aplicados nas quais o p-valor reduziu com o aumento do tamanho da amostra (de 10 para 20, de 10 para 30 e de 20 para 30) na aplicação do teste t em cada par de classificadores . 57

4.6 Porcentuais de resultados alterados com o aumento do tamanho da amostra no teste de Wilcoxon 59

A.1 P-valores obtidos com a aplicação do teste t para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 10 85

A.2 P-valores obtidos com a aplicação do teste t para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 20 86

A.3 P-valores obtidos com a aplicação do teste t para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 30 87

A.4 Medidas de tamanho do efeito obtidos com o $d'cohen$ para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 10 88

A.5 Medidas de tamanho do efeito obtidos com o $d'cohen$ para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 20 89

A.6 Medidas de tamanho do efeito obtidos com o $d'cohen$ para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 30 90

A.7 Poder dos testes t obtidos para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 10 91

A.8 Poder dos testes t obtidos para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 20 92

A.9 Poder dos testes t obtidos para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 30 93

B.1 P-valores obtidos com a aplicação do teste de Wilcoxon para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 10 95

B.2 P-valores obtidos com a aplicação do teste de Wilcoxon para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 20 96

B.3 P-valores obtidos com a aplicação do teste de Wilcoxon para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 30 97

B.4 Medidas de tamanho do efeito obtidos com o r para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 10 98

B.5 Medidas de tamanho do efeito obtidos com o r para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 20 99

B.6 Medidas de tamanho do efeito obtidos com o r para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 30 100

B.7 Poder dos testes de Wilcoxon obtidos para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 10 101

B.8 Poder dos testes de Wilcoxon obtidos para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 20 102

B.9 Poder dos testes de Wilcoxon obtidos para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 30 103

Lista de Abreviaturas e Siglas

ASA	:	American Statistical Association;
k-NN	:	k-Nearest Neighbors;
MASS	:	Modern Applied Statistics with S;
NB	:	Naive Bayes;
RF	:	Random Forest;
SVM	:	Support Vector Machine;
TLC	:	Teorema do Limite Central;
UCI	:	University of California, Irvine;

Sumário

1 Introdução **1**

2 Fundamentação Teórica **6**

2.1 Teste de Hipóteses 6

2.1.1 P-valor 15

2.1.2 Teste t de Student para Amostras Pareadas 15

2.1.3 Teste de Wilcoxon para Amostras Pareadas 17

2.2 Poder do Teste 20

2.2.1 Poder do Teste t de Student 22

2.2.2 Poder do Teste de Wilcoxon 22

2.3 Tamanho do Efeito 23

3 P-valor<0,05 é Suficiente?
Dois Estudos de Caso **26**

3.1 Caso com Teste t de Student 26

3.1.1 Conclusão com Base no P-valor 27

3.1.2 Avaliando o Tamanho do Efeito 29

3.1.3 Avaliando o Poder do Teste 29

3.2 Caso com Teste de Wilcoxon 31

3.2.1 Conclusão com Base no P-valor 33

3.2.2 Avaliando o Tamanho do Efeito 33

3.2.3 Avaliando o Poder do Teste 34

4 Análise Ampliada **36**

4.1 Bases de Dados 37

4.2 Análise com o Teste t de Student 39

4.2.1 Análise do Comportamento do P-valor 40

4.2.2 Avaliando o Tamanho do Efeito 44

4.2.3 Avaliando o Poder do Teste 50

4.3 Análise com o Teste de Wilcoxon 55

4.3.1 Análise do Comportamento do P-valor 56

4.3.2 Avaliando o Tamanho do Efeito 58

4.3.3 Avaliando o Poder do Teste 64

5 Dados Simulados **69**

5.1 Descrição da Simulação 69

5.2 Análise dos Resultados 70

6 Conclusão **79**

Referências **82**

Apêndice A – Resultados Obtidos no Estudo Empírico com o Teste t **84**

Apêndice B – Resultados Obtidos no Estudo Empírico com o Teste de Wilcoxon **94**

Capítulo 1

Introdução

Nas áreas de Aprendizado de Máquina e Mineração de Dados, uma das tarefas mais importantes é a de Classificação, que permite determinar à qual classe determinado elemento pertence a partir dos valores dos seus atributos. Um algoritmo de classificação busca "aprender" como classificar um novo elemento a partir de uma base de treinamento. Cada elemento dessa base é caracterizado pelos valores dos seus atributos e pelo valor da classe previamente conhecido. A tarefa de classificação é usada em diversas aplicações nas quais se deseja realizar alguma predição, como por exemplo: determinar se uma transação de cartão de crédito é fraudulenta, identificar quando um e-mail é spam ou até mesmo prever o resultado de um aluno em uma disciplina.

A busca pelo melhor desempenho na tarefa de classificação faz com que novos algoritmos sejam propostos e, assim, é de fundamental importância que os pesquisadores tenham as ferramentas adequadas para comparar conscientemente as novas abordagens com as estratégias já existentes [16].

O processo de comparação de classificadores é de extrema importância para o avanço da área, e uma ferramenta comumente utilizada nesse processo é a análise da significância estatística dos resultados. Esse tipo de análise é de grande importância para a ciência, já que estudos sobre populações inteiras são na maioria das vezes inviáveis, gerando a necessidade de inferir resultados de estudos amostrais sobre a população.

A significância estatística é verificada por meio de algum teste de hipóteses que busca evidências para rejeitar uma hipótese que se deseja colocar à prova [3], como por exemplo que dois classificadores têm resultados semelhantes. Porém, conforme destacado em [29], a significância estatística precisa ser muito bem compreendida e não deve ser a responsável por validar ou descartar pesquisas científicas.

O resultado de um teste de hipóteses é, na maioria das vezes, dado por meio do p-valor, uma medida que muitos pesquisadores desconhecem o significado, e percebe-se que muitos estão buscando cegamente a significância estatística por meio da condição p-valor<0,05. Ou seja, avaliando um nível de significância de 5%, sem uma compreensão exata do que esse valor representa na análise e sem dar a devida atenção a todos os elementos que a compõem. Como será visto neste estudo, esse ritual cego pode descartar resultados interessantes ou valorizar resultados não relevantes.

O tema é tão preocupante que pode ser visto em [29], que a ASA (American Statistical Association) se posiciona sobre o assunto, destacando que o p-valor é uma medida estatística útil, mas vem sendo utilizada abusivamente e mal interpretada. Ainda nesta publicação, eles aconselham que os pesquisadores evitem tirar conclusões científicas ou tomar decisões com base em p-valores isoladamente. Segundo o diretor executivo Ron Wasserstein, esta é a primeira vez em 177 anos de fundação que a ASA fez recomendações explícitas sobre um assunto tão fundamental em Estatística. Wasserstein acrescenta que os membros da ASA estão preocupados com que a má aplicação do p-valor lance dúvidas sobre as técnicas estatísticas em geral.

Outros conceitos de extrema importância, que têm sido ignorados pelos pesquisadores, são o poder do teste e o tamanho do efeito. O poder do teste é a probabilidade de rejeitar corretamente a hipótese nula quando esta deve ser rejeitada [13]. Desconsiderar essa probabilidade pode gerar um grave problema, especialmente nos casos em que não são obtidas diferenças estatisticamente significativas e é dado fim à investigação. Já o tamanho do efeito mede a força do resultado posto em teste, e ignorar essa medida é correr o risco de valorizar um resultado sem importância ou não considerar um resultado que poderia ser relevante.

Observa-se que a preocupação com essa questão está presente em diversas áreas. Na Biologia, por exemplo, em [20], conclui-se que o teste de significância é a abordagem predominante para análise de resultados e constata que, apesar disso, o teste não fornece informações importantes como a magnitude de um efeito de interesse e a precisão da estimativa desse efeito. Sendo assim, é sugerido que o p-valor venha acompanhado de uma medida de tamanho do efeito com seu respectivo intervalo de confiança.

Em [28], é observado que as análises estatísticas nas Ciências do Esporte não vão muito além da computação do p-valor e critica o fato de que o p-valor não fornece informações sobre a força real da relação entre as variáveis, e não permite ao pesquisador determinar o efeito de uma variável sobre outra. Indica-se ainda, nesse trabalho, que as medidas de

tamanho do efeito servem bem a esse propósito.

Já na Psicologia, em [24], informa-se que as práticas de análise de dados estão em mudança, o que é evidenciada pelo crescente número de periódicos que exigem informações sobre o tamanho do efeito. Nesse trabalho, é destacado uma afirmação de Gene V. Glass, estatístico e pesquisador americano que trabalha em Psicologia Educacional e Ciências Sociais: "A significância estatística é o que menos interessa em relação aos resultados. É importante que os resultados em termos de medidas de magnitude sejam apresentados, pois não se deve informar apenas que um tratamento afeta as pessoas, mas o quanto as afeta".

Na Medicina, em [26], destaca-se o problema do p-valor ser dependente do tamanho da amostra. Com uma amostra suficientemente grande, um teste estatístico quase sempre demonstrará uma diferença significativa a menos que não haja efeito algum, ou seja, quando o tamanho do efeito for exatamente zero. Afirma-se ainda que diferenças muito pequenas, mesmo que significativas, são muitas vezes sem importância. Assim, para que os leitores possam compreender completamente os resultados, não se deve relatar apenas o p-valor em uma análise.

Ainda em [26], apresenta-se um exemplo do problema do tamanho da amostra: um estudo sobre o uso da aspirina para prevenir o infarto do miocárdio, feito em mais de 22 mil indivíduos em um período de aproximadamente 5 anos, mostrou que a aspirina foi associada a uma redução do número de casos no infarto do miocárdio a partir da alta significância estatística observada nessa redução (com p-valor$<$0,00001). O estudo foi encerrado precocemente devido à evidência supostamente conclusiva e a aspirina foi recomendada para essa prevenção. No entanto, em um momento posterior foi verificado que o tamanho do efeito era extremamente pequeno. Dessa forma, muitas pessoas que foram aconselhadas a tomar aspirina não experimentaram benefício algum e ainda estavam em risco de efeitos adversos. Outros estudos encontraram efeitos ainda menores e a recomendação de usar aspirina desde então tem sido modificada.

Como visto acima, artigos de diversas áreas estão incentivando a apresentação de alguma medida de tamanho do efeito e, cada vez mais, esse tipo de abordagem vem sendo estimulada, em alguns casos até exigida, pelas publicações das áreas científicas.

Nesta dissertação foi feito um levantamento, no contexto de Aprendizado de Máquina e Mineração de Dados, considerando 11 artigos publicados em 2017 no periódico *Machine Learning*, que tratam de métodos de classificação e utilizam um ou mais testes de hipóteses para analisar seus resultados. Verificou-se que os pesquisadores da área estão simplificando

a análise, resumindo-a a busca pelo p-valor<0,05 e nenhum artigo apresentou alguma medida do tamanho do efeito nem relatou o poder do teste realizado.

Além disso, outros problemas também foram identificados nesse levantamento. Dois artigos não informam o teste que foi realizado, apenas que "a diferença foi significativa". Um deles informa o p-valor, porém pode-se dizer que é um "p-valor órfão" uma vez que o teste de hipóteses realizado não foi informado. Dos 11 artigos, apenas quatro apresentaram os p-valores dos testes aplicados. Outra informação que não foi apresentada em seis artigos foi o nível de significância.

Diante desse cenário, percebe-se a necessidade de se discutir sobre a metodologia utilizada para comparar classificadores. Nesta dissertação, então, é discutido e ilustrado como o p-valor, sozinho, não é suficiente para se realizar a comparação de classificadores. Além disso, são apresentadas medidas que acrescentam informação às análises realizadas, como o cálculo do tamanho do efeito e do poder do teste. Para evidenciar a importância dessas medidas, foram realizados estudos com diferentes classificadores muito utilizados atualmente: Random Forest [2], SVM [14], k-NN [7] e Naive Bayes [19], a partir de 50 bases de dados disponíveis no repositório da UCI [8]. O desempenho dos classificadores foi medido por meio das acurácias obtidas, e a significância estatística da diferença entre as médias das acurácias foi verificada por meio dos testes t de Student e do teste de Wilcoxon, juntamente com o cálculo do tamanho do efeito e do poder do teste.

Nos resultados, foram observados casos em que as três medidas (p-valor, tamanho do efeito e poder do teste) são concordantes. Por exemplo, p-valor evidenciando significância estatística com tamanho do efeito alto e poder do teste também alto. Porém, também foram encontrados casos em que essas medidas são discordantes, chamados aqui de "casos especiais", que terão destaque nesta dissertação.

Os casos em que a diferença dos resultados possui significância estatística, porém com um tamanho do efeito baixo, exigem uma atenção em relação à conclusão a ser tomada, já que existe o risco de valorizar um resultado sem real importância (como o exemplo da aspirina para o infarto). Já os casos opostos, com um tamanho do efeito alto porém sem significância estatística sobre as diferenças dos resultados, alertam para o problema de perder resultados interessantes caso a decisão seja tomada apenas com base no p-valor. Em algumas situações, o poder do teste pode ajudar a justificar o motivo de as duas medidas (p-valor e tamanho do efeito) estarem discordando, caso em que o teste pode não estar sendo utilizado em um ambiente ideal. Como, por exemplo, quando a amostra é muito pequena e não suficiente para indicar a significância do resultado.

Esta dissertação está organizada da seguinte forma. No Capítulo 2, apresenta-se o material teórico necessário para o desenvolvimento do estudo. Definem-se, principalmente, os testes t de Student e de Wilcoxon para dados pareados e as suas respectivas medidas de tamanho do efeito. No Capítulo 3, são apresentados dois estudos de casos, um para o teste t de Student e outro para o teste de Wilcoxon, onde o p-valor e o tamanho do efeito discordam, chamados casos especiais. No Capítulo 4, é realizado um estudo mais amplo, considerando-se quatro diferentes classificadores e 50 bases de dados, comparando-se os resultados por meio do teste t de Student e do teste de Wilcoxon. Será observado que os casos especiais ocorrem com frequência quando utilizados classificadores e bases de dados reais. Além disso, dados sintéticos que simulam os resultados de classificadores serão gerados e utilizados para ampliar a avaliação realizada e tornar os resultados mais conclusivos. Para finalizar, no Capítulo 6, são apresentadas as conclusões da dissertação e possíveis direções para trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Neste capítulo, a Seção 2.1 apresenta os principais conceitos de testes de hipóteses inserindo-os no contexto da comparação de classificadores. Em seguida, são definidos os testes que serão utilizados nesta dissertação: teste t de Student e teste de Wilcoxon. A Seção 2.2 define a função poder do teste e apresenta como ela calculada para os testes já apresentados. Finalizando o capítulo, na Seção 2.3, é apresentado o conceito de tamanho do efeito e definidas as medidas que serão utilizadas para calculá-lo.

2.1 Teste de Hipóteses

Em Aprendizado de Máquina e Mineração de Dados, a proposta de uma nova estratégia de classificação vem acompanhada pela comparação de seus resultados com o de estratégias já conhecidas, a fim se de defender sua eficácia. Ou seja, busca-se testar se os novos resultados são melhores que os já conhecidos. Assim, é de extrema importância que os pesquisadores tenham ferramentas adequadas para avaliar essas abordagens [16], e uma dessas ferramentas é o teste de hipóteses.

Para facilitar a compreensão dos conceitos abordados neste capítulo, a ideia de teste de hipóteses será introduzida por meio de um exemplo fictício que, partindo de uma situação simples, será gradualmente ampliado para atender à situação geral do teste de hipóteses.

Exemplo fictício: Suponha que um novo classificador B foi proposto e que o seu desenvolvedor afirma que ele tem resultados melhores do que o classificador A já existente, para uma determinada base de dados. Deseja-se verificar se o classificador B possui de fato melhores resultados do que o classificador A nessa base.

Uma das principais métricas utilizadas para avaliar o desempenho de um classificador

é a acurácia. Para calculá-la, basta verificar a porcentagem de acerto do classificador. Porém, em geral, não é possível obter a acurácia real (ou esperada) do classificador sem o conhecimento da verdadeira distribuição dos dados e suas classes [16].

Para estimar a acurácia, uma das abordagens mais populares no Aprendizado de Máquina é a validação cruzada com k partições (*k-fold cross validation*). Essa abordagem é uma forma de reamostragem e consiste em dividir o conjunto de dados em k subconjuntos de tamanhos aproximadamente iguais, onde cada subconjunto é chamado de partição. Então, o classificador é treinado em $k - 1$ dessas partições (juntas) e testado na partição restante. Esse procedimento é repetido k vezes com cada partição diferente sendo utilizada para teste. Assim, k estimativas de acurácias são obtidas através desse método.

Também é possível realizar múltiplas reamostragens para estimar a acurácia. Esse método consiste em calcular a média das k acurácias obtidas com a validação cruzada com k partições. E refazer todo o processo da validação cruzada o número de vezes desejadas. Porém, neste trabalho será utilizada apenas a comparação utilizando uma única execução do método de validação cruzada com k partições.

Em [18], é recomendada a utilização de validação cruzada com dez partições. Então, $k = 10$ será uma das opções utilizadas neste trabalho. Além de $k = 10$, também serão considerados os métodos de validação cruzada com 20 e 30 partições, ou seja, $k = 20$ e $k = 30$.

Considere que, para mostrar que seu novo classificador B tem resultados melhores do que o classificador A, o desenvolvedor aplicou esses classificadores na base de dados utilizando o método de validação cruzada com 10 partições. Assim, ele obteve 10 acurácias para o classificador A e 10 acurácias para o classificador B. Essas acurácias observadas também são chamadas de **amostras**.

A acurácia de um classificador é a característica da população que está sendo observada nesse exemplo, e em Estatística, essa característica é denominada **variável**. Usualmente, as variáveis são denotadas por letras maiúsculas. Considere, por exemplo, que a variável X seja a acurácia do classificador A e Y a acurácia do classificador B. As médias das acurácias dos classificadores são chamadas de **parâmetros**, que são características populacionais desconhecidas. Os parâmetros são usualmente representados por letras gregas. Considere por exemplo, μ_X a média populacional das acurácias do classificador A e μ_Y a média populacional das acurácias do classificador B.

Considere as acurácias obtidas pelo desenvolvedor através da aplicação do classifica-

dor A, ou seja, a amostra da característica X, como x_1,x_2, ...,x_{10}, e as acurácias observadas do classificador B, amostra da característica Y, como y_1,y_2, ...,y_{10}. Considere também que a média amostral das acurácias obtidas através do classificador A é $\overline{x}$ e que a média amostral das acurácias do classificador B é $\overline{y}$.

Como o objetivo é fazer uma afirmação sobre as médias das acurácias populacionais (os parâmetros) e elas são desconhecidas, é preciso utilizar uma função da amostra para estimar esses parâmetros, também chamada de **estimador**. Os valores numéricos observados para os estimadores, neste caso, as médias amostrais observadas, são chamadas de **estimativas pontuais**. Então, no exemplo em desenvolvimento, os estimadores dos parâmetros μ_X e μ_Y são $\overline{X}$ e $\overline{Y}$, e as estimativas são $\overline{x}$ e $\overline{y}$ (valores observados), respectivamente. A Figura 2.1 ilustra essas definições para a variável X.

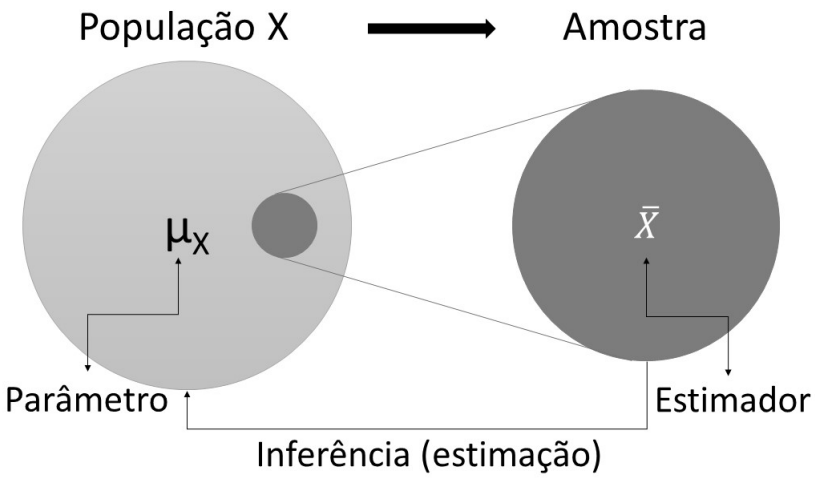


Figura 2.1: Esquema de estimação com a variável X

Suponha que seja aplicado o classificador A na base de dados em análise utilizando o método de validação cruzada com 10 partições, e a média das 10 acurácias observadas seja igual a 90 ($\overline{x}$). Se o mesmo experimento for realizado considerando outro particionamento (aleatório), a nova média observada não seria necessariamente igual a 90 novamente. Então, a média das acurácias ($\overline{X}$) é uma variável aleatória que possui uma distribuição de probabilidade.

Quando o tamanho da amostra (n) aumenta, independentemente da forma da distribuição da população, a distribuição amostral de $\overline{X}$ aproxima-se cada vez mais de uma distribuição normal. Esse resultado é conhecido como Teorema Limite Central (TLC), apresentado a seguir. Então, conforme o n vai aumentando, as observações de $\overline{X}$ tendem

a se concentrar em torno de uma média e a variância tende a diminuir. Assim, os casos mais extremos passam a ter baixa probabilidade.

Teorema 2.1.1. *TLC: Para amostras aleatórias simples* $(X_1, \ldots, X_n)$*, retiradas de uma população com média* μ *e variância* σ^2 *finita, a distribuição amostral da média* $\overline{X}$ *aproxima-se, para* n *grande, de uma distribuição normal, com média* μ *e variância* σ^2/n*.*

Portanto, com uma amostra observada e buscando testar afirmações sobre parâmetros, uma ferramenta muito utilizada nesses casos é o **teste de hipóteses**. No teste de hipóteses, são buscadas evidências para decidir entre uma hipótese conservadora e uma hipótese mais inovadora. Por exemplo, de que a média populacional das acurácias do classificador A é maior ou igual à média do classificador B, e de que a média populacional das acurácias do classificador B é maior do que a média do classificador A, respectivamente. Essas hipóteses complementares são chamadas de **hipótese nula** e **hipótese alternativa**, que são representadas por H_0 e H_1, respectivamente, e podem ser definidas da seguinte forma:

$$\begin{cases} H_0: & \mu_X \geq \mu_Y \\ H_1: & \mu_X < \mu_Y. \end{cases} \tag{2.1}$$

Entretanto, a definição das hipóteses poderia ser diferente para o exemplo abordado se o desenvolvedor tivesse afirmado que "os resultados do novo classificador B são **diferentes** dos resultados do classificador A", ao invés de "os resultados do novo classificador B são **melhores** do que os resultados do classificador A". Como não há informação sobre a direção da diferença entre as acurácias dos classificadores, devem ser consideradas as diferenças positivas e negativas. Então, nesses casos, as hipóteses são definidas da seguinte forma.

$$\begin{cases} H_0: & \mu_X = \mu_Y \\ H_1: & \mu_X \neq \mu_Y \end{cases} \tag{2.2}$$

A forma em que as hipóteses são definidas está ligada ao teste ser **unilateral** ou **bilateral**. Se existe conhecimento ou uma expectativa forte a priori sobre a direção em que a hipótese alternativa deve diferenciar-se da hipótese nula, deve ser usado o teste unilateral, como está sendo feito no exemplo abordado, já que o desenvolvedor afirmou que o resultado do classificador B é melhor do que o do classificador A.

É possível simplificar as hipóteses do teste definindo $D = X - Y$, ou seja, a variável D é a diferença entre a acurácia do classificador A e acurácia do classificador B. Consequentemente, o parâmetro μ_D é a diferença entre a média populacional das acurácias do

classificador A e a média populacional das acurácias do classificador B, e $\overline{D}$ é o estimador desse parâmetro. Além disso, as amostras coletadas também podem ser resumidas na diferença entre elas: d_1,d_2, ...,d_{10}, onde $d_i = x_i - y_i$. Assim, $\overline{d}$ é a média amostral da diferença entre as acurácias obtidas através dos classificadores A e B. Logo, as hipóteses definidas anteriormente em 2.1 e 2.2 podem ser redefinidas, respectivamente, como a seguir.

$$\begin{cases} H_0: & \mu_D \geq 0 \\ H_1: & \mu_D < 0 \end{cases} \qquad e \qquad \begin{cases} H_0: & \mu_D = 0 \\ H_1: & \mu_D \neq 0 \end{cases} \tag{2.3}$$

Vale destacar que essa simplificação só é possível quando se têm duas amostras dependentes, ou seja, quando cada observação da primeira amostra é pareada com a observação, respectiva, na segunda amostra e por isso $d_i = x_i - y_i$. Tal situação ocorre, por exemplo, na comparação de dois classificadores aplicados na mesma partição de uma base de dados. Em Estatística, os testes de hipóteses que devem ser utilizados em situações como essa são chamados de **testes de hipóteses para dados pareados**.

Se a média amostral das acurácias do classificador A for muito menor do que a média das acurácias da amostra do classificador B (por exemplo 70% e 90%, respectivamente), então existe um forte indício de que o desenvolvedor esteja certo, e que, de fato, o classificador B tem uma média das acurácias populacionais maior do que a média populacional das acurácias do classificador A.

Entretanto, se a diferença entre as médias das acurácias amostrais dos classificadores A e B não for tão extrema (por exemplo 80% e 81%), é mais difícil dizer que esse resultado indica que a média das acurácias populacionais do classificador B é maior do que a média populacional do classificador A, já que a diferença observada pode ser devido à amostra coletada e não ser realmente verificada na população.

Então, a partir de qual diferença entre as médias amostrais deve-se afirmar que o classificador B tem resultados melhores do que o A? Responder a essa pergunta é o mesmo que criar uma regra para decidir entre as hipóteses que estão em teste. Porém, não existe uma regra única para essa questão. Pessoas podem criar regras diferentes para tomar uma decisão. Uma pessoa poderia decidir, por exemplo, que se a diferença entre as médias amostrais das acurácias for menor que -3%, ela considera que o classificador B tem resultados melhores que o A, ou seja, rejeita a hipótese H_0. Já uma pessoa mais conservadora poderia afirmar que precisa de uma evidência mais forte para tomar essa decisão, então só afirmaria que o classificador B tem resultados melhores do que o A se

diferença entre as médias das acurácias amostrais for maior do que 8%, por exemplo.

Criar a regra de decisão é o mesmo que definir quais valores de médias de acurácias são favoráveis a hipótese alternativa (H_1). Em Estatística, esses valores formam a chamada **região crítica**. A primeira pessoa do exemplo anterior definiu como a região crítica, para a sua tomada de decisão, valores de diferença entre as médias amostrais das acurácias maiores do que 3%, ou seja, ela definiu $RC = \{\overline{d} : \overline{d} > 3\}$. É possível definir a região crítica de maneira geral, considerando v o valor de limite da região crítica, então:

$$RC = \{\overline{d} : \overline{d} > v\}. \tag{2.4}$$

Quando se trata de um teste bilateral, a região crítica é composta de duas partes:

$$RC = \{\overline{d} : \overline{d} > v_1 \ e \ \overline{d} < v_2\} \tag{2.5}$$

onde, v_1 e v_2 são os valores de limites da região crítica.

Independente da regra de decisão adotada, existe a probabilidade de se cometerem erros com a decisão tomada, já que se trata de uma decisão com base em uma amostra sobre a característica da população. Na situação que está sendo discutida, por exemplo, é possível cometer dois tipos de erro: o primeiro é afirmar que o classificador B tem resultados melhores do que o classificador A quando na verdade não tem, e o segundo é não afirmar que o classificador B tem resultados melhores do que o A e na verdade tem.

Em Estatística, o primeiro erro é chamado de **erro tipo I**, ou seja, quando a hipótese nula é rejeitada sendo que ela é verdadeira. Já o segundo erro é chamado de **erro tipo II**, ou seja, quando não se rejeita a hipótese nula e ela é falsa. A probabilidade de se cometer o erro tipo I, ou seja, de afirmar que o classificador B tem resultados melhores do que o classificador A quando na verdade não tem, é denotada por α. E a probabilidade de cometer o erro tipo II, ou seja, de não afirmar que o classificador B tem resultados melhores do que o A e na verdade tem, é denotado por β. Sendo assim:

$$\alpha = P(\text{erro tipo } I) = P(\text{rejeitar } H_0 | H_0 \text{ é verdadeira}), \tag{2.6}$$

$$\beta = P(\text{erro tipo } II) = P(\text{não rejeitar } H_0 | H_0 \text{ é falsa}). \tag{2.7}$$

Se H_0 é verdadeira, o valor de μ_D é 0. Então, nos cálculos envolvendo α, deve ser considerado que $\mu_D = 0$. Porém, se H_0 é falsa, não existe apenas um valor possível para μ_D. Então, β é na verdade uma função do real valor de μ_D. Logo, a Equação 2.7 pode

ser reescrita como:

$$\beta(\mu_D) = P(\text{não rejeitar } H_0|\mu_D). \tag{2.8}$$

Segundo [30], a Tabela 2.1 resume os erros contidos nos testes de hipóteses.

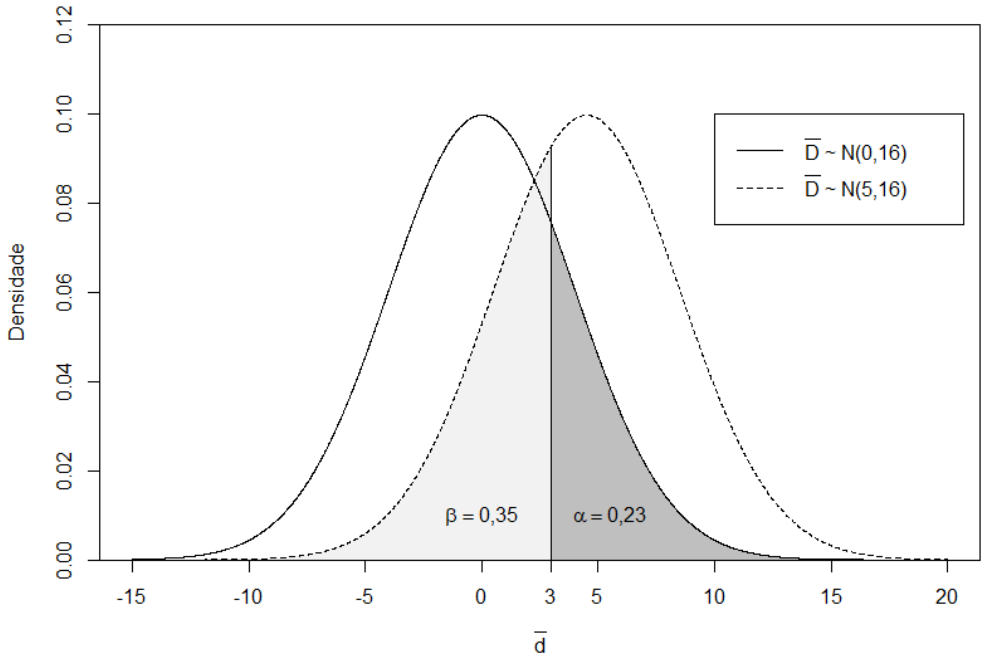
Tabela 2.1: Decisões e erros associados em testes estatísticos

Realidade (desconhecida)	Decisão do teste	
	Não rejeita H_0	Rejeita H_0
H_0 verdadeira	Decisão correta (1-α)	Erro tipo I (α)
H_0 falsa	Erro tipo II (β)	Decisão correta (1-β)

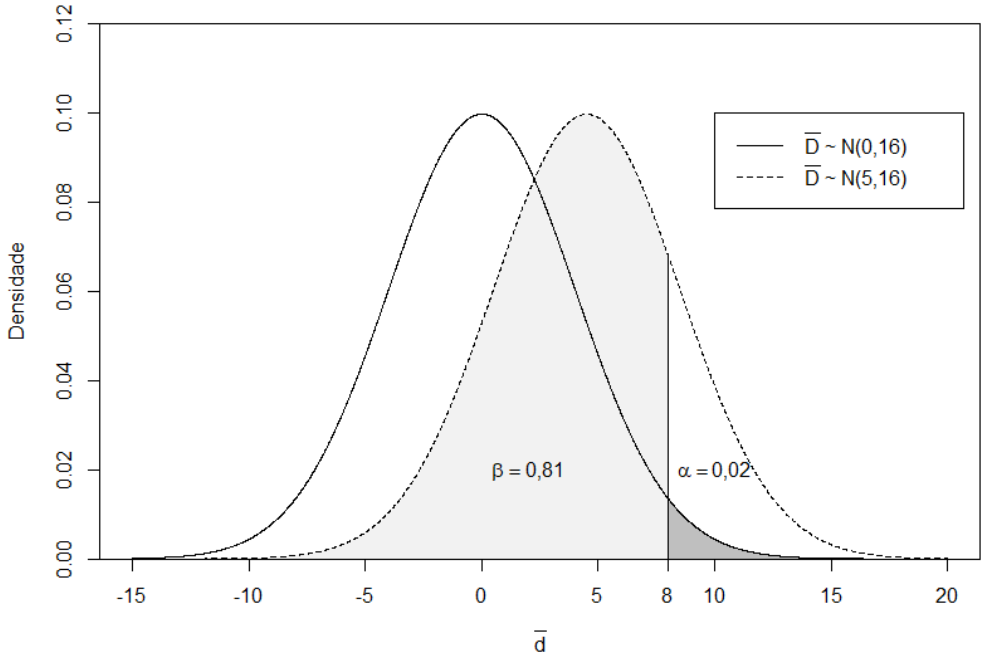
Na situação perfeita, a probabilidade seria nula nos dois tipos de erro, mas, na prática, elas não são. Então, como controlar esses erros? A pessoa que escolheu o valor 3 para ser o limite da sua região crítica tem uma probabilidade maior de cometer o erro do tipo I do que a pessoa que escolheu o 8, pois a chance de a diferença observada ter sido apenas por causa da amostra é maior. Já a pessoa que escolheu o valor limite de 8 tem maior chance de cometer o erro tipo II do que a pessoa que escolheu o limite de 3. Essa diferença entre as probabilidades dos erros pode ser vista na Figura 2.2.

A Figura 2.2 mostra as probabilidades dos erros para as regiões críticas utilizadas como exemplo. Para ilustrar H_0 verdadeira, foi considerado que a diferença entre as acurácias populacionais dos classificadores A e B tem distribuição normal com média 0 e variância 16, ou seja, $\overline{D} \sim N(0{,}16)$, representada pela linha sólida. E para ilustrar H_0 falsa, foi considerado que essa diferença tem distribuição normal com média 5 e variância 16, ou seja, $\overline{D} \sim N(5{,}16)$, representada pela linha pontilhada. Sendo assim, a probabilidade α de cometer o erro tipo I, rejeitar H_0 sendo ela verdadeira, é obtida através da área embaixo da curva gerada considerando H_0 verdadeira e para os valores pertencentes à região crítica definida. Então essa probabilidade de erro foi 23% para a região crítica definida com diferenças amostrais maiores que 3, e 2% para a região crítica definida para diferenças maiores que 8, representadas pelas áreas hachuradas mais escuras.

Já a probabilidade β de se cometer o erro tipo II, não rejeitar H_0 quando ela é falsa, é obtida através da área embaixo da curva gerada considerando H_0 falsa, assumindo um valor para a verdadeira diferença das médias, neste caso, 5. Assim, nessa situação, a probabilidade de afirmar que o classificador B não tem resultados melhores do que o classificador A para a base de dados em questão, quando na verdade ele tem resultados melhores, é 35% para a região crítica definida com diferenças amostrais maiores que 3, e 81% para a região crítica definida para diferenças maiores que 8, representadas pelas áreas



(a) $RC = \{\overline{d} : \overline{d} > 3\}$



(b) $RC = \{\overline{d} : \overline{d} > 8\}$

Figura 2.2: Probabilidades do erro tipo I (α) e do erro tipo II (β), para diferentes definições da região crítica, considerando que a diferença entre as acurácias tem distribuição normal com média 0 e variância 16 quando H_0 é verdadeira e a média dessa distribuição é 5 quando H_0 é falsa

hachuradas mais claras. Cabe ressaltar que esses valores foram observados para $\beta(5)$, ou seja, considerando $\mu_D = 5$.

Do exposto acima foi constatado que, escolhida uma região crítica, pode-se achar as probabilidades α e β de cometer cada tipo de erro. Mas também é possível proceder de modo inverso: fixar α e encontrar a regra de decisão que irá corresponder à probabilidade de erro de tipo I igual a α. Esse é um dos procedimentos que pode ser realizado no teste de hipóteses. Ou seja, quando o pesquisador vai utilizar um teste de hipóteses, ele precisa definir que probabilidade é aceitável para afirmar que o classificador B tem resultados melhores do que o classificador A quando na verdade não tem (α). E, a partir desse valor definido, definir a região crítica. Essa probabilidade de cometer o erro tipo I também é chamada de **nível de significância**, que é definido pelo pesquisador. Usualmente, esse valor de α é fixado em 5%, 1% ou 0,1%.

Para definir o nível de significância utilizado no teste, o pesquisador precisa analisar o seu problema. Não existe uma regra de qual é o melhor valor para ser adotado. Imagine um pesquisador que está testando se um novo tratamento é mais eficaz que um tratamento já conhecido e utilizado para tratar certa doença. O erro de ele afirmar que o novo tratamento é melhor do que o já utilizado quando na verdade não é, representa um erro que pode trazer graves consequências, já que trocar erradamente o tratamento poderia por em risco a vida dos pacientes. Nesse caso, utilizar o nível de significância 10% pode ser considerado uma imprudência, pois a probabilidade de cometer o erro tipo I deve ser o menor possível. Um valor bem utilizado seria por exemplo $\alpha = 0{,}1\%$.

