

UNIVERSIDADE FEDERAL DO RIO GRANDE  
CENTRO DE CIÊNCIAS COMPUTACIONAIS  
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO  
CURSO DE MESTRADO EM ENGENHARIA DE COMPUTAÇÃO

Dissertação de Mestrado

**Uma análise do impacto da diversidade sobre o resultado  
do empilhamento de classificadores supervisionados**

Mariele de Almeida Lanes

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal do Rio Grande, como requisito parcial para a obtenção do grau de Mestre em Engenharia de Computação

Orientador: Prof. Dr. Eduardo Borges Nunes

Rio Grande, 2017

### Ficha catalográfica

L267a Lanes, Mariele de Almeida.

Uma análise do impacto da diversidade sobre o resultado do empilhamento de classificadores supervisionados / Mariele de Almeida Lanes. – 2017.

70 f.

Dissertação (mestrado) – Universidade Federal do Rio Grande – FURG, Programa de Pós-graduação em Engenharia de Computação, Rio Grande/RS, 2017.

Orientador: Dr. Eduardo Nunes Borges.

1. Classificação 2. Combinação de classificadores  
3. Empilhamento 4. Diversidade I. Borges, Eduardo Nunes II. Título.

CDU 004.421

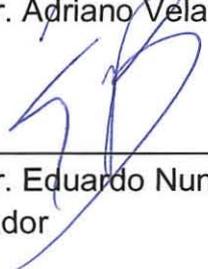
## ATA DE SESSÃO DE DEFESA DE DISSERTAÇÃO DE MESTRADO

Ata nº \_\_\_\_/2017

Na data de 29 de março de 2017, às 14 horas, ocorreu a Sessão de Defesa de Dissertação de Mestrado de Mariele de Almeida Lanes, que apresentou a dissertação intitulada "Uma análise do impacto da diversidade sobre o resultado do empilhamento de classificadores supervisionados", realizada sob a orientação do Prof. Dr. Eduardo Nunes Borges. A banca examinadora foi constituída pelos Profs. Drs. Renata de Matos Galante (UFRGS), Silvia Silva da Costa Botelho (FURG) e Adriano Velasque Werhli (FURG), sob a presidência do orientador. Após a apresentação do trabalho, a banca arguiu o candidato e, a seguir, deliberou pela

- aprovação da Dissertação
- aprovação da Dissertação, sugerindo modificações no texto
- reprovação da Dissertação

Rio Grande, 29 de março de 2017

\_\_\_\_\_  
Profa. Dra. Renata de Matos Galante\_\_\_\_\_  
Profa. Dra. Silvia Silva da Costa Botelho\_\_\_\_\_  
Prof. Dr. Adriano Velasque Werhli\_\_\_\_\_  
Prof. Dr. Eduardo Nunes Borges  
Orientador

## **AGRADECIMENTOS**

Início meus agradecimentos a DEUS, já que ele colocou pessoas tão especiais no meu caminho, sem as quais eu não teria conseguido concluir essa etapa.

Aos meus pais, pelo amor, incentivo e apoio incondicional.

Ao meu orientador, Professor Dr. Eduardo Nunes Borges, pela sua compreensão e paciência, além de sua dedicação, competência e especial atenção nas revisões e sugestões, fatores fundamentais para a conclusão deste trabalho.

A meus amigos do mestrado, pelos momentos divididos juntos, especialmente à Luisa, Narusci, Thiago e Plauto, que se demonstraram verdadeiros amigos e tornaram mais leve minha jornada de viagens a Rio Grande. Foi bom poder contar com vocês!

Ao Diretor do Campus Santo Antônio da Patrulha da Universidade Federal do Rio Grande – FURG, Professor Dr. Antônio Luis Schifino Valente, e ao Professor Dr. Luciano Silva da Silva, pelo apoio e confiança no momento em que autorizaram minha liberação para a realização do mestrado. E aos demais colegas de Campus, pela compreensão nos momentos de ausência.

A todos os professores do mestrado que de alguma forma contribuíram para minha formação.

Ao Centro de Ciências Computacionais da FURG, pelo apoio a minha participação no mestrado.

Aos familiares e amigos (novos e antigos) que sempre me incentivaram e apoiaram nessa jornada.

## RESUMO

LANES, Mariele de Almeida. **Uma análise do impacto da diversidade sobre o resultado do empilhamento de classificadores supervisionados**. 2017. 70 f. Dissertação (Mestrado) – Programa de Pós-Graduação em Computação. Universidade Federal do Rio Grande, Rio Grande.

Devido ao crescimento da pesquisa na área de reconhecimento de padrões, cada vez mais são testados os limites das técnicas utilizadas para a tarefa de classificação. Com isso, percebe-se que classificadores especializados e devidamente configurados são bastante eficazes. No entanto, não é uma tarefa trivial escolher o classificador mais adequado para tratar um determinado problema e configurá-lo corretamente. Além disso, não existe um algoritmo ideal para resolver todos os problemas de predição. Dessa forma, a fim de melhorar o resultado do processo de classificação, algumas técnicas combinam o conhecimento adquirido individualmente pelos algoritmos de aprendizagem visando descobrir novos padrões ainda não identificados. Entre estas técnicas, destaca-se a estratégia de empilhamento (*stacking*). Esta estratégia consiste na combinação dos resultados dos classificadores base, induzidos por vários algoritmos de aprendizado utilizando o mesmo conjunto de dados, por meio de outro classificador chamado de meta-classificador. O objetivo geral deste trabalho é avaliar o impacto da diversidade dos classificadores na qualidade do empilhamento, tendo como objetivos específicos estudar o método de empilhamento e a diversidade dos classificadores supervisionados. A abordagem proposta é baseada na afirmação de que quanto maior a diversidade dos padrões aprendidos pelos classificadores base, maior será a qualidade do empilhamento. Além disso, realizamos uma série de experimentos que mostram o impacto de múltiplas medidas de diversidade sobre o ganho de empilhamento, considerando muitos conjuntos de dados reais extraídos do repositório de aprendizado de máquina *UCI* e algumas bases de dados sintéticas com diferentes distribuições espaciais bidimensionais para auxiliar na validação por inspeção visual. A partir dos resultados desses experimentos, percebe-se que não existe uma relação significativa entre diversidade e qualidade do empilhamento.

**Palavras-chave:** Classificação, combinação de classificadores, empilhamento, diversidade.

## ABSTRACT

LANES, Mariele de Almeida. **An analysis of the impact of diversity on stacking supervised classifiers**. 2017. 70 f. Dissertação (Mestrado) – Programa de Pós-Graduação em Computação. Universidade Federal do Rio Grande, Rio Grande.

Due to the growth of research in pattern recognition area, the limits of the techniques used for the classification task are increasingly tested. Thus, it is clear that specialized and properly configured classifiers are quite effective. However, it is not a trivial task to choose the most appropriate classifier for deal with a particular problem and set it up properly. In addition, there is no optimal algorithm to solve all prediction problems. Thus, in order to improve the results of the classification process, some techniques combine the knowledge acquired individually by the learning algorithms in order to discover new patterns not yet identified. Among these techniques, there is the stacking strategy. This strategy consists in the combination of outputs of base classifiers, induced by several learning algorithms using the same dataset, by means of another classifier called meta-classifier. The main goal of this paper is to evaluate the impact of the classifiers diversity in the quality of stacking. The specific objectives are to study the stacking strategy and the diversity of supervised classifiers. The proposed approach is based on the assertion that the greater the diversity of patterns learned by base classifiers, the higher the quality of stacking. Moreover, we have performed a lot of experiments that show the impact of multiple diversity measures on the gain of stacking, considering many real datasets extracted from UCI machine learning repository, and some synthetic databases with different two-dimensional spatial distributions to aid visual inspection validation. From the results of these experiments, we can see that there is no significant relationship between diversity and stacking quality.

**Keywords:** classification, combining classifiers, stacking, diversity.

## LISTA DE FIGURAS

Figura 1	Exemplo de árvore de decisão aprendida a partir das informações da Tabela 1. . . . .	16
Figura 2	Representação do algoritmo de <i>Stacking</i> . . . . .	18
Figura 3	Metodologia do trabalho proposto. . . . .	42
Figura 4	Base de dados Spiral, R15 e D13 . . . . .	47
Figura 5	Erros de classificação do empilhamento usando configuração padrão dos algoritmos . . . . .	53
Figura 6	Relação entre medida de discordância e o ganho do empilhamento . .	58
Figura 7	Relação entre Variância <i>Kohavi-Wolpert</i> e ganho do empilhamento . .	59
Figura 8	Relação entre entropia e ganho do empilhamento . . . . .	59
Figura 9	Relação entre estatística $Q$ e o ganho do empilhamento . . . . .	59
Figura 10	Relação entre coeficiente de correlação e ganho do empilhamento . .	60
Figura 11	Relação entre concordância entre avaliadores e ganho do empilhamento	60
Figura 12	Relação entre falha dupla e o ganho do empilhamento . . . . .	60