Até este ponto, a região crítica foi definida em função de $\overline{d}$, mas ela também pode ser definida em função da **Estatística de Teste** T. A estatística de teste é uma função da amostra, cuja distribuição depende do parâmetro que está sendo posto em teste e é conhecida quando H_0 é verdadeira. A definição dessa função é específica para cada teste. Então, nas Seções 2.1.2 e 2.1.3, serão definidas as estatísticas de teste para os testes t e de Wilcoxon, respectivamente. De maneira geral, é preciso calcular a estatística de teste observada $t = T(d_1,d_2,\ldots,d_n)$ e determinar a distribuição de probabilidade de T sendo H_0 verdadeira [21]. Se t pertencer à região crítica, a hipótese nula é rejeitada.

O método de construção de um teste de hipóteses, descrito até aqui, parte da fixação do nível de significância α e a construção de uma região crítica. Outra maneira de se proceder, que leva à mesma tomada de decisão, consiste em apresentar o **p-valor** do teste. A tomada de decisão com base no p-valor é o método mais conhecido e mais utilizado por ser facilmente obtido em diversos softwares que realizam testes de hipóteses.

2.1.1 P-valor

O procedimento de tomada de decisão de um teste de hipóteses com base no p-valor é muito parecido (e leva ao mesmo resultado) com o procedimento já apresentado. O que se faz é indicar a probabilidade de ocorrer valores da estatística de teste mais extremos do que o observado, sob a hipótese de H_0 ser verdadeira [3]. Além disso, a apresentação do p-valor nas publicações científicas é defendida pois acrescenta mais informação do que apenas rejeitar ou não a hipótese nula através da região crítica, caso em que só seria informado o nível de significância α. Segundo [21], o p-valor fornece uma ideia do quão longe se está da veracidade ou da falsidade da hipótese nula.

P-valor é a probabilidade de se obter uma estatística de teste igual ou mais extrema quanto aquela observada em uma amostra, assumindo verdadeira a hipótese nula [11]. Sendo assim, o p-valor é calculado como

$$\text{p-valor} = P(T \geq t|H_0), \tag{2.9}$$

quando se trata de um teste de hipótese unilateral à direita. Quando se trata de um teste unilateral à esquerda, o p-valor é $P(T \leq t|H_0)$. Já quando se trata de um teste bilateral, o p-valor é $P(T \geq |t||H_0) + P(T \leq -|t||H_0)$.

Também é possível interpretar o p-valor como o menor valor do nível de significância para o qual é rejeitada H_0. Dessa forma, se o p-valor obtido for menor que o nível de significância α definido para o teste, a hipótese nula H_0 é rejeitada. Em termos gerais, um p-valor muito pequeno significa que a probabilidade de se obter um valor da estatística de teste como o observado é muito improvável, levando assim à rejeição da hipótese nula.

Para se obter um resultado eficaz em um teste de hipóteses, é imprescindível que se aplique um teste adequado àquele cenário e esse resultado seja interpretado corretamente. Embora se tenha à disposição uma ampla variedade de ferramentas com qualidade para realizar um teste de hipóteses, é importante que os pesquisadores estejam atentos a esses testes e, mais ainda, que compreendam a estrutura em que operam. Nas seções seguintes, são apresentados os testes de hipóteses t de Student e de Wilcoxon, ambos para dados pareados, pois são os testes utilizados no desenvolvimento deste trabalho.

2.1.2 Teste t de Student para Amostras Pareadas

O objetivo do teste t de Student, ou simplesmente teste t, é verificar se as médias das variáveis X e Y são iguais e, por ser um teste pareado, isso equivale a verificar se a

diferença $D = X - Y$ entre as médias das variáveis é igual a zero. O teste t é um **teste paramétrico** e por isso, para a sua aplicação, o pressuposto da distribuição normal precisa ser verificado.

Testes paramétricos devem ser aplicados a dados que obedecem a uma distribuição normal, ou seja, uma distribuição simétrica em volta da média (que coincide com a moda e mediana) e tem a forma de um sino. Existem diversos testes para verificar a suposição de normalidade dos dados, como Cramer-von Mises, Kolmogorov-Smirnov e Shapiro-Wilk, bem como recursos gráficos, como histograma e qqplot. Para a realização do estudo empírico deste trabalho, é utilizado o teste Kolmogorov-Smirnov, que consiste em observar a diferença entre a função de distribuição acumulada empírica dos dados da amostra com a distribuição esperada, no caso, a distribuição normal.

Para se realizar o teste t para amostras pareadas deve-se primeiramente estabelecer as hipóteses. Neste estudo será considerado o teste bilateral, então as hipóteses são definidas da seguinte forma.

$$\begin{cases} H_0: & \mu_D = 0 \\ H_1: & \mu_D \neq 0 \end{cases} \tag{2.10}$$

Vale lembrar que o parâmetro μ_D, diferença entre as médias populacionais das acurácias dos classificadores A e B, é estimado por $\overline{D}$, diferença entre as médias das n acurácias observadas nas aplicações desses classificadores.

A estatística do teste é calculada através da expressão

$$T = \frac{\overline{D} - \mu_D}{\sqrt{\frac{s_d^2}{n}}}, \tag{2.11}$$

onde n é o tamanho da amostra e s_d^2 é a variância das diferenças amostrais.

Sob H_0 verdadeira, ou seja, $\mu_D = 0$, T segue uma distribuição t de Student com $n-1$ graus de liberdade. Essa distribuição só é conhecida se D tem distribuição normal. Esse resultado é decorrente do seguinte teorema:

Teorema 2.1.2. *Sejam W e Z variáveis aleatórias independentes, W sendo normalmente distribuída com média 0 e variância 1, e Z tendo distribuição qui-quadrado com ν graus de liberdade. Então, a variável $T = \frac{W}{\sqrt{Z/\nu}}$ tem distribuição t de Student com ν graus de liberdade.*

Logo, quando se trata de um teste t bilaral, o p-valor é dado por

$$\text{p-valor} = P(T \geq |t||H_0) + P(T \leq -|t||H_0) = 2 \times P(T \geq |t||H_0). \tag{2.12}$$

Caso uma das condições do teste paramétrico não seja aceita, então o teste t para dados pareados não pode ser aplicado. O teste alternativo dado na literatura é o teste de Wilcoxon para dados pareados. Por ser um teste não-paramétrico, essas condições não precisam ser verificadas.

2.1.3 Teste de Wilcoxon para Amostras Pareadas

O teste de Wilcoxon para amostras pareadas é um teste não paramétrico dos mais populares. O objetivo desse teste é verificar se as distribuições das variáveis X e Y diferem em localização. Por ser um teste para amostras pareadas, é possível afirmar que o objetivo do teste é verificar se a mediana da diferença $D = X - Y$ é igual a zero.

Como os testes não paramétricos não necessitam de requisitos tão fortes como a normalidade, eles têm a desvantagem de não serem tão poderosos quanto os paramétricos, porém são indicados quando se opta por conclusões mais conservadoras, já que não é necessário fazer nenhuma suposição sobre a distribuição da variável aleatória. As suposições do teste são que os pares são selecionados aleatoriamente e que a população das diferenças tem distribuição aproximadamente simétrica.

Assim, será testado se as populações diferem (ou não) em localização utilizando a seguinte ideia: se a hipótese nula é aceita, tem-se que a mediana da diferença é nula, ou seja, as populações não diferem em localização. Já se a hipótese nula for rejeitada, ou seja, se a mediana da diferença não for nula, tem-se que as populações diferem em localização. Considerando δ_D a mediana de D, as hipóteses são definidas da seguinte forma.

$$\begin{cases} H_0: & \delta_D = 0 \\ H_1: & \delta_D \neq 0 \end{cases} \tag{2.13}$$

Para realizar o teste de Wilcoxon, deve-se atribuir postos a cada valor absoluto das diferenças entre os pares $|d_i|$, onde $d_i = x_i - y_i$. Ao menor $|d_i|$, é atribuído o posto 1, ao segundo maior, o posto 2 e assim por diante. A cada posto, deve-se atribuir o sinal da diferença, isto é, indicar quais postos decorrem de diferenças negativas e quais de diferenças positivas.

Eventualmente, os valores de dois pares serão iguais resultando em uma diferença

nula. Nesse caso, eles são excluídos da análise. Da mesma forma, o valor de n (tamanho da amostra) será reduzido na mesma quantidade de valores em que a diferença for nula.

Se duas ou mais diferenças têm o mesmo valor absoluto, atribui-se o mesmo posto aos empates. Este posto é a média dos postos que teriam sido atribuídos se as diferenças fossem diferentes. Por exemplo, se três pares acusam as diferenças: -1, -1 e $+1$, a cada par será atribuído o posto 2, que é a média entre 1, 2 e 3. O próximo valor, pela ordem, receberia o valor 4, porque já teriam sido utilizados os postos 1, 2 e 3.

Se X e Y são equivalentes, isto é, se H_0 é verdadeira, é de se esperar que algumas das maiores diferenças sejam positivas e outras negativas. Dessa forma, se forem somados os postos de cada sinal, deve-se esperar somas aproximadamente iguais. Se houver diferença entre essas duas somas, é sinal de que X e Y não se equivalem e deve-se então rejeitar a hipótese nula.

Seja a estatística de teste W a menor soma dos postos de mesmo sinal (negativos ou positivos), isto é, ou a soma dos postos positivos ou a soma dos postos negativos (a que for menor), então rejeita-se H_0 se W for "pequeno". A decisão também pode ser tomada através do p-valor que é calculado da seguinte forma:

$$p - valor = P(W \leq w \mid H_0) \tag{2.14}$$

Quando o teste de Wilcoxon é feito manualmente, a tomada de decisão geralmente é feita utilizando uma tabela com valores críticos tabelados para se comparar a estatística de teste W obtida, já que o cálculo do p-valor pode ser bastante extenso para grandes amostras. Porém, ele pode ser facilmente obtido em diversos softwares que realizam o teste de Wilcoxon para dados pareados. Para ilustrar a obtenção do p-valor, veja a seguir um exemplo inspirado em [3].

Considere os valores da Tabela 2.2 para as acurácias obtidas em 10 partições com a aplicação dos classificadores A e B numa dada base de dados.

Tabela 2.2: Amostras das acurácias obtidas através dos classificadores A e B em 10 partições de uma dada base, com postos para as diferenças

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
x	90	85	91	90	88	89	85	90	89	94
y	95	85	90	90	88	89	92	93	93	94
d	-5	0	1	0	0	0	-7	-3	-4	0
Posto de |d|	4		1				5	2	3	
Sinal	-		+				-	-	-	

Deseja-se testar a hipótese de que as acurácias de A e B são semelhantes contra a hipótese de que as acurácias de B são maiores. Ou, ainda, $H_0 : \delta_D = 0$ e $H_1 : \delta_D < 0$.

Nota-se que só há um posto positivo, +1, então a soma dos postos de sinal positivo é $W^+ = 1$, enquanto a soma dos postos de sinal negativo é $W^- = 4+5+2+3 = 14$. Logo, a estatística de teste, menor soma dos postos de mesmo sinal, é $W = W^+ = 1$. Note que $W^+ + W^- = 15$, que é a soma de todos os postos dos $\mid d_i \mid$, que, por sua vez, é $n(n+1)/2$, sendo $n = 5$ o número de pares com diferença não nula. Para se conduzir o teste, deve-se obter a distribuição dessa estatística, sob a hipótese nula H_0. Para isso, note que, se H_0 for verdadeira, cada posto tem a mesma probabilidade de ser associado com um sinal + ou com um sinal –. Logo, a sequência de postos sinalizados é uma de todas as possíveis combinações de $\pm 1, \pm 2, \ldots, \pm 5$. Há $2^5 = 32$ combinações, todas equiprováveis sob H_0, ou seja, com probabilidade 1/32. A Tabela 2.3 tem todas as possibilidades juntamente com o valor de W^+.

Tabela 2.3: Possíveis sinais para os postos

1	**2**	**3**	**4**	**5**	$\boldsymbol{W^+}$	**1**	**2**	**3**	**4**	**5**	$\boldsymbol{W^+}$
+	+	+	+	+	15	+	+	-	+	-	7
-	+	+	+	+	14	-	+	-	-	+	7
+	-	+	+	+	13	-	-	+	+	-	7
+	+	-	+	+	12	+	-	-	-	+	6
-	-	+	+	+	12	+	+	+	-	-	6
+	+	+	-	+	11	-	+	-	+	-	6
-	+	-	+	+	11	+	-	-	+	-	5
+	+	+	+	-	10	-	+	+	-	-	5
-	+	+	-	+	10	-	-	-	-	+	5
+	-	-	+	+	10	+	-	+	-	-	4
-	+	+	+	-	9	-	-	-	+	-	4
-	-	-	+	+	9	+	+	-	-	-	3
+	-	+	-	+	9	-	-	+	-	-	3
+	+	-	-	+	8	-	+	-	-	-	2
+	-	+	+	-	8	+	-	-	-	-	1
-	-	+	-	+	8	-	-	-	-	-	0

O p-valor é $P(W^+ \leq 1 \mid H_0) = 2/32 = 0{,}06$. Ou seja, ao nível de significância de 5%, não há indicação de que a acurácia do classificador B é maior do que a acurácia do classificador A. Observe que restaram poucos pares para a realização do teste e, apesar de ter sido observado somente um posto positivo, não foi suficiente para ter evidências para rejeitar a hipótese nula.

Na Seção 2.1, foram apresentados os conceitos relacionados a teste de hipóteses e os testes utilizados neste estudo: teste t e teste de Wilcoxon. Com essas ferramentas, é possível verificar se existe significância estatística entre a diferença dos resultados de dois classificadores em determinada base de dados. Na próxima seção, é apresentada a função poder do teste.

2.2 Poder do Teste

Em um teste de hipóteses, existe a probabilidade de se cometer dois tipos de erros, como já foi visto nas Equações 2.6 e 2.7. A probabilidade do erro tipo II (β) está relacionada com

o **poder do teste** estatístico $(1 - \beta)$. O poder de um teste estatístico é a probabilidade do teste rejeitar H_0 quando H_0 realmente é falsa. Sendo assim, no contexto considerado, o poder do teste é a probabilidade do teste afirmar que os classificadores A e B são diferentes quando realmente são.

O poder do teste é na verdade uma função, pois depende de alguns fatores: nível de significância α adotado, distância entre o valor "real" (desconhecido) do parâmetro (μ_D) e o considerado verdadeiro em H_0, a variância da população e o tamanho da amostra. É comum se argumentar que o teste deve ter poder de no mínimo 80%, como sugerido por [5], por exemplo. Porém, o valor do poder do teste deve fazer sentido no ambiente utilizado e, sobretudo, nas consequências do erro tipo II na decisão do investigador.

Dada a Equação 2.8, o poder do teste π é calculado da seguinte forma.

$$\pi(\mu_D) = 1 - \beta(\mu_D) = P(rejeitar\ H_0 \mid \mu_D) \tag{2.15}$$

Uma forma de se reduzir a probabilidade de se cometer o erro tipo II é aumentar o tamanho da amostra. Quanto maior o tamanho da amostra, maior a representatividade da mesma e, portanto, maior será o poder do teste. Isto é, maior será a probabilidade de rejeitar um H_0 falso. Na Figura 2.3, são apresentadas as curvas do poder do teste para um teste bilateral, cujas hipóteses são $H_0 : \mu_D = 0$ e $H_1 : \mu_D \neq 0$, e com dois tamanhos de amostras diferentes, $n = 10$ e $n = 30$. Supondo que $\mu_D = 2$, no teste com $n = 10$, o poder do teste é 44%, porém, aumentando o tamanho da amostra para $n = 30$, o poder passa para 94%.

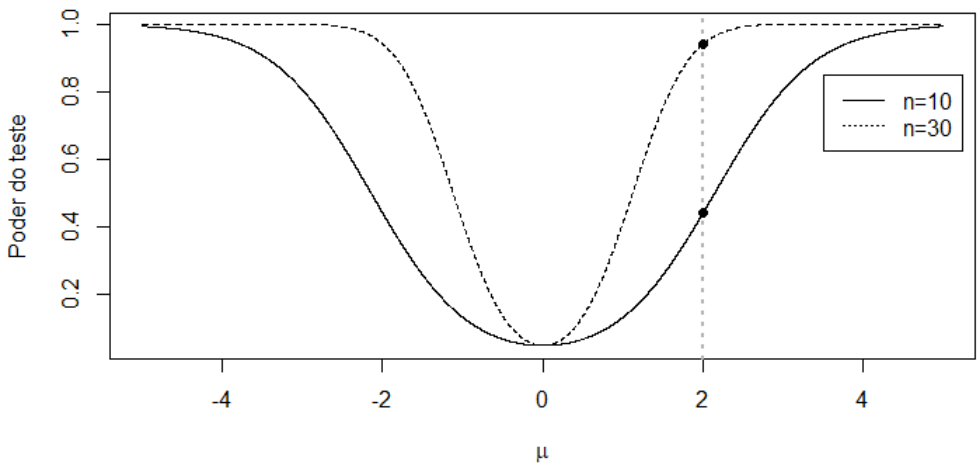


Figura 2.3: Função poder do teste para n=10 e n=30

Nas Subseções 2.2.1 e 2.2.2, são apresentadas as metodologias utilizadas neste trabalho para o cálculo do poder do teste para o teste t e o teste de Wilcoxon, respectivamente.

2.2.1 Poder do Teste t de Student

A partir da Equação 2.15, sabe-se que o poder do teste é a probabilidade de se rejeitar a hipótese nula para um dado μ_D. Considerando o teste t bilateral cujas hipóteses são definidas na Equação 2.10 e a estatística de teste definida na Equação 2.11, é possível calcular o poder do teste como:

$$\pi(\mu_D) = P\left(T > \left(\frac{v_1 - \mu_D}{\frac{s_d}{\sqrt{n}}}\right) \mid T \sim t_{(n-1)}\right) + P\left(T < \left(\frac{v_2 - \mu_D}{\frac{s_d}{\sqrt{n}}}\right) \mid T \sim t_{(n-1)}\right) \quad (2.16)$$

onde v_1 e v_2 são os valores limites da região crítica definida em (2.5). É possível simplificar essa equação quando a hipótese nula for definida como $\mu_D = 0$, já que assim, v_1 e v_2 serão simétricos ($v_1 = - \mid v_2 \mid$. Então, quando $H_0 : \mu_D = 0$, o poder do teste pode ser calculado como:

$$\pi(\mu_D) = 2 \times P\left(T > \left(\frac{v_1 - \mu_D}{\frac{s_d}{\sqrt{n}}}\right) \mid T \sim t_{(n-1)}\right) \quad (2.17)$$

2.2.2 Poder do Teste de Wilcoxon

Para o cálculo do poder do teste para o teste t, foi possível definir a equação pois a distribuição de t é conhecida. Porém, quando está sendo utilizado o teste de Wilcoxon, não existe informação sobre a distribuição da população. Uma vez que se trata de um teste não paramétrico, não foi necessário ser testada nenhuma pressuposição sobre essa distribuição. Nesse caso, o poder do teste é obtido através de simulação [4].

Neste estudo, o cálculo do poder do teste de Wilcoxon para a comparação das variáveis X e Y será feito com base na simulação de amostras de distribuições normais com média e desvio padrão definidos como as estimativas observadas: média e desvio padrão da amostra de X (μ_x e s_x) e de Y (μ_y e s_y). São gerados 1000 pares de amostras de mesmo tamanho da amostra que está sendo utilizada no teste. Assim, o poder do teste de Wilcoxon, considerando que a real diferença é igual a diferença amostral observada, é a proporção de vezes em que o teste de Wilcoxon rejeitou a hipótese nula para os 1000 pares gerados. Esse procedimento é descrito no Algoritmo 1.

Algoritmo 1 Simulação para obter o poder do teste de Wilcoxon

Entrada: μ_x, s_x, μ_y, s_y, n, α

Saída: $\pi(\mu_d)$

início

  $soma = 0$

  repita

    $amostra_x$ = amostra gerada de tamanho n de uma $N(\mu_x,s_x)$

    $amostra_y$ = amostra gerada de tamanho n de uma $N(\mu_y,s_y)$

    p-valor = p-valor (teste de Wilcoxon) na comparação das $amostra_x$ e $amostra_y$

    if *p-valor*$< \alpha$ **then**

      $soma = soma + 1$

  até *1000 vezes*;

  $\pi(\mu_d) = soma/1000$

fim

2.3 Tamanho do Efeito

Foi visto, até aqui, que os testes de hipóteses possuem a análise limitada à identificação da presença de uma diferença significativa entre os grupos. O p-valor informa somente se a diferença é estatisticamente significativa, ou seja, se a diferença observada é superior àquela que se esperaria encontrar por mero acaso ou por particularidade da amostra coletada. Nada foi dito sobre a dimensão ou magnitude dessa diferença. Para tanto, utiliza-se o **Tamanho do Efeito**.

O tamanho do efeito pode ser definido como o grau em que o fenômeno está presente na população, isto é, a diferença efetiva na população [5]. Assim, quanto maior for o tamanho do efeito, maior será a manifestação do fenômeno (diferença) na população.

Existem diversas medidas de tamanho do efeito e, segundo Joseph Edward [6], o uso dessas métricas é uma alternativa ao conceito de significância estatística, tratando de noções de significância prática específica [17], por exemplo, a significância clínica [15] e a significância educacional [27].

Além disso, outra motivação para o cálculo de uma medida do tamanho do efeito é que o p-valor é afetado por diversas características do estudo, sendo o tamanho da amostra o mais determinante [25], enquanto o tamanho do efeito não é afetado por essa característica. Assim, é mais provável obter um p-valor significativo com tamanhos grandes de amostras e, inversamente, em amostras pequenas, o p-valor pode não ser significativo [23]. Por isso,

cada vez mais a apresentação de uma medida do tamanho do efeito vem sendo estimulada. E, em alguns casos, até exigida pelas publicações da área cientifica.

O objetivo deste trabalho não inclui discutir sobre as diversas medidas de tamanho do efeito que existem. Por isso, serão apresentadas apenas as medidas aqui utilizadas para os testes t e de Wilcoxon, ambos para amostras pareadas. Para o Teste t será utilizado o ***d de Cohen*** que é calculado da seguinte maneira [5].

$$d'_{cohen} = \left| \frac{\overline{D}}{s_D} \right| \tag{2.18}$$

Pela Equação 2.11 (definição da estatística de teste do teste t), é possível reescrever o *d de Cohen* como:

$$d'_{cohen} = \left| \frac{\frac{\overline{D}}{\sqrt{n}}}{\frac{s_D}{\sqrt{n}}} \right| = \left| \frac{t}{\sqrt{n}} \right|, \tag{2.19}$$

onde t é a estatística de teste e n é a quantidade de pares em comparação (tamanho da amostra).

Não há consenso na literatura em relação ao que pode ser considerado tamanho do efeito grande ou pequeno. Cohen [5] propôs alguns pontos de corte para o *d de Cohen*, como descrito na Tabela 2.4, e estes serão utilizados neste trabalho.

Tabela 2.4: Valores para a interpretação da medida *d de Cohen* de tamanho do efeito

Insignificante	Pequeno	Médio	Grande	Muito grande
$< 0{,}19$	$0{,}2 - 0{,}49$	$0{,}5 - 0{,}79$	$0{,}8 - 1{,}29$	$> 1{,}3$

A maioria das medidas de tamanho do efeito assume que os dados têm uma distribuição normal. Quando testes não-paramétricos são realizados, por exemplo o teste de Wilcoxon, a significância pode ser calculada através da aproximação das distribuições das estatísticas de teste para a distribuição Z (Normal Padrão) quando os tamanhos das amostras não são muito pequenos. Pacotes estatísticos do R [22] e outros softwares já fornecem essa aproximação [12]. E através desse valor z é calculado um tamanho do efeito r, proposto por Cohen, para o Teste não-paramétrico de Wilcoxon para amostras pareadas[5] da seguinte forma:

$$r = \frac{z}{\sqrt{N}}, \tag{2.20}$$

onde N é a soma das amostras, ou seja, duas vezes a quantidade de pares. E para a classificação de r, Cohen propôs [5] os pontos de corte apresentados na Tabela 2.5.

Tabela 2.5: Valores para a interpretação da medida r de tamanho do efeito

Insignificante	Pequeno	Médio	Grande
$< 0{,}09$	$0{,}1 - 0{,}29$	$0{,}3 - 0{,}49$	$> 0{,}5$

Relatar o tamanho do efeito permite ao pesquisador julgar a magnitude das diferenças presentes entre os grupos, o que pode aumentar a capacidade de comparação entre os resultados de pesquisas recentes com pesquisas anteriores e julgar o significado prático dos resultados obtidos.

Capítulo 3

P-valor<0,05 é Suficiente? Dois Estudos de Caso

Este capítulo é composto de duas seções que visam, através de exemplos de aplicações reais, mostrar que a busca pelo p-valor menor do que 0,05 não é suficiente para uma análise estatística e pode ser perigosa para a ciência quando são ignorados possíveis resultados interessantes ou são valorizados resultados sem grande importância ou relevância real.

Os resultados dos exemplos abordados ilustram casos onde as medidas p-valor e tamanho do efeito levam a conclusões discordantes e como o cálculo do poder do teste pode ajudar a explicar tal discordância. A Seção 3.1 apresenta o primeiro exemplo de aplicação no qual é utilizado o Teste t. Na Seção 3.2, o Teste de Wilcoxon é utilizado no segundo exemplo. Em ambos os casos, deseja-se comparar os resultados dos classificadores k-NN com k=1 (1-NN) e k-NN com k=3 (3-NN), aplicados a duas bases de dados obtidas no repositório da UCI [8], utilizando o método de validação cruzada com 10 partições (para o caso utilizando o teste t) e 30 partições (para o caso com o Teste de Wilcoxon). Cabe ressaltar que as amostras utilizadas nesses exemplos têm tamanhos diferentes para serem abordados dois tipos de discordância entre o p-valor e o tamanho do efeito. Nos dois casos, é discutido como seriam as conclusões tomadas com base em três análises: apenas no p-valor, com o p-valor e o tamanho do efeito e com as três medidas combinadas – p-valor, tamanho do efeito e poder do teste.

3.1 Caso com Teste t de Student

Nesta seção, será explorado um caso em que o teste de hipótese não indica significância estatística, porém com um tamanho do efeito médio. Para tanto, será abordado um exem-

plo de aplicação do Teste t de Student para observações pareadas. Deseja-se verificar se os resultados dos classificadores 1NN e 3-NN são diferentes utilizando-se as amostras de resultados observados a partir da base de dados *Mammographic Mass*, obtida no repositório da UCI.

Para a realização desse experimento, foi utilizada a Ferramenta Weka [9], onde estão implementados os classificadores adotados. Foi utilizado o método de validação cruzada com 10 partições (*10-fold-cross-validation*). Em cada partição, foram aplicados o 1-NN e o 3-NN, ou seja, foram obtidas ao todo 10 pares de acurácias. Por esse motivo deve-se utilizar testes de hipóteses para dados pareados.

As acurácias obtidas estão apresentadas na Tabela 3.1. Esses valores são apresentados com apenas duas casas decimais por arredondamento. Porém as médias foram calculadas a partir dos valores não arredondados.

Tabela 3.1: Acurácias obtidas por partição para cada classificador

	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10
1-NN	77,32	71,88	72,92	73,96	71,88	70,83	78,12	72,92	81,25	81,25
3-NN	77,32	75,00	75,00	78,12	77,08	78,12	78,12	75,00	80,21	79,17

3.1.1 Conclusão com Base no P-valor

Sejam X a amostra de acurácias obtidas através da aplicação do 1-NN e Y a amostra de acurácias obtidas através da aplicação do 3-NN. Considere então (x_1,y_1), (x_2,y_2), ..., (x_n,y_n), onde $n = 10$, os dez pares de acurácias observadas e $d_i = x_i - y_i$ as diferenças entre esses pares de acurácias, onde $1 \leq i \leq n$. Inicialmente, é preciso ser verificada a suposição de normalidade das distribuições de X e Y para que o Teste t para dados pareados possa ser aplicado. Para isso, utiliza-se o teste de Kolmogorov-Smirnov para cada amostra, cujas hipóteses para cada classificador são as seguintes.

$$\begin{cases} H_0: & \textit{As acurácias têm distribuição normal} \\ H_1: & \textit{As acurácias não têm distribuição normal} \end{cases} \tag{3.1}$$

Sabe-se que as acurácias têm valores limitados entre 0 e 1, logo não têm distribuição normal. Mesmo com essa certeza, o teste de normalidade será realizado, pois a distribuição pode ser próxima da normal (simétrica com mesma média e mediana) e, com isso, não apresentar perdas ao aplicar o teste t. Além dos testes de normalidade, existem

outros maneiras de verificar se os dados seguem uma distribuição normal. Essa análise poderia ser feita, por exemplo, através da análise do gráfico da distribuição acumulada dos dados, comparando-o à distribuição acumulada da normal. Porém, devido ao grande número de comparações realizadas nos capítulos seguintes, será realizado apenas o teste de Kolmogorov para verificar a normalidade dos dados.

A aplicação do teste de Kolmogorov-Smirnov confirma a hipótese de normalidade da distribuição das acurácias do 1-NN e do 3-NN ao nível de significância de 5%, já que os p-valores obtidos são 0,68 e 0,83, respectivamente. Sendo assim, a hipótese nula (que representa a normalidade dos dados) não é rejeitada para ambos classificadores. Dessa forma, é possível realizar o Teste t para duas amostras pareadas e com variâncias desconhecidas.

Para definir as hipóteses a serem testadas, é necessário decidir se será um teste unilateral ou bilateral. Como não há o desejo de verificar se um classificador tem resultado melhor que o outro, porém apenas se eles têm resultados diferentes, será aplicado o teste bilateral. Dessa forma, a hipótese nula é de que a média populacional das acurácias obtidas pelo classificador 1-NN é igual à média populacional das acurácias obtidas pelo classificador 3-NN na base em análise. E a hipótese alternativa é de que a média das acurácias populacionais obtidas por esses classificadores são diferentes. Sendo assim, a hipótese nula pode considerar que que não existe diferença entre as médias das acurácias populacionais obtidas pelos classificadores 1-NN e 3-NN na base em análise, e a hipótese alternativa que essa diferença populacional é diferente de zero. Ou seja:

$$\begin{cases} H_0: & \mu_D = 0, \\ H_1: & \mu_D \neq 0, \end{cases} \tag{3.2}$$

onde μ_D é a diferença entre as médias das acurácias populacionais dos classificadores.

A diferença entre as médias observadas das acurácias amostrais é $\overline{d} = -2{,}08$. Deseja-se verificar se a diferença populacional é estatisticamente significante (se é diferente de 0), assim como foi observada na amostra, ou se essa diferença observada é particularidade daquelas amostras e não é uma característica da população.

Como visto na Equação 2.11, a estatística de teste t é calculada da seguinte maneira:

$$t = \frac{\overline{d} - 0}{\frac{s_d}{\sqrt{n}}} = \frac{-2{,}08}{\frac{2{,}95}{\sqrt{10}}} = -2{,}24, \tag{3.3}$$

já que

$$s_d = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \overline{d})^2}{n-1}} = \sqrt{\frac{78{,}12}{10-1}} = 2{,}95. \tag{3.4}$$

Calculando o p-valor, como visto em 2.12, obtém-se $2 \times P(t \geq |t||H_0) = 2 \times 0{,}0261 = 0{,}0522$. Como o p-valor é a probabilidade de se obter valores mais extremos do que o observado quando H_0 é verdadeira, então, basta somar as probabilidades de valores mais extremos do que a estatística de teste calculada para uma distribuição t com 9 graus de liberdade.

Logo, como o p-valor= 0,052 é maior que o nível de significância $\alpha = 0{,}05$, o teste t para amostras pareadas não obteve evidências para rejeitar a hipótese nula que é a hipótese conservadora de que a média das acurácias amostrais obtidas pelos classificadores 1-NN e 3-NN são iguais. Sendo assim, ao nível de significância de 5% não é possível afirmar que os classificadores 1-NN e 3-NN têm médias de acurácias populacionais diferentes.

3.1.2 Avaliando o Tamanho do Efeito

Uma medida de tamanho do efeito pode complementar a conclusão tirada com base no p-valor. Para o caso em questão, a medida d de Cohen, conforme definida na Seção 2.3, é calculada da seguinte forma $d'cohen = \frac{|t|}{\sqrt{n}} = \frac{2{,}24}{\sqrt{10}} = 0{,}71$, onde t é a estatística de teste e n é a quantidade de pares em comparação (tamanho da amostra).

Com o tamanho do efeito calculado, verifica-se na Tabela 2.4 que $d'cohen = 0{,}71$ representa um tamanho do efeito médio. Ou seja, o teste de hipóteses não indica diferença estatisticamente significativa entre as médias das acurácias dos classificadores 1-NN e 3-NN, porém a diferença entre essas médias tem tamanho do efeito médio indicando que a magnitude dessa diferença pode ser importante. Sendo assim, o p-valor e o tamanho do efeito indicam resultados diferentes. Nesse caso, o cálculo do poder do teste, como será visto na subseção seguinte, acrescenta informações necessárias para uma tomada de decisão mais fundamentada.

3.1.3 Avaliando o Poder do Teste

Como visto na Seção 2.2, o poder do teste representa a probabilidade do teste rejeitar H_0, ou seja, é a probabilidade do teste afirmar que os classificadores A e B são diferentes quando realmente são. Na realidade, o poder do teste é uma função pois, sendo $H_0 : \mu_D = 0$ falsa, não se sabe o valor verdadeiro para μ_D, sabe-se apenas que ele é diferente de zero,

logo o poder do teste é calculado para todos os valores possíveis de μ_D.

Na Figura 3.1, a curva referente ao valor $n = 10$ representa a função do poder do teste para os possíveis valores reais de diferença populacional entre as acurácias médias dos classificadores em análise, ou seja, para todos os possíveis valores de μ_D. Além da curva do exemplo em desenvolvimento, quando $n = 10$, também são apresentadas as curvas de como seria o poder do teste se as amostras fossem maiores, por exemplo, com $n = 25$ e $n = 50$. Nesses dois casos haveria 25 e 50 pares de acurácias, respectivamente. Na figura, o ponto marcado na curva com $n = 10$ refere-se ao valor do poder do teste se for considerado que a diferença real é $-2{,}08$, que foi a diferença amostral observada.

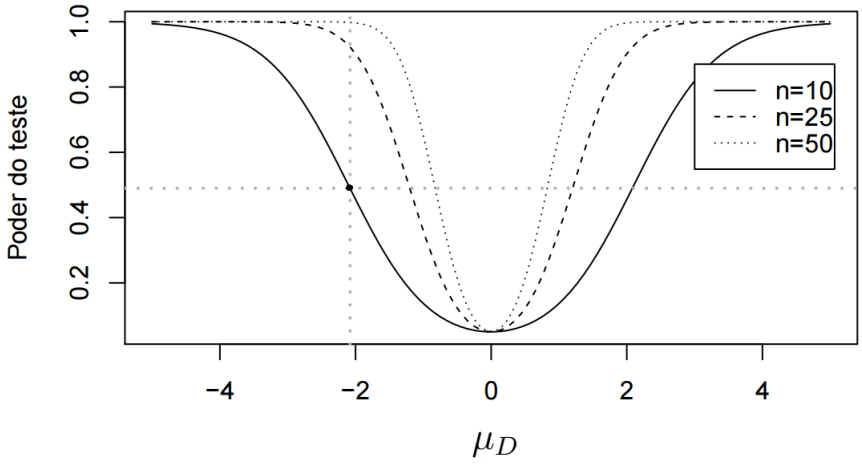


Figura 3.1: Curvas do poder do teste com diversos valores de n

Considerando que o valor real da diferença entre os classificadores é igual a diferença observada ($-2{,}08$), o cálculo do poder do teste, com base na Fórmula 2.15, resulta em 0,49 (ponto marcado na Figura 3.1). Ou seja, a probabilidade do teste afirmar que os classificadores são diferentes se a diferença real for $-2{,}08$ é de apenas 49%.

Além disso, na Figura 3.1, é possível perceber que para um mesmo valor de μ_D, o poder do teste aumenta conforme aumenta o tamanho da amostra (n). Como já foi dito, para $\mu_D = -2{,}08$ (diferença amostral) e $n = 10$, o poder do teste é 49% e, aumentando n para $n = 25$ e $n = 50$, o poder do teste seria 92% e 98%, respectivamente.

Conforme visto anteriormente, o teste t não obteve evidências para rejeitar a hipótese nula, já que p-valor=0,052, e assim não foi possível afirmar que as médias das acurácias populacionais dos classificadores 1-NN e 3-NN são diferentes ao nível de significância de 5%. Porém, o tamanho do efeito médio ($d'cohen = 0{,}71$) para a diferença entre as acurácias populacionais indica que a magnitude dessa diferença pode ser um resultado

importante. Ou seja, as duas medidas podem levar a conclusões distintas no que diz respeito aos classificadores terem ou não resultados diferentes.

O cálculo do poder do teste permite compreender o possível motivo das medidas anteriores terem resultados distintos. O poder do teste de apenas 49%, se a diferença real for $-2{,}08$, indica que o teste tem baixa probabilidade de afirmar que os classificadores são diferentes quando essa diferença realmente existe. Ou seja, o teste t foi aplicado mesmo sendo pouco poderoso para essa amostra. Portanto, esse exemplo ilustra não somente a importância do cálculo das três medidas, mas também o risco de tomar uma decisão equivocada se for utilizado apenas o p-valor. Além disso, ilustra o equívoco que é aplicar um teste de hipótese sem conhecer o poder desse teste para a amostra em análise.

Em casos como este, quando o teste realizado é pouco poderoso, o pesquisador tem a oportunidade de tentar aumentar o poder do teste buscando mais elementos da amostra, conforme visto na Figura 3.1. Ou seja, um pesquisador não deve desistir do seu estudo por não ter encontrado significância estatística por meio de um teste com poder baixo, pois pode representar a perda de um resultado importante.

3.2 Caso com Teste de Wilcoxon

Nesta seção, será explorado um caso em que o teste de hipótese indica significância estatística, porém com um tamanho do efeito pequeno. Para tanto, será abordado um exemplo de aplicação do Teste de Wilcoxon para observações pareadas. Deseja-se verificar se os resultados dos mesmos classificadores 1-NN e 3-NN são diferentes para uma outra base de dados, a base *Wholesale3*, também obtida no repositório da UCI. Outra diferença em relação ao exemplo anterior é que, no lugar do método de validação cruzada com 10 partições, foram utilizadas 30 partições, ou seja, a amostra nesse exemplo tem tamanho 30.

As acurácias obtidas pelos classificadores 1-NN e 3-NN são apresentadas na Tabela 3.2, onde também são apresentadas suas diferenças e seus respectivos postos.

Tabela 3.2: Acurácias observadas e o respectivo posto de cada par para o exemplo do Teste de Wilcoxon

1-NN	**3-NN**	***d***	**Sinal**	**Posto**
80	86,67	-6,67	-	8,5
93,33	93,33	0		
93,33	86,67	6,66	+	3,5
93,33	93,33	0		
86,67	93,33	-6,66	-	3,5
93,33	100	-6,67	-	8,5
86,67	93,33	-6,66	-	3,5
93,33	93,33	0		
80	86,67	-6,67	-	8,5
86,67	93,33	-6,66	-	3,5
93,33	93,33	0		
100	100	0		
80	100	-20	-	17
86,67	93,33	-6,66	-	3,5
86,67	93,33	-6,66	-	3,5
86,67	86,67	0		
73,33	80	-6,67	-	8,5
93,33	93,33	0		
73,33	86,67	-13,33	-	13
86,67	86,67	0		
92,86	92,86	0		
78,57	92,86	-14,29	-	15
85,71	78,57	7,14	+	11
71,43	78,57	-7,142	-	12
100	85,71	14,29	+	15
85,71	85,71	0		
92,86	92,86	0		
85,71	85,71	0		
78,57	92,86	-14,29	-	15
85,71	85,71	0		

3.2.1 Conclusão com Base no P-valor

Sejam X a amostra de acurácias obtidas através da aplicação do 1-NN e Y a amostra de acurácias obtidas através da aplicação do 3-NN. Considere então, (x_1,y_1), (x_2,y_2), ..., (x_n,y_n), onde $n = 30$, os 30 pares de acurácias observadas e as diferenças $d_i = x_i - y_i$, onde $1 \leq i \leq n$. Será aplicado o teste bilateral de Wilcoxon para amostras pareadas considerando nível de significância de 5%, cujas hipóteses são apresentadas em (3.5), onde δ_D é a mediana da diferença entre as acurácias populacionais dos classificadores 1-NN e 3-NN na base em análise.

$$\begin{cases} H_0: & \delta_D = 0 \\ H_1: & \delta_D \neq 0 \end{cases} \tag{3.5}$$

Dessa forma, a hipótese nula é de que não existe diferença (a diferença é igual a zero) entre as medianas das acurácias populacionais dos classificadores 1-NN e 3-NN na base em análise, ou seja, as medianas populacionais das acurácias obtidas por esses classificadores são iguais. E a hipótese alternativa é de que existe diferença (a diferença não é zero) entre as medianas das acurácias populacionais dos classificadores 1-NN e 3-NN na base em análise, ou seja, as medianas das acurácias populacionais dos classificadores são diferentes.

Inicialmente, são atribuídos postos para os valores absolutos das diferenças observadas ordenados de forma crescente, como mostra a Tabela 3.2. O valor de n é recalculado, já que existem pares com diferença nula, então $n = 17$. Como a estatística W é a menor soma dos postos de mesmo sinal (soma dos postos positivos ou soma dos postos negativos), ela será a soma dos postos positivos, neste caso. Então, $W = 3{,}5 + 11 + 15 = 29{,}5$.

Neste caso, o p-valor para o teste de Wilcoxon é 0,025 e, por ser menor que o nível de significância $\alpha = 0{,}05$, indica que há evidências para rejeitar a hipótese nula. Portanto, é possível afirmar, ao nível de significância de 5%, que as medianas das acurácias populacionais dos classificadores são estatisticamente diferentes.

3.2.2 Avaliando o Tamanho do Efeito

Uma medida de tamanho do efeito pode complementar a conclusão tirada com base no p-valor. Para o caso em questão, a medida r, conforme definida na Seção 2.3, é calculada da seguinte forma $r = \frac{|z|}{\sqrt{N}} = \frac{2{,}24}{\sqrt{60}} = 0{,}28$, onde z é a estatística do teste de Wilcoxon com aproximação pela distribuição normal e N é a soma dos tamanhos das amostras, ou seja, duas vezes a quantidade de pares em comparação ($N = 2 \times n$).

Com o tamanho do efeito calculado, verifica-se na Tabela 2.5 que $r = 0{,}28$ representa um tamanho do efeito pequeno. Ou seja, o teste de hipóteses indica que a diferença entre as medianas das acurácias populacionais dos classificadores 1-NN e 3-NN é estatisticamente significativa, porém o tamanho do efeito classificado como pequeno indica que a magnitude dessa diferença pode não ser um resultado importante ou relevante para o pesquisador. Sendo assim, o p-valor e o tamanho do efeito indicam resultados diferentes. Nesse caso, o cálculo do poder do teste, como será visto na subseção seguinte, acrescenta informações necessárias para uma tomada de decisão mais fundamentada.

3.2.3 Avaliando o Poder do Teste

Para realizar o Teste de Wilcoxon não foi necessário fazer nenhum pressuposto sobre a distribuição da população, uma vez que se trata de um teste não paramétrico. Porém, para o cálculo do poder do teste, é necessário que a distribuição da diferença entre as acurácias populacionais seja conhecida, o que torna esse cálculo diferente do que foi para o Teste t, uma vez que ele será obtido através de simulação.

Foram simulados 1000 pares de amostras com distribuição normal e de tamanho 30, com os parâmetros média e desvio padrão respectivos das amostras X e Y. Para cada par de amostra, foi realizado o Teste de Wilcoxon e feita a proporção de quantos foram significativos entre o total. Foram obtidos os seguintes resultados: dos 1000 pares de amostras, 433 foram significativos para o Teste de Wilcoxon, portanto, pode-se dizer que o poder do teste é de aproximadamente 43% para uma diferença real igual à diferença amostral de $\overline{d} = -3{,}36$. Ou seja, a probabilidade do teste afirmar que os classificadores são distintos, quando essa diferença for $-3{,}36$, é de aproximadamente 43%.

Conforme visto anteriormente, o teste de Wilcoxon obteve evidências para rejeitar a hipótese nula, já que p-valor= 0,025. Assim é possível afirmar que as acurácias populacionais dos classificadores 1-NN e 3-NN são diferentes em localização, ou seja, têm medianas diferentes, ao nível de significância de 5%. Porém, o tamanho do efeito pequeno ($r = 0{,}28$) para a diferença entre as medianas das acurácias populacionais indica que a magnitude dessa diferença é pequena e possivelmente pode ser um resultado sem relevância. Ou seja, as duas medidas podem levar a conclusões distintas no que diz respeito aos classificadores terem ou não resultados diferentes.

O cálculo do poder do teste permite compreender o possível motivo das medidas anteriores terem resultados distintos. O poder do teste de apenas 43%, quando a diferença populacional for igual a diferença amostral, indica que o teste tem baixa probabilidade de

afirmar que os classificadores são diferentes quando essa diferença realmente existe. Ou seja, o teste de Wilcoxon foi aplicado mesmo sendo pouco poderoso para essa amostra. Portanto, assim como no exemplo do teste t, esse exemplo ilustra não somente a importância do cálculo das três medidas, mas também o risco de tomar uma decisão equivocada se for utilizado apenas o p-valor. Além disso, ilustra o equívoco que é aplicar um teste de hipótese sem conhecer o poder desse teste para a amostra em análise.

Capítulo 4

Análise Ampliada

No Capítulo 3, foram abordados dois casos de comparação de classificadores, onde as decisões apenas com base no p-valor poderiam levar a conclusões equivocadas. Além disso, foi discutido como a análise do tamanho do efeito e do poder do teste colaboram para uma decisão mais fundamentada. Casos como esses, onde há discordância entre o p-valor e o tamanho do efeito, serão chamados, a partir deste ponto, de **casos especiais**.

Neste capítulo, as análises vistas anteriormente serão ampliadas. Para isso, será realizado um estudo empírico com o objetivo de mostrar o comportamento das três medidas (p-valor, tamanho do efeito e poder do teste) em diversas situações quando realizada a comparação de classificadores. Evidenciando, assim, que a quantidade de vezes em que os casos especiais ocorrem não deve ser ignorada. E, para fortalecer esse resultado, também são utilizados dados simulados que, quando comparados aos resultados dos dados reais, concordam com a conclusão encontrada.

O estudo empírico foi realizado utilizando-se 50 bases de dados do repositório da UCI. Considerando cada base de dados, foram aplicados os seguintes algoritmos de classificação: Random Forest com 100 árvores (RF100) e com 300 árvores (RF300), SVM, k-NN com k=1 (1-NN) e k=3 (3-NN), e Naive Bayes (NB), totalizando 6 classificadores e, consequentemente, 15 pares de classificadores que são enumerados a seguir:

1. RF100 e RF300
2. RF100 e SVM
3. RF100 e 1-NN
4. RF100 e 3-NN
5. RF100 e NB
6. RF300 e SVM

7. RF300 e 1-NN
8. RF300 e 3-NN
9. RF300 e NB
10. SVM e 1-NN
11. SVM e 3-NN
12. SVM e NB
13. 1-NN e 3-NN
14. 1-NN e NB
15. 3-NN e NB

Para tornar o experimento mais amplo, variou-se também a quantidade de partições do método de validação cruzada na avaliação dos classificadores. Para cada base, cada classificador foi avaliado utilizando 10, 20 e 30 partições. Os experimentos foram realizados utilizando-se a Ferramenta Weka.

Com as acurácias obtidas, foram aplicados os testes t de Student e de Wilcoxon, ambos para amostras pareadas. Cada teste foi aplicado em cada possível combinação de uma base de dados, um par de classificadores e um tamanho de amostra. Logo, o número total de cada teste que se deseja realizar é $15 \times 50 \times 3 = 2250$. Além dos resultados dos testes aplicados, medidos através dos p-valores, também são calculadas as medidas de tamanho do efeito e poder do teste. Os experimentos foram realizados utilizando-se o Software R.

Como estudo complementar, será realizada a simulação de valores de acurácias para a confirmação dos resultados em um número maior de casos. Na simulação, os valores de acurácia são gerados como amostras de distribuições normais bivariadas, onde são utilizados diversos valores de: parâmetros, diferenças entre as acurácias e tamanhos de amostras. Os mesmos testes utilizados para comparar as acurácias reais, serão aplicados para comparar os valores de acurácia simulados que buscam representar uma maior diversidade de valores.

As bases de dados utilizadas para a realização do estudo empírico são apresentadas na Seção 4.1. Em seguida, as Seções 4.2 e 4.3 apresentam os resultados obtidos considerando o teste t de Student e o teste de Wilcoxon, respectivamente.

4.1 Bases de Dados

Para a realização do estudo empírico, foram escolhidas 50 bases de dados da UCI que estão apresentadas na Tabela 4.1 com suas respectivas quantidades de instâncias, de atributos

e de classes. Uma vez que será utilizado o método de validação cruzada com até 30 partições, foram selecionadas bases com no mínimo 300 instâncias. Dessa forma, evitando a utilização de bases de teste muito pequenas.

Tabela 4.1: Bases utilizadas, quantidade de instâncias, atributos e classes em cada base

Número	Base	Instâncias	Atributos	Classes
1	Heart disease processed cleveland	303	14	5
2	Haberman	306	4	2
3	Ecoli	336	8	8
4	Leaf	340	16	30
5	Bupa	345	7	2
6	Ionosphere	351	34	2
7	Dermatology	366	35	6
8	Wholesale3	440	8	2
9	Arrhythmia	452	263	13
10	Thoracic Surgery	470	17	2
11	Dresses sales	500	13	2
12	Led7digit	500	8	10
13	Housing	506	14	2
14	Climate simulation craches	540	21	2
15	Monks1	556	7	2
16	Wdbc	569	31	2
17	Indian liver patient	583	11	2
18	Balance scale	625	5	3
19	Credit approval	690	16	2
20	Breast cancer wisconsin	699	11	2
21	Blood transfusion service	748	5	2
22	Energy efficiency y1	768	9	37
23	Pima indians diabetes	768	9	2
24	Anneling	898	39	5
25	Tictactoe	958	10	2
26	Mammographic mass	961	6	2
27	Cnae-9	1080	857	9
28	Messidor features	1151	20	2
29	Data banknote authentication	1372	5	2
30	Flare	1389	13	6
31	Contraceptive method choice	1473	10	3
32	Yeast	1484	9	10

Tabela 4.1: Bases utilizadas, quantidade de instâncias, atributos e classes em cada base

Número	Base	Instâncias	Atributos	Classes
33	Car evaluation	1728	7	4
34	Mfeat fourier	2000	77	10
35	Cardiotocographt 3class	2126	36	3
36	Ozone eighthr	2534	73	2
37	Seismic bumps	2584	19	2
38	Abalone 3class	4177	9	3
39	Bank marketing	4521	17	2
40	Spambase	4601	58	2
41	Wilt	4839	6	2
42	Banana	5300	3	2
43	Page blocks	5473	11	5
44	Phoneme	5404	6	2
45	Turkiye student evaluation	5820	33	5
46	First order theorem	6118	52	6
47	Artificial characters	10218	8	10
48	Pendigits	10992	17	10
49	Nursery	12960	9	5
50	Egg eye state	14980	15	2

4.2 Análise com o Teste t de Student

Nesta seção, serão apresentados os resultados encontrados pelas medidas p-valor, tamanho do efeito e poder do teste na comparação dos 15 pares de classificadores, utilizando o teste t considerando as 50 bases de dados.

Como o teste t de Student para amostras pareadas é um teste paramétrico, as amostras das acurácias devem ser provenientes de populações com distribuição normal. Então é utilizado o teste de Kolmogorov Smirnov para verificar a hipótese de normalidade em cada amostra de acurácias.

Além da verificação da suposição de normalidade, para este estudo também são feitas outras duas exigências para que seja mantida a comparação de um par de classificadores em determinada base de dados. Portanto, é necessário que se cumpram três critérios com as amostras que serão analisadas:

- deve ser verificada a suposição de normalidade para as duas amostras;
- a variância da diferença das amostras deve ser diferente de zero pois sem variabilidade das acurácias, não há necessidade de realizar o teste estatístico;
- os itens anteriores devem ser verificados nos três pares de diferentes tamanhos de amostras (10, 20 e 30) para que seja mantida a comparação do par de classificadores em determinada base de dados, ou seja, o teste será aplicado para os três tamanhos de amostras ou para nenhum. Esse critério é necessário, pois será realizada uma análise comparando os resultados obtidos com os diferentes tamanhos de amostras.

O número total de testes que se deseja realizar é $15 \times 50 \times 3 = 2250$. Porém, esse número foi reduzido para 1509 testes devido às amostras que não atenderam aos critérios definidos anteriormente.

A Tabela 4.2 apresenta, para cada combinação de uma base de dados e de um par de classificadores, um indicador de aplicação do teste t considerando os três tamanhos de amostras. Cada coluna da tabela é referente a um par de classificadores (a numeração dos pares foi apresentada no início deste capítulo) e cada linha é referente a uma base de dados (a numeração das bases foi apresentada na Tabela 4.1). No corpo da tabela, o valor 1 indica que será aplicado o teste t para comparar o par de classificadores na base de dados com amostras de tamanho 10, 20 e 30. Já o valor 0 indica que o teste não será realizado para nenhum tamanho de amostra, pois em algum caso foi violada a suposição de normalidade ou existência da variabilidade das acurácias.

Na utilização do teste t de Student bilateral para amostras pareadas, a hipótese nula é que a diferença média entre as acurácias populacionais (μ_D) dos classificadores em análise para a base em questão é igual a zero. Então, a hipótese alternativa é que essa diferença é diferente de zero. Logo, as hipóteses do teste são definidas como na Equação 3.2 (definida no Capítulo 2).

Os resultados a seguir são apresentados em três subseções. Na primeira, é analisado como o p-valor se comporta com a alteração do tamanho da amostra. Na segunda, é comparado o comportamento do p-valor e do tamanho do efeito e na terceira parte, é acrescentado o poder do teste nas análises.

4.2.1 Análise do Comportamento do P-valor

Uma das críticas ao p-valor é que ele é influenciado por diversas características do estudo, sendo o tamanho da amostra uma delas [25]. Então, nesta subseção, será analisado o

Tabela 4.2: Indicadores de aplicação do teste t para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 10, 20 e 30

Bases	Pares de Classificadores														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1	0	1	1	1	0	1	1	1	0	0	0	1	1	1
2	1	0	1	1	1	0	1	1	1	0	0	0	1	1	1
3	0	1	1	1	1	0	0	0	0	1	1	1	1	1	1
4	1	1	1	0	1	1	1	0	1	1	0	1	0	1	0
5	1	0	1	1	1	0	1	1	1	0	0	0	1	1	1
6	0	1	0	1	1	0	0	0	0	0	1	1	0	0	1
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
10	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
11	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
12	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
15	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	1	0	1	1	1	0	1	1	1	0	0	0	1	1	1
18	1	1	1	1	0	1	1	1	0	1	1	0	1	0	0
19	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
20	1	0	1	1	1	0	1	1	1	0	0	0	1	1	1
21	1	0	1	1	1	0	1	1	1	0	0	0	1	1	1
22	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
23	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
25	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
26	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
27	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
28	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
29	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
30	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
31	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
32	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
33	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
34	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36	1	0	1	1	1	0	1	1	1	0	0	0	1	1	1
37	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
38	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
39	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
40	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
41	1	0	1	1	1	0	1	1	1	0	0	0	1	1	1
42	1	0	1	1	1	0	1	1	1	0	0	0	1	1	1
43	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
44	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
45	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
46	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
47	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
48	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
49	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
50	1	0	1	1	0	0	1	1	0	0	0	0	1	0	0

impacto da variação dos p-valores para os diferentes tamanhos de amostras na avaliação de classificadores.

Deseja-se verificar que o p-valor é sensível em relação ao tamanho da amostra. Os resultados da comparação das acurácias obtidas entre os classificadores SVM e 1-NN na Base 31, por exemplo, ilustram essa sensibilidade. Com a amostra de tamanho 10 (método de validação cruzada com 10 partições) o p-valor na comparação desses classificadores foi 0,10. Já com amostras de tamanho 20 e 30, o p-valor foi 0,004 e 0,002, respectivamente. Ou seja, com 10 partições, o teste não teve significância para afirmar que os classificadores foram diferentes, mas aumentando a amostra para 20 ou 30, foi possível obter a significância estatística para afirmar diferença entre a médias das acurácias populacionais desses classificadores na base em questão.

A Tabela 4.3 mostra que essa sensibilidade (diminuição do p-valor com o aumento do tamanho da amostra) foi bastante frequente nos casos analisados. Nessa tabela, para cada par de classificadores, é apresentada a porcentagem de testes realizados em que o p-valor diminuiu com o aumento da amostra de 10 para 20, de 10 para 30 e de 20 para 30. Essa porcentagem representa a sensibilidade em questão, que pode ser observada, por exemplo, em aproximadamente 93% das bases analisadas na comparação entre SVM e 1-NN com o aumento do tamanho da amostra de 10 para 20, ou seja, 93% dos testes tiveram redução do p-valor com esse aumento da amostra. Todos os p-valores obtidos foram colocados no Apêndice A para não sobrecarregar o texto, nas Tabelas A.1, A.2 e A.3, para as amostras de tamanhos 10, 20 e 30, respectivamente.

Tabela 4.3: Quantidade porcentual de testes aplicados nas quais o p-valor reduziu com o aumento do tamanho da amostra (de 10 para 20, de 10 para 30 e de 20 para 30) na aplicação do teste t em cada par de classificadores

Classificadores	Aumento no tamanho da amostra		
	10 para 20	10 para 30	20 para 30
RF100 e RF300	54,29	57,14	37,14
RF100 e SVM	59,26	70,37	59,26
RF100 e 1-NN	69,44	72,22	58,33
RF100 e 3-NN	66,67	72,22	58,33
RF100 e NB	72,22	75,00	77,78
RF300 e SVM	65,38	76,92	84,62
RF300 e 1-NN	75,00	80,56	80,56
RF300 e 3-NN	71,43	82,86	80,00
RF300 e NB	79,41	76,47	82,35
SVM e 1-NN	92,86	78,57	60,71
SVM e 3-NN	71,43	75,00	71,43
SVM e NB	82,76	79,31	65,52
1-NN e 3-NN	54,05	48,65	48,65
1-NN e NB	73,68	76,32	63,16
3-NN e NB	69,23	79,49	74,36
Média	70,20	73,20	66,60

Os valores apresentados na Tabela 4.3 mostram que, na comparação de alguns pares de classificadores, o p-valor diminuiu com uma frequência maior que em outros pares. Porém, no geral, conforme visto na linha da média total (sem separação por pares de classificadores), o p-valor realmente tende a diminuir com o aumento do tamanho da amostra. De acordo com os valores obtidos, essa diminuição não é uma regra, já que, coletando outra amostra de tamanho maior, as novas estimativas observadas (como média e variância amostrais) podem ser diferentes e, consequentemente, também influenciar o novo p-valor observado.

Os resultados vistos até aqui, mostram que o p-valor realmente é sensível em relação ao tamanho da amostra. Agora, deseja-se analisar qual o porcentual de vezes em que a queda do p-valor alterou a conclusão tomada no teste de hipótese, ou seja, o resultado passou de não significativo para significativo.

As porcentagens de testes onde houve mudança na conclusão com o aumento do tama-

nho da amostra de 10 para 20, de 10 para 30 e de 20 para 30 são 8,7%, 11,96% e 10,81%, respectivamente. À primeira vista, esses valores podem parecer baixos se comparados às porcentagens de queda observadas, porém, ilustram que essa sensibilidade pode afetar um número significativo de resultados de pesquisas científicas.

Conforme visto nos resultados, a diminuição do p-valor com o aumento do tamanho da amostra é uma crítica ao p-valor. Outra crítica é que o p-valor não mede a importância (o peso) do resultado, apenas indica se o resultado tem significância estatística. Segundo [1], a utilização do p-valor como única medida pode levar o pesquisador a confundir a significância estatística com a significância prática ou científica. Um resultado com significância estatística não garante que a diferença seja grande ou importante, e isso deve estar claro na conclusão do teste estatístico.

Uma forma de complementar a análise é considerar uma medida de tamanho do efeito, já que tomar decisões com base em p-valores isolados pode prejudicar pesquisas científicas. A seguir são discutidos os resultados dos tamanhos do efeito para cada teste t aplicado.

4.2.2 Avaliando o Tamanho do Efeito

Nesta subseção, é acrescentada ao estudo a medida do tamanho do efeito $d'cohen$, que é calculada como visto na Equação 2.19, no Capítulo 2. Essa medida é calculada para todos os testes realizados, a fim de medir o grau da diferença entre as acurácias dos classificadores, na base de dados em questão. Assim, quanto maior for o tamanho do efeito, maior será a manifestação do fenômeno (diferença) na população.

Um dos motivos para que seja apresentada uma medida do tamanho do efeito acompanhando o resultado de um teste de hipóteses é que, ao contrário do p-valor, o tamanho do efeito é independente do tamanho da amostra. Essa questão também foi discutida em trabalhos de outras áreas, como em [10], [12], [26], entre outros.

Enquanto o p-valor indica se as amostras trazem evidências de que a diferença é realmente observada na população, o tamanho do efeito indica o grau em que essa diferença é observada na população. Portanto, o p-valor e o tamanho do efeito avaliam informações diferentes e podem indicar caminhos oposto: p-valor indicar significância estatística e tamanho do efeito pequeno ou insignificante, ou, p-valor não indicar significância estatística e tamanho do efeito médio ou maior. São os chamados casos especiais, discutidos no Capítulo 3.

Os valores calculados para as medidas de tamanho do efeito são apresentados no

Apêndice A, Tabelas A.4, A.5 e A.6, para as amostras de tamanhos 10, 20 e 30, respectivamente. Esses valores são classificados de acordo com os pontos de corte propostos por Cohen [5], que foram descritos na Tabela 2.4, apresentada no Capítulo 2.

A seguir, são realizadas duas análises: a primeira busca verificar o comportamento do tamanho do efeito com o aumento do tamanho da amostra, a fim de compará-lo com o comportamento do p-valor sob o mesmo aumento; e a segunda busca verificar a frequência dos chamados casos especiais (onde há discordância entre o p-valor e tamanho do efeito) nas análises realizadas.

Dando início à primeira análise, a Tabela 4.4 apresenta os porcentuais de testes em que os resultados do p-valor e do tamanho do efeito foram sensíveis e mudaram a conclusão com o aumento do tamanho da amostra. Na primeira linha é apresentado o porcentual de testes que não eram significativos e com o aumento do tamanho da amostra passaram a ser significativos, e na segunda linha é apresentado o porcentual de testes que tinham tamanho do efeito pequeno ou insignificante e passaram para médio ou grande com o aumento do tamanho da amostra de 10 para 20, de 10 para 30 e de 20 para 30, respectivamente.

Tabela 4.4: Porcentuais de resultados alterados com o aumento do tamanho da amostra no teste t

	Aumento no tamanho da amostra		
	10 para 20	10 para 30	20 para 30
p-valor	8,70%	11,96%	10,81%
d'cohen	4,29%	1,43%	3,13%

Como pode ser visto na Tabela 4.4, o porcentual de testes em que o tamanho do efeito passou de pequeno ou insignificante para médio ou grande foi consideravelmente inferior ao porcentual de testes em que o p-valor mudou o resultado de não significativo para significativo, para as três possíveis situações de aumento do tamanho da amostra. Vale destacar que, apesar de não ser sensível ao tamanho da amostra, os valores do tamanho do efeito podem ter aumentado (ou diminuído) com o aumento do tamanho da amostra devido às estimativas observadas nas novas amostras (média e variância) serem diferentes das anteriores, e não devido ao fato da amostra ser maior.

Para ilustrar essa diferença entre o comportamento do p-valor e do tamanho do efeito em relação ao aumento do tamanho da amostra, é apresentado um caso específico que já foi discutido anteriormente: a comparação das acurácias obtidas pelos classificadores SVM e 1-NN na Base 31. Com amostra de tamanho 10, o p-valor foi maior que 0,05 e, com

20 e 30 foi menor que 0,05. Entretanto, a classificação do tamanho do efeito não sofreu alteração com o aumento do tamanho da amostra. O tamanho do efeito foi 0,59, 0,72 e 0,60, para amostras de tamanho 10, 20 e 30, respectivamente, ou seja, todos classificados como tamanho do efeito médio.

Agora, será realizada a segunda análise: serão verificadas as frequências dos casos especiais (onde há discordância entre o p-valor e tamanho do efeito). Neste estudo empírico, foram encontrados casos onde a diferença entre as médias das acurácias populacionais dos classificadores é estatisticamente significante, porém o tamanho do efeito dessa diferença é pequeno. Também foram encontrados casos onde não há significância estatística para essa diferença e o tamanho do efeito é médio. E claro, também existem casos em que essas duas medidas concordam: com significância estatística e tamanho do efeito médio ou maior; e sem significância estatística com tamanho do efeito pequeno ou insignificante.

Voltando ao exemplo anterior (comparação das acurácias obtidas entre os classificadores SVM e 1-NN na Base 31), além da ilustração sobre a sensibilidade em relação ao tamanho da amostra, também é possível verificar que para as amostras de tamanho 20 e 30, as conclusões das duas medidas são concordantes: há significância estatística para afirmar que existe diferença entre os acurácias populacionais e tamanho do efeito médio para essa diferença. Já com amostra de tamanho 10, a diferença entre as médias das acurácias populacionais dos classificadores não é estatisticamente significante, porém o tamanho do efeito para essa diferença é médio, o que neste trabalho é considerado um caso especial.

A Figura 4.1 apresenta o comportamento do p-valor e do tamanho do efeito para cada tamanho de amostra. Nessa figura, as áreas hachuradas delimitam os casos especiais e cada ponto representa um teste realizado. Estão apresentados todos os 1509 testes feitos. Cada curva representa um tamanho de amostra diferente.

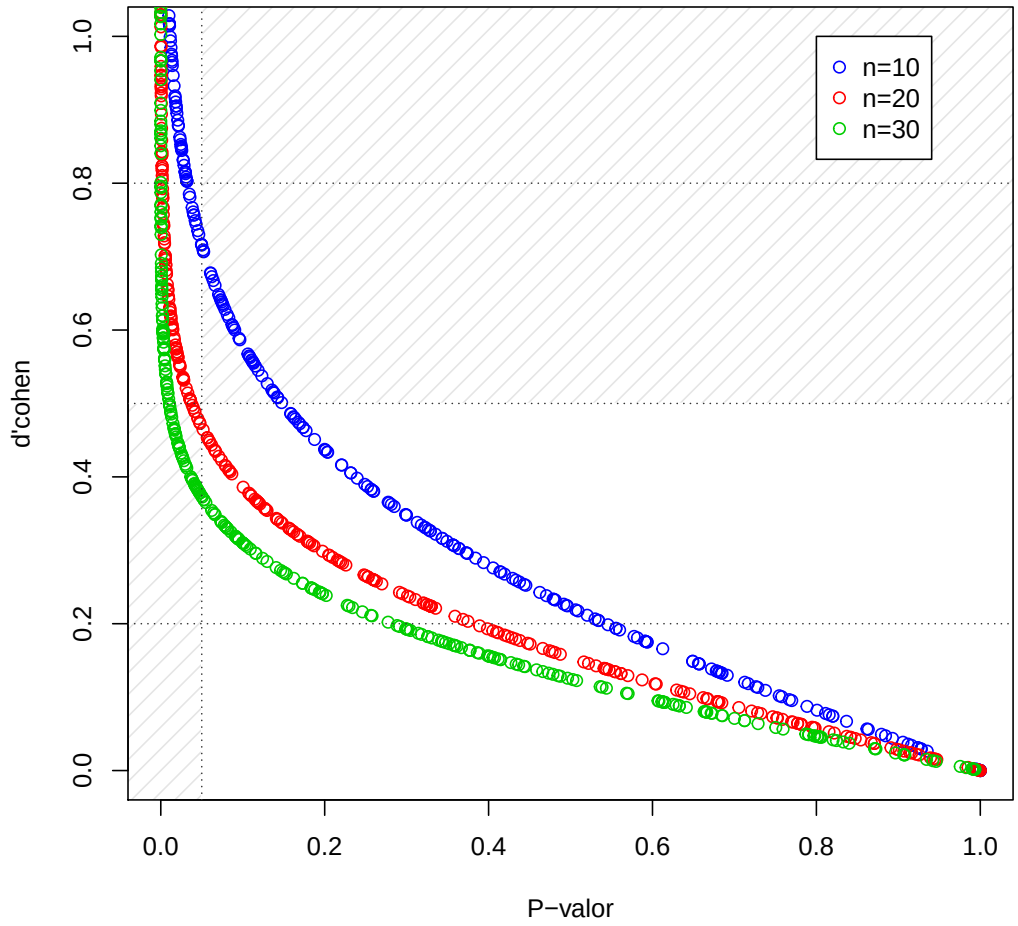


Figura 4.1: Representação dos resultados dos p-valores dos testes t aplicados com tamanhos de amostras 10, 20 e 30 e suas respectivas medidas de tamanho do efeito obtidas

É possível perceber que nos três tamanhos de amostras, na grande maioria dos casos o p-valor e o tamanho do efeito concordam. São casos onde: o p-valor indica significância estatística (p-valor<0,05) e o d'cohen indica tamanho do efeito médio ou maior (d'cohen≥0,5); ou o p-valor não indica significância estatística (p-valor≥0,05) e o d'cohen indica tamanho do efeito pequeno ou insignificante (d'cohen<0,5). Porém, para os três tamanhos de amostras, também foram encontrados casos especiais, onde essas medidas não concordam.

Esses casos especiais estão na área hachurada do gráfico e é possível perceber que os testes realizados com tamanho de amostra 10 tiveram um determinado tipo de discordância e os com tamanho 20 e 30 outro tipo. Com tamanho de amostra 10, foram observados alguns casos onde o teste t não obteve significância (através do p-valor) para afirmar que

os classificadores têm resultados diferentes na base em questão, porém o d'cohen para esses casos indicou que o tamanho do efeito dessa diferença é médio. Já nos testes com tamanho de amostra 20 e 30, foram encontrados alguns casos onde o teste t obteve evidências para rejeitar a hipótese nula (p-valor<0,05) e afirmar que os classificadores têm resultados diferentes na base em questão, porém o tamanho do efeito foi pequeno.

Agora, será analisada a frequência desses casos especiais em cada par de classificadores. Para isso, a Figura 4.2 apresenta as frequências dos casos especiais, quando não tem significância estatística no teste t e tem tamanho do efeito médio, por par de classificadores e tamanho da amostra. E a Figura 4.3 apresenta as frequências dos casos especiais, quando tem significância estatística no teste t e tem tamanho do efeito pequeno, por par de classificadores e tamanho da amostra.

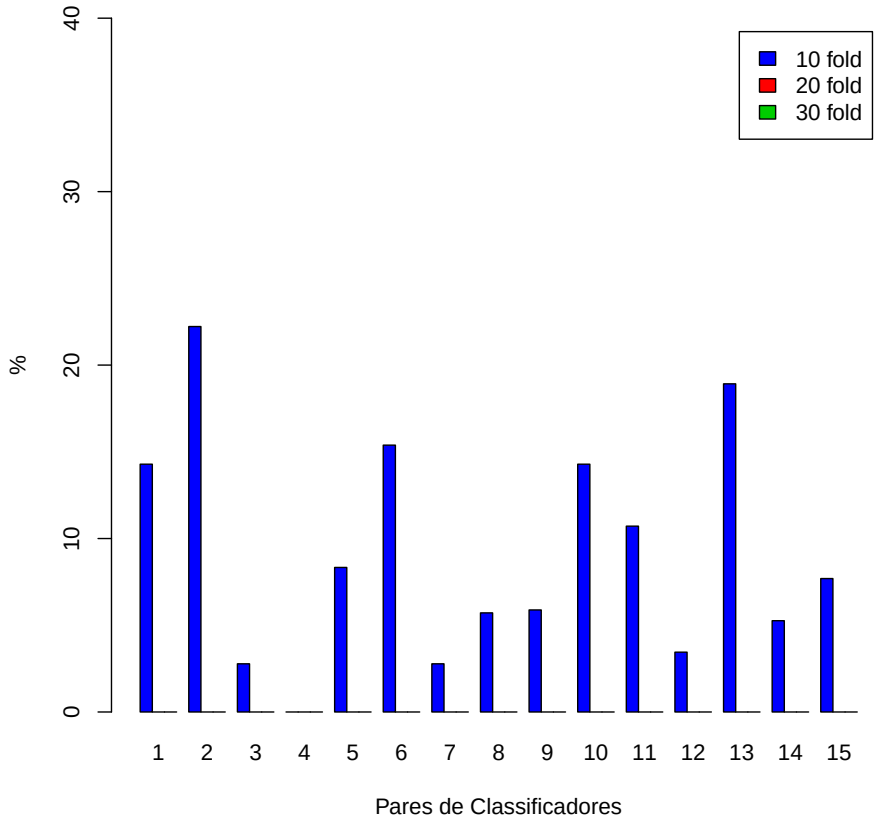


Figura 4.2: Porcentual de bases sem significância estatística no teste t e com tamanho do efeito médio, dentre as bases analisadas em cada um dos 15 pares de classificadores, por tamanho da amostra

Na Figura 4.2, fica claro que só foram encontrados casos sem significância estatística

no teste t e com tamanho do efeito médio para 10 partições. Além disso, é possível ver, por exemplo, 14,29% dos testes t realizados na comparação dos resultados do 1-NN e 3-NN (par 1) não indicaram significância estatística porém com tamanho do efeito médio. No geral, para todos os pares de classificadores, esse valor foi 8,8%, ou seja, 8,8% dos testes t realizados com amostras de tamanho 10 não obtiveram significância estatística para afirmar que os resultados de classificadores em uma determinada base de dados são diferentes, porém, o tamanho do efeito médio indica que essa diferença tem magnitude relevante.