## LISTA DE TABELAS

Tabela 1	Informações de um solicitante de empréstimo junto com o rótulo de classe indicando se o empréstimo deve ser concedido ou não. . . . .	16
Tabela 2	Conjunto de regras para o problema de classificação da Tabela 1. . . . .	16
Tabela 3	Predições dos algoritmos <i>C4.5</i> e <i>RIPPER</i> para a classe empréstimo. . . . .	18
Tabela 4	Instâncias do conjunto de treinamento do meta-classificador usando 2 classificadores de nível básico. . . . .	19
Tabela 5	Predições do meta-classificador <i>RF</i> . . . . .	19
Tabela 6	Matriz de relacionamento <i>R</i> entre um par de classificadores $C_a$ e $C_b$ . . . . .	20
Tabela 7	Resumo das sete medidas de diversidade. . . . .	22
Tabela 8	Exemplo de classificadores com baixa diversidade. . . . .	22
Tabela 9	Exemplo de classificadores com alta diversidade. . . . .	22
Tabela 10	Medidas de diversidade calculadas para os exemplos apresentados. . . . .	22
Tabela 11	Principais características dos trabalhos analisados. . . . .	31
Tabela 12	Análise do uso de medidas de diversidade de acordo com os trabalhos estudados. . . . .	40
Tabela 13	Conjunto de dados do <i>UCI</i> usados na avaliação experimental. . . . .	46
Tabela 14	Conjunto de dados sintéticos usados na avaliação experimental. . . . .	47
Tabela 15	Configuração de parâmetros para cada algoritmo. . . . .	49
Tabela 16	Medidas de diversidade e os melhores e piores resultados do empilhamento para as bases do <i>UCI</i> . . . . .	51
Tabela 17	Medidas de diversidade e o resultado do empilhamento variando entre -1 e 1 para as bases do <i>UCI</i> . . . . .	54
Tabela 18	Medidas de diversidade e o resultado do empilhamento para as bases de dados sintéticas. . . . .	55
Tabela 19	Teste estatístico aplicado ao ganho do empilhamento para as bases do <i>UCI</i> . . . . .	56
Tabela 20	Resultado do teste-t para as bases sintéticas . . . . .	57
Tabela 21	Avaliação dos modelos de regressão. . . . .	62

## LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizagem de Máquina
BN	Bayes Net
DL	Digital Library
DT	Decision Trees
GP	Programação Genética
KDE	Kernel Density Estimation
K-NN	k-Nearest Neighbors
MLP	Multi Layer Perceptron
MLR	Multi-response Linear Regression
NB	Naive Bayes
PSO	Particle Swarm Optimization
RF	Random Forest
RI	Recuperação de Informações
ROC	Receiver Operator Characteristic Curve
RRSE	Raiz do Erro Quadrático Relativo
SIG	Sistema de Informação Geográfica
SMO	Sequential Minimal Optimization
SVM	Support Vector Machine
STAV	Vetor Empilhado de Marcador de Afinidade
STBV	Vetor Empilhado Binário de Marcador de Afinidade
TAV	Vetor Marcador de Afinidade
TBV	Vetor Marcador Binário
UCI	Machine Learning Repository
VP	Voted Perceptron

# SUMÁRIO

<b>1</b>	<b>INTRODUÇÃO</b>	11
1.1	Objetivo Geral e Específicos	12
1.2	Organização do Texto	12
<b>2</b>	<b>FUNDAMENTAÇÃO TEÓRICA</b>	14
2.1	Classificação de Dados Supervisionados	14
2.2	Combinando Classificadores com o Empilhamento ( <i>Stacking</i> )	17
2.3	Medidas de Diversidade	19
2.3.1	Falha dupla $df$	20
2.3.2	Medida de discordância $Dis$	20
2.3.3	Estatística $Q$	20
2.3.4	Coefficiente de correlação $\rho$	21
2.3.5	Variância Kohavi-Wolpert $KW$	21
2.3.6	Concordância entre avaliadores $k$	21
2.3.7	Entropia $E$	21
2.3.8	Comparação e classificação das medidas	21
2.3.9	Exemplo	22
<b>3</b>	<b>TRABALHOS RELACIONADOS</b>	23
3.1	Técnicas de Empilhamento	23
3.1.1	(WOLPERT, 1992)	23
3.1.2	(BREIMAN, 1996)	24
3.1.3	(TING; WITTEN, 1999)	24
3.1.4	(DZEROSKI; ZENKO, 2004)	25
3.1.5	(NESS et al., 2009)	25
3.1.6	(EBRAHIMPOUR et al., 2010)	26
3.1.7	(LARIOS et al., 2011)	26
3.1.8	(GARCÍA-GUTIÉRREZ; MATEOS-GARCÍA; RIQUELME-SANTOS, 2012)	27
3.1.9	(BORGES, 2013)	27
3.1.10	(PEPPOLONI et al., 2014)	28
3.1.11	(ALI; MAJID, 2015)	29
3.1.12	Análise Comparativa das Aplicações de Empilhamento	29
3.2	Aplicações das Medidas de Diversidade	32
3.2.1	(SHIPP; KUNCHEVA, 2002)	32
3.2.2	(KUNCHEVA; WHITAKER, 2003)	32
3.2.3	(DYMITR; BOGDAN, 2005)	33

3.2.4	(OLIVEIRA, 2008) . . . . .	34
3.2.5	(MUHAMMAD; JIM, 2010) . . . . .	35
3.2.6	(MAKHTAR et al., 2012) . . . . .	35
3.2.7	(WHALEN; PANDEY, 2013) . . . . .	36
3.2.8	(FARIA et al., 2014) . . . . .	36
3.2.9	(SLUBAN; LAVRAC, 2015) . . . . .	37
3.2.10	Análise Comparativa do uso de Medidas de Diversidade . . . . .	38
<b>4</b>	<b>ABORDAGEM PROPOSTA</b> . . . . .	<b>41</b>
<b>4.1</b>	<b>Objetivo</b> . . . . .	<b>41</b>
<b>4.2</b>	<b>Método Proposto</b> . . . . .	<b>41</b>
4.2.1	Classificadores . . . . .	42
4.2.2	Diversidade . . . . .	43
4.2.3	Analisando o impacto da diversidade . . . . .	43
<b>5</b>	<b>AVALIAÇÃO EXPERIMENTAL</b> . . . . .	<b>45</b>
<b>5.1</b>	<b>Bases de Dados</b> . . . . .	<b>45</b>
<b>5.2</b>	<b>Configuração dos Experimentos</b> . . . . .	<b>48</b>
<b>5.3</b>	<b>Diversidade e Resultados do Empilhamento</b> . . . . .	<b>49</b>
5.3.1	Teste Estatístico . . . . .	56
5.3.2	Interpretação gráfica da relação entre diversidade e qualidade do empilhamento . . . . .	57
5.3.3	Modelos de Regressão . . . . .	61
<b>6</b>	<b>CONCLUSÃO</b> . . . . .	<b>63</b>
	<b>REFERÊNCIAS</b> . . . . .	<b>65</b>

# 1 INTRODUÇÃO

Com a disponibilidade atual de grandes quantidades de dados e de recursos computacionais, técnicas de reconhecimento de padrões têm recebido esforços significativos da comunidade científica. Estas técnicas tornam-se cada vez mais precisas para a descoberta de conhecimento a partir da análise de dados.

Neste contexto, a Aprendizagem de Máquina (AM) tem se destacado, principalmente porque suas técnicas podem realizar o reconhecimento de padrões através do aprendizado supervisionado. Esses métodos de aprendizagem podem construir, a partir de padrões disponíveis em modelos de conjuntos de dados de treinamento, funções capazes de classificar novos padrões, isto é, predizer a que classe de dados uma nova instância pertence (BERNARDINI, 2002).

No entanto, a qualidade dos resultados da classificação dependerá substancialmente da qualidade e do volume das amostras de dados utilizadas na fase de treinamento. Também são fatores determinantes a seleção das características e a configuração dos parâmetros (KUNCHEVA; WHITAKER, 2003).

Embora alguns classificadores forneçam individualmente soluções consideradas eficazes, a avaliação experimental realizada por DIETTERICH (2000) mostra uma queda na qualidade quando existem grandes conjuntos de padrões e/ou um número significativo de amostras incompletas de dados ou características irrelevantes. Ou seja, tais classificadores não conseguem reconhecer padrões de uma forma eficaz e/ou eficiente em problemas complexos.

Com o intuito de melhorar os resultados da classificação, técnicas de combinação de classificadores têm sido utilizadas, visando aproveitar diversos esquemas de classificação, onde as saídas de cada classificador podem ser combinadas em uma decisão final que melhora a capacidade de generalização. Logo, essas técnicas combinam vários modelos em um geralmente mais preciso do que o melhor de seus componentes (SENI; ELDER, 2010).

Dentre essas técnicas, destaca-se o empilhamento como uma forma de combinar classificadores que consiste em utilizar um algoritmo de aprendizado de segundo nível para combinar, de forma ótima, uma coleção de previsões feitas por diferentes mode-

los. Usando múltiplos níveis de representações de dados, o algoritmo de aprendizado torna-se capaz de aprender conceitos abstratos e possivelmente relacionamentos de interdependência entre as amostras (WOLPERT, 1992).

No método de empilhamento a escolha dos algoritmos base é muito importante. É desejável que existam diferentes soluções para o problema a ser resolvido, isto é, é importante que seja obtida uma diversidade considerável entre os resultados encontrados por esses classificadores. De acordo com OPITZ; MACLIN (1999), o desempenho do método de empilhamento depende fortemente da precisão e diversidade dos resultados dos classificadores utilizados em sua composição. Para verificar essa diversidade, existem várias medidas baseadas na concordância e/ou discordância dos classificadores (KUNCHEVA; WHITAKER, 2003).

Portanto, o uso de algoritmos base com diferentes detalhes é ideal, uma vez que os padrões aprendidos tendem a não serem iguais. Desse modo, mesmo classificadores de baixa acurácia, quando combinados, podem gerar um classificador forte, proporcionando ganho para o empilhamento. Caso contrário, quando vários classificadores concordam na grande maioria das respostas (sem diversidade), a combinação possivelmente terá o mesmo resultado, sem melhora na qualidade do empilhamento.

## **1.1 Objetivo Geral e Específicos**

O objetivo desta dissertação é avaliar o impacto da diversidade dos classificadores na qualidade do empilhamento, tendo como objetivos específicos estudar o método de empilhamento, e a diversidade dos classificadores supervisionados.

A questão de pesquisa é baseada na ideia de que quanto maior a diversidade dos padrões aprendidos pelos classificadores base, maior a qualidade do empilhamento.

Como contribuição científica demonstra-se a relação entre múltiplas medidas de diversidade e o ganho do empilhamento, através de experimentos realizados com bases de dados reais e sintéticas de diferentes áreas do conhecimento.

## **1.2 Organização do Texto**

O restante deste trabalho está organizado conforme mencionado a seguir.