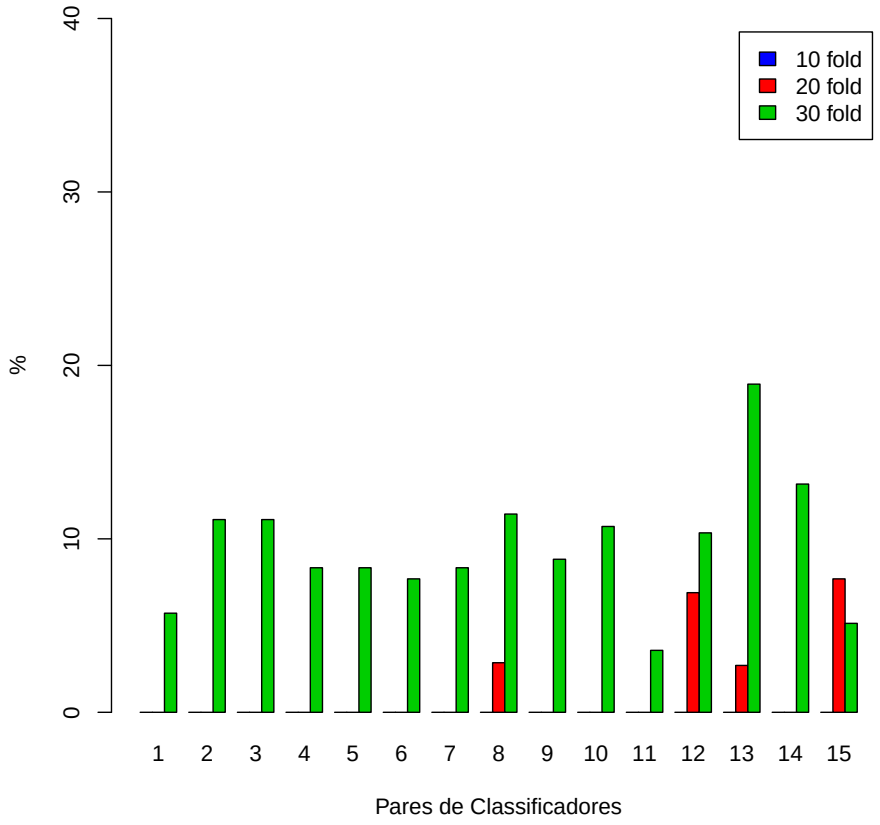


Figura 4.3: Porcentual de bases com significância estatística no teste t e com tamanho do efeito pequeno, dentre as bases analisadas em cada um dos 15 pares de classificadores, por tamanho da amostra

Já na Figura 4.3, é reforçado que só foram encontrados casos com significância estatística no teste t e com tamanho do efeito pequeno para testes com amostras de tamanho 20 e 30. Além disso, em todos os pares de classificadores com amostras de tamanho 30, foram obtidos esse tipo de caso especial. Já nos testes com amostras de tamanho 20, foram observados esses casos especiais em apenas 4 pares de classificadores e com frequências

abaixo de 10%. No geral, as frequências de casos especiais com significância estatística e tamanho do efeito pequeno para amostras de tamanho 20 e 30 foram 1,4% e 9,6%, respectivamente. De fato, os testes realizados em amostras de tamanho 20 foram os que apresentaram o menor porcentual de casos especiais.

Vale ressaltar que os casos especiais não indicam necessariamente um erro, pois o p-valor e o tamanho do efeito medem informações diferentes, mas é um sinal de que os dados merecem mais atenção e, se fosse utilizado apenas o p-valor, a conclusão poderia ser equivocada. O poder do teste é outra medida que deve ser calculada em todo estudo que utilize teste de significância estatística, para avaliar a adequação do teste ao ambiente em que está sendo aplicado. O problema é que essa medida tão importante não é muito utilizada em algumas áreas, ou por ser pouco conhecida ou pouco compreendida.

4.2.3 Avaliando o Poder do Teste

Nesta subseção, é acrescentado ao estudo o poder dos testes t realizados, que é calculado como visto na Equação 2.17, no Capítulo 2. O poder do teste é, na verdade, uma função e será calculada para todos os testes realizados, considerando que o valor real da diferença entre os classificadores é igual a diferença observada.

A aplicação de um teste de hipóteses sempre deve vir acompanhada do cálculo do poder do teste, para medir a eficiência do teste na amostra em que está sendo aplicado, já que o poder do teste é a probabilidade de rejeitar a hipótese nula para um dado μ_D. Sendo assim, no contexto da avaliação de classificadores, o poder do teste é a probabilidade do teste afirmar que os classificadores A e B são diferentes dada a diferença real (estimada pela diferença amostral observada).

Será verificado o comportamento do poder dos testes em conjunto com o p-valor e o tamanho do efeito. Para isso, será analisado o comportamento do poder do teste para os casos onde o p-valor e o tamanho do efeito concordam, e para os casos especiais. Os valores calculados para o poder dos testes realizados são apresentados no Apêndice A, Tabelas A.7, A.8 e A.9, para as amostras de tamanhos 10, 20 e 30, respectivamente.

Inicialmente, será verificado o poder do teste para o caso específico que vem sendo discutido nas subseções anteriores (a comparação das acurácias obtidas pelos os classificadores SVM e 1-NN na Base 31), onde o p-valor foi 0,10, 0,004 e 0,002, e o d'cohen foi 0,59, 0,72 e 0,60, para amostras de tamanho 10, 20 e 30, respectivamente. Com as amostras de tamanho 10, não foi possível afirmar que a diferença entre as acurácias dos

classificadores é estatisticamente significativa, porém tem tamanho do efeito médio. O poder do teste calculado foi 0,34, ou seja, um poder baixo (o recomendado deve acima de 0,8) que indica que o teste não está sendo aplicado no ambiente ideal, ou seja, a amostra é pequena. Já com amostras de tamanho 20 e 30, o teste t obteve significância estatística para a diferença entre os classificadores e o d'cohen indica que essa diferença tem tamanho do efeito médio. Reforçando essas conclusões, o poder do teste alto de 0,87, 0,89, respectivamente, indica que o teste tem grande probabilidade de rejeitar a hipótese nula quando ela realmente é falsa.

Para facilitar as análises seguintes, os resultados dos testes de significância estatística e os valores do d'cohen são separados em 4 grupos, a fim de verificar como o poder do teste se comporta em cada grupo.

- Grupo 1: Com significância estatística e tamanho do efeito médio ou grande
- Grupo 2: Sem significância estatística e tamanho do efeito pequeno ou insignificante
- Grupo 3: Sem significância estatística e tamanho do efeito médio ou grande
- Grupo 4: Com significância estatística e tamanho do efeito pequeno ou insignificante

Os dois primeiros grupos têm resultados concordantes entre o resultado do teste t e a medida do tamanho do efeito, ambos indicam ou não diferença entre os grupos. Já nos dois últimos grupos, essas medidas indicam conclusões opostas, são os casos especiais.

A Figura 4.4 contém o boxplot dos valores obtidos na função poder do teste para esses grupos separados por tamanho da amostra (10, 20 e 30). O boxplot é um gráfico utilizado para avaliar a distribuição empírica de um conjunto de valores, onde a caixa central é formada pelo primeiro, segundo (mediana) e terceiro quartis, Q_1, Q_2 e Q_3, respectivamente. As hastes inferiores e superiores se estendem, respectivamente, do quartil inferior até o menor valor não excedente ao limite inferior $(Q_1 - 1{,}5(Q_3 - Q_1))$ e do quartil superior até o maior valor não excedente ao limite superior $(Q_3 + 1{,}5(Q_3 - Q_1))$. Os pontos fora destes limites são considerados *outliers*, aqui representados por círculos.

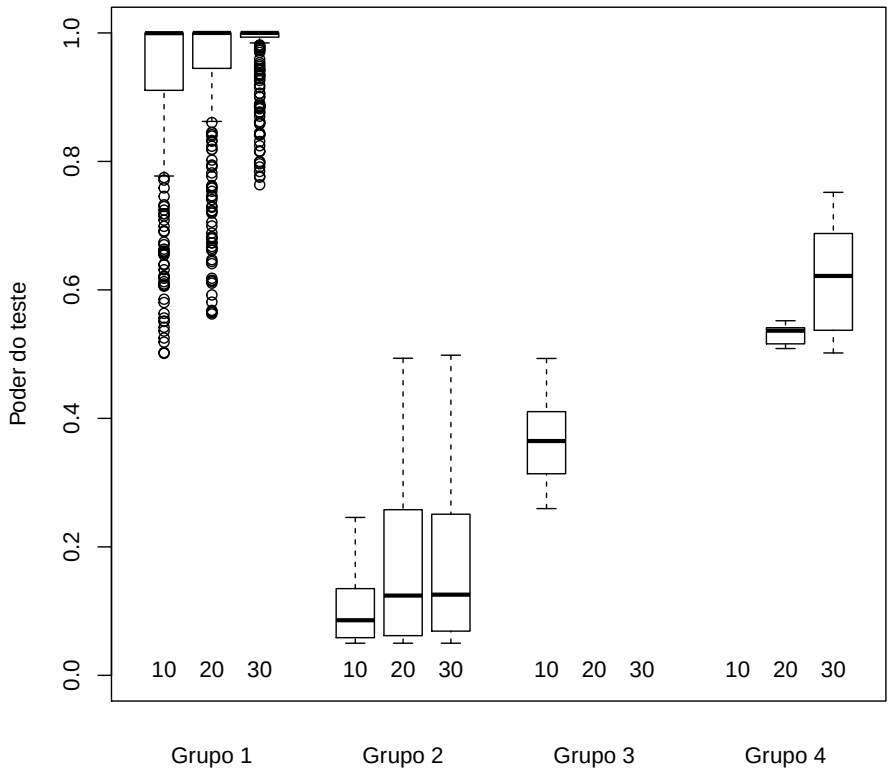


Figura 4.4: Boxplot dos valores obtidos na função poder do teste t por grupos e tamanhos das amostras

Na Figura 4.4, é possível perceber que no primeiro grupo, com significância estatística e tamanho do efeito médio ou grande, em geral, o poder do teste t foi alto, apenas com alguns outliers com valores abaixo de 80%, para os três tamanhos de amostras. Nesse cenário, as três medidas se reforçam: o poder do teste alto dá credibilidade ao resultado do teste, que foi afirmar que existe diferença estatisticamente significativa entre os classificadores e o tamanho do efeito acrescenta que a magnitude dessa diferença é grande ou importante.

Já no segundo grupo, é encontrada a situação oposta. O teste t não rejeitou a hipótese nula e o tamanho do efeito é pequeno, concordando com o resultado do teste. Porém, o poder do teste é baixo indicando que a probabilidade do teste rejeitar a hipótese nula quando realmente deve ser rejeitada é baixa. Independente de as duas primeiras medidas estarem concordando, o poder do teste baixo sugere que a amostra deve ser ampliada para tornar o teste mais poderoso. Na Figura 2.3, apresentada no Capítulo 2, foi ilustrado como

o poder do teste aumenta com o aumento do tamanho da amostra.

Ainda na Figura 4.4, percebe-se que para o terceiro grupo, onde só existem casos com amostras de tamanho 10, o poder do teste foi baixo, indicando que o tamanho da amostra (10) é pequeno para a realização do teste t. No quarto grupo, onde só existem casos com tamanho da amostra 20 e 30, o poder do teste (entre 50% e 60% para amostras de tamanho 20 e entre 50% e 70% para amostras de tamanho 30, aproximadamente) indica que a chance de o teste rejeitar a hipótese nula quando realmente deve rejeitar está em torno de apenas 60% para esses casos. Como o tamanho do efeito é baixo, é possível fazer uma analogia ao exemplo da aspirina apresentado na introdução: a diferença é estatisticamente significativa mas pode não ter significância prática.

Vale destacar que todos os casos com poder do teste alto ($> 0{,}8$) pertencem ao Grupo 1, ou seja, obtiveram significância estatística no teste t e tamanho do efeito médio ou grande, conforme visto na Figura 4.4. Para auxiliar a compreensão desse resultado, cabe relembrar que, conforme já apresentado na Subseção 2.2, o parâmetro μ_D (diferença média populacional) é um dos fatores que influenciam a função do poder do teste. Como é utilizada a estimativa μ_d (diferença média amostral) para o cálculo dessa função, ela será analisada na Figura 4.5. Essa figura apresenta o boxplot dos valores observados das diferenças médias amostrais por grupos e por tamanho da amostra. Para acrescentar mais informação para a análise, também é apresentada a Figura 4.6 que contém o boxplot dos valores dos desvios padrões observados nas diferenças amostrais por grupos e por tamanho da amostra.

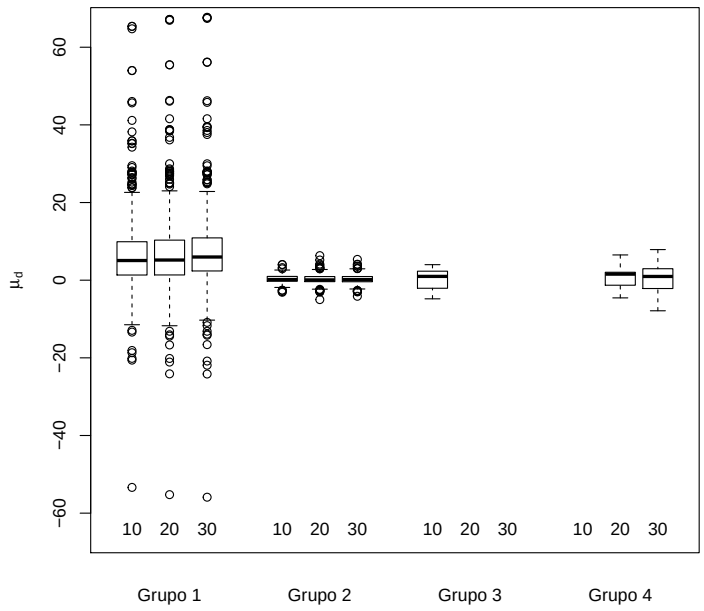


Figura 4.5: Boxplot das diferenças médias amostrais observadas por grupos do teste t e tamanhos das amostras

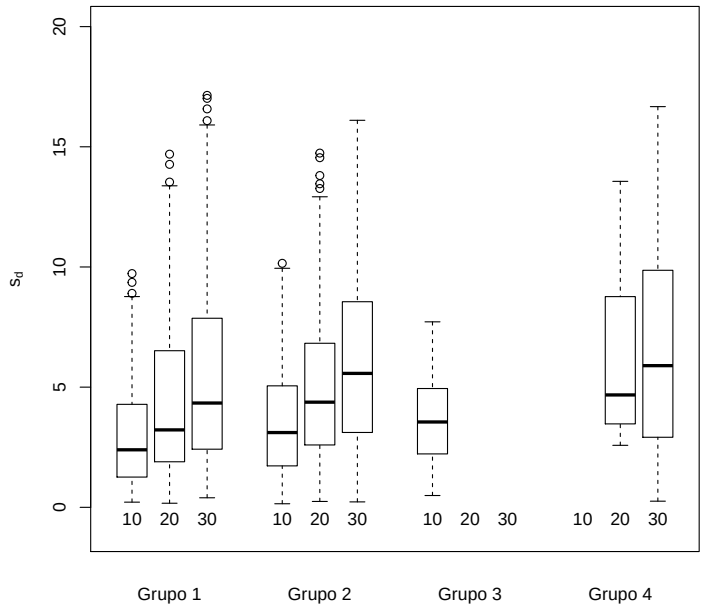


Figura 4.6: Boxplot dos desvios padrões das diferenças amostrais observadas por grupos do teste t e tamanhos das amostras

Na Figura 4.5, é possível perceber que os maiores valores de diferenças amostrais observadas pertencem ao Grupo 1. Para esses valores maiores, é mais provável o teste obter significância estatística, o tamanho do efeito ser médio ou maior e o poder do teste ser alto. E realmente essas três medidas foram observadas com esses resultados. Ao contrário, com valores de diferenças amostrais pequenas, é menos provável essas três medidas obterem esses resultados, porém, não é impossível.

No Grupo 3, por exemplo, mesmo com as diferenças amostrais pequenas, o tamanho do efeito foi médio ou grande. Este resultado pode ser explicado pelo fato de o desvio padrão para a diferença entre as acurácias amostrais também ser pequeno, como pode ser observado na Figura 4.6. Já no Grupo 4, mesmo com as diferenças amostrais pequenas, o teste t obteve significância estatística para afirmar que existe diferença entre os resultados dos classificadores.

Agora, serão realizadas análises semelhantes às feitas nesta seção, porém para os resultados obtidos através da aplicação do teste de Wilcoxon.

4.3 Análise com o Teste de Wilcoxon

Nesta seção, serão apresentados os resultados encontrados pelas medidas p-valor, tamanho do efeito e poder do teste na comparação dos 15 pares de classificadores, utilizando o teste de Wilcoxon considerando as 50 bases de dados.

Como o Teste de Wilcoxon é um teste não paramétrico, não é necessário realizar o teste de normalidade. Portanto, para que seja mantida a comparação de um par de classificadores em determinada base de dados, é necessário apenas que as amostras com 10, 20 e 30 partições não tenham variância igual a zero.

No teste de Wilcoxon bilateral para amostras pareadas, a hipótese nula é que a mediana das diferenças das acurácias é igual a zero, ou seja, os dois classificadores em questão geram resultados similares na base que está sendo utilizada. A hipótese alternativa é que a mediana da diferença não é zero, ou seja, os resultados diferem em posição. As hipóteses são definidas como visto na Equação 2.13, definida no Capítulo 2.

Os resultados a seguir são apresentados em três subseções. Na primeira, é analisado como o p-valor se comporta com a alteração do tamanho da amostra. Na segunda, é comparado o comportamento do p-valor e do tamanho do efeito e, na terceira parte, é acrescentado o poder do teste nas análises.

4.3.1 Análise do Comportamento do P-valor

Nesta subseção, será analisado o impacto da variação dos p-valores obtidos com o teste de Wilcoxon para os diferentes tamanhos de amostras na avaliação dos classificadores, assim como foi feito na Subseção 4.2.1 para os p-valores obtidos com o teste t.

Assim como feito nas subseções da Seção 4.2, será utilizado um caso específico para ilustrar as discussões apresentadas nas subseções com os resultados do teste de Wilcoxon. Para tanto, será utilizado o mesmo caso específico que foi utilizado com os resultados do teste t, que é a comparação das acurácias obtidas entre os classificadores SVM e 1-NN na Base 31. O primeiro resultado que se deseja ilustrar é a sensibilidade do p-valor em relação ao tamanho da amostra. Com a amostra de tamanho 10 (método de validação cruzada com 10 partições) o p-valor na comparação desses classificadores foi 0,08. Já com amostras de tamanho 20 e 30, o p-valor foi 0,003 e 0,004, respectivamente. Ou seja, com 10 partições, o teste não teve significância para afirmar que os classificadores foram diferentes, mas aumentando a amostra para 20 ou 30, foi possível obter a significância estatística para afirmar diferença entre a médias das acurácias populacionais desses classificadores na base em questão, utilizando o teste de Wilcoxon.

A Tabela 4.5 mostra que essa sensibilidade (diminuição do p-valor com o aumento do tamanho da amostra) foi bastante frequente nos casos analisados. Nessa tabela, para cada par de classificadores, é apresentada a porcentagem de testes realizados em que o p-valor diminuiu com o aumento da amostra de 10 para 20, de 10 para 30 e de 20 para 30. Essa porcentagem representa a sensibilidade em questão, que pode ser observada, por exemplo, em aproximadamente 82% das bases analisadas na comparação entre SVM e 1-NN com o aumento do tamanho da amostra de 10 para 20, ou seja, 82% dos testes tiveram redução do p-valor com esse aumento da amostra. Todos os p-valores obtidos foram colocados no Apêndice B para não sobrecarregar o texto, nas Tabelas B.1, B.2 e B.3, para as amostras de tamanhos 10, 20 e 30, respectivamente.

Tabela 4.5: Quantidade porcentual de testes de Wilcoxon aplicados nas quais o p-valor reduziu com o aumento do tamanho da amostra (de 10 para 20, de 10 para 30 e de 20 para 30) na aplicação do teste t em cada par de classificadores

Classificadores	Aumento no tamanho da amostra		
	10 para 20	10 para 30	20 para 30
RF100 e RF300	51,11	52,08	43,48
RF100 e SVM	70,00	70,00	66,00
RF100 e KNN1	82,00	80,00	64,00
RF100 e 3-NN	82,00	82,00	62,00
RF100 e NB	80,00	84,00	82,00
RF300 e SVM	68,00	79,59	73,47
RF300 e KNN1	82,00	80,00	78,00
RF300 e 3-NN	70,00	80,00	86,00
RF300 e NB	80,00	78,00	78,00
SVM e 1-NN	82,00	78,00	66,00
SVM e 3-NN	68,00	72,00	70,00
SVM e NB	82,00	78,00	70,00
1-NN e 3-NN	60,00	68,89	53,33
1-NN e NB	76,00	82,00	66,00
3-NN e NB	74,00	78,00	80,00
Média	74,05	76,28	69,46

Os valores apresentados na Tabela 4.5 mostram que, na comparação de alguns pares de classificadores, o p-valor diminuiu com uma frequência maior que em outros pares. Porém, no geral, conforme visto na linha da média total (sem separação por pares de classificadores), o p-valor realmente tende a diminuir com o aumento do tamanho da amostra. De acordo com os valores obtidos, essa diminuição não é uma regra, já que, coletando outra amostra de tamanho maior, as novas estimativas observadas (como média e variância amostrais) podem ser diferentes e, consequentemente, também influenciar o novo p-valor observado.

Assim como nas análises dos resultados do teste t, os resultados vistos até aqui mostram que o p-valor (obtido com o teste de Wilcoxon) realmente é sensível em relação ao tamanho da amostra. Agora, deseja-se analisar qual o porcentual de vezes em que a queda do p-valor alterou a conclusão tomada no teste de hipótese, ou seja, o resultado passou de não significativo para significativo.

As porcentagens de testes de Wilcoxon onde houve mudança na conclusão com o aumento do tamanho da amostra de 10 para 20, de 10 para 30 e de 20 para 30 são 14,52%, 18,15% e 14,69%, respectivamente. À primeira vista, esses valores podem parecer baixos se comparados às porcentagens de queda observadas, porém, ilustram que essa sensibilidade pode afetar um número significativo de resultados de pesquisas científicas.

Uma forma de complementar a análise é considerar uma medida de tamanho do efeito, já que tirar decisões com base em p-valores isolados pode levar a decisões equivocadas. A seguir são discutidos os resultados dos tamanhos do efeito para cada teste de Wilcoxon aplicado.

4.3.2 Avaliando o Tamanho do Efeito

Nesta subseção, é acrescentada ao estudo a medida do tamanho do efeito r, que é calculada como visto na Equação 2.20, no Capítulo 2. Essa medida é calculada para todos os testes realizados a fim de medir o grau da diferença entre as acurácias dos classificadores, na base de dados em questão. Assim, quanto maior for o tamanho do efeito, maior será a manifestação do fenômeno (diferença) na população.

Os valores calculados para a medida de tamanho do efeito são apresentados no Apêndice B, Tabelas B.4, B.5 e B.6, para as amostras de tamanhos 10, 20 e 30, respectivamente. Esses valores são classificados de acordo com os pontos de corte propostos por Cohen [5], que foram descritos na Tabela 2.5, apresentada no Capítulo 2.

A seguir, são realizadas duas análises: a primeira busca verificar o comportamento do tamanho do efeito com o aumento do tamanho da amostra, a fim de compará-lo com o comportamento do p-valor sob o mesmo aumento; a segunda busca verificar a frequência dos chamados casos especiais (onde há discordância entre o p-valor e tamanho do efeito) nas análises realizadas com o teste de Wilcoxon.

Dando início à primeira análise, a Tabela 4.6 apresenta os porcentuais de testes em que os resultados do p-valor e do tamanho do efeito foram sensíveis e mudaram a conclusão com o aumento do tamanho da amostra. Na primeira linha, é apresentado o porcentual de testes que não eram significativos e com o aumento do tamanho da amostra passaram a ser significativos, e na segunda linha é apresentado o porcentual de testes que tinham tamanho do efeito pequeno ou insignificante e passaram para médio ou grande com o aumento do tamanho da amostra de 10 para 20, de 10 para 30 e de 20 para 30, respectivamente.

Tabela 4.6: Porcentuais de resultados alterados com o aumento do tamanho da amostra no teste de Wilcoxon

	Aumento no tamanho da amostra		
	10 para 20	10 para 30	20 para 30
p-valor	14,52%	18,15%	14,69%
r	5,03%	1,11%	5,45%

Como pode ser visto na Tabela 4.6, o porcentual de testes em que o tamanho do efeito passou de pequeno ou insignificante para médio ou grande foi consideravelmente inferior ao porcentual de testes em que o p-valor mudou o resultado de não significativo para significativo, para as três possíveis situações de aumento do tamanho da amostra. Vale destacar que, apesar de não ser sensível ao tamanho da amostra, os valores do tamanho do efeito podem ter aumentado (ou diminuído) com o aumento do tamanho da amostra devido às estimativas observadas nas novas amostras (média e variância) serem diferentes das anteriores, e não devido ao fato de a amostra ser maior.

Para ilustrar essa diferença entre o comportamento do p-valor e do tamanho do efeito em relação ao aumento do tamanho da amostra, é apresentado um caso específico que já foi discutido anteriormente: a comparação das acurácias obtidas pelos classificadores SVM e 1-NN na Base 31. Com amostra de tamanho 10, o p-valor para o teste de Wilcoxon foi maior que 0,05 e, com 20 e 30, foi menor que 0,05. Entretanto, a classificação do tamanho do efeito não sofreu alteração com o aumento do tamanho da amostra. A medida de tamanho do efeito r foi 0,39, 0,45 e 0,36, para amostras de tamanho 10, 20 e 30, respectivamente, ou seja, todos classificados como tamanho do efeito médio.

Agora, será realizada a segunda análise: serão verificadas as frequências dos casos especiais (onde há discordância entre o p-valor e tamanho do efeito). Neste estudo empírico com o teste de Wilcoxon, assim com o teste t, foram encontrados casos onde a diferença entre as médias das acurácias populacionais dos classificadores é estatisticamente significante, porém o tamanho do efeito dessa diferença é pequeno. Também foram encontrados casos onde não há significância estatística para essa diferença e o tamanho do efeito é médio. E claro, também existem casos em que essas duas medidas concordam: com significância estatística e tamanho do efeito médio ou maior, e sem significância estatística com tamanho do efeito pequeno ou insignificante.

Voltando ao exemplo anterior (comparação das acurácias obtidas entre os classificadores SVM e 1-NN na Base 31), além da ilustração sobre a sensibilidade em relação ao

tamanho da amostra, também é possível verificar que, para as amostras de tamanho 20 e 30, as conclusões das duas medidas são concordantes: há significância estatística para afirmar que existe diferença entre os acurácias populacionais e tamanho do efeito médio para essa diferença. Já com amostra de tamanho 10, a diferença entre as médias das acurácias populacionais dos classificadores não é estatisticamente significante, porém o tamanho do efeito para essa diferença é médio, o que neste trabalho é considerado um caso especial.

A Figura 4.7 apresenta o comportamento do p-valor e do tamanho do efeito para cada tamanho de amostra. Nessa figura, as áreas hachuradas delimitam os casos especiais e cada ponto representa um teste realizado. Cada curva representa um tamanho de amostra diferente.

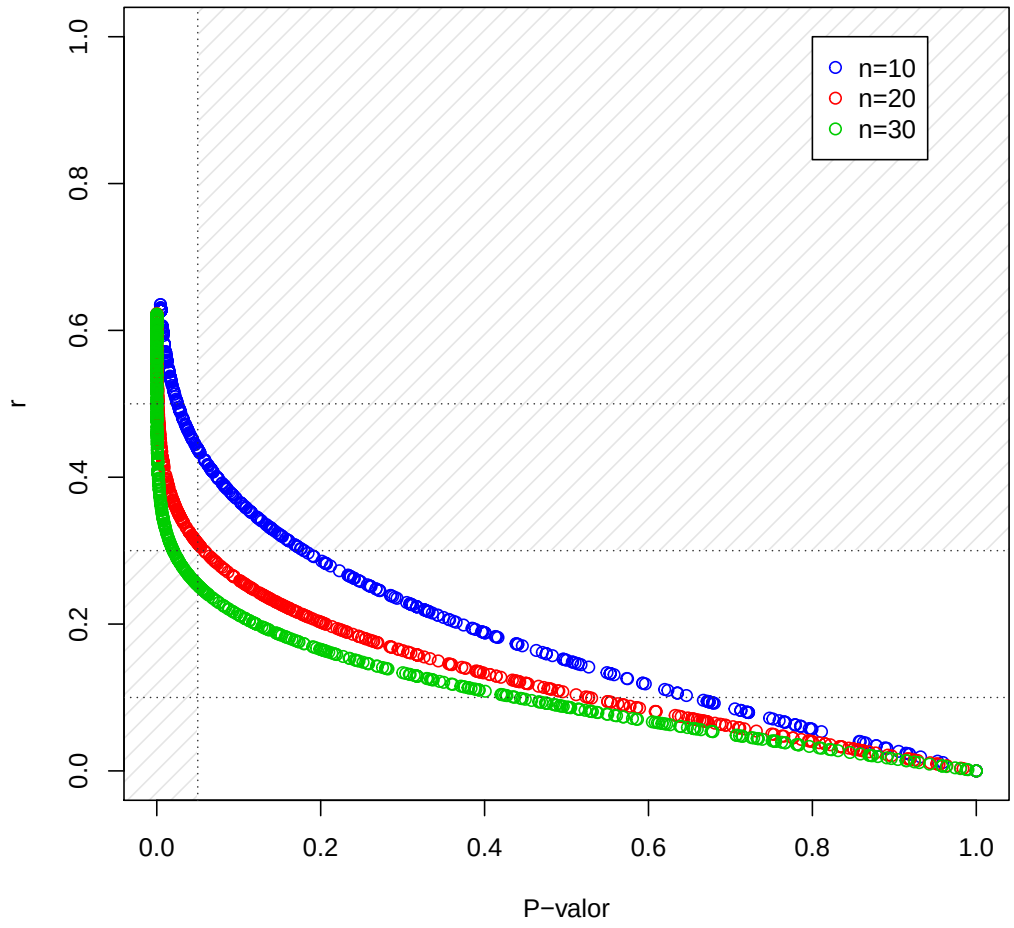


Figura 4.7: Representação dos resultados dos p-valores dos testes de Wilcoxon aplicados com tamanhos de amostras 10, 20 e 30 e suas respectivas medidas de tamanho do efeito obtidas

É possível perceber que nos três tamanhos de amostras, na grande maioria dos casos, o p-valor e o tamanho do efeito concordam. São casos onde: o p-valor indica significância estatística (p-valor<0,05) e a medida r indica tamanho do efeito médio ou maior ($r \geq 0{,}3$); ou o p-valor não indica significância estatística (p-valor≥0,05) e a medida r indica tamanho do efeito pequeno ou insignificante ($r < 0{,}3$). Porém, para os três tamanhos de amostras, também foram encontrados casos especiais, onde essas medidas não concordam.

Esses casos especiais estão na área hachurada do gráfico e é possível perceber que os testes realizados com os tamanhos de amostras 10 e 20 tiveram um determinado tipo de discordância e os testes com tamanho 30, outro tipo. Com tamanhos de amostras 10 e 20, foram observados alguns casos onde o teste de Wilcoxon não obteve significância (através do p-valor) para afirmar que os classificadores têm resultados diferentes na base em questão, porém a medida r para esses casos indicou que o tamanho do efeito dessa diferença é médio. Já nos testes com tamanho de amostra 30, foram encontrados alguns casos onde o teste de Wilcoxon obteve evidências para rejeitar a hipótese nula (p-valor<0,05) e afirmar que os classificadores têm resultados diferentes na base em questão, porém o tamanho do efeito foi pequeno.

Agora, será analisada a frequência desses casos especiais em cada par de classificadores. Para isso, a Figura 4.8 apresenta as frequências dos casos especiais, que não têm significância estatística no teste de Wilcoxon e têm tamanho do efeito médio, por par de classificadores e tamanho da amostra. E a Figura 4.9 apresenta as frequências dos casos especiais, que têm significância estatística no teste de Wilcoxon e tem tamanho do efeito pequeno, por par de classificadores e tamanho da amostra.

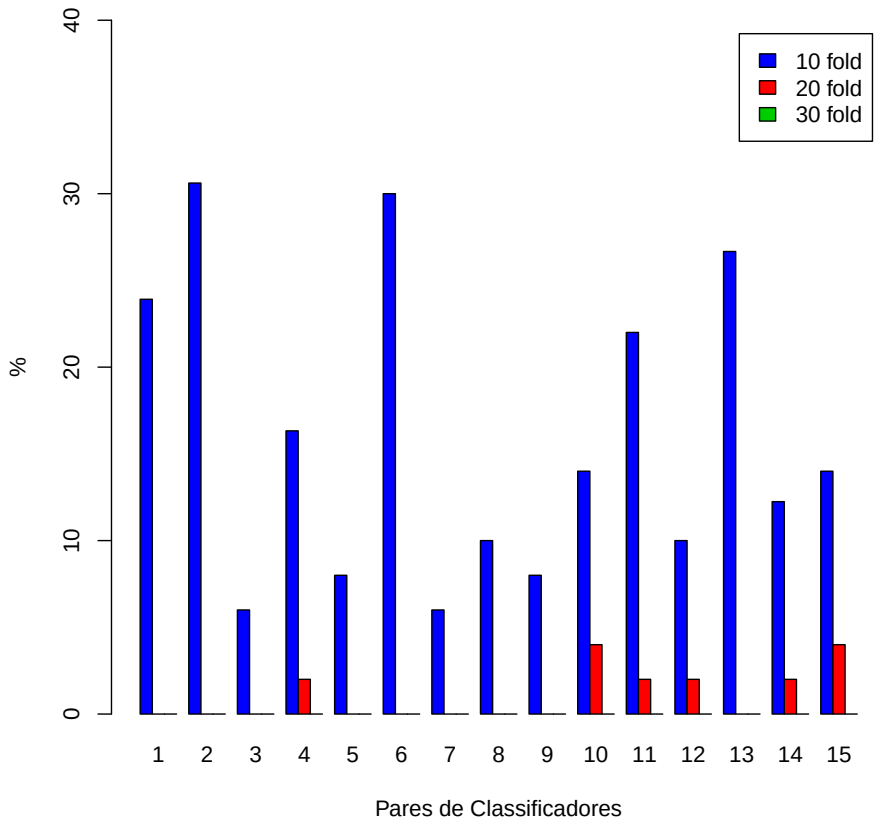


Figura 4.8: Porcentual de bases sem significância estatística no teste de Wilcoxon e com tamanho do efeito médio, dentre as bases analisadas em cada um dos 15 pares de classificadores, por tamanho da amostra

Na Figura 4.8, fica claro que só foram encontrados casos sem significância estatística no teste de Wilcoxon e com tamanho do efeito médio para amostras de tamanhos 10 e 20. Além disso, é possível ver, por exemplo, que 23,93% dos testes de Wilcoxon realizados na comparação dos resultados do 1-NN e 3-NN (par 1) não indicaram significância estatística, porém apresentaram tamanho do efeito médio, com tamanho de amostra 10. No geral, para todos os pares de classificadores, esses valores foram 15,72% e 1,08% para amostras de tamanho 10 e 20, respectivamente.

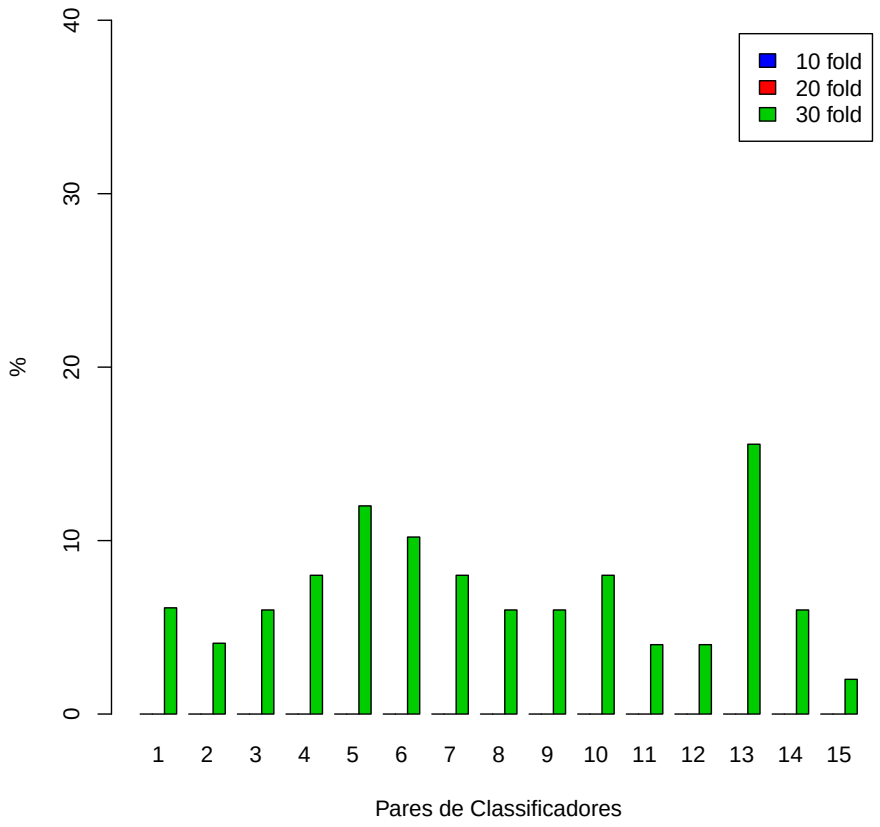


Figura 4.9: Porcentual de bases com significância estatística no teste de Wilcoxon e com tamanho do efeito pequeno, dentre as bases analisadas em cada um dos 15 pares de classificadores, por tamanho da amostra

Já na Figura 4.9, é reforçado que só foram encontrados casos com significância estatística no teste de Wilcoxon e com tamanho do efeito pequeno para testes com amostras de tamanho 30. Além disso, em todos os pares de classificadores com amostras de tamanho 30, foram obtidos esse tipo de caso especial. No geral, a frequência de casos especiais com significância estatística e tamanho do efeito pequeno para amostras de tamanho 30 foi 7,0%.

Vale ressaltar que, equivalente ao encontrado nas análises dos resultados do teste t, os testes de Wilcoxon realizados em amostras de tamanho 20 foram os que apresentaram o menor porcentual de casos especiais.