O capítulo 2 apresenta a fundamentação teórica onde são mostrados os conceitos fundamentais sobre empilhamento de classificadores e medidas de diversidade. A Seção 2.1 define a tarefa de classificação de dados e apresenta os classificadores supervisionados usados nos experimentos. Nas Seções 2.2 e 2.3 são apresentados, respectivamente, o método de empilhamento como estratégia de combinação de classificadores e as medidas de diversidade usadas para verificar a concordância e/ou discordância entre os classificadores.

Em seguida, o capítulo 3 apresenta um estudo sobre trabalhos que abordam a técnica de empilhamento, e uso de medidas de diversidade. Primeiramente, são analisadas diversas técnicas de empilhamento, através de trabalhos que mostram diferentes abordagens de aplicação (Seção 3.1). Na sequência é apresentada uma comparação das principais características desses trabalhos. Já a Seção 3.2 mostra aplicações de medidas de diversidade no contexto de construção de diferentes esquemas de combinação (*ensemble*). Ao final também é apresentada uma análise comparativa entre essas aplicações.

No capítulo 4 é apresentada a abordagem proposta nesta dissertação. A Seção 4.1 retoma os objetivos do trabalho. O método proposto é descrito em todas as suas etapas na Seção 4.2.

Na sequência, o capítulo 5 apresenta a avaliação experimental da pesquisa, onde são descritas as bases de dados utilizadas (Seção 5.1), a configuração dos experimentos (Seção 5.2) e a discussão dos resultados sobre a relação entre diversidade e ganho do empilhamento (Seção 5.3). Sobre os resultados é aplicado um teste para verificar a significância estatística do ganho do empilhamento e são apresentadas graficamente as relações entre cada medida de diversidade e o incremento qualidade.

Por fim, o capítulo 6 sintetiza as considerações finais e os trabalhos futuros.

## 2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta conceitos importantes sobre a tarefa de classificação de dados, classificadores supervisionados, empilhamento de classificadores e medidas de diversidade. A Seção 2.1 apresenta os principais conceitos relacionados a tarefa de classificação de dados e os classificadores supervisionados usados para a realização dos experimentos. Na Seção 2.2 é apresentado o método de empilhamento como estratégia de combinação de classificadores. Já a Seção 2.3 apresenta as medidas de diversidade usadas no decorrer do trabalho.

### 2.1 Classificação de Dados Supervisionados

A classificação é a tarefa mais comum entre as tarefas de mineração de dados. Baseia-se na descoberta de regras de previsão que auxiliam na tomada de decisões. Geralmente, essa tarefa é utilizada quando há grandes quantidades de registros em uma base de dados, que possuem diversos atributos, e é necessário extrair desta base algum conhecimento relevante com capacidade de previsão.

De acordo com TAN; STEINBACH; KUMAR (2005), a classificação pode ser definida como o processo de encontrar, através da aprendizagem supervisionada, um modelo ou função que descreve diferentes classes de dados. A finalidade da classificação é rotular, automaticamente, novas instâncias da base de dados com uma determinada classe aplicando o modelo ou a função previamente aprendida. Este modelo é baseado no valor dos atributos das instâncias de treinamento.

Os algoritmos de classificação podem ser organizados em diferentes tipos de acordo com as características técnicas que usam na aprendizagem. Cada tipo é mais adequado para um determinado conjunto de dados.

Alguns algoritmos de classificação, como o *C4.5* (QUINLAN, 1993), e o *Random Forest (RF)* (BREIMAN, 2001) utilizam árvores de decisão para classificar registros. Uma árvore de decisão, ou *Decision Tree (DT)* é composta por nós intermediários que representam atributos. As arestas definem um conjunto de valores que cada atributo pode assumir, e os nós folha indicam a classe de dados utilizada para rotular uma instância. As regras

de classificação são extraídas a partir de todos os possíveis caminhos entre o nó raiz e as folhas (BORGES, 2013).

Já o algoritmo de classificação *RIPPER* é baseado em regras, onde os atributos de entrada  $A_j$  e seus valores  $v_j$  são combinados por operadores relacionais *op* ( $=, \neq, <, >, \leq, \geq$ ) e usados em expressões condicionais para formar um conjunto de regras  $r_i: (A_1 \text{ op } v_1) \wedge (A_2 \text{ op } v_2) \wedge \dots \wedge (A_k \text{ op } v_k) \rightarrow y_j$ , em que  $y_j$  é a classe predita pela regra  $r_i$ . As regras são induzidas sequencialmente e para uma classe de cada vez. A última regra classifica todas as instâncias remanescentes (COHEN, 1995).

Um terceiro tipo de classificador baseia-se em redes neurais artificiais. *Multilayer Perceptron (MLP)* (HAYKIN, 2007) é uma rede que pode conter, além das camadas de entrada e saída do *perceptron*, camadas de nós intermediárias denominadas camadas ocultas. Além disso, implementa o algoritmo *back propagation* (HECHT-NIELSEN, 1989) para atualizar os pesos da rede.

Entre outros classificadores baseados em função, destaca-se o *Support Vector Machine (SVM)* (BOSER; GUYON; VAPNIK, 1992), como sendo um algoritmo de classificação binária que traça um hiperplano ótimo que maximiza a margem de separação entre duas classes de dados. A etapa principal do algoritmo é descobrir os vetores de suporte que são as instâncias equidistantes do hiperplano. Quando os dados não são linearmente separáveis, o espaço de entrada é transformado aplicando uma função de núcleo que eleva o número de dimensões até que seja encontrado um espaço passível de separação linear. Já o *Sequential Minimal Optimization (SMO)* (PLATT, 1999a) é uma variação do *SVM* otimizada que utiliza uma quantidade de memória linear em relação ao tamanho do conjunto de treinamento.

O *Naive Bayes (NB)* é considerado um modelo probabilístico que computa a probabilidade  $P(c | r)$  de um registro  $r$  pertencer a uma determinada classe  $c$  a partir da probabilidade a priori  $P(c)$  de um registro ser desta classe e das probabilidades condicionais  $P(v_k | c)$  de cada valor  $v_k$  de atributo ocorrer em um registro da mesma classe. O objetivo do algoritmo é encontrar a melhor classe para um registro maximizando a probabilidade a posteriori (JOHN; LANGLEY, 1995).

A Tabela 1 apresenta um exemplo de conjunto de dados contendo informações de um solicitante de empréstimo junto com o rótulo de classe indicando se o empréstimo deve ser concedido ou não. A seguir é mostrado um exemplo de aplicação, para os algoritmos *C4.5* e *RIPPER*, baseado nesses dados. Na sequência, a Figura 1 mostra um exemplo de árvore de decisão aprendida a partir dessas informações, e a Tabela 2 mostra o exemplo de um conjunto de regras para o referido problema de classificação.

Tabela 1: Informações de um solicitante de empréstimo junto com o rótulo de classe indicando se o empréstimo deve ser concedido ou não.

ID	Montante	Idade	Salário	Conta	Empréstimo
1	médio	sênior	baixo	sim	Não
2	médio	sênior	baixo	não	Não
3	baixo	sênior	baixo	sim	Sim
4	alto	média	baixo	sim	Sim
5	alto	jovem	alto	sim	Sim
6	alto	jovem	alto	não	Não
7	baixo	jovem	alto	não	Sim
8	médio	média	baixo	sim	Não
9	médio	jovem	alto	sim	Sim
10	alto	média	alto	sim	Sim
11	médio	média	alto	não	Sim
12	baixo	jovem	baixo	não	Sim
13	baixo	sênior	alto	sim	Sim
14	alto	média	baixo	não	Não

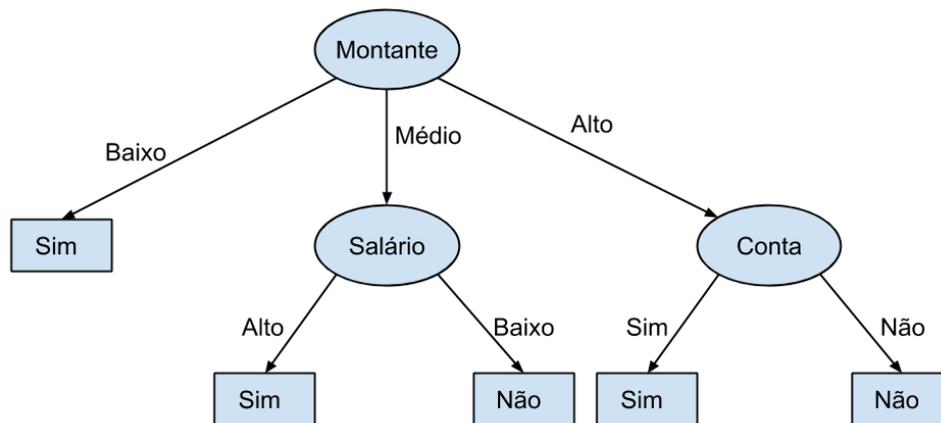


Figura 1: Exemplo de árvore de decisão aprendida a partir das informações da Tabela 1.

Tabela 2: Conjunto de regras para o problema de classificação da Tabela 1.

ID	Regra
$r_1$	$(\text{salário} = \text{baixo}) \wedge (\text{montante} = \text{médio}) \rightarrow \text{empréstimo} = \text{Não}$
$r_2$	$(\text{conta} = \text{não}) \wedge (\text{montante} = \text{alto}) \rightarrow \text{empréstimo} = \text{Não}$
$r_3$	$(\text{montante} = \text{médio}) \wedge (\text{salário} = \text{alto}) \rightarrow \text{empréstimo} = \text{Sim}$
$r_4$	$(\text{montante} = \text{baixo}) \rightarrow \text{empréstimo} = \text{Sim}$

## 2.2 Combinando Classificadores com o Empilhamento (*Stacking*)

Classificadores que implementam algoritmos diferentes potencialmente fornecem características diferentes sobre os padrões a serem classificados. Por esse motivo, não é necessário a manipulação do conjunto de treinamento, podendo assim ser utilizado os mesmos dados para todos os algoritmos, pois a combinação das predições realizadas por estes pode resultar em uma classificação mais exata, ou seja, com maior precisão.

Nesse contexto, o *stacking* (TING; WITTEN, 1999) é um método amplamente utilizado para combinar vários classificadores gerados a partir de diferentes algoritmos de aprendizagem aplicados no mesmo conjunto de dados. É também conhecido na literatura como *stacked generalization* (DZEROSKI; ZENKO, 2004) and (WOLPERT, 1992).

O método de empilhamento combina vários classificadores base, também chamados de classificadores de nível 0, treinados usando diferentes algoritmos de aprendizagem  $L$  em um único conjunto de dados  $S$ , por meio de um meta-classificador, também conhecido como classificador de nível 1 (MERZ, 1999) and (KOTSIANTIS; PINTELAS, 2004). Cada amostra de treinamento  $s_j = (X_j, y_j)$  é um par composto por vetores de características  $X_j$  e o rótulo da classe  $y_j$ .

Uma característica importante do empilhamento é a livre escolha, por parte do usuário, dos modelos para compor o nível 0. Eles podem ser derivados de um único algoritmo de aprendizagem, usando diferentes configurações, ou de uma variedade de diferentes algoritmos. Além disso, o empilhamento é ideal para computação paralela, pois a construção de cada modelo de nível 0 procede de forma independente (TING; WITTEN, 1999).

O empilhamento pode ser descrito em dois níveis distintos, como mostrado na Figura 2. O primeiro nível, ou nível 0, é composto por um conjunto de  $N$  classificadores base, onde  $C_i = L_i(S) | 1 \leq i \leq N$ . Os classificadores do nível 0 são treinados e testados usando o procedimento de validação cruzada ou de exclusão. O conjunto de dados de saída  $D$  usado para treinar o meta-classificador é composto por exemplos  $((y_j^1, \dots, y_j^i), y_j)$ , ou seja, um vetor de predições para cada classificador base  $y_j^i = C_i(X_j)$  e o mesmo rótulo da classe original  $y_j$  (DZEROSKI; ZENKO, 2004). No segundo nível, ou nível 1, o meta-classificador combina as saídas  $D$  dos classificadores base em uma predição final  $y_j^f$ . O pseudocódigo do empilhamento pode ser visto no Algoritmo 1.

A Tabela 3 representa a saída gerada na linha 5 do Algoritmo 1, e mostra as predições obtidas pelos algoritmos *C4.5* e *RIPPER*, para os dados de ID 1 e 4 da Tabela 1, referente a classe empréstimo. Esses classificadores foram treinados conforme a linha 4 do Algoritmo 1.

Já a Tabela 4 representa os dados gerados na linha 7 do Algoritmo 1, onde são mostradas as instâncias do conjunto de treinamento do meta-classificador, usando dois classificadores de nível básico (*C4.5*; *RIPPER*), construídos a partir dos dados originais, mostrados na Tabela 1. O vetor de características  $X = (\text{Sim/Não})$ , mostrados na 2ª e 3ª coluna da Ta-

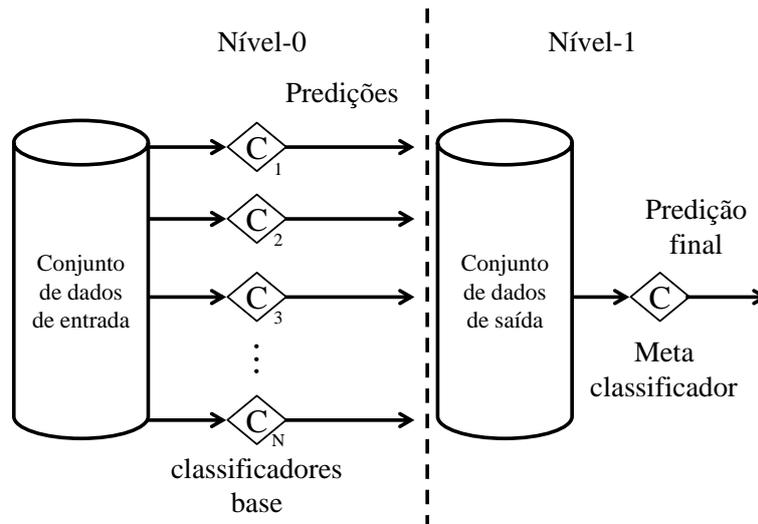


Figura 2: Representação do algoritmo de *Stacking*.

---

**Algoritmo 1:** Combinando classificadores com o *Stacking*

---

**Entrada:** Amostras de treinamento  $s_j \in S$

**Saída:** Predições finais  $y_j^f$

```

1 início
2   Selecionar  $N$  algoritmos de aprendizagem  $(L_1, L_2, \dots, L_N)$ ;
3   para  $i = 1, 2, \dots, N$  faça
4     Treinar  $C_i = L_i(S)$  usando validação cruzada;
5      $y_j^i = C_i(X_j)$ ;
6   fim
7   Criar um novo conjunto de dados  $D$  combinando todas as predições  $y_j^i$ ;
8   Treinar  $M = L(D)$  usando validação cruzada;
9    $y_j^f = M(D)$ ;
10 fim
11 retorna  $y_j^f$ 

```

---

Tabela 3: Predições dos algoritmos *C4.5* e *RIPPER* para a classe empréstimo.

ID	<i>C4.5</i>	<i>RIPPER</i>
1	Não	Não
4	Não	Sim

bela 4, correspondem às predições feitas pelos classificadores para as instâncias originais.

Em seguida, o classificador *RF* é selecionado como meta-classificador, sendo treinado a partir do conhecimento adquirido pelos classificadores de nível básico, e finalmente é usado para inferir a classe das instâncias do conjunto de teste.

A Tabela 5 mostra a saída do meta-classificador com as predições para a classe empréstimo.

Tabela 4: Instâncias do conjunto de treinamento do meta-classificador usando 2 classificadores de nível básico.

ID	<i>C4.5</i>	<i>RIPPER</i>	Empréstimo
1	não	não	Não
2	não	não	Não
3	não	sim	Sim
4	não	sim	Sim
5	sim	sim	Sim
6	sim	sim	Não
7	sim	sim	Sim
8	sim	não	Não
9	sim	sim	Sim
10	sim	sim	Sim
11	não	sim	Sim
12	não	sim	Sim
13	sim	sim	Sim
14	sim	sim	Não

Tabela 5: Predições do meta-classificador *RF*.

ID	Empréstimo
1	Sim
2	Sim
3	Sim
4	Sim
5	Sim
6	Sim
7	Sim
8	Sim
9	Sim
10	Sim
11	Sim
12	Sim
13	Sim
14	Não

### 2.3 Medidas de Diversidade

A diversidade das predições é uma questão-chave na combinação de classificadores. (KUNCHEVA; WHITAKER, 2003) define várias medidas de diversidade e as relaciona com a qualidade do sistema de classificação. Estas medidas baseiam-se no acordo ou desacordo dos classificadores utilizados no conjunto.

Seja  $n$  o número de instâncias avaliadas por um par de classificadores  $C_a$  e  $C_b$  e  $R$  uma matriz de relacionamento entre eles, contendo o número de instâncias em que cada classificador acerta (1) e/ou erra (0) a predição da classe de dados (Tabela 6). Por exemplo,  $n^{01}$  representa o número de instâncias identificadas erroneamente por  $C_a$  e corretamente

por  $C_b$ . A diagonal principal mostra o número de instâncias igualmente rotuladas por ambos os classificadores. Já a diagonal secundária mostra o número de registros em que os classificadores discordam na classe predita. A soma de todas as células é o número total de instâncias avaliadas pelos classificadores analisados.

Tabela 6: Matriz de relacionamento  $R$  entre um par de classificadores  $C_a$  e  $C_b$ .

	Acertos $C_b$	Erros $C_b$
Acertos $C_a$	$n^{11}$	$n^{10}$
Erros $C_a$	$n^{01}$	$n^{00}$
$n = n^{11} + n^{10} + n^{01} + n^{00}$		

As subseções seguintes apresentam várias medidas de diversidade aplicadas em *ensembles* de classificadores (KUNCHEVA; WHITAKER, 2003) utilizadas na avaliação experimental desta dissertação.

### 2.3.1 Falha dupla $df$

A medida falha dupla  $df$  é definida pela Equação 1 e representa a proporção de instâncias simultaneamente classificadas erroneamente por um par de classificadores (GIACINTO; ROLI, 2001).  $df$  retorna valores no intervalo fechado  $[0, 1]$  e é inversamente proporcional à diversidade entre os classificadores.

$$df = \frac{n^{00}}{n^{00} + n^{01} + n^{10} + n^{11}} \quad (1)$$

### 2.3.2 Medida de discordância $Dis$

A medida de discordância  $Dis$  é definida pela Equação 2 e representa a razão entre o número de instâncias em que os classificadores discordam e o total de instâncias (HO, 1998).  $Dis$  varia no intervalo fechado  $[0, 1]$  e é diretamente proporcional à diversidade entre os classificadores.

$$Dis = \frac{n^{01} + n^{10}}{n^{00} + n^{01} + n^{10} + n^{11}} \quad (2)$$

### 2.3.3 Estatística $Q$

A estatística  $Q$  é uma medida de diversidade que opera sobre a saída de um par de classificadores, definida pela Equação 3 (AFIFI; AZEN, 2014). Esta medida retorna valores no intervalo fechado  $[-1, 1]$ , sendo inversamente proporcional à diversidade entre os classificadores.

$$Q = \frac{n^{11}n^{00} - n^{01}n^{10}}{n^{11}n^{00} + n^{01}n^{10}} \quad (3)$$

### 2.3.4 Coeficiente de correlação $\rho$

O coeficiente de correlação entre dois classificadores é definido pela Equação 4 (SNEATH; SOKAL, 1973). Como a estatística  $Q$ , ela retorna valores no intervalo  $[-1, 1]$ .  $\rho$  também é inversamente proporcional à diversidade entre os classificadores.

$$\rho = \frac{n^{11}n^{00} - n^{01}n^{10}}{\sqrt{(n^{11} + n^{10})(n^{01} + n^{00})(n^{11} + n^{01})(n^{10} + n^{00})}} \quad (4)$$