O poder do teste é outra medida que deve ser calculada em todo estudo que utilize teste de significância estatística, para avaliar a adequação do teste ao ambiente em que está sendo aplicado. Na subseção a seguir, serão analisados os resultados encontrados

para essa medida.

4.3.3 Avaliando o Poder do Teste

Nesta subseção, é acrescentado ao estudo o poder dos testes de Wilcoxon realizados, que é calculado com base em simulação de amostras, conforme descrito no Algoritmo 1, no Capítulo 2. O poder do teste é na verdade uma função, e será calculada para todos os testes realizados, considerando que o valor real da diferença entre os classificadores é igual a diferença observada.

Assim como analisado para os resultados obtidos através do teste t, será analisado o comportamento do poder do teste de Wilcoxon para os casos onde o p-valor e o tamanho do efeito concordam, e para os casos especiais. Os valores calculados para o poder dos testes realizados são apresentados no Apêndice B, Tabelas B.7, B.8 e B.9, para as amostras de tamanhos 10, 20 e 30, respectivamente.

Inicialmente, será verificado o poder do teste para o caso específico que vêm sendo discutido nas subseções anteriores (a comparação das acurácias obtidas pelos classificadores SVM e 1-NN na Base 31), onde o p-valor no teste de Wilcoxon foi 0,08, 0,004 e 0,004, e a medida do tamanho do efeito r foi 0,39, 0,45 e 0,36, para amostras de tamanho 10, 20 e 30, respectivamente. Com as amostras de tamanho 10, não foi possível afirmar que a diferença entre as acurácias dos classificadores é estatisticamente significativa, porém tem tamanho do efeito médio. O poder do teste calculado foi 0,45, ou seja, um poder baixo (o recomendado deve ser acima de 0,8), que indica que o teste não está sendo aplicado no ambiente ideal, ou seja, a amostra é pequena. Já com amostras de tamanho 20 e 30, o teste de Wilcoxon obteve significância estatística para a diferença entre os classificadores e a medida r indica que essa diferença tem tamanho do efeito médio. Reforçando essas conclusões, o poder do teste alto de 0,84 e 0,88, respectivamente, indica que o teste tem grande probabilidade de rejeitar a hipótese nula quando ela realmente é falsa (quando a diferença real é igual à observada nas amostras).

As análises seguintes utilizam os mesmos grupos definidos na Subseção 4.2.3, onde os resultados dos testes de significância estatística e os valores do d'cohen são separados em 4 grupos. Lembrando que os dois primeiros grupos têm resultados concordantes entre o resultado do teste de Wilcoxon e a medida do tamanho do efeito, e os dois últimos grupos são os casos especiais. A Figura 4.10 contém o boxplot dos valores obtidos na função poder do teste de Wilcoxon para os grupos separados por tamanho da amostra (10, 20 e 30).

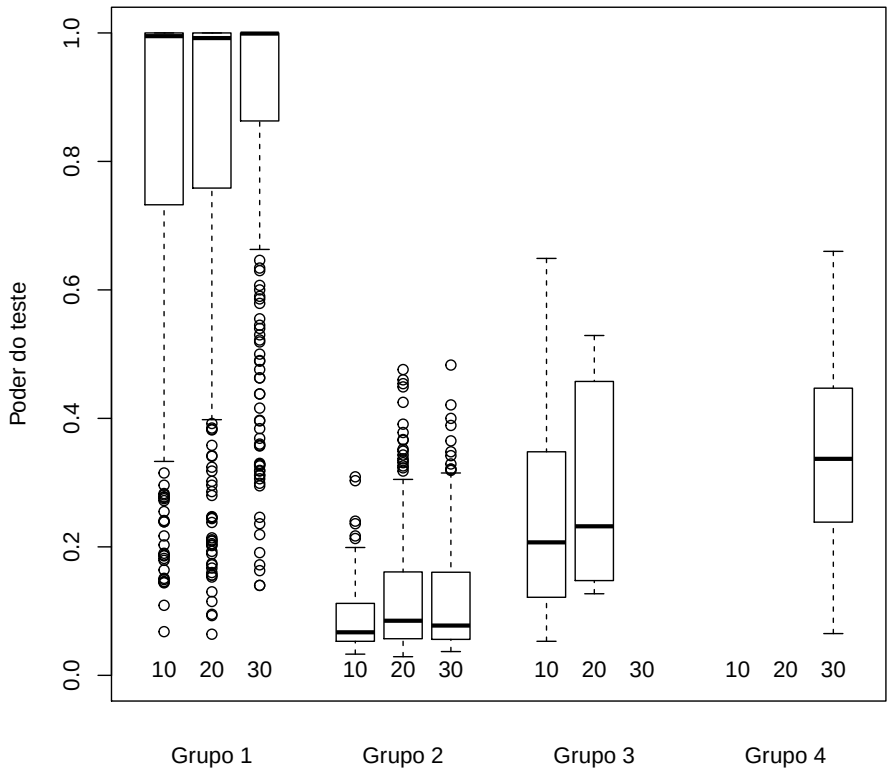


Figura 4.10: Boxplot dos valores obtidos na função poder do teste de Wilcoxon por grupos e tamanhos das amostras

Na Figura 4.10, é possível perceber que no primeiro grupo, com significância estatística e tamanho do efeito médio ou grande, a maioria dos casos apresentaram o poder do teste de Wilcoxon alto. Nesse cenário, as três medidas se reforçam: o poder do teste alto dá credibilidade ao resultado do teste, que foi afirmar que existe diferença estatisticamente significativa entre os classificadores e o tamanho do efeito acrescenta que a magnitude dessa diferença é relevante. Porém, nesse mesmo grupo, é observado que alguns casos tiveram poder do teste baixo, com maior frequência para amostras de tamanho 10 do que de tamanho 30, como era esperado já que o poder do teste tende a aumentar com o aumento do tamanho da amostra. Nesse cenário, apesar de o p-valor e tamanho do efeito concordarem, o poder do teste baixo indica que o teste de Wilcoxon realizado não é muito confiável, já que a probabilidade de o teste rejeitar a hipótese nula, quando a diferença real é igual à observada, é baixa.

Já no segundo grupo, é encontrada a situação oposta. O teste de Wilcoxon não

rejeitou a hipótese nula e o tamanho do efeito é pequeno, concordando com o resultado do teste. Porém, o poder do teste é baixo indicando que a probabilidade de o teste rejeitar a hipótese nula quando realmente deve ser rejeitada é baixa. Independente de as duas primeiras medidas estarem concordando, o poder do teste baixo sugere que a amostra deve ser ampliada para tornar o teste mais poderoso.

Ainda na Figura 4.10, percebe-se que para o terceiro grupo, onde os testes não obtiveram significância estatística porém o tamanho do efeito é médio ou grande, o poder do teste de Wilcoxon foi baixo. O baixo poder indica que o teste não é confiável, já que possui baixa probabilidade de rejeitar a hipótese nula quando a diferença real é igual à diferença amostral observada. Nesses casos, o tamanho do efeito alto juntamente com o poder do teste baixo dão indícios de que o teste não obteve evidências para rejeitar a hipótese nula pois a amostra era pequena para fornecer essa evidência.

O poder do teste também foi baixo para os casos do quarto grupo, onde os testes de Wilcoxon obtiveram significância estatística para a diferença entre os classificadores porém o tamanho do efeito para essa diferença é pequeno ou insignificante. Nesses casos, deve-se tomar cuidado para não valorizar a diferença, uma vez que o teste não é confiável pois tem poder baixo e o tamanho do efeito dessa diferença também é pequeno.

Assim como observado com os resultados do teste t, os resultados do teste de Wilcoxon mostram que todos os casos com poder do teste alto ($> 0{,}8$) pertencem ao Grupo 1, conforme visto na Figura 4.10. Ou seja, os casos com poder do teste alto obtiveram significância estatística no teste de Wilcoxon e tamanho do efeito médio ou grande. Para auxiliar a compreensão deste e de outros resultados, são apresentadas as Figuras 4.11 e 4.12. A Figura 4.11 contém o boxplot dos valores observados das diferenças médias amostrais por grupos e por tamanho da amostra. E a Figura 4.12 contém o boxplot dos valores dos desvios padrões observados nas diferenças amostrais por grupos e por tamanho da amostra.

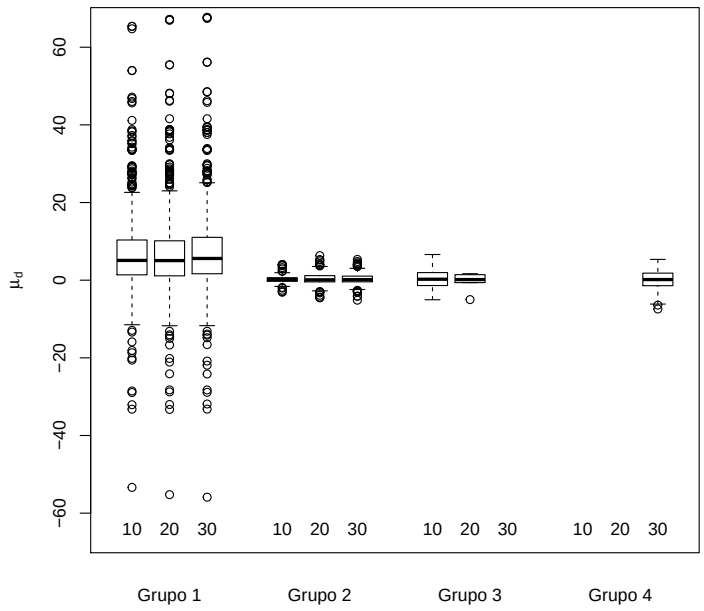


Figura 4.11: Boxplot das diferenças médias amostrais observadas por grupos do teste de Wilcoxon e tamanhos das amostras

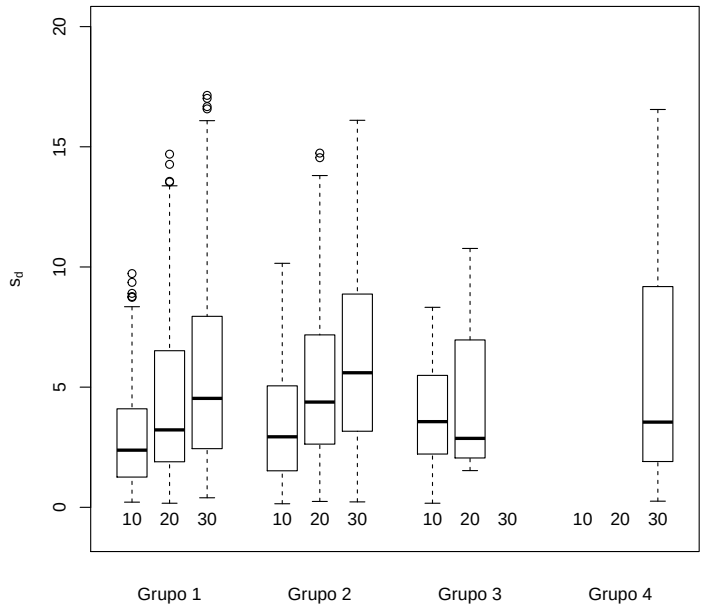


Figura 4.12: Boxplot dos desvios padrões das diferenças amostrais observadas por grupos do teste de Wilcoxon e tamanhos das amostras

Na Figura 4.11, é possível perceber que os maiores valores de diferenças médias amostrais observadas pertencem ao Grupo 1. E como era esperado, nos casos com grandes diferenças observadas, o teste obteve significância estatística, o tamanho do efeito foi médio ou maior e o poder do teste foi alto. Porém, não é apenas em casos com grandes diferenças médias observadas que estes resultados são observados.

No Grupo 3, por exemplo, mesmo com as diferenças amostrais pequenas, o tamanho do efeito foi médio ou grande. Já no Grupo 4, mesmo com as diferenças amostrais pequenas, o teste de Wilcoxon obteve significância estatística para afirmar que existe diferença entre os resultados dos classificadores.

A Figura 4.12 mostra que o Grupo 4 tem o desvio padrão amostral um pouco maior que o observado no Grupo 3. Apesar disso e de terem médias observadas muito pequenas e distribuições parecidas, o teste de Wilcoxon obteve evidências para rejeitar a hipótese nula no Grupo 4 e não obteve no Grupo 3. O que pode explicar a diferença do resultado do teste é o tamanho da amostra. Apesar de o desvio padrão ser um pouco maior no Grupo 4, o tamanho da amostra também é maior do que no Grupo 3, por isso, esse resultado foi observado.

Agora, serão realizadas análises semelhantes às feitas neste capítulo, porém para os resultados obtidos através dos testes aplicados aos dados simulados.

Capítulo 5

Dados Simulados

Neste capítulo, busca-se simular amostras de acurácias para ampliar (ou generalizar) algumas análises realizadas no capítulo anterior. Serão simulados diversos pares de amostras, representando acurácias de pares de classificadores em uma base de dados, nas quais serão aplicados os testes t e de Wilcoxon para avaliar as três medidas em destaque neste estudo: p-valor, tamanho do efeito e poder do teste.

Toda simulação e análise são realizadas utilizando o Software R. Na Seção 5.1 será descrito como essa simulação foi realizada e na Seção 5.2 serão analisados os resultados.

5.1 Descrição da Simulação

Para realizar a simulação de um par de amostras de acurácias obtidas pela aplicação de dois classificadores, A e B, por exemplo, em uma base de dados será utilizada a distribuição normal bivariada, ou distribuição binormal, que é uma distribuição conjunta para duas variáveis (X e Y) normais e dependentes. Essa distribuição é necessária, pois serão realizados testes para amostras pareadas e um teste paramétrico.

Para realizar a simulação de amostras com distribuição normal bivariada, é utilizada a função *mvrnorm* do pacote *MASS* do R. Essa função necessita do tamanho das amostras que serão geradas (referente ao método que se deseja simular: validação cruzada com 10, 20 ou 30 partições) e dos parâmetros: vetor de médias de X e Y e matriz de variâncias e covariâncias entre X e Y.

As médias de X e Y são definidas para representar as médias populacionais das acurácias dos classificadores A e B na base de dados. Entretanto, se esses valores forem muito próximos de 100, é provável que na amostra tenham valores maiores que 100, o que

não representa valores de acurácia. Então, para diminuir a chance de ter valores fora do intervalo de 0 a 100, as médias estarão sempre em torno do valor 50 com um intervalo de diferença entre elas, onde essa diferença é um valor uniformemente obtido entre 0,001 e 10. Além disso, é sorteada qual amostra será a maior.

Para gerar a matriz de variância e covariâncias, inicialmente são gerados dois valores uniformes entre 0,01 e 500 para representar as variâncias de X e Y e um valor uniforme entre -1 e 1 para representar a covariância entre X e Y. O valor máximo para a variância 500 foi definido como um valor superior ao valor da variância máxima observada nos dados reais, que foi aproximadamente 400. Essa margem de 100 foi dada para ampliar a possibilidade de casos observados.

Então, com os parâmetros da normal bivariada definidos, são gerados 4 pares de amostras. Cada par de amostras tem tamanho 10, 20, 30 e 100. Os 3 primeiros pares de amostras representam as acurácias obtidas com a aplicação de um par de classificadores em determinada base de dados através do método de validação cruzada com 10, 20 e 30 partições, respectivamente. E o par de amostras de tamanho 100 foi acrescentado para ampliar a análise, e assim, observar o comportamento das medidas de interesse para amostras maiores.

A partir de cada par de amostras, é aplicado o teste t e são obtidos o p-valor, o tamanho do efeito (através da medida $d'cohen$) e o poder do teste. Também é aplicado o teste de Wilcoxon e são obtidos o p-valor, o tamanho do efeito (através da medida r) e o poder do teste.

Porém, para cada conjunto de parâmetros definidos (médias e matriz de variância e covariância) são gerados dez pares de amostras para cada tamanho (10, 20, 30 e 100) desejado, a partir da normal bivariada com esses parâmetros, a fim de explorar diversas possíveis amostras de cada população. Além disso, nesta simulação, foram gerados 100 conjuntos de parâmetros, a fim explorar diversas possíveis populações com parâmetros diferentes. Então, são simulados 10.000 pares de amostras de cada tamanho desejado. Na próxima seção, são apresentadas as análises realizadas com os resultados dos testes aplicados a essas amostras simuladas.

5.2 Análise dos Resultados

Nesta seção, a fim de complementar os resultados obtidos no estudo empírico com os testes t e de Wilcoxon, é realizada a análise dos resultados obtidos através da aplicação

dos mesmos testes em dados simulados que buscam representar valores de acurácias de forma mais ampla.

Inicialmente, é possível verificar que o comportamento dos pontos que representam os resultados dos p-valores e das medidas de tamanho do efeito é semelhante ao comportamento observado com os dados reais. A Figura 5.1 apresenta os resultados observados para os p-valores e tamanhos do efeito para o teste t. Já a Figura 5.2 apresenta os mesmos resultados observados para o teste de Wilcoxon.

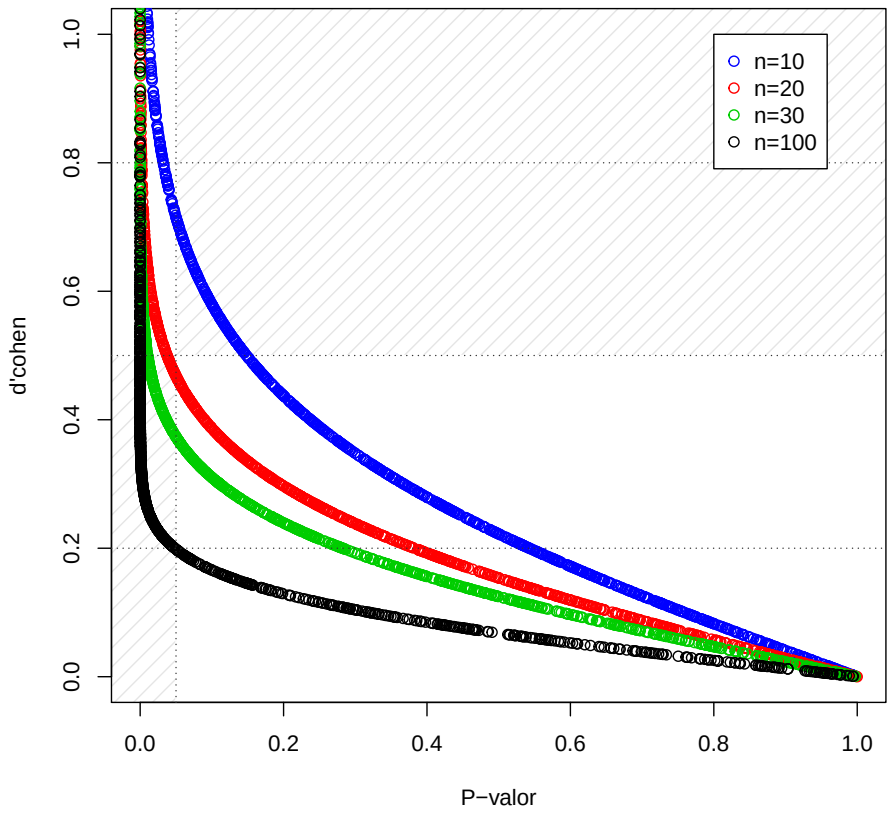


Figura 5.1: Representação dos resultados dos p-valores dos testes t aplicados aos dados simulados e suas respectivas medidas de tamanho do efeito obtidas por tamanho de amostra

Como já era esperado, as Figuras 5.1 e 5.2 são semelhantes às Figuras 4.1 e 4.7, respectivamente. Observam-se casos especiais em todos os tamanhos de amostras para ambos os testes. Então, para ampliar a análise, também é observado o comportamento do p-valor e tamanho do efeito para amostras de tamanho 100.

Apesar de não ser um tamanho de amostra observado no contexto de comparação

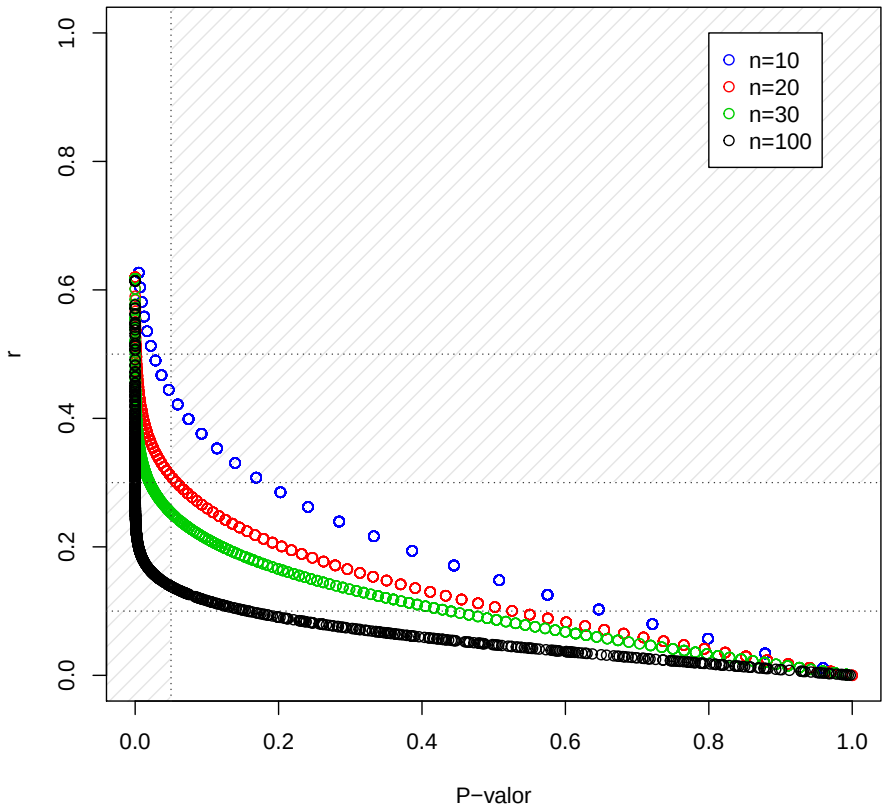


Figura 5.2: Representação dos resultados dos p-valores dos testes de Wilcoxon aplicados aos dados simulados e suas respectivas medidas de tamanho do efeito obtidas por tamanho de amostra

de classificadores, já que não é comum utilizar o método de validação cruzada com 100 partições, existe um motivo para a escolha desse valor. Com os tamanhos de amostras 10, 20 e 30 não foi observado nenhum caso onde o p-valor é menor que 0,05 e o tamanho do efeito é insignificante (menor que 0,2 para o d'cohen e menor que 0,1 para o r). Um caso como esse representaria uma discordância extrema, com significância estatística e tamanho do efeito insignificante, entre essas duas medidas e foi encontrado em amostras de tamanho 100.

Utilizando o teste de Wilcoxon e a medida r, não foi possível encontrar um caso com significância estatística e tamanho do efeito insignificante nem com tamanho de amostras 100, como pode ser visto na Figura 5.2.

As análises seguintes utilizam os mesmos grupos definidos na Subseção 4.2.3 para avaliar o poder dos testes realizados. Lembrando que os dois primeiros grupos têm resultados

concordantes entre o resultado do teste e do tamanho do efeito. O primeiro grupo tem significância estatística e tamanho do efeito médio ou grande, e o segundo grupo não tem significância estatística e tem tamanho do efeito pequeno ou insignificante. Já os dois últimos grupos são os casos especiais. O terceiro grupo não tem significância estatística e tem tamanho do efeito médio ou grande, e o quarto grupo tem significância estatística e tamanho do efeito pequeno ou insignificante.

A Figura 5.3 contém o boxplot dos valores obtidos na função poder do teste para os testes t aplicados aos dados simulados, separados por grupos e tamanhos das amostras (10, 20, 30 e 100).

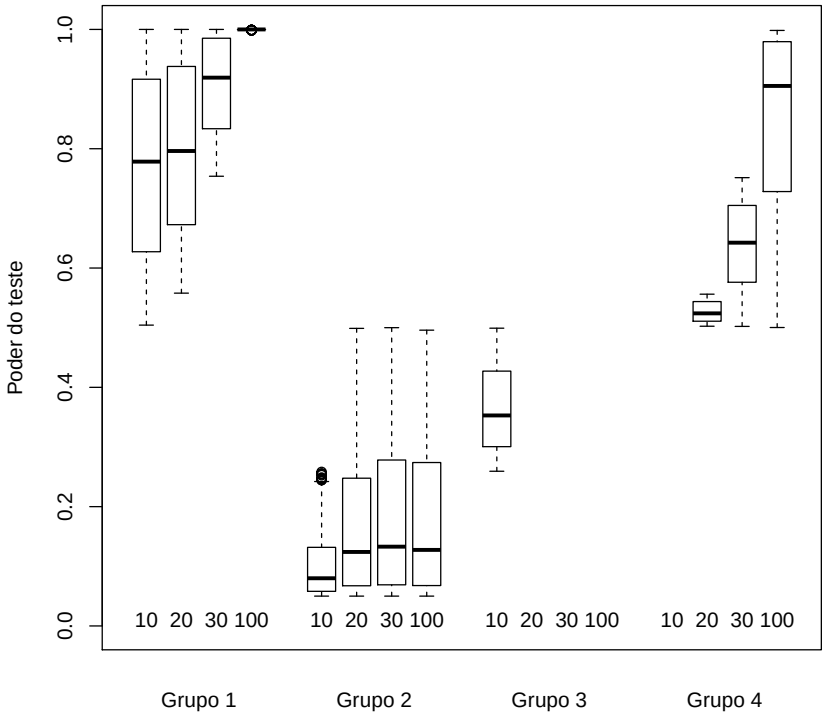


Figura 5.3: Boxplot dos valores obtidos na função poder do teste t por grupos e tamanhos das amostras simuladas

Comparando a Figura 5.3 à Figura 4.4, é possível verificar que o comportamento dessas medidas foi bastante semelhante. Então, os dados simulados reforçam as conclusões obtidas na Subseção 4.2.3.

No primeiro grupo, as três medidas se reforçam: o poder do teste alto dá credibilidade ao resultado do teste t, que foi afirmar que existe diferença estatisticamente significativa

entre os classificadores e o tamanho do efeito acrescenta que a magnitude dessa diferença é grande ou importante. Já no segundo grupo, independentemente de o p-valor e o tamanho do efeito estarem concordando, o poder do teste baixo sugere que a amostra deve ser ampliada para tornar o teste mais poderoso.

Para os casos especiais, Grupos 3 e 4, o poder do teste baixo indica que a probabilidade de o teste t rejeitar a hipótese nula, quando a diferença real é igual a diferença observada, é baixa. O poder baixo pode ser devido a uma diferença observada realmente muito pequena ou devido ao tamanho da amostra ser muito pequeno. No terceiro grupo, é possível que, aumentando o tamanho da amostra, o teste t consiga obter significância estatística para a diferença, já que o tamanho do efeito indica que a magnitude da diferença entre os classificadores é alta.

A fim de realizar análises semelhantes às feitas com o poder do teste t para dados simulados, a Figura 5.4 apresenta o boxplot dos valores obtidos na função poder do teste para os testes de Wilcoxon aplicados aos dados simulados, separados por grupos e tamanhos das amostras (10, 20, 30 e 100).

Comparando a Figura 5.4 à Figura 4.10, é possível verificar que o comportamento do poder do teste de Wilcoxon para dados reais e para dados simulados não foi tão semelhante quanto no teste t com resultados reais e simulados. As maiores diferenças são observadas nos Grupos 1 e 4. No Grupo 1, a grande maioria dos casos reais possuem poder alto, enquanto nos dados simulados, a maioria dos valores do poder do teste estão abaixo de 0,8, ou seja, poder do teste baixo. Já no Grupo 4, o poder do teste para os casos reais sempre foi baixo, enquanto, para os dados simulados, alguns casos obtiveram poder alto.

A diferença observada entre o Grupo 4 com dados reais e com dados simulados foi muito pequena e pode ser explicada por ter uma maior quantidade e variedade de casos nos dados simulados. Porém, apesar de a diferença ter sido muito pequena, a simulação de amostras representando acurácias encontrou um novo resultado que não tinha sido observado nos testes com dados reais.

Esse novo resultado são casos onde o teste de Wilcoxon obteve significância estatística, tamanho do efeito baixo e poder do teste alto. Nesses casos, apesar de o p-valor e tamanho do efeito serem discordantes, o poder do teste confirma que esse resultado é confiável. Casos discordantes realmente podem acontecer e, ao concluir sua pesquisa, o pesquisador deve deixar claro os valores dessas medidas.

Já no Grupo 1, apesar de a diferença da distribuição do poder do teste para os dados

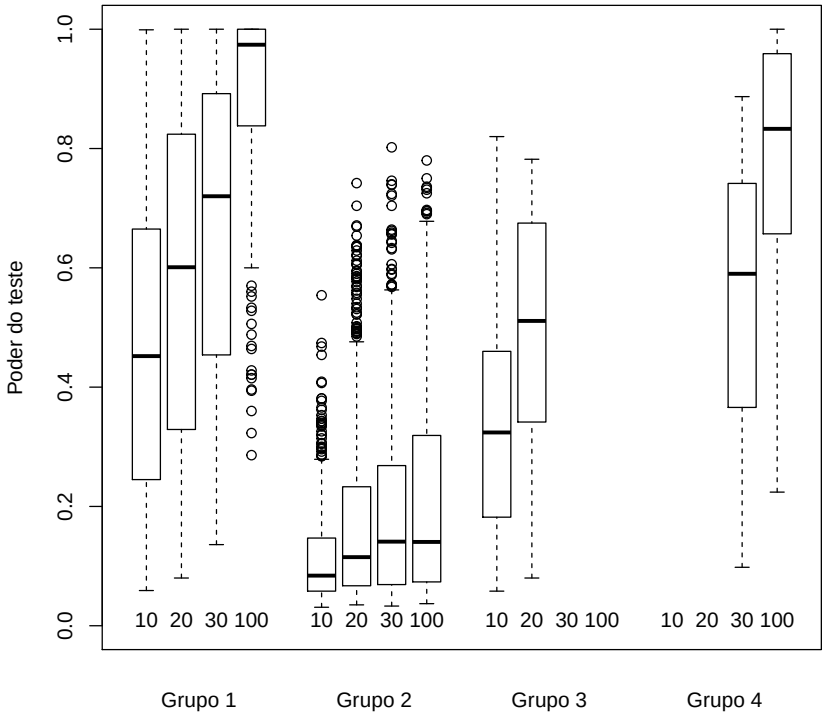


Figura 5.4: Boxplot dos valores obtidos na função poder do teste de Wilcoxon por grupos e tamanhos das amostras simuladas

simulados ter sido bastante diferente da distribuição do poder para os dados reais, não foi observado nenhum caso novo. Nos dados reais, a maioria dos casos tiveram poder do teste alto, porém alguns casos tiveram o poder do teste baixo. E nos dados simulados, a maioria dos casos tiveram poder do teste baixo e alguns tiveram poder do teste alto.

Os resultados observados nos Grupos 2 e 3 para dados simulados têm o mesmo comportamento dos resultados observados com os dados reais. Então, os dados simulados reforçam as conclusões obtidas na Subseção 4.3.3 para esses grupos. No segundo grupo, independentemente de o p-valor e o tamanho do efeito estarem concordando, o poder do teste baixo sugere que a amostra deve ser ampliada para tornar o teste mais poderoso. No Grupo 3, o tamanho do efeito alto juntamente com o poder do teste baixo dão indícios de que o teste não obteve evidências para rejeitar a hipótese nula pois a amostra era muito pequena para fornecer essa evidência.

Através das Figuras 5.3 e 5.4, é possível verificar que o poder do teste realmente é maior em amostras maiores. Em geral, as amostras de tamanho 100 apresentam os

maiores poderes observados em cada grupo.

Para colaborar com as análises que estão sendo realizadas, são apresentados os boxplots das médias e dos desvios padrões das diferenças amostrais observadas em cada teste, separadas por grupo e tamanho das amostras. As Figuras 5.5 e 5.6 apresentam os boxplot das diferenças médias amostrais por grupos e tamanhos das amostras simuladas, para os testes t e de Wilcoxon, respectivamente. E as Figuras 5.7 e 5.8 apresentam os boxplot dos desvios padrões das diferenças amostrais por grupos e tamanhos das amostras simuladas, para os testes t e de Wilcoxon, respectivamente.

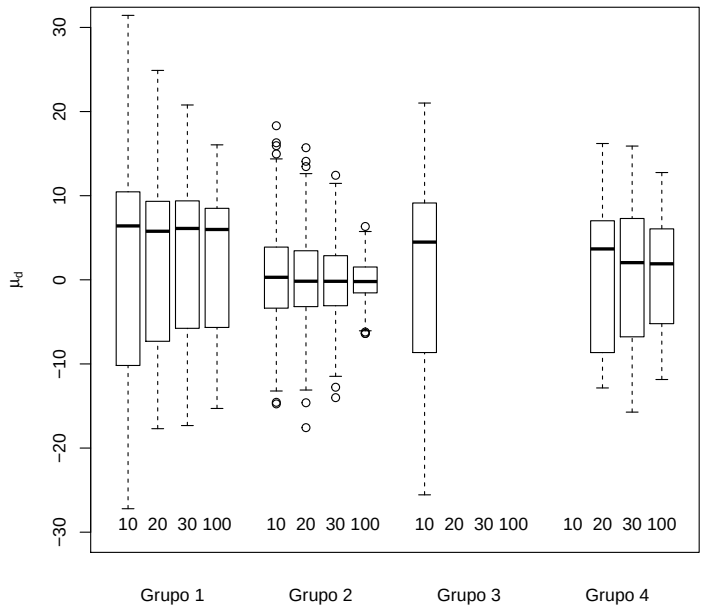


Figura 5.5: Boxplot das diferenças médias amostrais por grupos do teste t e tamanhos das amostras simuladas

Pode ser visto, nas Figuras 5.5 e 5.6, que as médias das diferenças amostrais para os testes realizados com dados simulados não tiveram a mesma distribuição entre os grupos, como foi observado nos resultados com dados reais (Figuras 4.5 e 4.11). Com os dados reais, as maiores diferenças pertenciam apenas ao Grupo 1 (em ambos os testes). Porém, com os dados simulados, existem casos com grandes diferenças em todos os grupos.

Analisando as Figuras 5.7 e 5.8, é possível ver que os desvios padrões dos casos simulados são maiores do que os observados com os casos reais. Já que, como apresentado na Seção 5.1, isso foi feito para ampliar a possibilidade de casos observados. Logo, esta diferença pode auxiliar a compreensão dos novos casos observados com os dados simulados.

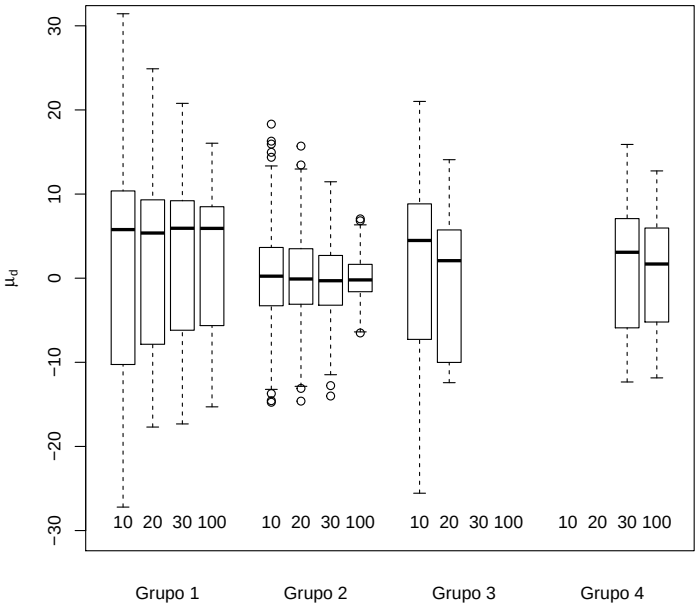


Figura 5.6: Boxplot das diferenças médias amostrais por grupos do teste de Wilcoxon e tamanhos das amostras simuladas

Levando-se em conta os resultados observados obtidos a partir dos dados simulados, que reforçam os resultados a partir de dados reais, fica clara a importância da análise de uma medida de tamanho do efeito e da função poder do teste sempre que for realizado um teste de hipóteses. Tendo conhecimento das três medidas abordadas neste trabalho, a tomada de decisão de um teste de hipóteses é mais fundamentada e responsável do que quando analisado apenas o p-valor isoladamente. E assim, reduz-se o risco de valorizar um resultado sem grande importância ou, até mesmo, deixar de reconhecer um resultado que pode ser importante para o desenvolvimento de algum estudo.