### 2.3.5 Variância Kohavi-Wolpert $KW$

A variância de Kohavi-Wolpert mede a diversidade entre um conjunto de  $N$  classificadores (KOHAVI; WOLPERT et al., 1996). Retorna valores no intervalo  $[0, 1/2]$  e é diretamente proporcional à diversidade.  $KW$  diverge da média da medida de discordância  $Dis_{avg}$  por um coeficiente, de acordo com a Equação 5.

$$KW = \frac{N - 1}{2N} Dis_{avg} \quad (5)$$

### 2.3.6 Concordância entre avaliadores $k$

A concordância entre avaliadores  $k$  é definida pela Equação 6, onde  $\bar{p}$  denota a precisão média de classificação individual (DIETTERICH, 2000). Esta medida opera sobre as predições de um conjunto de  $N$  classificadores e retorna um valor no intervalo fechado  $[-1, 1]$ . É inversamente proporcional à diversidade entre os classificadores.

$$k = 1 - \frac{N}{(N - 1)\bar{p}(1 - \bar{p})} KW \quad (6)$$

### 2.3.7 Entropia $E$

A entropia opera sobre a saída de um conjunto de  $N$  classificadores e é definida pela Equação 7, onde  $n$  é o número de instâncias e  $l(s_j)$  é o número de classificadores que rotularam corretamente a instância  $s_j$  (CUNNINGHAM; CARNEY, 2000).  $E$  varia no intervalo  $[0, 1]$  e é diretamente proporcional à diversidade entre os classificadores.

$$E = \frac{1}{n} \sum_{j=1}^n \frac{\min[l(s_j), N - l(s_j)]}{N - [N/2]} \quad (7)$$

### 2.3.8 Comparação e classificação das medidas

A Tabela 7 mostra um resumo das sete medidas de diversidade apresentadas. A seta indica se a medida é diretamente ( $\uparrow$ ) ou inversamente ( $\downarrow$ ) proporcional à diversidade entre os classificadores. Além disso, existem medidas que operam sobre um par de classificadores e outras que operam sobre um conjunto de  $N$  classificadores.

Tabela 7: Resumo das sete medidas de diversidade.

Nome		direta/inversa	Par	Conjunto
Falha dupla	$df$	↓	•	
Medida de discordância	$Dis$	↑	•	
Estatística	$Q$	↓	•	
Coefficiente de correlação	$\rho$	↓	•	
Variância Kohavi-Wolpert	$KW$	↑		•
Concordância entre avaliadores	$k$	↓		•
Entropia	$E$	↑		•

### 2.3.9 Exemplo

As Tabelas 8 e 9 exemplificam o cálculo das medidas de diversidade apresentadas. Essas tabelas apresentam matrizes de relacionamento entre classificadores com baixa e alta diversidade (BORGES, 2013). Os valores resultantes das medidas que operam sobre pares de classificadores estão apresentados na Tabela 10. Essas medidas poderiam ser utilizadas para um conjunto de classificadores considerando a média dos valores entre todas as combinações de pares.

Tabela 8: Exemplo de classificadores com baixa diversidade.

	Acertos $C4.5$	Erros $C4.5$
Acertos $SMO$	70	5
Erros $SMO$	15	10
$n = 100$		

Tabela 9: Exemplo de classificadores com alta diversidade.

	Acertos $RIPPER$	Erros $RIPPER$
Acertos $MLP$	25	50
Erros $MLP$	20	5
$n = 100$		

Tabela 10: Medidas de diversidade calculadas para os exemplos apresentados.

Diversidade	$df$	$Q$	$\rho$	$Dis$
Alta	0,05	-0,77	-0,406	0,7
Baixa	0,1	0,806	0,404	0,2

## 3 TRABALHOS RELACIONADOS

Este capítulo apresenta uma série de trabalhos que exemplificam técnicas de empilhamento, e o uso de medidas de diversidade. A Seção 3.1 discute diferentes abordagens de aplicação do método de empilhamento e apresenta, ao final, uma análise comparativa das principais características desses trabalhos. Já a Seção 3.2, apresenta exemplos de aplicações de medidas de diversidade, mostrando ao final uma análise sobre a relevância do uso dessas, em cada trabalho.

### 3.1 Técnicas de Empilhamento

#### 3.1.1 (WOLPERT, 1992)

WOLPERT (1992) introduz uma nova abordagem para combinar múltiplos classificadores conhecida como *stacked generalization*. Essa abordagem visa minimizar a taxa de erro de generalização de um ou mais classificadores, tendo como ideia principal aprender um classificador de meta nível (ou nível 1) com base na saída de classificadores de nível básico (ou nível 0). Após a introdução da referida metodologia e da justificativa de sua utilização, são apresentados dois experimentos de uso da mesma.

O primeiro demonstra como *stacked generalization* melhora o desempenho de um algoritmo de ajuste de superfície. Para este experimento, o espaço de entrada do classificador de nível 0 foi unidimensional. O mesmo tem por função ligar linearmente os pontos do conjunto de aprendizagem para fazer uma superfície entrada-saída que serve como uma estimativa para a função principal. Já o espaço de entrada do classificador de nível 1 foi bidimensional, sendo o mesmo baseado na métrica *HERBIE* que funciona através do envio de uma soma ponderada normalizada das saídas dos  $p$  vizinhos mais próximos quando aplicada a casos com características de valor real e que usa uma métrica de *Hamming* para recursos simbólicos.

Já o segundo experimento foi baseado no problema *Nettalk*, onde a função principal tem como entrada sete letras (devidamente codificadas), e como saída a suposição de um vetor num espaço de 21 dimensões. Este vetor é então convertido numa suposição de fonema, que seria dublado por um alto-falante em Inglês ao se encontrar a letra do meio,

se todas as sete letras tivessem ocorrido no meio de um texto que o orador estava lendo em voz alta. O classificador de nível 1 foi baseado na métrica *HERBIE* usando uma métrica *Hamming* sobre o espaço de entrada tri dimensional para o nível 1.

### 3.1.2 (BREIMAN, 1996)

De acordo com BREIMAN (1996), o empilhamento pode ser usado com sucesso para formar combinações lineares de diferentes preditores para uma melhor precisão da predição. Neste caso, o modelo de nível zero são árvores de regressão de diferentes tamanhos, usados em uma simulação de empilhamento de subconjuntos de regressões lineares com diferentes números de variáveis. Para a formação do classificador de nível um, seguindo a restrição de que todos os coeficientes na combinação sejam não negativos, é utilizado sob os subconjuntos de regressões o método de validação cruzada e dos mínimos quadrados.

Segundo QUININO et al. (2013), o método dos mínimos quadrados é o procedimento de estimação dos parâmetros de um modelo de regressão por meio da minimização da soma dos quadrados das diferenças entre os valores observados da variável resposta em uma amostra e seus valores preditos pelo modelo. Para BREIMAN (1996), a restrição da não negatividade é considerada crucial para garantir que a precisão das predições seja melhor do que o alcançado selecionando um único melhor preditor.

### 3.1.3 (TING; WITTEN, 1999)

Para TING; WITTEN (1999) o empilhamento pode ser usado de forma confiável em tarefas de classificação, utilizando como meta-classificador uma adaptação de regressão linear denominada *Multi-response Linear Regression (MLR)*, como sendo uma técnica estatística para construção de modelos lineares. Esses modelos associam um conjunto de uma ou mais variáveis respostas à um conjunto de preditores (JOHNSON; WICHERN, 2002).

Durante a predição, os atributos usados como entrada para as funções de regressão do *MLR* são as probabilidades de cada um dos valores de classe retornados por cada um dos algoritmos de nível básico, sendo eles *C4.5*, *IB1*, e *NB*. O algoritmo *IB1* é uma variante de um algoritmo de aprendizagem preguiçosa (*AHA*; KIBLER; ALBERT, 1991), que emprega o método *p-Nearest Neighbors*, usando uma métrica de diferença de valor modificada para atributos nominais e binários (COST; SALZBERG, 1993).

Assim, para um número de classes de dados originais distintas  $K$ , o conjunto de dados do meta-classificador consiste de  $n$  instâncias  $(p^{C_i}(k | x_j), y_j)$ , em que o vetor de características  $p^{C_i}(k | x_j)$  é formado pelas probabilidades preditas para cada classe de dados  $K$   $| 1 \leq K \leq K$ , e  $y_j$  é a classe original de dados.

Os autores comparam a utilização de quatro diferentes algoritmos como meta-classificador: os de nível básico, mencionados anteriormente, e o *MLR*, sendo esse consi-

derado o mais adequado, e que apresentou melhor desempenho. Além disso, argumentam que essa técnica permite utilizar não só as predições, mas também a confiança dos classificadores de nível básico.

### 3.1.4 (DZEROSKI; ZENKO, 2004)

Como extensão do método de empilhamento *MLR*, apresentado anteriormente, é proposto por DZEROSKI; ZENKO (2004) um modelo de árvores multi resposta no meta-classificador. O conjunto de treinamento utilizado no aprendizado do meta-classificador é composto pelos seguintes atributos: a distribuição de probabilidades de cada classe, a distribuição de probabilidade da classe multiplicada pela probabilidade máxima considerando todas as classes, e a entropia da distribuição de probabilidade para cada classificador, sendo esses dois últimos atributos aplicados a todas as instâncias  $j$  do vetor de características  $X_j$ .

Um experimento foi realizado com dois conjuntos de classificadores de nível básico, sendo o primeiro conjunto composto pelos algoritmos *C4.5*, *IBk* e *NB*. O algoritmo de aprendizagem *IBk* é baseado em instâncias, tendo como objetivo maximizar a acurácia sobre novas instâncias do problema de classificação (AHA; KIBLER; ALBERT, 1991).