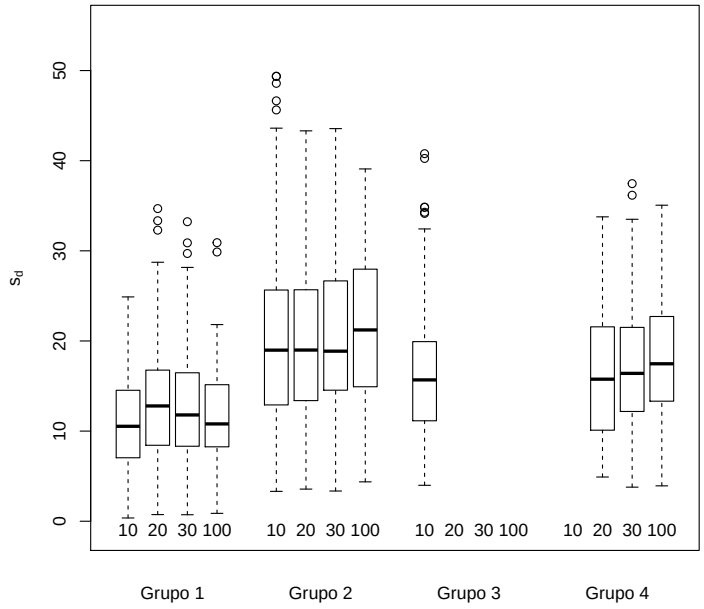


Figura 5.7: Boxplot dos desvios padrões das diferenças amostrais por grupos do teste t e tamanhos das amostras simuladas

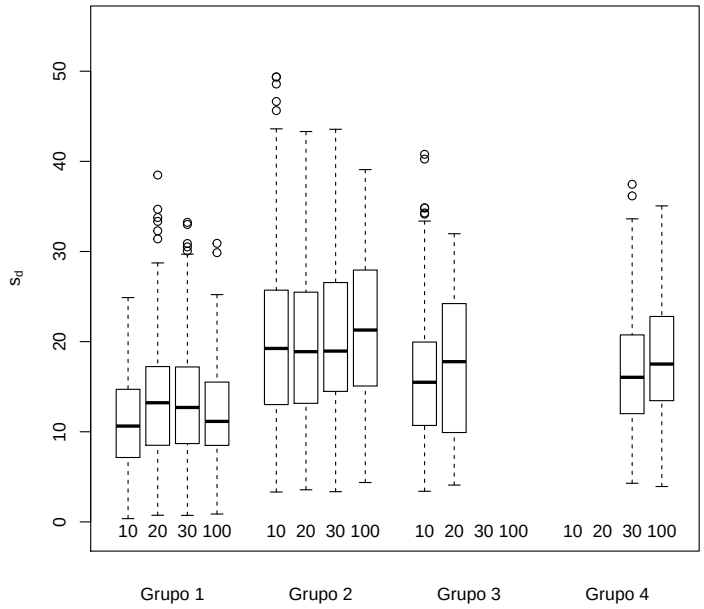


Figura 5.8: Boxplot dos desvios padrões das diferenças amostrais por grupos do teste de Wilcoxon e tamanhos das amostras simuladas

Capítulo 6

Conclusão

A avaliação de classificadores através da análise da significância estatística é de extrema importância para as áreas de Aprendizado de Máquina e Mineração de Dados, visto que, através dela, é possível verificar se um classificador tem, de fato, resultados melhores do que outro. A significância estatística é verificada por meio de algum teste de hipóteses que busca evidências para rejeitar uma hipótese conservadora, como por exemplo que dois classificadores têm resultados semelhantes.

Entretanto, apesar de ser uma ótima ferramenta, se mal realizado, o teste de hipóteses pode levar a conclusões equivocadas. A análise do p-valor isoladamente é um problema discutido em diversas áreas ([20], [24], [26], [28]), já que a busca pelo p-valor<0,05 é cada vez mais cegamente realizada. No levantamento feito nesta dissertação, apresentado no Capítulo 1, foi visto que todos os 11 artigos que utilizam teste de hipóteses para avaliar classificadores tiveram a decisão embasada apenas em p-valores.

Dado que a má utilização do p-valor é um problema presente na comparação de classificadores, é possível que alguns pesquisadores estejam valorizando resultados sem grande importância real ou, até mesmo, deixando de reconhecer resultados que podem contribuir com o avanço de pesquisas. A fim de ilustrar possíveis problemas e soluções, nesta dissertação foram realizados um estudo empírico e um estudo com dados simulados.

Os resultados obtidos através dos estudos (empírico e com dados simulados) reforçam que alguns problemas, já levantados em outras áreas, podem afetar a conclusão na comparação de classificadores. Através dos testes de hipóteses t de Student e de Wilcoxon, aplicados em amostras de diferentes tamanhos (obtidas através do método de validação cruzada com 10, 20 e 30 partições), foi verificado que o p-valor é sensível ao aumento do tamanho da amostra. Por outro lado, foi verificado que o tamanho do efeito tende a não

sofrer alteração no resultado com o aumento da amostra.

Além disso, cabe ressaltar que o termo significância estatística não quer dizer significância prática. Por isso, foram verificados casos em que o p-valor indica que a diferença entre os classificadores é estatisticamente significante, porém o tamanho do efeito para essa diferença é pequeno. Também foram encontrados casos com situação oposta, sem significância estatística para a diferença entre os classificadores, porém com tamanho do efeito médio. Esses casos, onde o p-valor e o tamanho do efeito indicam resultados diferentes, foram chamados de casos especiais.

Os casos especiais não representam necessariamente um erro, eles merecem atenção especial pois ilustram que se um pesquisador tivesse verificado apenas o p-valor, estaria ignorando uma informação importante que poderia mudar a sua conclusão. No estudo empírico, foi verificado que a frequência de vezes em que esses casos especiais ocorreram não deve ser ignorada. Nos testes t aplicados em amostras de tamanhos 10, 20 e 30, os casos especiais representam 8,8%, 1,4% e 9,6% do total de teste realizados em cada tamanho de amostra, respectivamente. Já nos testes de Wilcoxon, esses valores foram 15,72%, 1,08% e 7,0%, respectivamente.

Além de destacar a importância do cálculo do tamanho do efeito, os valores apresentados anteriormente ilustraram que os casos especiais ocorrem com menor frequência em amostras de tamanho 20, tanto para os dados reais quanto para os dados simulados. Portanto, se o pesquisador deseja que haja uma maior chance de obter similaridade entre as significâncias estatística e prática, poderia optar por realizar o método de validação cruzada com 20 partições. Vale destacar que ao fazer essa escolha (de trabalhar com amostra de tamanho 20), não está descartada a importância do cálculo de uma medida de tamanho do efeito.

Outro problema de realizar um teste de hipóteses e calcular apenas o p-valor é ignorar o poder do teste. O poder do teste é a probabilidade de o teste rejeitar a hipótese nula (afirmar que os classificadores são diferentes) dado o real valor dessa diferença. Então, quando realizado um teste de hipóteses com poder baixo, a sua conclusão não é confiável.

Os resultados dos estudos realizados nesta dissertação mostram a importância do cálculo do poder do teste. Primeiro, deve-se destacar que no contexto de comparação de classificadores, geralmente são utilizadas amostras pequenas para realizar um teste de hipóteses. Por isso, foram encontrados muitos casos com poder do teste baixo. Então, esse resultado evidencia que ignorar o poder do teste neste contexto pode fazer com que muitos pesquisadores obtenham resultados não confiáveis.

O cálculo do poder do teste também forneceu mais informações para os casos especiais. O poder do teste baixo em um caso sem significância estatística e com tamanho do efeito médio pode sugerir uma pesquisa futura com maior poder, ou seja, com uma amostra maior. Com isso, é possível concluir que um pesquisador não deve desistir do seu estudo por não ter encontrado significância estatística por meio de um teste com poder baixo, pois pode representar a perda de um resultado importante, como indicado pelo tamanho do efeito.

Entretanto, também é possível obter o poder do teste alto em um caso com significância estatística e com tamanho do efeito baixo. Com esses resultados, é possível concluir que provavelmente o teste foi realizado em uma amostra grande, já que foi possível obter evidências para rejeitar a hipótese nula mesmo com uma pequena diferença. Vale lembrar o caso real da aspirina apresentado no Capítulo 1, obtido em [26]. Casos como esse, reforçam que o teste de significância estatística não deve ser o responsável em validar ou descartar pesquisas científicas [29]. O resultado do teste indica um caminho, mas o pesquisador deve ter outras medidas, como o tamanho do efeito, para não correr o risco de valorizar um resultado sem importância.

Como trabalhos futuros, serão apontadas três linhas de trabalho: (i) ampliar o estudo realizado para outros testes de hipóteses paramétricos e não paramétricos, como por exemplo o teste de Friedman e ANOVA, que teriam como objetivo, no contexto deste estudo, comparar o desempenho de diferentes classificadores para um conjunto de bases de dados; (ii) utilizar outra forma de amostragem das acurácias que serão utilizadas nos testes de hipóteses, como por exemplo multiplas validações cruzadas; e (iii) realizar um estudo teórico para justificar o fato de o tamanho de amostra 20 ter gerado um número significativamente menor de casos especiais, o que reforçaria a recomendação desse tamanho de amostra quando a concordância entre significância estatística e prática for um aspecto importante da análise.

Dadas as evidências apresentadas e discutidas ao longo de toda a dissertação, espera-se estimular a apresentação do poder do teste e de alguma medida de tamanho do efeito nas pesquisas das áreas de Aprendizado de Máquina e Mineração de Dados, a fim de tornar as conclusões estatísticas mais fundamentadas.

Referências

[1] Berben, L.; Sereika, S. M.; Engberg, S. Effect size estimation: methods and examples. *International Journal of Nursing Studies 49*, 8 (2012), 1039–1047.

[2] Breiman, L. Random forests. *Machine Learning 45*, 1 (2001), 5–32.

[3] Bussab, W. O.; Morettin, P. *Estatística Básica, Editora Saraiva, 6a. edição.* 2010.

[4] Coelho Barros, E. A.; Mazucheli, J. Um estudo sobre o tamanho e poder dos testes t-student e wilcoxon. *Acta Scientiarum: Technology 27*, 1 (2005), 23–32.

[5] Cohen, J. *Statistical power analysis for the behavioral sciences*, 2nd ed. Hillsdale, NJ: erlbaum, 1988.

[6] Conboy, J. E. Algumas medidas típicas univariadas da magnitude do efeito. *Análise Psicológica 21*, 2 (2012), 145–158.

[7] Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE transactions on information theory 13*, 1 (1967), 21–27.

[8] Dheeru, D.; Karra Taniskidou, E. UCI machine learning repository, 2017.

[9] Eibe Frank, Mark A. Hall, and Ian H. Witten. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*, 4th ed. Morgan Kaufmann, 2016.

[10] Fern, E. F.; Monroe, K. B. Effect-size estimates: Issues and problems in interpretation. *Journal of Consumer Research 23*, 2 (1996), 89–105.

[11] Fisher, R. A. *Statistical methods for research workers.* Springer, 1925.

[12] Fritz, C. O.; Morris, P. E.; Richler, J. J. Effect size estimates: current use, calculations, and interpretation. *Journal of Experimental Psychology: General 141*, 1 (2012), 2–18.

[13] Hair, J. F.; Black, W. C.; Babin, B. J.; Anderson, R. E.; Tatham, R. L. *Análise multivariada de dados.* Bookman Editora, 2009.

[14] Hearst, M. A.; Dumais, S. T.; Osuna, E.; Platt, J.; Scholkopf, B. Support vector machines. *IEEE Intelligent Systems and Their Applications 13*, 4 (1998), 18–28.

[15] Jacobson, N. S.; Truax, P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology 59*, 1 (1991), 12.

[16] Japkowicz, N.; Shah, M. *Evaluating learning algorithms: a classification perspective.* Cambridge University Press, 2011.

[17] Kirk, R. E. Practical significance: A concept whose time has come. *Educational and Psychological Measurement 56*, 5 (1996), 746–759.

[18] Kohavi, R., et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI* (1995), vol. 14, Montreal, Canada, pp. 1137–1145.

[19] Langley, P.; Iba, W.; Thompson, K., et al. An analysis of bayesian classifiers. In *Aaai* (1992), vol. 90, pp. 223–228.

[20] Nakagawa, S.; Cuthill, I. C. Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews 82*, 4 (2007), 591–605.

[21] Pinheiro, J.; Gomes, G.; Carvajal, S.; Cunha, S. *Probabilidade e estatística: quantificando a incerteza.* Elsevier Brasil, 2013.

[22] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2018.

[23] Santo, H. E.; Daniel, F. B. Calcular e apresentar tamanhos do efeito em trabalhos científicos (1): As limitações do p<0,05 na análise de diferenças de médias de dois grupos. *Revista Portuguesa de Investigação Comportamental e Social 1*, 1 (2015), 3–16.

[24] Sharpe, D. Beyond significance testing: Reforming data analysis methods in behavioral research. 317–319.

[25] Snyder, P.; Lawson, S. Evaluating results using corrected and uncorrected effect size estimates. *The Journal of Experimental Education 61*, 4 (1993), 334–349.

[26] Sullivan, G. M.; Feinn, R. Using effect size-or why the p-value is not enough. *Journal of Graduate Medical Education 4*, 3 (2012), 279–282.

[27] Tallmadge, G. K. The joint dissemination review panel ideabook.

[28] Tomczak, M.; Tomczak, E. The need to report effect size estimates revisited. an overview of some recommended measures of effect size. *Trends in Sport Sciences 21*, 1 (2014), 19–25.

[29] Wasserstein, R. L.; Lazar, N. A. The asa's statement on p-values: context, process, and purpose. *The American Statistician* (2016).

[30] Wonnacott, R. J.; Wonnacott, T. H. *Fundamentos de Estatística: descobrindo o poder da estatística.* Livros Técnicos e Científicos, 1985.

APÊNDICE A – Resultados Obtidos no Estudo Empírico com o Teste t

Tabela A.1: P-valores obtidos com a aplicação do teste t para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 10

Bases	Pares de Classificadores														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.66		0.77	0.48	0.91		0.92	0.30	0.74				0.31	0.69	0.67
2	0.18		0.86	0.16	0.01		0.61	0.58	0.01				0.19	0.02	0.05
3		0.11	0.01	0.17	0.08					0.13	0.51	0.53	0.05	0.05	0.99
4	1.00	0.00	0.00		0.14	0.00	0.00		0.16	0.00		0.00		0.01	
5	0.71		0.00	0.00	0.00		0.00	0.00	0.00				0.30	0.02	0.03
6		0.03		0.00	0.00						0.07	0.04			0.16
7															
8															
9	0.56	0.14	0.00	0.00	0.00	0.11	0.00	0.00	0.01	0.00	0.00	0.00	0.03	0.01	0.28
10													0.00	0.68	0.24
11	0.11	0.59	0.07	1.00	0.04	0.25	0.26	0.58	0.01	0.14	0.68	0.02	0.14	0.00	0.06
12	1.00	0.12	0.50	0.26	0.73	0.07	0.52	0.33	0.68	0.08	0.42	0.01	0.02	1.00	0.16
13															
14													0.01		
15										0.00	0.00	0.34	0.42	0.00	0.00
16															
17	0.36		0.01	0.00	0.00		0.03	0.01	0.00				0.93	0.02	0.00
18	0.08	0.01	0.00	0.00		0.01	0.00	0.00		0.36	0.36				
19	0.34	0.17	0.01	0.41	0.00	0.30	0.00	0.56	0.00	0.02	0.88	0.00	0.00	0.05	0.00
20	0.59		0.44	0.32	0.51		0.56	0.37	0.75				0.76	0.68	0.47
21	0.45		0.03	0.33	0.09		0.03	0.26	0.06				0.00	0.01	0.73
22	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.04	0.00	0.02	0.28
23	1.00	0.11	0.02	0.17	0.48	0.08	0.01	0.07	0.53	0.00	0.01	0.34	0.03	0.02	0.11
24															0.00
25					0.00										
26	0.59	0.94	0.01	0.03	0.48	0.86	0.01	0.02	0.43	0.02	0.12	0.32	0.05	0.09	0.43
27	0.02	0.39	0.00	0.00	0.50	0.89	0.00	0.00	0.15	0.00	0.00	0.20	0.05	0.00	0.00
28	0.59	0.38	0.00	0.00	0.00	0.37	0.00	0.00	0.00	0.00	0.03	0.00	0.06	0.01	0.00
29												0.00			
30	0.28	0.68	0.01	0.05	0.00	0.88	0.12	0.41	0.00	0.20	0.35	0.03	0.26	0.33	0.03
31	0.32	0.09	0.00	0.00	0.66	0.11	0.00	0.00	0.80	0.10	0.16	0.01	0.44	0.03	0.04
32	0.14	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.09	0.68	0.02	0.00	0.01
33	0.10	0.07	0.65	0.65	0.00	0.03	0.36	0.36	0.00	0.20	0.20	0.00		0.00	0.00
34	0.92	0.84	0.00	0.03	0.00	0.91	0.00	0.02	0.00	0.00	0.02	0.00	0.11	0.00	0.00
35															
36	1.00		0.00	0.46	0.00		0.00	0.53	0.00				0.00	0.00	0.00
37													0.00	0.01	0.00
38	0.93	0.23	0.00	0.00	0.00	0.22	0.00	0.00	0.00	0.00	0.01	0.00	0.03	0.82	0.06
39	0.82	0.20	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.49	0.01
40	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.22	0.55	0.00	0.11	0.00	0.00
41	0.28		0.00	0.00	0.00		0.00	0.00	0.00				0.11	0.00	0.00
42	0.79		0.00	0.23	0.00		0.00	0.07	0.00				0.00	0.00	0.00
43	0.81	0.00	0.04	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.00
44	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.66	0.00	0.00
45	0.82	0.03	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.92	0.02	0.04
46	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
47	0.72	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
48						0.00	0.04	0.01	0.00	0.00	0.00	0.00	0.77	0.00	0.00
49	0.37	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00
50	0.00		0.00	0.00			0.00	0.00					0.03		

Tabela A.2: P-valores obtidos com a aplicação do teste t para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 20

Bases	Pares de Classificadores														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.84		0.49	0.21	0.32		0.52	0.26	0.45				0.75	0.90	0.63
2	0.76		0.87	0.19	0.00		0.99	0.26	0.00				0.20	0.01	0.01
3		0.22	0.01	0.29	0.67					0.12	0.57	0.47	0.05	0.04	0.77
4	0.06	0.00	0.00		0.11	0.00	0.00		0.01	0.00		0.00		0.00	
5	0.08		0.00	0.00	0.00		0.00	0.00	0.00				0.93	0.07	0.05
6		0.00		0.00	0.00						0.14	0.02			0.17
7															
8															
9	0.37	0.08	0.00	0.00	0.01	0.12	0.00	0.00	0.01	0.00	0.00	0.00	0.03	0.03	0.32
10													0.00	0.25	0.06
11	0.11	0.20	0.27	0.95	0.03	0.54	0.02	0.48	0.10	0.06	0.30	0.31	0.37	0.00	0.05
12	0.77	0.01	0.67	0.12	0.87	0.02	0.48	0.13	0.68	0.02	0.25	0.00	0.00	0.87	0.12
13															
14													0.00		
15										0.00	0.00	0.33	0.54	0.00	0.00
16															
17	0.66		0.00	0.00	0.00		0.01	0.00	0.00				0.91	0.00	0.00
18	0.42	0.00	0.00	0.00		0.00	0.00	0.00		0.78	0.78				
19	0.43	0.07	0.00	0.25	0.00	0.02	0.00	0.12	0.00	0.02	0.64	0.00	0.00	0.01	0.00
20	0.33		0.33	0.43	0.63		0.59	0.71	1.00				0.80	0.54	0.68
21	0.80		0.01	0.18	0.16		0.02	0.16	0.17				0.00	0.01	0.92
22	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.02	0.00	0.00	0.55
23	0.13	0.39	0.01	0.08	1.00	0.14	0.01	0.22	0.54	0.00	0.01	0.25	0.03	0.01	0.16
24															0.00
25					0.00										
26	0.26	1.00	0.01	0.11	0.21	0.68	0.00	0.04	0.11	0.01	0.07	0.04	0.15	0.22	0.79
27	0.33	0.41	0.00	0.00	0.68	0.78	0.00	0.00	0.42	0.00	0.00	0.23	0.30	0.00	0.00
28	0.26	0.82	0.00	0.00	0.00	0.47	0.00	0.00	0.00	0.00	0.00	0.00	0.40	0.02	0.00
29												0.00			
30	0.73	0.94	0.29	0.94	0.18	0.95	0.18	0.92	0.15	0.45	0.99	0.05	0.32	0.41	0.17
31	0.12	0.41	0.00	0.00	1.00	0.16	0.00	0.00	0.65	0.00	0.09	0.16	0.03	0.00	0.04
32	0.85	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.44	0.01	0.00	0.02
33	0.26	0.14	0.84	0.84	0.00	0.06	0.56	0.56	0.00	0.18	0.18	0.00		0.00	0.00
34	0.78	1.00	0.00	0.01	0.00	0.90	0.00	0.01	0.00	0.00	0.01	0.00	0.01	0.00	0.00
35															
36	0.60		0.00	1.00	0.00		0.00	0.75	0.00				0.00	0.00	0.00
37													0.00	0.00	0.00
38	0.36	0.07	0.00	0.00	0.00	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.72	0.02
39	0.53	0.85	0.00	0.01	0.00	0.60	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.68	0.04
40	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.22	0.55	0.00	0.04	0.00	0.00
41	0.82		0.00	0.00	0.00		0.00	0.00	0.00				0.11	0.00	0.00
42	0.55		0.00	0.02	0.00		0.00	0.07	0.00				0.00	0.00	0.00
43	0.89	0.00	0.02	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00
44	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.75	0.00	0.00
45	0.78	0.04	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.73	0.04	0.01
46	0.34	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
47	0.91	0.00	0.08	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
48						0.00	0.13	0.02	0.00	0.00	0.00	0.00	0.52	0.00	0.00
49	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00
50	0.18		0.00	0.00			0.00	0.00					0.04		

Tabela A.3: P-valores obtidos com a aplicação do teste t para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 30

Bases	Pares de Classificadores														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.15		0.73	0.36	0.51		0.33	0.11	0.23				0.61	0.76	0.99
2	0.37		0.50	0.61	0.04		0.83	0.35	0.02				0.20	0.01	0.04
3		0.41	0.04	0.80	0.38					0.20	0.39	0.09	0.05	0.02	0.28
4	0.84	0.00	0.00		0.07	0.00	0.00		0.04	0.00		0.00		0.00	
5	0.09		0.00	0.00	0.00		0.00	0.00	0.00				0.69	0.01	0.02
6		0.00		0.00	0.00						0.13	0.05			0.15
7															
8															
9	0.98	0.08	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.25
10													0.00	0.19	0.18
11	0.67	0.48	0.18	0.23	0.02	0.38	0.30	0.34	0.02	0.11	0.14	0.00	0.99	0.00	0.01
12	0.10	0.01	0.80	0.18	0.09	0.00	0.44	0.03	0.38	0.02	0.19	0.00	0.03	0.17	0.01
13															
14													0.00		
15										0.00	0.00	0.33	0.48	0.00	0.00
16															
17	0.79		0.01	0.00	0.00		0.00	0.00	0.00				0.57	0.00	0.00
18	0.80	0.00	0.00	0.00		0.01	0.00	0.00		0.87	0.87				
19	0.71	0.35	0.00	0.54	0.00	0.30	0.00	0.41	0.00	0.01	0.67	0.00	0.00	0.02	0.00
20	0.33		0.44	0.39	0.63		0.79	0.67	0.99				0.80	0.82	0.70
21	0.18		0.01	0.47	0.47		0.00	0.68	0.61				0.01	0.02	0.75
22	0.36	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.29	0.02	0.00	0.00	0.12
23	0.98	0.54	0.00	0.01	0.94	0.61	0.00	0.02	0.93	0.00	0.00	0.49	0.05	0.01	0.10
24															0.00
25					0.00										
26	0.80	0.63	0.00	0.01	0.06	0.61	0.00	0.01	0.06	0.04	0.09	0.00	0.29	0.40	0.87
27	0.41	0.91	0.00	0.00	0.40	0.61	0.00	0.00	0.19	0.00	0.00	0.41	0.23	0.00	0.00
28	0.99	0.30	0.00	0.00	0.00	0.34	0.00	0.00	0.00	0.00	0.01	0.00	0.43	0.00	0.00
29												0.00			
30	0.64	0.82	0.09	0.50	0.05	0.95	0.12	0.67	0.07	0.44	0.80	0.04	0.41	0.30	0.08
31	0.32	0.34	0.00	0.00	0.90	0.15	0.00	0.00	0.80	0.00	0.03	0.20	0.09	0.00	0.01
32	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.57	0.04	0.00	0.06
33	0.26	0.04	0.91	0.91	0.00	0.02	0.61	0.61	0.00	0.07	0.07	0.00		0.00	0.00
34	0.54	0.40	0.00	0.02	0.00	0.63	0.00	0.01	0.00	0.00	0.00	0.00	0.08	0.00	0.00
35															
36	0.81		0.00	0.16	0.00		0.00	0.31	0.00				0.00	0.00	0.00
37													0.00	0.00	0.00
38	0.79	0.08	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.91	0.15
39	0.67	0.36	0.00	0.00	0.00	0.26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.63	0.01
40	0.36	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.71	0.00	0.05	0.00	0.00
41	0.66		0.00	0.00	0.00		0.00	0.00	0.00				0.03	0.00	0.00
42	0.12		0.00	0.09	0.00		0.00	0.03	0.00				0.00	0.00	0.00
43	0.19	0.00	0.05	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
44	0.66	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.40	0.00	0.00
45	0.17	0.04	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.46	0.10	0.07
46	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
47	0.10	0.00	0.08	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
48						0.00	0.04	0.01	0.00	0.00	0.00	0.00	0.43	0.00	0.00
49	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00
50	0.30		0.00	0.00			0.00	0.00					0.05		

Tabela A.4: Medidas de tamanho do efeito obtidos com o *d'cohen* para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 10

Bases	Pares de Classificadores														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.15		0.10	0.23	0.04		0.03	0.35	0.11				0.34	0.13	0.14
2	0.46		0.06	0.48	1.06		0.17	0.18	1.01				0.45	0.85	0.72
3		0.56	0.96	0.47	0.62					0.53	0.22	0.21	0.74	0.73	0.00
4	0.00	3.90	2.02		0.52	2.89	1.90		0.48	1.39		2.35		1.03	
5	0.12		1.96	1.62	2.31		2.30	1.82	1.92				0.35	0.86	0.81
6		0.85		1.26	1.85						0.66	0.76			0.48
7															
8															
9	0.19	0.51	2.31	1.78	1.20	0.56	2.47	1.85	0.98	2.23	1.70	1.37	0.85	1.05	0.37
10													1.46	0.13	0.40
11	0.56	0.18	0.65	0.00	0.75	0.39	0.38	0.18	1.12	0.52	0.14	0.91	0.51	1.19	0.67
12	0.00	0.55	0.22	0.38	0.11	0.64	0.21	0.33	0.13	0.63	0.27	0.99	0.91	0.00	0.49
13															
14													0.97		
15										2.71	3.05	0.32	0.27	2.77	2.97
16															
17	0.31		1.00	1.74	2.93		0.82	1.02	3.28				0.03	0.93	1.39
18	0.62	1.17	1.90	1.90		1.11	1.68	1.68		0.31	0.31				
19	0.32	0.47	1.17	0.27	1.79	0.35	1.24	0.19	1.80	0.88	0.05	1.42	1.28	0.71	1.53
20	0.18		0.25	0.33	0.22		0.19	0.30	0.10				0.10	0.13	0.24
21	0.25		0.81	0.33	0.61		0.79	0.38	0.68				1.17	1.14	0.11
22	0.60	6.23	7.93	6.29	7.85	7.14	7.54	6.76	7.01	1.91	0.84	0.76	1.62	0.89	0.36
23	0.00	0.56	0.86	0.47	0.23	0.63	1.10	0.65	0.20	1.55	1.14	0.32	0.81	0.91	0.56
24															3.37
25					8.64										
26	0.17	0.03	1.05	0.82	0.23	0.06	1.07	0.92	0.26	0.92	0.55	0.33	0.71	0.61	0.26
27	0.85	0.28	2.94	4.13	0.22	0.04	3.16	4.53	0.50	3.58	5.34	0.44	0.71	2.41	2.87
28	0.18	0.29	2.13	1.65	2.82	0.30	2.42	1.79	2.72	1.22	0.83	1.98	0.67	1.15	1.54
29												5.74			
30	0.36	0.14	1.06	0.72	1.26	0.05	0.54	0.28	1.21	0.43	0.31	0.83	0.38	0.33	0.80
31	0.33	0.60	1.88	2.33	0.15	0.57	1.86	2.90	0.08	0.59	0.49	0.97	0.26	0.80	0.74
32	0.51	1.07	3.55	1.95	1.31	1.10	3.86	1.84	1.32	1.26	0.60	0.13	0.90	1.66	1.02
33	0.59	0.64	0.15	0.15	3.12	0.81	0.30	0.30	3.26	0.44	0.44	3.23		3.33	3.33
34	0.03	0.07	1.52	0.80	3.13	0.04	1.18	0.88	3.47	2.12	0.95	2.81	0.55	1.82	2.12
35															
36	0.00		1.94	0.24	8.73		1.77	0.20	8.59				2.04	7.58	8.67
37													2.12	1.10	2.53
38	0.03	0.41	4.33	2.16	2.08	0.42	3.39	2.09	2.15	1.30	1.02	2.91	0.85	0.08	0.68
39	0.07	0.43	2.54	1.41	2.39	0.39	3.11	1.48	2.17	2.40	1.24	1.81	3.48	0.23	0.96
40	0.00	6.09	8.53	6.24	7.51	6.19	8.12	5.83	8.36	0.42	0.20	5.10	0.56	5.06	4.68
41	0.36		3.93	5.27	5.23		3.81	5.42	5.14				0.56	5.22	4.11
42	0.09		1.39	0.41	12.36		1.63	0.64	12.58				2.06	11.65	13.46
43	0.08	9.85	0.77	2.20	6.53	8.39	0.72	2.12	6.82	8.48	7.57	0.64	2.45	7.28	5.88
44	0.00	5.87	2.97	2.73	3.84	5.49	3.16	2.62	3.87	2.92	3.51	1.24	0.15	2.65	2.53
45	0.07	0.82	2.29	2.16	4.21	0.98	2.14	2.28	4.16	1.90	3.60	3.83	0.03	0.90	0.76
46	0.47	9.34	2.66	5.12	22.75	9.96	2.51	6.10	22.30	7.28	4.63	27.35	1.41	27.53	17.47
47	0.12	34.50	1.15	30.37	78.46	33.14	1.14	30.26	76.50	30.39	11.30	6.88	39.10	86.17	27.03
48						2.47	0.78	1.06	10.40	3.62	3.44	12.36	0.10	10.88	11.40
49	0.30	8.01	2.49	2.49	10.54	7.67	2.58	2.58	9.89	6.61	6.61	3.68		9.94	9.94
50	1.67		4.19	4.35			4.43	4.60					0.80		

Tabela A.5: Medidas de tamanho do efeito obtidos com o *d'cohen* para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 20

Bases	Pares de Classificadores														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.05		0.16	0.29	0.23		0.15	0.26	0.17				0.07	0.03	0.11
2	0.07		0.04	0.31	0.87		0.00	0.26	0.84				0.29	0.66	0.62
3		0.29	0.68	0.24	0.10					0.37	0.13	0.16	0.47	0.50	0.07
4	0.45	2.26	1.01		0.37	2.10	1.06		0.69	1.29		1.87		0.72	
5	0.41		0.93	0.95	1.39		0.86	0.79	1.30				0.02	0.43	0.48
6		0.78		1.18	1.23						0.34	0.57			0.32
7															
8															
9	0.21	0.42	1.20	0.96	0.63	0.37	1.29	1.13	0.70	1.47	1.37	0.92	0.53	0.53	0.23
10													0.97	0.26	0.45
11	0.38	0.30	0.25	0.02	0.54	0.14	0.55	0.16	0.39	0.45	0.24	0.23	0.20	0.81	0.46
12	0.07	0.62	0.10	0.37	0.04	0.56	0.16	0.35	0.09	0.58	0.27	0.72	0.82	0.04	0.36
13															
14													0.78		
15										1.76	2.41	0.22	0.14	1.78	2.44
16															
17	0.10		0.75	0.84	2.01		0.64	0.82	2.03				0.02	1.08	1.19
18	0.18	1.03	1.40	1.40		1.05	1.47	1.47		0.06	0.06				
19	0.18	0.42	0.93	0.26	1.49	0.55	0.95	0.36	1.64	0.59	0.11	1.12	0.92	0.61	1.35
20	0.22		0.22	0.18	0.11		0.12	0.09	0.00				0.06	0.14	0.09
21	0.06		0.70	0.31	0.33		0.58	0.33	0.32				0.74	0.68	0.02
22	0.23	3.01	4.30	4.05	3.42	2.97	4.36	3.89	3.22	1.43	0.37	0.57	1.82	1.04	0.14
23	0.36	0.20	0.70	0.41	0.00	0.34	0.61	0.29	0.14	1.03	0.63	0.27	0.53	0.70	0.33
24															2.43
25					3.62										
26	0.26	0.00	0.65	0.38	0.29	0.09	0.77	0.49	0.37	0.63	0.43	0.49	0.34	0.28	0.06
27	0.22	0.19	2.26	2.20	0.09	0.06	2.14	2.04	0.18	2.39	2.28	0.28	0.24	1.71	1.72
28	0.26	0.05	0.94	0.82	1.82	0.17	1.12	1.05	2.03	0.93	0.84	1.80	0.19	0.57	0.76
29												4.73			
30	0.08	0.02	0.24	0.02	0.31	0.02	0.31	0.02	0.34	0.17	0.00	0.47	0.23	0.19	0.32
31	0.37	0.19	1.62	1.06	0.00	0.33	1.76	1.16	0.10	0.73	0.40	0.33	0.51	0.78	0.49
32	0.04	0.82	2.30	1.66	0.91	0.74	1.99	1.41	0.81	0.93	0.41	0.18	0.60	1.08	0.56
33	0.26	0.34	0.05	0.05	3.11	0.45	0.13	0.13	3.04	0.31	0.31	2.86		3.85	3.85
34	0.06	0.00	1.28	0.60	2.07	0.03	1.26	0.61	2.06	1.31	0.68	1.87	0.66	0.89	1.34
35															
36	0.12		0.99	0.00	4.93		1.06	0.07	4.97				1.43	3.96	4.74
37													1.53	0.84	1.94
38	0.21	0.44	1.92	1.61	1.81	0.36	1.73	1.60	1.60	1.10	0.77	1.77	0.52	0.08	0.55
39	0.14	0.04	1.29	0.62	0.95	0.12	1.49	0.66	1.03	1.46	0.69	0.95	1.16	0.09	0.50
40	0.22	2.96	2.68	3.49	6.97	3.10	2.63	3.41	6.69	0.28	0.13	4.64	0.51	4.02	3.99
41	0.05		3.48	3.05	4.10		3.34	3.09	4.28				0.37	2.84	3.07
42	0.13		1.67	0.57	10.79		1.31	0.44	11.21				1.52	9.89	9.00
43	0.03	6.72	0.57	1.93	5.82	7.08	0.60	2.25	6.00	7.81	5.55	0.56	0.99	6.22	5.09
44	0.22	3.70	1.48	1.56	3.11	3.79	1.57	1.56	3.08	2.32	2.64	1.14	0.07	2.18	2.14
45	0.06	0.50	2.21	1.78	2.65	0.53	2.01	1.76	2.26	1.45	1.77	1.73	0.08	0.51	0.65
46	0.22	7.93	2.02	3.83	16.01	7.23	2.23	4.45	16.32	6.19	3.61	16.49	1.28	16.67	11.55
47	0.03	24.25	0.41	18.68	38.11	24.66	0.44	17.47	38.28	23.44	5.31	5.73	19.05	37.29	12.82
48						2.11	0.36	0.59	9.08	1.91	2.23	8.60	0.15	9.18	9.23
49	0.72	5.18	2.52	2.52	8.43	5.35	3.23	3.23	9.15	4.01	4.01	2.71		7.42	7.42
50	0.31		4.37	4.99			4.59	5.26					0.51		

Tabela A.6: Medidas de tamanho do efeito obtidos com o *d′cohen* para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 30