O segundo conjunto foi composto por, além dos três já mencionados,  $K^*$ , *KDE*, *DT*, e *MLR*. O *Kernel Density Estimation (KDE)* é um método não-paramétrico para se estimar a função de densidade de probabilidade de uma variável aleatória (PARZEN, 1962). Já o  $K^*$  é um algoritmo que utiliza medidas baseadas em instâncias (CLEARY; TRIGG, 1995).

No meta-classificador foram avaliados o desempenho de seis técnicas diferentes para combinar classificadores, cada uma aplicada com os dois conjuntos diferentes de algoritmos do nível básico. Nesse contexto, o meta-classificador baseado no modelo de árvores multi resposta apresentou melhor desempenho. De acordo com os autores, o referido modelo é uma boa escolha para a aprendizagem do meta-classificador, independentemente da escolha dos classificadores de nível básico.

### 3.1.5 (NESS et al., 2009)

NESS et al. (2009) descrevem como o empilhamento pode ser utilizado para melhorar o desempenho de um sistema automático de anotação de *tags*, que utiliza extração de características de áudio. Para cada faixa da coleção de áudio é calculado um vetor de características que é usado para treinar um classificador de distribuição, um *SVM* linear com saídas probabilísticas, que gera como saída um vetor marcador de afinidade (TAV) para cada *tag*. O TAV pode ser utilizado diretamente para a recuperação e armazenamento ou convertido num vetor marcador binário (TBV) por algum método de limiar. Para transformar em probabilidades as saídas do referido classificador, foi utilizada a técnica proposta por PLATT (1999b), onde os parâmetros de uma função sigmóide adicional são treinados para mapear as saídas *SVM*.

Quando o empilhamento é utilizado, visando explorar possíveis correlações entre as *tags*, o TAV é usado como um vetor de características semânticas, a fim de treinar um segundo nível do classificador *SVM*, que produz um vetor empilhado de marcador de afinidade (STAV) e um vetor empilhado binário de marcador de afinidade (STBV). A afinidade prevista resultante e o vetor binário podem ser utilizados para avaliar a eficácia do sistema de recuperação de informações (RI), usando métricas como a curva *Receiver Operator Characteristic Curve (ROC)* para o TAV, e medidas de recuperação de informação (MANNING et al., 2008) para o TBV.

Assim, o sistema concentra-se na anotação de *tag* automática de faixas de áudio em que o sistema de recuperação aprende uma relação entre as características acústicas e palavras a partir de um conjunto de dados de faixas de áudio anotadas. O modelo treinado resultante pode recuperar faixas de áudio com base em listas de *tags* e pode anotar faixas de áudio não marcados com *tags*.

### 3.1.6 (EBRAHIMPOUR et al., 2010)

Visando enfrentar o problema de reconhecimento de imagens de face com baixa resolução, EBRAHIMPOUR et al. (2010) apresenta uma solução adequada usando combinação de classificadores. O modelo proposto baseia-se na extração de vetores de características das imagens de face de baixa resolução usando três extratores de recursos, com o objetivo de reduzir a dimensão das imagens originais.

Para classificar as amostras das imagens de face, foi utilizada a técnica de empilhamento, onde primeiramente são usados três classificadores de nível básico *MLP*, um para cada extrator, e em seguida, um meta-classificador *MLP* que combina as saídas dos classificadores de nível básico. Segundo os autores, utilizando o empilhamento como técnica de combinação gera melhorias no reconhecimento de imagem de face com baixa resolução, tornando a taxa de desempenho mais elevada do que utilizando o melhor classificador.

### 3.1.7 (LARIOS et al., 2011)

LARIOS et al. (2011) propõem um método de reconhecimento de classe de objeto que combina características locais diferentes geradas por detectores e descritores discriminativos de estrutura. Nesse contexto, a abordagem de empilhamento quantifica de forma eficiente os dados de entrada e aumenta a precisão da classificação, enquanto permite a utilização espacial da informação. Este método de classificação é aplicado à tarefa de identificação automática de insetos em espécies para fins de biomonitoramento. A estrutura de classificação é composta de três etapas principais:

1. Classificação das características locais, usando o algoritmo *RF*;
2. Criação do conjunto de histogramas espaciais de escores por imagem, que constitui os atributos de treinamento do meta-classificador;

3. Identificação das espécies da amostra, utilizando como meta-classificador um *SVM* espacial de *kernel* piramidal.

Segundo LAZEBNIK; SCHMID; PONCE (2006), o classificador espacial *SVM* possibilita a representação de um local em vários níveis de resolução, por meio da combinação de *kernels* baseado em pirâmides, construindo assim pirâmides no espaço de características.

### 3.1.8 (GARCÍA-GUTIÉRREZ; MATEOS-GARCÍA; RIQUELME-SANTOS, 2012)

GARCÍA-GUTIÉRREZ; MATEOS-GARCÍA; RIQUELME-SANTOS (2012) propuseram um método chamado *EVOR-STACK* como sendo uma nova aplicação na área de inteligência artificial híbrida visando melhorar a precisão de mapas temáticos. Este método combina a aplicação de *ensembles* em sensoriamento remoto, que tira proveito de informações contextuais a partir de várias fontes de dados obtidas pelo sistema LIDAR <sup>1</sup> e imagens aéreas, e a utilização da computação evolutiva para melhorar a separabilidade de pixels para cada rótulo.

O método funciona em dois níveis de processamento:

1. São calculadas estatísticas a partir da fusão de dados LIDAR e imagens aéreas, seguindo uma estratégia baseada em características orientada a pixel.
2. Um algoritmo evolutivo é usado para obter uma matriz de ponderação multi-rótulo que transforma o espaço de características, atribuindo pesos diferentes a cada atributo, dependendo da classe.

Em seguida, essa matriz de pesos é usada para melhorar o método chamado *R-STACK* (GARCIA-GUTIERREZ; MATEOS-GARCIA; RIQUELME-SANTOS, 2010), baseado no empilhamento de um *SVM* e múltiplos classificadores *k-Nearest Neighbors (k-NN)* (COVER; HART, 1967), e dar origem ao método *EVOR-STACK*.

Além disso, os autores compararam a precisão do classificador *SVM*, do empilhamento original (*R-STACK*) e do seu método melhorado (*EVOR-STACK*), sendo esse o que apresentou melhor resultado.

### 3.1.9 (BORGES, 2013)

Uma nova abordagem de aplicação do empilhamento é apresentada por BORGES (2013) através do desenvolvimento de um método efetivo e automático para identificar metadados bibliográficos duplicados usando a combinação do aprendizado de múltiplos classificadores supervisionados. Para treinar os diferentes classificadores de nível básico

---

<sup>1</sup>Tecnologia ótica de sensoriamento remoto que mede propriedades da luz refletida de modo a obter pontos tridimensionais que contém diversas informações

(*MLP*, *Voted Perceptron (VP)*, *SMO*, *NB*, *Bayes Net (BN)*, *RIPPER*, *C4.5*) são utilizados como dados de entrada os escores produzidos por funções de similaridade aplicadas sobre os metadados bibliográficos do conjunto de treinamento. Os classificadores aprendidos são combinados através da estratégia de empilhamento para gerar predições binárias (réplica ou não réplica) para cada um dos pares de registros bibliográficos contidos no conjunto de teste.

Mantendo a mesma classe de dados, foram definidos três conjuntos de instâncias do meta nível, seguindo abordagens distintas:

1. O vetor de características na primeira é composto pelo rótulo das classes preditas pelos modelos treinados;
2. Na segunda é composto pelo valor das predições de cada um dos modelos treinados;
3. Na terceira é composto tanto pelo rótulo como pelo valor das predições de cada um dos modelos treinados.

Por fim, o classificador do meta nível é treinado utilizando qualquer algoritmo de classificação a partir do conhecimento adquirido pelos classificadores de nível básico e é finalmente usado para inferir a classe das instâncias do conjunto de teste.

Segundo o autor, a avaliação experimental mostra que o empilhamento de classificadores supervisionados pode aumentar a qualidade da identificação de registros duplicados quando comparado a outras estratégias de combinação de classificadores.

### **3.1.10 (PEPPOLONI et al., 2014)**

PEPPOLONI et al. (2014) apresentam o método do empilhamento como uma nova técnica para detecção e reconhecimento de objetos conhecidos em uma cena, adquiridos com um sensor RGB-D. Os dados do sensor são submetidos a três fases, que consistem em pré-processamento, detecção de objetos e classificação, tendo como saída a identificação de cada objeto detectado. No pré-processamento, a nuvem de pontos, detectada pelo sensor, é segmentada em *clusters*, que representam os objetos detectados na cena. Em seguida, é realizada a extração do vetor de características para cada *cluster*, e por fim, a classificação de cada *cluster*, sendo que cada classe corresponde a um objeto a ser reconhecido.

Na fase de classificação, as características são passadas como entrada para uma combinação de múltiplos níveis de classificadores *SVM*, onde um primeiro conjunto desses classificadores (nível-0) é treinado para classificar o objeto detectado usando apenas um subconjunto do vetor de características. As predições de saídas do nível-0 são utilizadas como entradas para um classificador *SVM* adicional, onde cada camada é ponderada seletivamente, a fim de adaptar o processo de classificação para as diferentes condições das cenas possíveis.

Esta abordagem também é chamada de esquema de fusão, onde todas as características extraídas são agregadas em um vetor global usado como entrada para um único classificador *SVM*.

### 3.1.11 (ALI; MAJID, 2015)

Com o objetivo de prever as sequências de aminoácidos associados com o câncer de mama, ALI; MAJID (2015) propuseram o desenvolvimento de um novo sistema de empilhamento baseado no *ensemble* evolutivo *Can-Evo-Ens*.

Primeiramente, foram selecionados quatro tipos diferentes de algoritmos de aprendizagem como classificadores base, sendo eles: *NB*, *K-NN*, *SVM*, e *RF*. Esses classificadores são treinados individualmente em diferentes espaços de características usando as propriedades físico-químicas dos aminoácidos, tendo como saída um conjunto de previsões em forma de valores numéricos.

Em seguida, a Programação Genética (GP) é utilizada para desenvolver um meta-classificador que combina as previsões dos classificadores base. O desempenho do melhor indivíduo evoluído no final das simulações de GP é calculado usando o limite ótimo obtido a partir do algoritmo *Particle Swarm Optimization (PSO)*. No final desse processo a melhor função *Evo-Ens* numérica é desenvolvida.

Segundo os autores, os experimentos demonstraram a robustez do sistema *Can-Evo-Ens* para um conjunto de dados de validação independente.

### 3.1.12 Análise Comparativa das Aplicações de Empilhamento

A Tabela 11 apresenta uma análise comparativa dos trabalhos estudados. Primeiramente, são apresentados os critérios utilizados para comparação. Em seguida, é feita a análise de cada critério.

Os seguintes critérios foram utilizados para realizar a comparação dos trabalhos relacionados:

- Objetivo do trabalho;
- Área de aplicação;
- Algoritmos utilizados como classificadores base - *nível 0*;
- Algoritmos utilizados como meta-classificadores - *nível 1*;
- Estratégia para composição do vetor de características do meta-nível - *atributos empilhados*.

Os primeiros quatro trabalhos são bem focados no problema do empilhamento. A referida técnica é introduzida por WOLPERT (1992) como sendo uma nova abordagem para combinar múltiplos classificadores. BREIMAN (1996) mostra que o empilhamento pode

ser usado com sucesso para formar combinações lineares de diferentes preditores, além de poder ser usado de forma confiável em tarefas de classificação, utilizando o *MLR* como meta-classificador (TING; WITTEN, 1999). Além disso, DZEROSKI; ZENKO (2004) propõem um modelo de árvores multi resposta no meta-classificador, como extensão do método de empilhamento *MLR*, visando melhorar a precisão da classificação.

Os demais trabalhos estudados são aplicações do empilhamento para resolver um problema de outras áreas: recuperação de informações (NESS et al., 2009), visão (EBRAHIMPOUR et al., 2010) e (PEPPOLONI et al., 2014), biomonitoramento (Biomonit.) (LARIOS et al., 2011), sistema de informação geográfica (SIG) (GARCÍA-GUTIÉRREZ; MATEOS-GARCÍA; RIQUELME-SANTOS, 2012), *Digital Library* (DL)(BORGES, 2013), e *healthcare* (ALI; MAJID, 2015).

Os trabalhos analisados utilizam como técnica de aprendizagem de máquina a tarefa de classificação, onde os tipos de classificadores variam consideravelmente, tanto os aplicados no nível básico, quanto no meta-classificador.

As técnicas de classificação mais utilizadas no nível 0 são as baseadas em árvores de decisão (*C4.5*, *RF*), em função (*MLP*, *SVM*), e no teorema de *Bayes* (NB). Já no nível 1 destaca-se o uso da técnica baseada em função (*MLP*, *SVM*). Como exceção ao uso de técnicas de classificação no nível 1, cita-se o trabalho apresentado por ALI; MAJID (2015) que utiliza GP para desenvolver um meta-classificador.

Quanto aos atributos empilhados, percebe-se que a maioria dos trabalhos estudados apresenta como vetor de características do meta-nível dados indicando predições, rótulos, ou rótulos e predições.

Tabela 11: Principais características dos trabalhos analisados.

Trabalho	Área	Nível 0	Nível 1	Atributos empilhados
<i>WOLPERT</i> (1992) apresenta a técnica de empilhamento	AM	Qualquer classificador	Qualquer classificador	Valores numéricos de 2 ou 3 dimensões
<i>BREIMAM</i> (1996) usa o empilhamento na formação de combinações lineares de diferentes preditores	AM	Árvores de regressão	Árvore de regressão	Subconjunto de regressões lineares
<i>TING; WITTEN</i> (1999) mostram o uso do empilhamento em tarefas de classificação	AM	<i>C4.5, NB, IB1</i>	<i>C4.5, NB, IB1, MLR</i>	Probab. preditas
<i>DZEROSKI; ZENKO</i> (2004) propõem um modelo de árvores multi resposta no meta classificador	AM	<i>C4.5, IBk, NB, K*, KDE, DT, MLR</i>	Árvores Multi Resposta	Probab. preditas; Probab. máxima; Entropia das probab.
<i>NESS et al.</i> (2009) utilizam o empilhamento para melhorar o desempenho de um sistema automático de anotação de <i>tags</i>	RI	<i>SVM</i>	<i>SVM</i>	TAV
<i>EBRAHIMPOUR et al.</i> (2010) usam o empilhamento para enfrentar o problema de reconhecimento de imagens de face com baixa resolução	Visão	<i>MLP</i>	<i>MLP</i>	Predições
<i>LARIOS et al.</i> (2011) aplicam o empilhamento à tarefa de identificação automática de espécies de insetos	Biomonit.	<i>RF</i>	<i>SVM</i>	Histogramas espaciais
<i>GARCÍA-G.; MATEOS-G.; RIQUELME-S.</i> (2012) propuseram o método <i>EVOR-STACK</i> visando melhorar a precisão de mapas temáticos	SIG	<i>SVM, K-NN</i>	<i>SVM, K-NN</i>	Predições e Matriz de pesos
<i>BORGES</i> (2013) identifica metadados bibliográficos duplicados	DL	<i>MLP, VP, SMO, NB, BN, RIPPER, C4.5</i>	Qualquer classificador	Rótulos; Predições; Rótulos e Predições
<i>PEPPOLONI et al.</i> (2014) apresentam o empilhamento como uma técnica para detecção e reconhecimento de objetos conhecidos em uma cena	Visão	<i>SVM</i>	<i>SVM</i>	Predições
<i>ALI; MAJID</i> (2015) propuseram o sistema <i>Can-Evo-Ens</i> , baseado no empilhamento, para prever as sequências de aminoácidos associados com o câncer de mama	Healthcare	<i>NB, K-NN, RF, SVM</i>	GP	Predições numéricas

## 3.2 Aplicações das Medidas de Diversidade

Esta seção apresenta um conjunto de trabalhos que utilizam diferentes medidas de diversidade para composição de comitês de classificação.

### 3.2.1 (SHIPP; KUNCHEVA, 2002)

SHIPP; KUNCHEVA (2002) analisaram as relações entre diferentes métodos de combinação de classificadores e diferentes medidas de diversidade. Nesse estudo foi verificada a correlação entre os métodos de combinação, a correlação entre as medidas de diversidade e a correlação cruzada entre os métodos de combinação e as medidas de diversidade.

Para a realização dos experimentos foram considerados dois conjuntos de dados e 10 métodos de combinação: voto da maioria, máximo, mínimo, média, produto, *naive bayes*, *Behavior-knowledge space*, método de *Wernicke*, modelos de decisão, oráculo. Também foram utilizadas 10 medidas de diversidade: estatística  $Q$ , medida de discordância  $Dis$ , coeficiente de correlação  $\rho$ , falha dupla  $df$ , variância *kohavi-Wolpert*  $KW$ , entropia  $E$ , concordância entre avaliadores  $k$ , medida de dificuldade  $\theta$ , diversidade generalizada  $GD$ , diversidade de falhas coincidentes  $CFD$ .

Verificando a correlação entre os métodos de combinação e as medidas de diversidade separadamente, percebe-se que o voto da maioria está altamente correlacionado com o *Naive Bayes*, e a  $GD$  está fortemente correlacionada com  $k$  e o  $\rho$ .

Entretanto, quanto ao resultado da correlação entre os métodos e as medidas, a única correlação positiva foi que a medida de diversidade  $df$  e a  $\theta$  apresentaram correlação razoável com as combinações voto de maioria e *Naive Bayes*, para ambos os conjuntos de dados. Os experimentos não detectaram uma relação consistente entre as outras medidas de diversidade e os 10 métodos de combinação.

Segundo os autores, as correlações mostram que muitos dos métodos de combinação e medidas de diversidade são mesmo independentes, existindo pouca evidência de qualquer relação entre a precisão do método de combinação e o valor da medida de diversidade. Isso significa que essas medidas de diversidade dificilmente podem ser usadas como um indicador na concepção de conjuntos de classificadores.

### 3.2.2 (KUNCHEVA; WHITAKER, 2003)

KUNCHEVA; WHITAKER (2003) definem uma série de medidas de diversidade e as relacionam com a qualidade de um sistema de classificação. Os autores investigaram respostas para as seguintes questões:

1. Como se pode definir e medir a diversidade?
2. Como as várias medidas de diversidade estão relacionadas entre si?

3. Como as medidas estão relacionadas com a precisão do *ensemble*?
4. Existe uma medida que é melhor para o desenvolvimento de comitês que minimizam o erro?
5. Como se pode usar as medidas na concepção do *ensemble* final?

Foram estudadas as mesmas medidas de diversidade mencionadas por SHIPP; KUNCHEVA (2002), que podem medir a diversidade entre as saídas do classificador binário (voto correto ou incorreto para o rótulo da classe). Além disso, tentando responder às perguntas 2 e 3, foram utilizadas quatro configurações experimentais.

Os dois primeiros experimentos mostraram as medidas de diversidade fortemente correlacionadas entre si. Já, no terceiro experimento, utilizando a base de dados *Wisconsin breast cancer (UCI)*, os resultados revelaram a possível inadequação das medidas de diversidade para prever uma melhora na acurácia individual. Nesses experimentos foram utilizadas as seguintes técnicas de combinação: voto da maioria, *naive bayes*, máximo, mínimo, média e produto dos métodos simples, *Behavior-knowledge space* e modelos de decisão.

Além disso, esta deficiência foi confirmada no quarto experimento em que nove classificadores foram utilizados com dois conjuntos de dados: *Phoneme (UCI)* e *Conetorus (KUNCHEVA, 2000)*. Nesse caso, foram aplicados os métodos de *ensemble bagging* e métodos de classificação fracos randômicos.

Em relação às respostas para as perguntas 4 e 5, os autores sugerem que a escolha da medida de diversidade pode ser baseada na facilidade de interpretação, sendo nesse caso recomendado o uso da estatística  $Q_{av}$ . Quanto ao uso de medidas de diversidade para aprimorar o uso de *ensembles* de classificadores (questão 5) ainda é uma questão em aberto devido a falta de uma relação definitiva entre as medidas e a melhoria da acurácia.