Bases	Pares de Classificadores														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.27		0.06	0.17	0.12		0.18	0.30	0.22				0.09	0.06	0.00
2	0.17		0.12	0.09	0.39		0.04	0.17	0.46				0.24	0.50	0.39
3		0.15	0.39	0.05	0.16					0.24	0.16	0.32	0.37	0.47	0.20
4	0.04	1.76	0.88		0.35	1.83	0.87		0.40	1.27		1.57		0.58	
5	0.32		0.67	0.58	1.04		0.75	0.66	1.13				0.07	0.47	0.46
6		0.59		0.90	1.14						0.28	0.37			0.27
7															
8															
9	0.00	0.33	1.49	1.34	0.56	0.31	1.34	1.17	0.58	1.59	1.35	0.80	0.46	0.51	0.22
10													0.79	0.25	0.25
11	0.08	0.13	0.25	0.22	0.45	0.16	0.19	0.18	0.46	0.31	0.28	0.56	0.00	0.57	0.51
12	0.31	0.48	0.05	0.25	0.33	0.60	0.14	0.41	0.16	0.45	0.24	0.94	0.41	0.26	0.52
13															
14													0.62		
15										1.76	2.01	0.18	0.13	1.76	2.01
16															
17	0.05		0.49	0.60	1.52		0.56	0.70	1.52				0.11	0.64	1.18
18	0.05	0.60	1.05	1.05		0.54	0.97	0.97		0.03	0.03				
19	0.07	0.17	0.91	0.11	1.22	0.19	0.84	0.15	1.23	0.53	0.08	0.74	0.80	0.44	0.97
20	0.18		0.14	0.16	0.09		0.05	0.08	0.00				0.05	0.04	0.07
21	0.25		0.50	0.13	0.13		0.60	0.07	0.10				0.54	0.46	0.06
22	0.17	2.45	3.86	3.36	3.76	2.59	4.15	3.52	4.03	1.11	0.20	0.44	1.18	0.87	0.30
23	0.01	0.11	0.60	0.50	0.01	0.09	0.63	0.45	0.02	0.69	0.60	0.13	0.38	0.52	0.32
24															2.27
25					4.29										
26	0.05	0.09	0.61	0.50	0.37	0.09	0.59	0.49	0.35	0.40	0.32	0.67	0.20	0.16	0.03
27	0.15	0.02	1.54	1.74	0.16	0.09	1.63	1.80	0.24	1.40	1.70	0.15	0.23	1.22	1.55
28	0.00	0.19	0.97	0.74	1.65	0.18	1.09	0.79	1.67	0.65	0.54	1.24	0.15	0.56	0.68
29												2.45			
30	0.09	0.04	0.32	0.13	0.38	0.01	0.30	0.08	0.35	0.14	0.05	0.40	0.15	0.19	0.33
31	0.19	0.18	1.11	0.67	0.02	0.27	1.25	0.88	0.05	0.60	0.43	0.24	0.32	0.68	0.55
32	0.00	0.57	1.38	1.00	0.59	0.66	1.47	1.07	0.64	0.56	0.31	0.10	0.39	0.65	0.35
33	0.21	0.39	0.02	0.02	1.79	0.44	0.09	0.09	1.82	0.34	0.34	2.05		2.05	2.05
34	0.11	0.16	0.66	0.43	1.40	0.09	0.76	0.51	1.47	0.77	0.67	1.69	0.33	0.61	1.03
35															
36	0.05		1.29	0.26	6.14		1.13	0.19	5.74				0.95	5.15	6.37
37													1.17	0.60	1.34
38	0.05	0.33	1.94	1.15	1.39	0.38	2.10	1.24	1.57	1.03	0.62	1.29	0.39	0.02	0.27
39	0.08	0.17	1.30	0.73	1.17	0.21	1.29	0.77	1.11	1.28	0.74	1.07	0.85	0.09	0.53
40	0.17	2.67	2.08	2.37	4.41	3.12	2.37	2.62	4.41	0.43	0.07	3.42	0.37	3.01	2.89
41	0.08		2.20	3.05	3.23		2.17	3.00	3.23				0.42	2.22	2.32
42	0.29		1.09	0.32	6.11		1.22	0.43	6.23				1.02	6.00	6.27
43	0.24	4.80	0.38	1.03	4.91	4.86	0.44	1.05	4.60	5.64	4.04	0.49	0.75	4.92	4.00
44	0.08	3.45	1.35	1.67	2.91	3.50	1.33	1.68	2.92	1.83	2.26	1.10	0.16	2.24	2.20
45	0.25	0.38	1.45	1.08	1.44	0.55	1.45	1.25	1.62	0.87	0.86	1.10	0.14	0.31	0.34
46	0.38	4.83	1.19	2.75	12.30	5.34	1.49	2.91	13.16	5.74	3.35	13.09	1.24	16.77	12.66
47	0.31	22.74	0.33	12.23	27.41	22.73	0.47	12.25	27.41	20.76	4.77	4.05	12.34	25.74	7.99
48						1.56	0.40	0.53	8.13	2.04	1.91	7.07	0.14	8.00	8.55
49	0.42	5.13	1.68	1.68	7.07	5.48	2.07	2.07	7.02	4.09	4.09	2.71		5.56	5.56
50	0.19		4.84	5.11			4.72	4.97					0.37		

Tabela A.7: Poder dos testes t obtidos para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 10

Bases	Pares de Classificadores														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.06		0.06	0.09	0.05		0.05	0.14	0.06				0.14	0.06	0.06
2	0.22		0.05	0.24	0.85		0.07	0.07	0.82				0.21	0.66	0.50
3		0.32	0.77	0.23	0.38					0.28	0.08	0.08	0.53	0.52	0.05
4	0.05	1.00	1.00		0.27	1.00	1.00		0.24	0.97		1.00		0.83	
5	0.06		1.00	0.99	1.00		1.00	1.00	1.00				0.14	0.67	0.61
6		0.66		0.94	1.00						0.43	0.55			0.24
7															
8															
9	0.08	0.27	1.00	1.00	0.92	0.32	1.00	1.00	0.78	1.00	0.99	0.97	0.66	0.84	0.15
10													0.98	0.06	0.17
11	0.32	0.07	0.42	0.05	0.54	0.17	0.16	0.07	0.88	0.28	0.06	0.72	0.27	0.92	0.44
12	0.05	0.30	0.09	0.16	0.06	0.41	0.08	0.13	0.06	0.40	0.10	0.80	0.72	0.05	0.25
13															
14													0.78		
15										1.00	1.00	0.12	0.10	1.00	1.00
16															
17	0.12		0.80	0.99	1.00		0.62	0.82	1.00				0.05	0.75	0.97
18	0.39	0.91	1.00	1.00		0.88	0.99	0.99		0.12	0.12				
19	0.12	0.23	0.91	0.10	1.00	0.14	0.93	0.08	1.00	0.69	0.05	0.97	0.95	0.49	0.98
20	0.07		0.10	0.13	0.08		0.08	0.11	0.06				0.06	0.06	0.09
21	0.10		0.61	0.13	0.37		0.59	0.16	0.45				0.91	0.89	0.06
22	0.37	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.65	0.56	0.99	0.70	0.15
23	0.05	0.31	0.67	0.23	0.09	0.40	0.87	0.42	0.08	0.99	0.89	0.13	0.62	0.72	0.32
24															1.00
25					1.00										
26	0.07	0.05	0.84	0.62	0.09	0.05	0.85	0.73	0.10	0.73	0.31	0.13	0.49	0.37	0.10
27	0.66	0.11	1.00	1.00	0.09	0.05	1.00	1.00	0.26	1.00	1.00	0.20	0.49	1.00	1.00
28	0.07	0.11	1.00	0.99	1.00	0.11	1.00	1.00	1.00	0.93	0.64	1.00	0.45	0.90	0.99
29												1.00			
30	0.15	0.06	0.85	0.50	0.94	0.05	0.30	0.11	0.92	0.20	0.12	0.64	0.16	0.13	0.61
31	0.13	0.37	1.00	1.00	0.06	0.33	1.00	1.00	0.05	0.35	0.25	0.78	0.10	0.61	0.54
32	0.27	0.86	1.00	1.00	0.95	0.87	1.00	1.00	0.95	0.94	0.36	0.06	0.72	0.99	0.82
33	0.35	0.41	0.06	0.06	1.00	0.62	0.12	0.12	1.00	0.20	0.20	1.00		1.00	1.00
34	0.05	0.05	0.98	0.61	1.00	0.05	0.91	0.69	1.00	1.00	0.76	1.00	0.31	1.00	1.00
35															
36	0.05		1.00	0.09	1.00		1.00	0.08	1.00				1.00	1.00	1.00
37													1.00	0.87	1.00
38	0.05	0.18	1.00	1.00	1.00	0.19	1.00	1.00	1.00	0.95	0.82	1.00	0.66	0.05	0.45
39	0.05	0.20	1.00	0.97	1.00	0.17	1.00	0.98	1.00	1.00	0.93	1.00	1.00	0.09	0.77
40	0.05	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.19	0.08	1.00	0.32	1.00	1.00
41	0.15		1.00	1.00	1.00		1.00	1.00	1.00				0.32	1.00	1.00
42	0.06		0.97	0.18	1.00		0.99	0.41	1.00				1.00	1.00	1.00
43	0.05	1.00	0.56	1.00	1.00	1.00	0.50	1.00	1.00	1.00	1.00	0.40	1.00	1.00	1.00
44	0.05	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.94	0.06	1.00	1.00
45	0.05	0.63	1.00	1.00	1.00	0.79	1.00	1.00	1.00	1.00	1.00	1.00	0.05	0.71	0.55
46	0.23	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00
47	0.06	1.00	0.90	1.00	1.00	1.00	0.89	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
48						1.00	0.58	0.85	1.00	1.00	1.00	1.00	0.06	1.00	1.00
49	0.11	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00		1.00	1.00
50	0.99		1.00	1.00			1.00	1.00					0.61		

Tabela A.8: Poder dos testes t obtidos para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 20

Bases	Pares de Classificadores														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.05		0.10	0.22	0.15		0.09	0.18	0.11				0.06	0.05	0.07
2	0.06		0.05	0.24	0.96		0.05	0.18	0.94				0.22	0.79	0.74
3		0.21	0.82	0.17	0.07					0.33	0.08	0.10	0.51	0.56	0.06
4	0.47	1.00	0.99		0.34	1.00	0.99		0.83	1.00		1.00		0.86	
5	0.39		0.97	0.98	1.00		0.95	0.92	1.00				0.05	0.44	0.52
6		0.91		1.00	1.00						0.29	0.67			0.26
7															
8															
9	0.13	0.41	1.00	0.98	0.76	0.33	1.00	1.00	0.84	1.00	1.00	0.97	0.61	0.61	0.15
10													0.98	0.19	0.47
11	0.35	0.23	0.18	0.05	0.62	0.09	0.65	0.10	0.36	0.46	0.16	0.16	0.13	0.93	0.49
12	0.06	0.75	0.07	0.33	0.05	0.66	0.10	0.31	0.07	0.69	0.19	0.86	0.94	0.05	0.33
13															
14													0.91		
15										1.00	1.00	0.15	0.09	1.00	1.00
16															
17	0.07		0.88	0.94	1.00		0.78	0.93	1.00				0.05	0.99	1.00
18	0.11	0.99	1.00	1.00		0.99	1.00	1.00		0.06	0.06				
19	0.11	0.42	0.97	0.19	1.00	0.64	0.98	0.32	1.00	0.70	0.07	1.00	0.97	0.74	1.00
20	0.15		0.15	0.11	0.07		0.08	0.06	0.05				0.06	0.09	0.07
21	0.06		0.84	0.25	0.27		0.68	0.27	0.26				0.88	0.82	0.05
22	0.15	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.33	0.67	1.00	0.99	0.08
23	0.31	0.12	0.84	0.40	0.05	0.29	0.73	0.22	0.09	0.99	0.76	0.19	0.61	0.85	0.27
24															1.00
25					1.00										
26	0.18	0.05	0.78	0.35	0.22	0.07	0.90	0.54	0.33	0.76	0.43	0.54	0.29	0.21	0.06
27	0.15	0.12	1.00	1.00	0.07	0.06	1.00	1.00	0.11	1.00	1.00	0.21	0.16	1.00	1.00
28	0.18	0.05	0.98	0.94	1.00	0.10	1.00	0.99	1.00	0.97	0.94	1.00	0.12	0.67	0.89
29												1.00			
30	0.06	0.05	0.16	0.05	0.25	0.05	0.24	0.05	0.28	0.11	0.05	0.51	0.15	0.12	0.26
31	0.33	0.12	1.00	0.99	0.05	0.27	1.00	1.00	0.07	0.87	0.39	0.27	0.58	0.91	0.54
32	0.05	0.93	1.00	1.00	0.97	0.88	1.00	1.00	0.93	0.97	0.41	0.11	0.72	0.99	0.66
33	0.18	0.29	0.05	0.05	1.00	0.47	0.08	0.08	1.00	0.25	0.25	1.00		1.00	1.00
34	0.06	0.05	1.00	0.72	1.00	0.05	1.00	0.73	1.00	1.00	0.82	1.00	0.79	0.96	1.00
35															
36	0.08		0.98	0.05	1.00		0.99	0.06	1.00				1.00	1.00	1.00
37													1.00	0.94	1.00
38	0.14	0.45	1.00	1.00	1.00	0.31	1.00	1.00	1.00	0.99	0.90	1.00	0.59	0.06	0.64
39	0.09	0.05	1.00	0.75	0.98	0.08	1.00	0.80	0.99	1.00	0.83	0.98	1.00	0.07	0.55
40	0.15	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.21	0.08	1.00	0.57	1.00	1.00
41	0.05		1.00	1.00	1.00		1.00	1.00	1.00				0.34	1.00	1.00
42	0.08		1.00	0.68	1.00		1.00	0.44	1.00				1.00	1.00	1.00
43	0.05	1.00	0.68	1.00	1.00	1.00	0.72	1.00	1.00	1.00	1.00	0.66	0.98	1.00	1.00
44	0.15	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.06	1.00	1.00
45	0.06	0.56	1.00	1.00	1.00	0.61	1.00	1.00	1.00	1.00	1.00	1.00	0.06	0.57	0.79
46	0.14	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
47	0.05	1.00	0.40	1.00	1.00	1.00	0.46	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
48						1.00	0.31	0.71	1.00	1.00	1.00	1.00	0.09	1.00	1.00
49	0.87	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00		1.00	1.00
50	0.25		1.00	1.00			1.00	1.00					0.57		

Tabela A.9: Poder dos testes t obtidos para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 30

Bases	Pares de Classificadores														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.29		0.06	0.14	0.10		0.15	0.35	0.21				0.08	0.06	0.05
2	0.14		0.10	0.08	0.53		0.05	0.14	0.69				0.24	0.75	0.54
3		0.12	0.53	0.06	0.13					0.23	0.13	0.38	0.49	0.69	0.18
4	0.05	1.00	1.00		0.45	1.00	0.99		0.55	1.00		1.00		0.87	
5	0.39		0.94	0.86	1.00		0.98	0.94	1.00				0.07	0.70	0.69
6		0.88		1.00	1.00						0.32	0.50			0.29
7															
8															
9	0.05	0.42	1.00	1.00	0.84	0.37	1.00	1.00	0.86	1.00	1.00	0.99	0.68	0.78	0.20
10													0.99	0.25	0.25
11	0.07	0.10	0.25	0.21	0.65	0.13	0.17	0.15	0.68	0.36	0.30	0.84	0.05	0.86	0.78
12	0.37	0.72	0.06	0.25	0.40	0.89	0.11	0.59	0.13	0.67	0.24	1.00	0.58	0.26	0.79
13															
14													0.91		
15										1.00	1.00	0.15	0.10	1.00	1.00
16															
17	0.06		0.73	0.88	1.00		0.84	0.96	1.00				0.08	0.92	1.00
18	0.06	0.89	1.00	1.00		0.82	1.00	1.00		0.05	0.05				
19	0.06	0.14	1.00	0.09	1.00	0.17	0.99	0.12	1.00	0.80	0.07	0.97	0.99	0.64	1.00
20	0.15		0.11	0.13	0.07		0.06	0.07	0.05				0.06	0.05	0.06
21	0.25		0.75	0.11	0.10		0.89	0.07	0.08				0.82	0.68	0.06
22	0.14	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.17	0.64	1.00	0.99	0.34
23	0.05	0.09	0.89	0.75	0.05	0.08	0.91	0.65	0.05	0.95	0.89	0.10	0.51	0.78	0.38
24															1.00
25					1.00										
26	0.06	0.07	0.90	0.75	0.48	0.08	0.88	0.73	0.45	0.55	0.38	0.94	0.17	0.13	0.05
27	0.12	0.05	1.00	1.00	0.12	0.08	1.00	1.00	0.24	1.00	1.00	0.12	0.21	1.00	1.00
28	0.05	0.17	1.00	0.97	1.00	0.15	1.00	0.98	1.00	0.93	0.82	1.00	0.12	0.84	0.95
29												1.00			
30	0.07	0.06	0.38	0.10	0.51	0.05	0.34	0.07	0.45	0.11	0.06	0.55	0.12	0.17	0.41
31	0.16	0.15	1.00	0.94	0.05	0.29	1.00	1.00	0.06	0.89	0.61	0.24	0.38	0.95	0.83
32	0.05	0.86	1.00	1.00	0.88	0.93	1.00	1.00	0.93	0.84	0.36	0.08	0.54	0.93	0.46
33	0.19	0.53	0.05	0.05	1.00	0.64	0.08	0.08	1.00	0.42	0.42	1.00		1.00	1.00
34	0.09	0.12	0.94	0.63	1.00	0.07	0.98	0.76	1.00	0.98	0.94	1.00	0.41	0.90	1.00
35															
36	0.06		1.00	0.27	1.00		1.00	0.16	1.00				1.00	1.00	1.00
37													1.00	0.88	1.00
38	0.06	0.41	1.00	1.00	1.00	0.51	1.00	1.00	1.00	1.00	0.91	1.00	0.53	0.05	0.28
39	0.07	0.14	1.00	0.97	1.00	0.19	1.00	0.98	1.00	1.00	0.97	1.00	0.99	0.07	0.80
40	0.14	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.62	0.06	1.00	0.50	1.00	1.00
41	0.07		1.00	1.00	1.00		1.00	1.00	1.00				0.59	1.00	1.00
42	0.32		1.00	0.38	1.00		1.00	0.62	1.00				1.00	1.00	1.00
43	0.24	1.00	0.52	1.00	1.00	1.00	0.64	1.00	1.00	1.00	1.00	0.74	0.98	1.00	1.00
44	0.07	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.12	1.00	1.00
45	0.26	0.52	1.00	1.00	1.00	0.82	1.00	1.00	1.00	0.99	0.99	1.00	0.11	0.36	0.43
46	0.52	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
47	0.38	1.00	0.41	1.00	1.00	1.00	0.69	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
48						1.00	0.56	0.80	1.00	1.00	1.00	1.00	0.11	1.00	1.00
49	0.60	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00		1.00	1.00
50	0.16		1.00	1.00			1.00	1.00					0.49		

APÊNDICE B – Resultados Obtidos no Estudo Empírico com o Teste de Wilcoxon

Tabela B.1: P-valores obtidos com a aplicação do teste de Wilcoxon para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 10

Bases	Pares de Classificadores														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.67	0.07	0.45	0.50	0.95	0.06	0.50	0.24	0.80	0.26	0.07	0.06	0.16	0.64	0.72
2	0.10	0.02	0.55	0.24	0.03	0.11	0.72	0.78	0.04	0.05	0.06	0.40	0.14	0.04	0.05
3		0.16	0.03	0.11	0.14	0.16	0.03	0.11	0.14	0.19	0.55	0.20	0.08	0.07	0.87
4	0.83	0.01	0.00	0.01	0.21	0.01	0.01	0.01	0.12	0.01	0.09	0.01	0.04	0.02	0.01
5	0.31	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.05	0.11	0.33	0.11	0.03	0.04
6	0.71	0.04	0.01	0.01	0.01	0.03	0.01	0.01	0.01	0.19	0.04	0.06	0.88	0.21	0.15
7	0.32	0.79	0.21	0.68	0.89	1.00	0.21	1.00	0.71	0.08	0.92	1.00	0.13	0.09	0.59
8	0.26	0.05	0.02	0.07	0.03	0.02	0.01	0.02	0.01	1.00	0.57	0.51	0.26	0.62	0.77
9	0.71	0.14	0.01	0.01	0.01	0.10	0.01	0.01	0.02	0.01	0.01	0.01	0.04	0.02	0.44
10	0.41	0.08	0.01	0.31	0.15	0.16	0.01	0.25	0.08	0.01	0.17	0.03	0.01	0.51	0.23
11	0.14	0.77	0.08	0.87	0.06	0.50	0.27	0.33	0.01	0.20	0.86	0.02	0.13	0.01	0.05
12	0.92	0.13	0.47	0.27	0.75	0.06	0.56	0.34	0.37	0.09	0.42	0.02	0.02	0.94	0.15
13	0.11	0.02	0.38	0.92	0.01	0.06	0.12	0.48	0.01	0.06	0.08	0.00	0.27	0.02	0.01
14	0.32	1.00	0.01	0.07	0.03	0.32	0.01	0.03	0.02	0.01	0.03	0.02	0.01	0.09	0.10
15		0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.32	0.37	0.01	0.01
16	0.06	0.05	0.89	0.15	0.03	0.26	0.29	0.86	0.02	0.02	0.35	0.02	0.25	0.10	0.02
17	0.44	0.91	0.02	0.01	0.01	0.40	0.04	0.02	0.01	0.01	0.01	0.00	0.68	0.03	0.01
18	0.10	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.31	0.31	0.01		0.02	0.02
19	0.52	0.17	0.01	0.53	0.01	0.27	0.01	0.44	0.01	0.00	0.67	0.01	0.01	0.05	0.00
20	0.79	0.18	0.67	0.29	0.39	0.14	0.67	0.40	0.60	0.04	0.04	0.11	0.91	0.81	0.55
21	0.29	0.08	0.03	0.39	0.07	0.08	0.04	0.24	0.04	0.01	0.33	0.41	0.02	0.01	0.55
22	0.12	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.04	0.04	0.01	0.03	0.36
23	0.92	0.09	0.03	0.12	0.67	0.07	0.02	0.06	0.57	0.01	0.01	0.31	0.05	0.03	0.10
24	0.32	0.01	0.34	0.01	0.01	0.01	0.27	0.01	0.01	0.01	0.40	0.01	0.01	0.01	0.01
25	0.34	0.05	0.01	0.01	0.01	0.14	0.01	0.01	0.00	0.09	0.09	0.01		0.01	0.01
26	0.59	0.72	0.02	0.03	0.51	0.65	0.02	0.02	0.31	0.02	0.14	0.29	0.05	0.09	0.36
27	0.04	0.67	0.01	0.01	0.46	0.72	0.00	0.00	0.13	0.01	0.01	0.19	0.04	0.01	0.01
28	0.62	0.48	0.01	0.01	0.01	0.51	0.01	0.01	0.01	0.01	0.03	0.01	0.05	0.02	0.01
29	0.18	0.05	0.07	0.07	0.01	0.02	0.07	0.07	0.01	0.01	0.01	0.01		0.01	0.01
30	0.23	0.51	0.01	0.05	0.01	0.78	0.18	0.44	0.01	0.24	0.33	0.05	0.33	0.29	0.05
31	0.33	0.09	0.01	0.00	0.68	0.09	0.01	0.01	0.72	0.08	0.14	0.03	0.41	0.02	0.04
32	0.22	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.11	0.68	0.03	0.01	0.02
33	0.11	0.07	0.64	0.64	0.01	0.04	0.33	0.33	0.01	0.15	0.15	0.01		0.01	0.01
34	0.78	0.88	0.00	0.05	0.01	0.86	0.01	0.03	0.00	0.01	0.02	0.01	0.11	0.01	0.01
35	0.32	0.17	0.71	0.09	0.01	0.21	0.50	0.07	0.01	0.18	0.01	0.01	0.25	0.01	0.01
36	0.75	0.07	0.00	0.48	0.01	0.12	0.01	0.48	0.01	0.02	0.36	0.01	0.01	0.01	0.01
37	0.18	0.93	0.01	0.01	0.01	0.17	0.01	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.01
38	0.96	0.17	0.01	0.01	0.01	0.14	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.77	0.07
39	0.80	0.26	0.01	0.01	0.01	0.24	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.30	0.01
40	0.80	0.00	0.01	0.01	0.01	0.01	0.00	0.01	0.01	0.24	0.40	0.01	0.11	0.01	0.01
41	0.14	0.00	0.01	0.01	0.01	0.00	0.01	0.01	0.01	0.03	0.01	0.01	0.08	0.01	0.01
42	0.77	0.00	0.01	0.40	0.01	0.00	0.01	0.07	0.01	0.01	0.01	0.01	0.01	0.01	0.01
43	0.67	0.01	0.05	0.01	0.01	0.01	0.04	0.01	0.01	0.01	0.01	0.07	0.01	0.01	0.01
44	0.67	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.55	0.01	0.01
45	0.89	0.03	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.76	0.03	0.05
46	0.16	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
47	0.72	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
48	0.79	0.01	0.07	0.03	0.01	0.01	0.03	0.02	0.01	0.01	0.01	0.01	0.80	0.01	0.01
49	0.55	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01		0.01	0.01
50	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.07	0.01	0.01

Tabela B.2: P-valores obtidos com a aplicação do teste de Wilcoxon para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 20

Bases	Pares de Classificadores														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.93	0.85	0.55	0.15	0.70	0.44	0.71	0.26	0.66	0.45	0.08	0.12	0.48	0.98	0.85
2	0.86	0.01	0.94	0.18	0.00	0.05	0.92	0.25	0.00	0.05	0.07	0.18	0.27	0.02	0.02
3	0.58	0.23	0.02	0.14	0.67	0.18	0.01	0.11	0.53	0.13	0.67	0.64	0.09	0.09	0.61
4	0.03	0.00	0.00	0.00	0.14	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.07	0.01	0.00
5	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.12	0.57	0.83	0.09	0.03
6	0.65	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.32	0.22	0.03	0.48	0.08	0.19
7	0.32	0.79	0.15	0.48	0.95	0.85	0.20	0.73	0.80	0.16	0.61	0.84	0.39	0.10	0.52
8	0.32	0.03	0.00	0.09	0.07	0.04	0.00	0.14	0.11	0.25	0.20	0.55	0.02	0.13	0.53
9	0.44	0.07	0.00	0.00	0.01	0.07	0.00	0.00	0.01	0.00	0.00	0.00	0.03	0.03	0.45
10	0.16	0.53	0.00	0.12	0.02	0.36	0.00	0.23	0.03	0.00	0.11	0.01	0.00	0.18	0.06
11	0.12	0.11	0.38	0.49	0.03	0.44	0.02	0.38	0.08	0.09	0.22	0.31	0.50	0.00	0.05
12	0.76	0.02	0.71	0.12	0.87	0.03	0.48	0.14	0.78	0.02	0.29	0.01	0.00	0.69	0.13
13	0.32	0.29	0.13	0.96	0.00	0.25	0.17	0.88	0.00	0.16	0.44	0.00	0.27	0.01	0.00
14	0.32	0.65	0.00	0.05	0.01	0.32	0.00	0.04	0.00	0.00	0.04	0.00	0.01	0.07	0.21
15		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.32	0.65	0.00	0.00
16	0.32	0.26	0.23	0.51	0.01	0.36	0.15	0.67	0.00	0.00	0.47	0.00	0.05	0.16	0.00
17	0.55	0.39	0.01	0.00	0.00	0.49	0.01	0.00	0.00	0.01	0.00	0.00	0.78	0.00	0.00
18	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.69	0.69	0.00		0.00	0.00
19	0.67	0.06	0.00	0.25	0.00	0.03	0.00	0.13	0.00	0.03	0.31	0.00	0.00	0.03	0.00
20	0.27	0.15	0.24	0.31	0.64	0.09	0.45	0.75	0.82	0.02	0.04	0.05	0.63	0.37	0.56
21	0.92	0.01	0.03	0.28	0.17	0.02	0.03	0.20	0.21	0.00	0.41	0.22	0.01	0.01	0.86
22	0.34	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.02	0.00	0.00	0.68
23	0.05	0.39	0.01	0.06	0.86	0.16	0.03	0.17	0.61	0.00	0.01	0.30	0.06	0.01	0.17
24		0.00	0.41	0.00	0.00	0.00	0.41	0.00	0.00	0.00	0.89	0.00	0.00	0.00	0.00
25	0.04	0.00	0.00	0.00	0.00	0.05	0.02	0.02	0.00	0.05	0.05	0.00		0.00	0.00
26	0.25	0.85	0.01	0.20	0.20	0.80	0.00	0.05	0.11	0.01	0.11	0.05	0.12	0.23	0.67
27	0.25	0.29	0.00	0.00	0.80	0.69	0.00	0.00	0.36	0.00	0.00	0.26	0.41	0.00	0.00
28	0.39	0.85	0.00	0.00	0.00	0.42	0.00	0.00	0.00	0.00	0.00	0.00	0.39	0.02	0.01
29		0.02	0.04	0.04	0.00	0.02	0.04	0.04	0.00	0.00	0.00	0.00		0.00	0.00
30	0.80	0.95	0.33	0.72	0.21	0.86	0.16	0.82	0.18	0.57	0.83	0.08	0.20	0.29	0.14
31	0.09	0.26	0.00	0.00	0.71	0.13	0.00	0.00	0.36	0.00	0.10	0.18	0.09	0.00	0.04
32	0.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.33	0.02	0.00	0.02
33	0.40	0.14	0.82	0.82	0.00	0.19	0.59	0.59	0.00	0.15	0.15	0.00		0.00	0.00
34	0.75	0.95	0.00	0.03	0.00	0.81	0.00	0.02	0.00	0.00	0.01	0.00	0.02	0.00	0.00
35		0.17	0.45	0.03	0.00	0.17	0.45	0.03	0.00	0.09	0.00	0.00	0.07	0.00	0.00
36	0.44	0.02	0.00	0.95	0.00	0.04	0.00	0.66	0.00	0.01	0.15	0.00	0.00	0.00	0.00
37	1.00	0.16	0.00	0.01	0.00	0.05	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00
38	0.39	0.07	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.77	0.03
39	0.55	0.92	0.00	0.01	0.00	0.66	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.65	0.05
40	0.43	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.18	0.47	0.00	0.04	0.00	0.00
41	0.92	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.14	0.00	0.00
42	0.57	0.00	0.00	0.02	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00
43	0.84	0.00	0.02	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
44	0.23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.84	0.00	0.00
45	0.90	0.05	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.98	0.05	0.01
46	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
47	0.61	0.00	0.10	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
48	0.67	0.00	0.13	0.02	0.00	0.00	0.16	0.03	0.00	0.00	0.00	0.00	0.59	0.00	0.00
49	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00
50	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00

Tabela B.3: P-valores obtidos com a aplicação do teste de Wilcoxon para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 30

Bases	Pares de Classificadores														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.09	0.75	0.64	0.22	0.62	0.73	0.38	0.06	0.22	0.54	0.28	0.19	0.59	0.88	0.87
2	0.35	0.16	0.43	0.62	0.04	0.04	0.76	0.30	0.02	0.02	0.06	0.56	0.16	0.01	0.04
3	0.17	0.55	0.06	0.50	0.54	0.24	0.02	0.15	0.96	0.34	0.48	0.14	0.07	0.03	0.28
4	0.37	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.06	0.01	0.00
5	0.22	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.05	0.09	0.57	0.71	0.02	0.01
6	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.56	0.10	0.08	0.72	0.14	0.20
7	0.32	1.00	0.04	0.38	0.78	0.65	0.04	0.24	0.96	0.02	0.38	0.78	0.05	0.04	0.39
8	0.27	0.00	0.01	0.11	0.02	0.01	0.01	0.21	0.02	0.71	0.05	0.39	0.03	0.17	0.15
9	0.83	0.02	0.00	0.00	0.01	0.02	0.00	0.00	0.01	0.00	0.00	0.00	0.03	0.02	0.34
10	0.18	0.23	0.00	0.23	0.04	0.49	0.00	0.13	0.01	0.00	0.12	0.00	0.00	0.20	0.26
11	0.51	0.65	0.20	0.11	0.02	0.44	0.38	0.31	0.02	0.09	0.12	0.02	0.82	0.00	0.01
12	0.17	0.02	0.87	0.12	0.13	0.01	0.48	0.05	0.49	0.06	0.21	0.00	0.02	0.09	0.02
13	0.32	0.02	0.81	0.58	0.00	0.01	0.92	0.48	0.00	0.06	0.11	0.00	0.51	0.00	0.00
14	0.65	0.65	0.00	0.06	0.04		0.00	0.03	0.06	0.00	0.03	0.06	0.01	0.13	0.73
15		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.32	0.44	0.00	0.00
16	0.33	0.02	0.82	0.05	0.03	0.04	0.91	0.10	0.01	0.02	0.42	0.00	0.12	0.05	0.01
17	0.43	0.92	0.01	0.00	0.00	0.60	0.00	0.00	0.00	0.00	0.00	0.00	0.61	0.00	0.00
18	1.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.87	0.87	0.00		0.00	0.00
19	1.00	0.21	0.00	0.37	0.00	0.26	0.00	0.46	0.00	0.01	0.71	0.00	0.00	0.03	0.00
20	0.45	0.31	0.47	0.53	0.64	0.10	0.62	0.68	0.85	0.12	0.15	0.17	0.76	0.66	0.81
21	0.18	0.20	0.01	0.49	0.51	0.31	0.00	0.77	0.66	0.00	0.16	0.26	0.01	0.02	0.73
22	0.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.03	0.00	0.00	0.12
23	0.50	0.50	0.01	0.01	0.94	0.53	0.00	0.02	0.96	0.00	0.01	0.36	0.05	0.01	0.08
24	0.32	0.00	0.74	0.00	0.00	0.00	0.73	0.00	0.00	0.00	0.98	0.00	0.00	0.00	0.00
25	0.33	0.06	0.00	0.00	0.00	0.24	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00
26	0.80	0.47	0.01	0.02	0.04	0.52	0.01	0.02	0.06	0.06	0.05	0.00	0.28	0.44	0.62
27	0.43	0.57	0.00	0.00	0.30	0.49	0.00	0.00	0.36	0.00	0.00	0.55	0.28	0.00	0.00
28	0.78	0.39	0.00	0.00	0.00	0.39	0.00	0.00	0.00	0.00	0.01	0.00	0.47	0.00	0.00
29	1.00	0.02	0.04	0.04	0.00	0.01	0.03	0.03	0.00	0.00	0.00	0.00		0.00	0.00
30	0.63	0.68	0.05	0.47	0.07	0.75	0.15	0.77	0.08	0.42	0.99	0.05	0.40	0.52	0.09
31	0.20	0.21	0.00	0.00	0.90	0.14	0.00	0.00	0.71	0.00	0.02	0.25	0.15	0.00	0.01
32	0.90	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.13	0.61	0.07	0.00	0.06
33	0.34	0.03	0.99	0.99	0.00	0.03	0.53	0.53	0.00	0.08	0.08	0.00		0.00	0.00
34	0.50	0.39	0.00	0.01	0.00	0.48	0.00	0.01	0.00	0.00	0.00	0.00	0.10	0.00	0.00
35	0.16	0.11	0.68	0.11	0.00	0.03	0.96	0.24	0.00	0.15	0.00	0.00	0.19	0.00	0.00
36	0.86	0.01	0.00	0.42	0.00	0.00	0.00	0.33	0.00	0.00	0.38	0.00	0.00	0.00	0.00
37	0.01	0.02	0.00	0.01	0.00	0.76	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
38	0.66	0.05	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.97	0.11
39	0.93	0.49	0.00	0.00	0.00	0.45	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.58	0.01
40	0.34	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.73	0.00	0.04	0.00	0.00
41	0.89	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.02	0.00	0.00
42	0.05	0.00	0.00	0.05	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00
43	0.32	0.00	0.05	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
44	0.78	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.27	0.00	0.00
45	0.15	0.03	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.61	0.12	0.06
46	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
47	0.17	0.00	0.11	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
48	0.75	0.00	0.07	0.03	0.00	0.00	0.05	0.02	0.00	0.00	0.00	0.00	0.50	0.00	0.00
49	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00
50	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00

Tabela B.4: Medidas de tamanho do efeito obtidos com o r para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 10

Bases	Pares de Classificadores														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.10	0.41	0.17	0.15	0.01	0.42	0.15	0.27	0.06	0.25	0.41	0.42	0.31	0.11	0.08
2	0.36	0.51	0.13	0.26	0.49	0.36	0.08	0.06	0.47	0.44	0.42	0.19	0.33	0.47	0.44
3		0.31	0.50	0.35	0.33	0.31	0.50	0.35	0.33	0.29	0.13	0.29	0.40	0.41	0.04
4		0.63	0.63	0.63	0.28	0.63	0.60	0.63	0.35	0.63	0.38	0.63	0.46	0.52	0.63
5	0.23	0.60	0.60	0.60	0.63	0.61	0.63	0.63	0.63	0.44	0.36	0.22	0.35	0.49	0.45
6	0.08	0.47	0.60	0.57	0.63	0.49	0.56	0.60	0.60	0.29	0.45	0.42	0.03	0.28	0.32
7	0.22	0.06	0.28	0.09	0.03	0.00	0.28	0.00	0.08	0.39	0.02	0.00	0.34	0.38	0.12
8	0.25	0.44	0.50	0.41	0.49	0.54	0.60	0.53	0.55	0.00	0.13	0.15	0.25	0.11	0.07
9	0.08	0.33	0.63	0.63	0.55	0.37	0.63	0.63	0.54	0.63	0.63	0.59	0.45	0.54	0.17
10	0.18	0.39	0.60	0.23	0.32	0.32	0.60	0.26	0.39	0.57	0.31	0.47	0.60	0.15	0.27
11	0.33	0.07	0.39	NA	0.42	0.15	0.25	0.22	0.56	0.28	0.04	0.54	0.34	0.56	0.43
12		0.33	0.16	0.25	0.07	0.42	0.13	0.21	0.20	0.38	0.18	0.51	0.51		0.32
13	0.36	0.52	0.19	0.02	0.63	0.42	0.34	0.16	0.63	0.42	0.39	0.63	0.25	0.52	0.60
14	0.22		0.57	0.40	0.49	0.22	0.61	0.48	0.53	0.57	0.47	0.53	0.56	0.38	0.36
15		0.64	0.63	0.63	0.63	0.64	0.63	0.63	0.63	0.63	0.63	0.22	0.20	0.63	0.63
16	0.42	0.44	0.03	0.32	0.49	0.25	0.24	0.04	0.54	0.53	0.21	0.52	0.26	0.37	0.53
17	0.17	0.03	0.54	0.63	0.63	0.19	0.46	0.53	0.63	0.58	0.60	0.63	0.09	0.48	0.60
18	0.37	0.55	0.60	0.60	0.60	0.55	0.60	0.60	0.60	0.23	0.23	0.56		0.50	0.50
19	0.14	0.31	0.60	0.14	0.63	0.25	0.60	0.17	0.63	0.63	0.09	0.60	0.60	0.43	0.63
20	0.06	0.30	0.09	0.24	0.19	0.33	0.09	0.19	0.12	0.45	0.46	0.36	0.02	0.05	0.13
21	0.24	0.39	0.48	0.19	0.40	0.39	0.46	0.26	0.45	0.57	0.22	0.18	0.54	0.56	0.13
22	0.35	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.47	0.46	0.60	0.49	0.21
23	0.02	0.38	0.48	0.34	0.09	0.40	0.54	0.42	0.13	0.60	0.56	0.23	0.43	0.49	0.37
24	0.22	0.56	0.21	0.60	0.63	0.56	0.25	0.60	0.63	0.56	0.19	0.63	0.60	0.63	0.63
25	0.22	0.45	0.56	0.56	0.63	0.33	0.54	0.54	0.63	0.38	0.38	0.63		0.63	0.63
26	0.12	0.08	0.54	0.48	0.15	0.10	0.54	0.50	0.23	0.52	0.33	0.24	0.44	0.38	0.21
27	0.45	0.09	0.63	0.63	0.16	0.08	0.63	0.63	0.34	0.63	0.63	0.29	0.46	0.63	0.63
28	0.11	0.16	0.63	0.60	0.63	0.15	0.63	0.60	0.63	0.60	0.49	0.63	0.44	0.54	0.58
29	0.30	0.44	0.41	0.41	0.63	0.53	0.41	0.41	0.63	0.60	0.60	0.63		0.63	0.63
30	0.27	0.15	0.56	0.44	0.58	0.06	0.30	0.17	0.55	0.27	0.22	0.43	0.22	0.24	0.43
31	0.22	0.38	0.63	0.63	0.09	0.38	0.63	0.63	0.08	0.39	0.33	0.48	0.18	0.50	0.47
32	0.27	0.56	0.63	0.60	0.58	0.56	0.63	0.60	0.58	0.58	0.35	0.09	0.49	0.60	0.54
33	0.36	0.41	0.11	0.11	0.63	0.45	0.22	0.22	0.63	0.32	0.32	0.63		0.63	0.63
34	0.06	0.03	0.63	0.44	0.63	0.04	0.57	0.49	0.63	0.63	0.52	0.63	0.36	0.63	0.63
35	0.22	0.31	0.08	0.38	0.63	0.28	0.15	0.40	0.63	0.30	0.56	0.63	0.26	0.63	0.63
36	0.07	0.41	0.63	0.16	0.63	0.34	0.63	0.16	0.63	0.51	0.20	0.63	0.63	0.63	0.63
37	0.30	0.02	0.63	0.60	0.63	0.31	0.63	0.60	0.63	0.63	0.63	0.63	0.63	0.56	0.63
38	0.01	0.31	0.63	0.63	0.63	0.33	0.63	0.63	0.63	0.60	0.56	0.63	0.57	0.07	0.41
39	0.06	0.25	0.63	0.57	0.63	0.26	0.63	0.60	0.63	0.63	0.56	0.60	0.63	0.23	0.56
40	0.06	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.26	0.19	0.63	0.35	0.63	0.63
41	0.33	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.49	0.63	0.63	0.39	0.63	0.63
42	0.07	0.63	0.60	0.19	0.63	0.63	0.60	0.41	0.63	0.63	0.63	0.63	0.63	0.63	0.63
43	0.09	0.63	0.44	0.63	0.63	0.63	0.47	0.63	0.63	0.63	0.63	0.40	0.63	0.63	0.63
44	0.09	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.60	0.13	0.63	0.63
45	0.03	0.49	0.63	0.63	0.63	0.53	0.63	0.63	0.63	0.63	0.63	0.63	0.07	0.49	0.44
46	0.31	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.60	0.63	0.63
47	0.08	0.63	0.57	0.63	0.63	0.63	0.52	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63
48	0.06	0.63	0.41	0.49	0.63	0.63	0.49	0.54	0.63	0.63	0.63	0.63	0.06	0.63	0.63
49	0.13	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63		0.63	0.63
50	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.60	0.41	0.63	0.63

Tabela B.5: Medidas de tamanho do efeito obtidos com o r para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 20

Bases	Pares de Classificadores														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.01	0.03	0.09	0.23	0.06	0.12	0.06	0.18	0.07	0.12	0.28	0.25	0.11	0.00	0.03
2	0.03	0.42	0.01	0.21	0.45	0.31	0.02	0.18	0.45	0.32	0.29	0.21	0.18	0.36	0.37
3	0.09	0.19	0.37	0.24	0.07	0.21	0.42	0.25	0.10	0.24	0.07	0.07	0.26	0.27	0.08
4	0.34	0.60	0.51	0.59	0.23	0.61	0.52	0.55	0.42	0.60	0.48	0.61	0.28	0.43	0.52
5	0.26	0.57	0.49	0.52	0.60	0.57	0.48	0.47	0.59	0.25	0.25	0.09	0.03	0.27	0.34
6	0.07	0.51	0.54	0.54	0.54	0.48	0.51	0.52	0.54	0.16	0.19	0.34	0.11	0.27	0.21
7	0.16	0.04	0.23	0.11	0.01	0.03	0.20	0.05	0.04	0.22	0.08	0.03	0.14	0.26	0.10
8	0.16	0.35	0.52	0.27	0.28	0.32	0.51	0.24	0.25	0.18	0.20	0.09	0.38	0.24	0.10
9	0.12	0.29	0.55	0.51	0.40	0.29	0.57	0.55	0.42	0.60	0.59	0.46	0.34	0.35	0.12
10	0.22	0.10	0.52	0.24	0.37	0.14	0.51	0.19	0.35	0.54	0.26	0.44	0.52	0.21	0.30
11	0.25	0.25	0.14	0.11	0.34	0.12	0.36	0.14	0.27	0.26	0.19	0.16	0.11	0.45	0.31
12	0.05	0.38	0.06	0.25	0.03	0.35	0.11	0.23	0.04	0.37	0.17	0.42	0.46	0.06	0.24
13	0.16	0.17	0.24	0.01	0.52	0.18	0.22	0.02	0.51	0.22	0.12	0.55	0.17	0.40	0.45
14	0.16	0.07	0.50	0.31	0.39	0.16	0.52	0.33	0.46	0.50	0.32	0.45	0.43	0.28	0.20
15		0.62	0.62	0.61	0.62	0.62	0.62	0.61	0.62	0.60	0.61	0.16	0.07	0.60	0.61
16	0.16	0.18	0.19	0.10	0.43	0.15	0.23	0.07	0.45	0.45	0.11	0.46	0.32	0.22	0.48
17	0.09	0.14	0.42	0.47	0.60	0.11	0.43	0.48	0.62	0.44	0.56	0.61	0.04	0.55	0.54
18	0.20	0.51	0.57	0.57	0.61	0.52	0.57	0.57	0.59	0.06	0.06	0.54		0.47	0.47
19	0.07	0.30	0.47	0.18	0.60	0.34	0.49	0.24	0.60	0.34	0.16	0.53	0.50	0.35	0.61
20	0.17	0.23	0.19	0.16	0.07	0.27	0.12	0.05	0.04	0.36	0.33	0.31	0.08	0.14	0.09
21	0.02	0.39	0.35	0.17	0.22	0.36	0.34	0.20	0.20	0.50	0.13	0.19	0.42	0.39	0.03
22	0.15	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.59	0.24	0.37	0.60	0.53	0.06
23	0.31	0.13	0.40	0.29	0.03	0.22	0.34	0.22	0.08	0.55	0.39	0.16	0.30	0.42	0.22
24		0.57	0.13	0.55	0.62	0.57	0.13	0.55	0.62	0.53	0.02	0.62	0.53	0.62	0.62
25	0.32	0.50	0.54	0.54	0.62	0.31	0.38	0.38	0.62	0.31	0.31	0.62		0.62	0.62
26	0.18	0.03	0.40	0.20	0.20	0.04	0.46	0.32	0.26	0.39	0.25	0.31	0.25	0.19	0.07
27	0.18	0.17	0.62	0.62	0.04	0.06	0.62	0.62	0.15	0.61	0.62	0.18	0.13	0.61	0.62
28	0.14	0.03	0.52	0.46	0.60	0.13	0.55	0.52	0.62	0.53	0.47	0.59	0.13	0.36	0.42
29	NA	0.36	0.33	0.33	0.62	0.36	0.33	0.33	0.62	0.50	0.50	0.62		0.62	0.62
30	0.04	0.01	0.15	0.06	0.20	0.03	0.22	0.04	0.21	0.09	0.03	0.28	0.20	0.17	0.23
31	0.26	0.18	0.61	0.52	0.06	0.24	0.60	0.53	0.14	0.46	0.26	0.21	0.27	0.45	0.32
32	0.07	0.54	0.62	0.61	0.60	0.46	0.62	0.55	0.49	0.49	0.35	0.16	0.37	0.52	0.37
33	0.13	0.23	0.04	0.04	0.62	0.21	0.09	0.09	0.62	0.23	0.23	0.62		0.62	0.62
34	0.05		0.59	0.34	0.61	0.04	0.58	0.37	0.62	0.56	0.40	0.62	0.38	0.50	0.59
35		0.22	0.12	0.35	0.62	0.22	0.12	0.35	0.62	0.26	0.47	0.62	0.29	0.62	0.62
36	0.12	0.38	0.53	0.01	0.62	0.32	0.55	0.07	0.62	0.42	0.23	0.62	0.61	0.62	0.62
37	0.00	0.22	0.60	0.43	0.62	0.31	0.59	0.39	0.62	0.59	0.44	0.62	0.59	0.49	0.62
38	0.13	0.29	0.62	0.62	0.60	0.25	0.62	0.62	0.59	0.57	0.47	0.60	0.34	0.05	0.35
39	0.09	0.02	0.58	0.42	0.52	0.07	0.57	0.39	0.55	0.61	0.40	0.53	0.55	0.07	0.31
40	0.12	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.21	0.12	0.62	0.32	0.62	0.62
41	0.02	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.31	0.56	0.62	0.23	0.62	0.62
42	0.09	0.62	0.60	0.36	0.62	0.62	0.57	0.31	0.62	0.62	0.62	0.62	0.58	0.62	0.62
43	0.03	0.62	0.36	0.60	0.62	0.62	0.39	0.62	0.62	0.62	0.62	0.41	0.52	0.62	0.62
44	0.19	0.62	0.59	0.62	0.62	0.62	0.59	0.62	0.62	0.60	0.62	0.61	0.03	0.62	0.62
45	0.02	0.31	0.62	0.61	0.62	0.35	0.62	0.61	0.60	0.61	0.60	0.60	0.00	0.31	0.40
46	0.17	0.62	0.62	0.62	0.62	0.62	0.61	0.62	0.62	0.62	0.62	0.62	0.60	0.62	0.62
47	0.08	0.62	0.26	0.62	0.62	0.62	0.32	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62
48	0.07	0.61	0.24	0.37	0.62	0.61	0.22	0.35	0.62	0.60	0.62	0.62	0.09	0.62	0.62
49	0.36	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62		0.62	0.62
50	0.23	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.32	0.62	0.62

Tabela B.6: Medidas de tamanho do efeito obtidos com o r para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 30

Bases	Pares de Classificadores														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.22	0.04	0.06	0.16	0.06	0.04	0.11	0.24	0.16	0.08	0.14	0.17	0.07	0.02	0.02
2	0.12	0.18	0.10	0.06	0.26	0.26	0.04	0.13	0.30	0.30	0.24	0.08	0.18	0.33	0.26
3	0.18	0.08	0.25	0.09	0.08	0.15	0.29	0.19	0.01	0.12	0.09	0.19	0.23	0.28	0.14
4	0.12	0.60	0.48	0.54	0.23	0.61	0.49	0.57	0.25	0.53	0.42	0.59	0.24	0.34	0.50
5	0.16	0.50	0.37	0.34	0.51	0.51	0.40	0.37	0.52	0.25	0.22	0.07	0.05	0.31	0.33
6	0.13	0.37	0.41	0.48	0.53	0.40	0.42	0.49	0.55	0.08	0.21	0.23	0.05	0.19	0.16
7	0.13		0.27	0.11	0.04	0.06	0.27	0.15	0.01	0.29	0.11	0.04	0.26	0.27	0.11
8	0.14	0.43	0.36	0.21	0.30	0.34	0.34	0.16	0.29	0.05	0.25	0.11	0.29	0.18	0.19
9	0.03	0.31	0.57	0.54	0.34	0.30	0.57	0.52	0.34	0.60	0.58	0.45	0.28	0.31	0.12
10	0.17	0.16	0.47	0.15	0.27	0.09	0.47	0.19	0.32	0.45	0.20	0.38	0.46	0.16	0.15
11	0.09	0.06	0.17	0.20	0.29	0.10	0.11	0.13	0.29	0.22	0.20	0.30	0.03	0.38	0.32
12	0.18	0.30	0.02	0.20	0.19	0.35	0.09	0.25	0.09	0.24	0.16	0.49	0.29	0.22	0.31
13	0.13	0.30	0.03	0.07	0.48	0.32	0.01	0.09	0.48	0.24	0.20	0.56	0.08	0.42	0.50
14	0.06	0.06	0.40	0.24	0.27		0.43	0.29	0.25	0.43	0.29	0.25	0.34	0.20	0.04
15		0.62	0.59	0.60	0.62	0.62	0.59	0.60	0.62	0.60	0.62	0.13	0.10	0.60	0.62
16	0.12	0.30	0.03	0.26	0.29	0.26	0.01	0.21	0.33	0.30	0.10	0.40	0.20	0.25	0.36
17	0.10	0.01	0.32	0.39	0.59	0.07	0.37	0.46	0.59	0.37	0.43	0.59	0.07	0.39	0.54
18	0.00	0.36	0.50	0.50	0.54	0.35	0.50	0.50	0.53	0.02	0.02	0.50		0.41	0.41
19	0.00	0.16	0.48	0.11	0.55	0.15	0.46	0.10	0.55	0.33	0.05	0.42	0.44	0.28	0.52
20	0.10	0.13	0.09	0.08	0.06	0.21	0.06	0.05	0.03	0.20	0.19	0.18	0.04	0.06	0.03
21	0.17	0.17	0.33	0.09	0.08	0.13	0.38	0.04	0.06	0.41	0.18	0.15	0.34	0.31	0.04
22	0.04	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.53	0.16	0.28	0.56	0.46	0.20
23	0.09	0.09	0.36	0.32	0.01	0.08	0.39	0.31	0.01	0.42	0.34	0.12	0.25	0.33	0.22
24	0.13	0.44	0.04	0.48	0.62	0.38	0.05	0.49	0.62	0.41	0.00	0.62	0.45	0.62	0.62
25	0.13	0.24	0.43	0.43	0.62	0.15	0.40	0.40	0.62	0.37	0.37	0.62		0.62	0.62
26	0.03	0.09	0.36	0.30	0.26	0.08	0.35	0.31	0.24	0.24	0.26	0.39	0.14	0.10	0.06
27	0.10	0.07	0.59	0.61	0.13	0.09	0.60	0.61	0.12	0.56	0.59	0.08	0.14	0.58	0.59
28	0.04	0.11	0.50	0.44	0.59	0.11	0.52	0.45	0.60	0.39	0.35	0.56	0.09	0.37	0.40
29	0.00	0.31	0.27	0.27	0.62	0.33	0.29	0.29	0.62	0.46	0.46	0.62		0.62	0.62
30	0.06	0.05	0.25	0.09	0.23	0.04	0.19	0.04	0.23	0.10	0.00	0.26	0.11	0.08	0.22
31	0.17	0.16	0.53	0.38	0.02	0.19	0.57	0.47	0.05	0.36	0.30	0.15	0.18	0.39	0.36
32	0.02	0.36	0.58	0.49	0.39	0.40	0.60	0.52	0.42	0.35	0.20	0.07	0.23	0.42	0.25
33	0.12	0.28	0.00	0.00	0.62	0.28	0.08	0.08	0.61	0.22	0.22	0.61		0.60	0.60
34	0.09	0.11	0.42	0.32	0.57	0.09	0.46	0.34	0.58	0.46	0.40	0.60	0.21	0.41	0.52
35	0.18	0.21	0.05	0.21	0.62	0.29	0.01	0.15	0.62	0.19	0.37	0.62	0.17	0.62	0.62
36	0.02	0.32	0.58	0.10	0.62	0.36	0.57	0.13	0.62	0.48	0.11	0.62	0.48	0.62	0.62
37	0.34	0.31	0.60	0.35	0.59	0.04	0.61	0.46	0.59	0.60	0.45	0.60	0.56	0.37	0.58
38	0.06	0.25	0.61	0.55	0.58	0.27	0.61	0.57	0.61	0.53	0.40	0.56	0.28	0.00	0.20
39	0.01	0.09	0.56	0.42	0.54	0.10	0.57	0.46	0.53	0.55	0.43	0.52	0.46	0.07	0.33
40	0.12	0.62	0.61	0.62	0.62	0.62	0.61	0.62	0.62	0.26	0.04	0.62	0.27	0.62	0.62
41	0.02	0.62	0.61	0.62	0.62	0.62	0.61	0.62	0.62	0.17	0.48	0.62	0.29	0.61	0.61
42	0.26	0.62	0.54	0.25	0.62	0.62	0.58	0.27	0.62	0.62	0.62	0.59	0.52	0.62	0.62
43	0.13	0.62	0.26	0.52	0.62	0.62	0.30	0.51	0.62	0.62	0.62	0.32	0.44	0.62	0.62
44	0.04	0.62	0.57	0.59	0.62	0.62	0.56	0.59	0.62	0.61	0.61	0.57	0.14	0.61	0.62
45	0.19	0.28	0.58	0.54	0.59	0.35	0.59	0.56	0.60	0.47	0.48	0.53	0.07	0.20	0.24
46	0.25	0.62	0.56	0.61	0.62	0.62	0.58	0.62	0.62	0.62	0.62	0.62	0.57	0.62	0.62
47	0.18	0.62	0.21	0.62	0.62	0.62	0.30	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62
48	0.04	0.59	0.24	0.29	0.62	0.61	0.25	0.30	0.62	0.62	0.62	0.62	0.09	0.62	0.62
49	0.28	0.62	0.61	0.61	0.62	0.62	0.61	0.61	0.62	0.62	0.62	0.62		0.62	0.62
50	0.15	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.25	0.62	0.62

Tabela B.7: Poder dos testes de Wilcoxon obtidos para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 10

Bases	Pares de Classificadores														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.06	0.40	0.06	0.09	0.05	0.33	0.04	0.12	0.07	0.17	0.48	0.27	0.11	0.05	0.06
2	0.08	0.77	0.04	0.17	0.78	0.49	0.06	0.08	0.57	0.63	0.43	0.11	0.13	0.66	0.50
3		0.26	0.64	0.10	0.09	0.28	0.64	0.10	0.10	0.30	0.08	0.08	0.38	0.42	0.04
4	0.04	1.00	0.99	1.00	0.15	1.00	0.97	0.99	0.16	0.91	0.17	1.00	0.30	0.80	0.94
5	0.04	1.00	0.80	0.91	0.99	1.00	0.71	0.84	0.97	0.49	0.35	0.14	0.06	0.45	0.38
6	0.06	0.69	0.93	0.89	1.00	0.69	0.94	0.92	0.99	0.17	0.15	0.65	0.06	0.31	0.36
7	0.06	0.05	0.20	0.06	0.04	0.05	0.13	0.04	0.05	0.18	0.04	0.05	0.16	0.24	0.07
8	0.06	0.43	0.57	0.33	0.28	0.60	0.71	0.57	0.42	0.07	0.10	0.05	0.18	0.09	0.05
9	0.06	0.09	1.00	0.99	0.62	0.16	1.00	1.00	0.59	1.00	1.00	0.72	0.78	0.74	0.18
10	0.09	0.44	0.80	0.12	0.31	0.31	0.86	0.17	0.34	0.94	0.40	0.46	0.56	0.07	0.18
11	0.07	0.07	0.11	0.05	0.30	0.13	0.07	0.07	0.43	0.24	0.07	0.27	0.13	0.66	0.31
12	0.07	0.11	0.05	0.07	0.05	0.09	0.05	0.08	0.06	0.13	0.06	0.15	0.07	0.05	0.08
13	0.09	0.50	0.08	0.05	0.91	0.32	0.16	0.07	0.96	0.43	0.28	0.98	0.07	0.79	0.88
14	0.07	0.05	0.82	0.33	0.64	0.07	0.85	0.42	0.71	0.81	0.33	0.63	0.42	0.26	0.11
15		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.03	0.10	1.00	1.00
16	0.11	0.48	0.05	0.14	0.38	0.24	0.12	0.05	0.59	0.49	0.13	0.83	0.13	0.42	0.60
17	0.05	0.05	0.69	0.91	1.00	0.09	0.49	0.68	1.00	0.81	1.00	1.00	0.05	0.81	0.96
18	0.05	0.92	0.79	0.75	1.00	0.88	0.72	0.72	1.00	0.14	0.12	0.69		0.90	0.90
19	0.05	0.09	0.88	0.08	1.00	0.07	0.87	0.06	1.00	0.66	0.04	0.99	0.67	0.42	0.98
20	0.05	0.10	0.07	0.10	0.07	0.11	0.06	0.06	0.06	0.19	0.22	0.14	0.04	0.06	0.06
21	0.04	0.45	0.24	0.12	0.18	0.46	0.18	0.13	0.19	0.94	0.20	0.12	0.57	0.68	0.06
22	0.07	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	0.15	0.34	0.91	0.69	0.10
23	0.05	0.16	0.61	0.19	0.07	0.21	0.71	0.25	0.08	0.90	0.52	0.07	0.16	0.62	0.27
24	0.08	0.98	0.12	0.98	1.00	0.99	0.21	0.99	1.00	0.88	0.11	1.00	0.91	1.00	1.00
25	0.13	0.59	0.96	0.96	1.00	0.29	0.77	0.77	1.00	0.18	0.16	1.00		1.00	1.00
26	0.04	0.04	0.73	0.55	0.10	0.05	0.78	0.66	0.10	0.58	0.27	0.06	0.24	0.30	0.09
27	0.11	0.07	1.00	1.00	0.06	0.05	1.00	1.00	0.18	1.00	1.00	0.12	0.32	1.00	1.00
28	0.05	0.08	0.89	0.78	1.00	0.09	0.86	0.73	1.00	0.66	0.45	1.00	0.11	0.68	0.92
29	0.07	0.62	0.52	0.50	1.00	0.72	0.42	0.41	1.00	0.98	0.97	1.00		1.00	1.00
30	0.05	0.06	0.28	0.14	0.58	0.06	0.15	0.07	0.37	0.22	0.13	0.64	0.09	0.12	0.32
31	0.06	0.42	0.99	0.94	0.06	0.35	0.97	0.88	0.05	0.46	0.28	0.16	0.09	0.75	0.59
32	0.05	0.60	1.00	0.96	0.72	0.70	1.00	0.98	0.80	0.83	0.18	0.05	0.51	0.97	0.28
33	0.07	0.25	0.07	0.07	1.00	0.34	0.13	0.13	1.00	0.22	0.22	1.00		1.00	1.00
34	0.05	0.06	0.79	0.35	1.00	0.06	0.69	0.28	1.00	0.89	0.36	1.00	0.20	0.89	0.98
35	0.05	0.13	0.05	0.18	1.00	0.12	0.05	0.20	1.00	0.18	0.45	1.00	0.13	1.00	1.00
36	0.05	0.36	0.88	0.06	1.00	0.34	0.90	0.05	1.00	0.78	0.13	1.00	0.74	1.00	1.00
37	0.07	0.06	1.00	0.98	1.00	0.16	1.00	0.99	1.00	1.00	1.00	1.00	1.00	0.86	1.00
38	0.04	0.20	1.00	0.98	1.00	0.19	1.00	0.99	1.00	0.93	0.68	0.98	0.28	0.06	0.38
39	0.05	0.14	1.00	0.98	1.00	0.17	1.00	1.00	1.00	1.00	0.84	0.99	1.00	0.11	0.73
40	0.05	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.09	0.06	1.00	0.12	1.00	1.00
41	0.06	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.54	1.00	1.00	0.21	1.00	1.00
42	0.05	1.00	0.78	0.14	1.00	1.00	0.82	0.14	1.00	1.00	1.00	1.00	0.60	1.00	1.00
43	0.05	1.00	0.42	0.97	1.00	1.00	0.39	0.97	1.00	1.00	1.00	0.32	0.60	1.00	1.00
44	0.03	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.89	0.06	1.00	1.00
45	0.05	0.19	0.86	0.98	0.99	0.26	0.94	1.00	1.00	0.78	0.95	1.00	0.04	0.14	0.19
46	0.08	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.93	1.00	1.00
47	0.05	1.00	0.20	1.00	1.00	1.00	0.18	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
48	0.06	1.00	0.42	0.52	1.00	1.00	0.47	0.53	1.00	1.00	1.00	1.00	0.04	1.00	1.00
49	0.08	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00		1.00	1.00
50	0.24	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.07	1.00	1.00

Tabela B.8: Poder dos testes de Wilcoxon obtidos para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 20

Bases	Pares de Classificadores														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.05	0.05	0.10	0.16	0.10	0.06	0.09	0.17	0.08	0.11	0.24	0.12	0.06	0.05	0.06
2	0.05	0.83	0.04	0.15	0.86	0.68	0.04	0.11	0.75	0.60	0.48	0.18	0.10	0.71	0.57
3	0.05	0.11	0.38	0.08	0.06	0.17	0.45	0.10	0.06	0.18	0.07	0.08	0.27	0.26	0.04
4	0.09	1.00	0.94	1.00	0.14	1.00	0.98	1.00	0.30	0.96	0.62	1.00	0.28	0.72	0.96
5	0.08	1.00	0.88	0.88	1.00	1.00	0.81	0.82	0.99	0.34	0.45	0.11	0.05	0.38	0.41
6	0.04	0.70	0.79	0.87	0.99	0.67	0.79	0.86	0.99	0.13	0.16	0.70	0.06	0.33	0.29
7	0.05	0.04	0.20	0.08	0.04	0.04	0.18	0.06	0.05	0.17	0.05	0.04	0.12	0.21	0.08
8	0.04	0.54	0.79	0.24	0.39	0.55	0.76	0.21	0.37	0.14	0.19	0.07	0.48	0.23	0.11
9	0.06	0.16	1.00	0.95	0.43	0.11	1.00	0.98	0.55	1.00	1.00	0.78	0.48	0.55	0.14
10	0.07	0.06	0.98	0.20	0.75	0.13	0.96	0.12	0.67	0.99	0.32	0.84	0.84	0.16	0.42
11	0.09	0.17	0.11	0.05	0.48	0.09	0.29	0.09	0.27	0.45	0.14	0.13	0.08	0.73	0.39
12	0.05	0.21	0.06	0.08	0.04	0.20	0.07	0.09	0.06	0.24	0.09	0.19	0.12	0.05	0.10
13	0.06	0.15	0.12	0.05	0.96	0.20	0.09	0.05	0.96	0.29	0.14	1.00	0.08	0.83	0.93
14	0.06	0.06	0.97	0.30	0.54	0.05	0.97	0.41	0.58	0.97	0.30	0.56	0.65	0.35	0.09
15		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.06	0.08	1.00	1.00
16	0.06	0.21	0.27	0.06	0.65	0.18	0.33	0.04	0.73	0.77	0.15	0.94	0.44	0.30	0.78
17	0.05	0.09	0.78	0.89	1.00	0.07	0.74	0.89	1.00	0.82	0.99	1.00	0.05	0.97	1.00
18	0.06	0.96	0.97	0.95	1.00	0.97	0.97	0.97	1.00	0.07	0.05	0.89		0.74	0.74
19	0.06	0.14	0.82	0.08	1.00	0.15	0.88	0.09	1.00	0.49	0.07	0.98	0.52	0.49	0.99
20	0.06	0.10	0.08	0.06	0.05	0.13	0.05	0.05	0.05	0.24	0.17	0.16	0.06	0.07	0.05
21	0.06	0.60	0.21	0.10	0.16	0.55	0.20	0.11	0.13	0.98	0.14	0.17	0.47	0.58	0.05
22	0.06	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.22	0.36	0.98	0.94	0.08
23	0.06	0.10	0.75	0.17	0.05	0.16	0.63	0.10	0.07	0.96	0.42	0.10	0.33	0.78	0.21
24		1.00	0.12	1.00	1.00	1.00	0.10	1.00	1.00	1.00	0.05	1.00	1.00	1.00	1.00
25	0.17	0.55	0.86	0.86	1.00	0.16	0.41	0.43	1.00	0.13	0.14	1.00		1.00	1.00
26	0.07	0.06	0.49	0.25	0.18	0.05	0.55	0.32	0.22	0.41	0.22	0.15	0.10	0.16	0.04
27	0.06	0.06	1.00	1.00	0.06	0.07	1.00	1.00	0.06	1.00	1.00	0.12	0.12	1.00	1.00
28	0.06	0.06	0.90	0.87	1.00	0.07	0.96	0.93	1.00	0.88	0.82	1.00	0.08	0.49	0.74
29		0.83	0.45	0.48	1.00	0.85	0.45	0.45	1.00	0.99	0.98	1.00		1.00	1.00
30	0.04	0.05	0.06	0.05	0.21	0.04	0.09	0.03	0.25	0.09	0.05	0.26	0.09	0.12	0.24
31	0.08	0.13	0.98	0.66	0.04	0.27	1.00	0.88	0.07	0.84	0.34	0.10	0.20	0.95	0.57
32	0.04	0.89	1.00	1.00	0.90	0.84	1.00	0.99	0.86	0.82	0.25	0.05	0.34	0.96	0.44
33	0.06	0.20	0.04	0.07	1.00	0.28	0.06	0.08	1.00	0.19	0.20	1.00		1.00	1.00
34	0.05	0.06	0.90	0.39	1.00	0.04	0.92	0.43	1.00	0.95	0.44	1.00	0.32	0.80	0.98
35		0.13	0.05	0.16	1.00	0.11	0.05	0.17	1.00	0.18	0.40	1.00	0.12	1.00	1.00
36	0.06	0.50	0.95	0.05	1.00	0.66	0.97	0.04	1.00	0.89	0.37	1.00	0.93	1.00	1.00
37	0.04	0.21	1.00	0.65	1.00	0.34	1.00	0.63	1.00	1.00	0.87	1.00	0.98	0.84	1.00
38	0.06	0.46	1.00	1.00	1.00	0.35	1.00	1.00	1.00	0.99	0.89	1.00	0.38	0.07	0.55
39	0.06	0.06	1.00	0.72	0.97	0.07	1.00	0.78	0.99	1.00	0.70	0.98	0.79	0.07	0.53
40	0.06	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.20	0.08	1.00	0.28	1.00	1.00
41	0.06	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.53	1.00	1.00	0.32	1.00	1.00
42	0.05	1.00	0.93	0.24	1.00	1.00	0.91	0.21	1.00	1.00	1.00	1.00	0.60	1.00	1.00
43	0.05	1.00	0.38	0.96	1.00	1.00	0.40	0.97	1.00	1.00	1.00	0.45	0.62	1.00	1.00
44	0.05	1.00	0.97	0.98	1.00	1.00	0.98	0.99	1.00	1.00	1.00	1.00	0.06	1.00	1.00
45	0.05	0.20	0.97	0.98	1.00	0.19	0.96	0.97	1.00	0.86	0.87	1.00	0.06	0.17	0.25
46	0.06	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00
47	0.05	1.00	0.10	1.00	1.00	1.00	0.10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
48	0.04	1.00	0.26	0.55	1.00	1.00	0.20	0.45	1.00	1.00	1.00	1.00	0.06	1.00	1.00
49	0.21	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00		1.00	1.00
50	0.08	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.13	1.00	1.00

Tabela B.9: Poder dos testes de Wilcoxon obtidos para cada combinação de uma base de dados e de um par de classificadores considerando amostras de tamanho 30

Bases	Pares de Classificadores														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.08	0.07	0.05	0.11	0.06	0.04	0.08	0.19	0.15	0.10	0.16	0.13	0.07	0.06	0.07
2	0.05	0.35	0.05	0.06	0.45	0.54	0.06	0.09	0.66	0.70	0.32	0.10	0.11	0.75	0.42
3	0.07	0.08	0.26	0.05	0.06	0.16	0.38	0.07	0.06	0.16	0.07	0.21	0.26	0.44	0.10
4	0.04	1.00	0.94	1.00	0.15	1.00	0.94	1.00	0.18	0.96	0.72	1.00	0.22	0.73	0.97
5	0.08	1.00	0.63	0.71	1.00	1.00	0.76	0.81	1.00	0.45	0.26	0.21	0.06	0.61	0.48
6	0.05	0.78	0.95	0.98	0.99	0.79	0.97	0.98	0.99	0.08	0.14	0.48	0.08	0.36	0.29
7	0.06	0.05	0.26	0.06	0.06	0.06	0.33	0.10	0.05	0.24	0.07	0.06	0.15	0.35	0.10
8	0.06	0.56	0.83	0.28	0.49	0.46	0.73	0.18	0.35	0.08	0.18	0.06	0.44	0.20	0.14
9	0.05	0.16	1.00	1.00	0.63	0.14	1.00	1.00	0.60	1.00	1.00	0.82	0.47	0.58	0.15
10	0.06	0.11	0.92	0.10	0.49	0.08	0.96	0.16	0.59	0.98	0.28	0.70	0.78	0.20	0.28
11	0.06	0.08	0.16	0.14	0.47	0.10	0.12	0.12	0.48	0.33	0.27	0.30	0.05	0.84	0.73
12	0.05	0.17	0.05	0.07	0.08	0.25	0.04	0.11	0.05	0.18	0.08	0.33	0.06	0.07	0.14
13	0.05	0.33	0.04	0.06	0.98	0.44	0.05	0.07	0.98	0.29	0.18	1.00	0.07	0.94	0.98
14	0.05	0.05	0.88	0.27	0.28		0.89	0.22	0.29	0.88	0.25	0.28	0.44	0.24	0.06
15		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.05	0.11	1.00	1.00
16	0.08	0.30	0.05	0.22	0.34	0.18	0.06	0.12	0.53	0.28	0.05	0.80	0.16	0.42	0.73
17	0.05	0.04	0.74	0.71	1.00	0.07	0.85	0.85	1.00	0.91	0.94	1.00	0.07	0.90	0.98
18	0.05	0.89	0.90	0.92	1.00	0.83	0.86	0.86	1.00	0.06	0.05	0.94		0.95	0.93
19	0.06	0.08	0.79	0.07	1.00	0.09	0.78	0.06	1.00	0.52	0.06	0.98	0.59	0.40	0.97
20	0.07	0.11	0.08	0.08	0.06	0.19	0.05	0.06	0.04	0.26	0.28	0.19	0.05	0.05	0.05
21	0.05	0.31	0.31	0.08	0.11	0.21	0.32	0.06	0.07	0.94	0.23	0.13	0.54	0.68	0.06
22	0.05	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.15	0.54	0.97	0.96	0.17
23	0.07	0.05	0.84	0.31	0.06	0.05	0.79	0.33	0.04	0.92	0.49	0.07	0.31	0.70	0.25
24	0.05	0.99	0.10	0.99	1.00	0.96	0.07	0.96	1.00	0.95	0.05	1.00	0.95	1.00	1.00
25	0.10	0.40	0.94	0.94	1.00	0.16	0.74	0.72	1.00	0.36	0.33	1.00		1.00	1.00
26	0.05	0.06	0.38	0.31	0.26	0.06	0.40	0.32	0.25	0.34	0.23	0.19	0.07	0.11	0.06
27	0.05	0.04	1.00	1.00	0.08	0.06	1.00	1.00	0.14	1.00	1.00	0.09	0.10	1.00	1.00
28	0.06	0.15	0.99	0.94	1.00	0.12	0.99	0.95	1.00	0.88	0.72	1.00	0.10	0.84	0.95
29	0.06	0.78	0.48	0.45	1.00	0.77	0.55	0.56	1.00	0.99	0.99	1.00		1.00	1.00
30	0.05	0.05	0.10	0.07	0.31	0.07	0.08	0.05	0.22	0.09	0.06	0.26	0.08	0.10	0.19
31	0.08	0.13	0.98	0.86	0.06	0.24	1.00	0.95	0.04	0.88	0.55	0.14	0.24	0.96	0.78
32	0.05	0.73	1.00	0.98	0.70	0.77	1.00	1.00	0.74	0.69	0.24	0.06	0.28	0.85	0.39
33	0.05	0.30	0.04	0.07	1.00	0.42	0.07	0.07	1.00	0.32	0.32	1.00		1.00	1.00
34	0.04	0.10	0.82	0.46	1.00	0.06	0.86	0.52	1.00	0.96	0.72	1.00	0.22	0.84	1.00
35	0.06	0.12	0.05	0.16	1.00	0.22	0.04	0.07	1.00	0.16	0.40	1.00	0.12	1.00	1.00
36	0.04	0.66	1.00	0.11	1.00	0.50	1.00	0.10	1.00	0.99	0.12	1.00	0.96	1.00	1.00
37	0.22	0.42	1.00	0.79	1.00	0.06	1.00	0.96	1.00	1.00	0.99	1.00	1.00	0.70	1.00
38	0.05	0.32	1.00	0.99	1.00	0.34	1.00	0.99	1.00	0.99	0.77	0.99	0.22	0.06	0.29
39	0.04	0.12	1.00	0.90	0.99	0.17	1.00	0.94	1.00	1.00	0.86	0.99	0.98	0.08	0.75
40	0.06	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.29	0.06	1.00	0.30	1.00	1.00
41	0.04	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.28	0.99	1.00	0.26	1.00	1.00
42	0.07	1.00	0.93	0.14	1.00	1.00	0.98	0.25	1.00	1.00	1.00	1.00	0.69	1.00	1.00
43	0.06	1.00	0.38	0.99	1.00	1.00	0.45	0.99	1.00	1.00	1.00	0.37	0.65	1.00	1.00
44	0.05	1.00	0.98	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	0.07	1.00	1.00
45	0.08	0.19	0.98	0.98	1.00	0.36	1.00	1.00	1.00	0.90	0.89	0.99	0.06	0.13	0.27
46	0.08	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00
47	0.06	1.00	0.15	1.00	1.00	1.00	0.24	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
48	0.04	1.00	0.24	0.37	1.00	1.00	0.26	0.36	1.00	1.00	1.00	1.00	0.07	1.00	1.00
49	0.11	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00		1.00	1.00
50	0.07	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.25	1.00	1.00