De acordo com os autores, embora haja conexões comprovadas entre diversidade e acurácia em alguns casos especiais, os resultados levantaram algumas dúvidas sobre a utilidade das medidas de diversidade na construção de conjuntos de classificadores em problemas de reconhecimento de padrões da vida real.

### 3.2.3 (DYMITR; BOGDAN, 2005)

DYMITR; BOGDAN (2005) apresentam uma revisão da metodologia de seleção de classificadores e avalia a aplicabilidade prática das medidas de diversidade no contexto da combinação de classificadores por voto da maioria.

Uma série de algoritmos de busca são implementados de forma flexível, o que permitiu a aplicação de uma série de critérios de seleção diferentes, incluindo erro de voto da maioria e várias medidas de diversidade (falha dupla  $df$ , entropia  $E$ , estatística  $Q$ , medida de discordância  $Dis$ , falha da maioria  $FM$ , concordância entre avaliadores  $k$ , coeficiente

de correlação  $\rho$ , medida de dificuldade  $\theta$ , diversidade generalizada  $GD$ , diversidade de falhas coincidentes  $CFD$ , variância *kohavi–Wolpert*  $KW$ ).

Um trabalho experimental foi realizado, onde todos os algoritmos de busca foram executados sobre os resultados de 15 classificadores diferentes aplicados a 27 conjuntos de dados. Isto foi repetido em muitas versões correspondentes a diferentes critérios de seleção. Esses algoritmos utilizaram um vetor binário de incidências classificadoras, indicando exclusão (0) ou inclusão (1) do classificador na combinação, como representação da solução de seleção. A partir disso, o voto da maioria foi aplicado às melhores combinações devolvidas pelos algoritmos e forneceu a base para a avaliação de diferentes medidas de diversidade usadas como critérios de seleção.

De acordo com os autores, obteve-se melhores resultados utilizando o erro de voto da maioria como critério de seleção, embora uma seleção baseada em medidas de  $df$  e  $FM$ , bem correlacionadas, também tenham proporcionado bons resultados. Além disso, os resultados indicam inapropriado o uso de medidas de diversidade como critérios de seleção. Isso se deve ao fato de que a maioria delas não é invariante ao tamanho das combinações, o que pode induzir ao erro o algoritmo de busca.

### 3.2.4 (OLIVEIRA, 2008)

O trabalho desenvolvido por OLIVEIRA (2008) tem por objetivo investigar o comportamento de multiclassificadores heterogêneos mediante o dilema diversidade–acurácia.

Para realização dessa investigação foram utilizados dois algoritmos genéticos (um mono–objetivo e um multiobjetivo) na escolha dos componentes para compor o comitê de classificadores. Assim, foi possível ser realizada uma análise do impacto de se escolher classificadores base utilizando-se: apenas a acurácia ou apenas a diversidade (mono–objetivo); e, conjuntamente, a acurácia e a diversidade (multiobjetivo).

Foram utilizados três tipos diferentes de algoritmos de aprendizado para construir os classificadores base:  $k$ - $NN$ ,  $MLP$  e  $SVM$ . Para cada um desses algoritmos, foram treinados 10 classificadores base.

Os experimentos foram realizados com três bases de dados retiradas do *UCI*. A cada experimento envolvendo a diversidade, este foi executado três vezes, alterando-se em sua configuração a medida de diversidade usada, para uma das seguintes: Estatística  $Q$ , Coeficiente de Correlação  $\rho$  ou Concordância  $k$ . Além disso, foram executados utilizando-se os métodos de combinação Soma (KITTLER; ALKOOT, 2003) e o classificador *Naive Bayes*.

Para o autor, combinar a diversidade com a acurácia pode conduzir a criação de comitês mais acurados do que os gerados com essas medidas isoladamente. Além disso, o autor considera que a diversidade possui um papel importante na geração de multiclassificadores.

### 3.2.5 (MUHAMMAD; JIM, 2010)

Com o objetivo de melhorar o desempenho do classificador *Nearest-Neighbour (INN)*, MUHAMMAD; JIM (2010) propuseram uma nova técnica de *ensemble* que combina vários classificadores *INN*, onde cada um usa uma métrica de distância diferente, e um subconjunto diferente de vetor de características.

O *ensemble* proposto *DF-TS3-INN* é avaliado utilizando vários conjuntos de dados de referência do *UCI* e uma aplicação do mundo real. Os resultados indicam um aumento significativo no desempenho quando comparado com diferentes métodos de criação de classificadores individuais ou *ensembles*, sendo eles: *C4.5*, *Random Forest*, *Naive Bayes*, *Bagging*, *AdaBoost*, *Random Sub-Space*.

Além disso, foram usadas as medidas de diversidade “Discordância *Dis*” e “Entropia *E*” para avaliar o impacto da diversidade na melhoria da precisão da classificação utilizando o *ensemble*.

Segundo os autores, os valores das duas medidas são altos para todos os conjuntos de dados ao selecionar subconjuntos de características para todos os classificadores simultaneamente (*DF-TS3-INN*). Assim, a diversidade desempenha um papel importante no aumento da precisão de classificação de vários conjuntos de dados usando a técnica de *ensemble* proposta.

### 3.2.6 (MAKHTAR et al., 2012)

MAKHTAR et al. (2012) propõem uma abordagem de *ensemble* que tem a diversidade calculada usando a medida de discordância *Dis* em relação às saídas da classificação. Um sistema de *ranking* de classificação é introduzido para a seleção de classificadores relevantes e diversos para formarem o *ensemble* final. Também é proposto um método de otimização de *ensemble* como sendo uma técnica que se aplica à seleção dos classificadores.

A previsão é feita selecionando os classificadores relevantes de uma coleção de classificadores existentes. Para obter melhores resultados o método propõe que os classificadores possam ser selecionados com base em suas medidas de desempenho, tais como precisão, taxa de falso negativo e taxa de falso positivo. Esses classificadores serão então combinados utilizando a votação por maioria simples.

O algoritmo do método de otimização começa montando um *ranking* de classificação para todos os classificadores da coleção. O modelo que apresentar melhor valor será atribuído como um modelo base. Assim, para encontrar o classificador mais diverso em relação ao modelo base, é calculada a diversidade entre o modelo e os demais classificadores relevantes, usando a medida *Dis*.

Na realização dos experimentos foram utilizados os algoritmos *IBk*, *J48*, *JRip*, *Naive Bayes*, *MultilayerPerceptron*. Os modelos preditivos foram aplicados a diversos dados toxicológicos.

De acordo com os autores, o método de otimização de *ensemble* supera quatro outros métodos de *ensemble* (*bagging*, *stacking*, *Bayes*, e *boosting*). Além disso, a medida *Dis* tem um importante papel na formação do *ensemble* final.

### 3.2.7 (WHALEN; PANDEY, 2013)

WHALEN; PANDEY (2013) abordam a questão de como diferentes métodos de *ensemble* utilizam a diversidade para aumentar a precisão usando conjuntos de dados complexos. Assim, foram utilizados conjuntos de dados genômicos do mundo real para analisar e comparar o desempenho de métodos de *ensembles* para dois importantes problemas nesta área: predição de funções de proteína e a predição de interações genéticas. Para a realização dos experimentos foram treinados 27 tipos de classificadores heterogêneos.

Esses métodos incluem agregação simples, meta-aprendizagem, meta-aprendizagem baseada em *cluster* e seleção de conjunto usando classificadores heterogêneos. Além disso, foi estabelecida uma nova conexão entre seleção de conjunto e meta-aprendizagem, demonstrando como ambos os métodos díspares estabelecem um equilíbrio entre diversidade de *ensemble* e desempenho.

Para avaliar a diversidade foi usada a medida estatística  $Q$  criando primeiro os rótulos preditos a partir de probabilidades classificadoras limiares, produzindo 1 para valores maiores que 0,5 e 0 caso contrário, ou seja, um classificador binário.

Os experimentos verificando o desempenho em função da diversidade, para todas as combinações por pares dos 27 classificadores base, mostram uma relação ruim, que se mantém para cada um dos conjuntos de dados, pois combinações mais diversas tipicamente resultam em um menor desempenho, exceto para classificadores de alto desempenho, como *Random Forest* e modelos de regressão generalizada.

Além disso, analisando a diversidade do *ensemble* e o desempenho em função do número de iteração tanto para seleção gulosa simples e seleção de *ensemble* nos dois conjuntos de dados, os experimentos mostram que a primeira não melhora a diversidade e o desempenho diminui com o tamanho do conjunto, enquanto a seleção de *ensemble* explora com êxito a compensação entre diversidade e desempenho e atinge um equilíbrio ao longo do tempo.

Para os autores, a diversidade deve ser equilibrada com a precisão, pois embora melhorar o desempenho dependa da diversidade, o tipo errado de diversidade limita o desempenho do *ensemble*.

### 3.2.8 (FARIA et al., 2014)

FARIA et al. (2014) propõem um *framework* para a seleção e fusão de classificadores. O método busca combinar automaticamente os classificadores mais discriminativos usando (*SVM*), bem como explorar o uso de medidas de diversidade para selecionar os classificadores menos correlacionados, porém eficazes.

Foram realizados experimentos com quatro conjuntos de dados diferentes, onde foram utilizados sete métodos de aprendizagem, sendo eles: *Naive Bayes*, Árvore de Decisão, *Simple Logistic*, *Naive Bayes Tree*, e *KNN*, usando  $k=1$ ,  $k=3$ , e  $k=5$ .

Para determinar quais métodos de aprendizagem são mais adequados para serem combinados foram utilizadas as medidas de diversidade coeficiente de correlação  $\rho$ , falha dupla  $df$ , medida de discordância  $Dis$ , concordância entre avaliadores  $k$  e estatística  $Q$ .

Na realização dos experimentos foram comparadas seis técnicas de fusão: o *framework* proposto usando *SVM* (*FSVM-KERNEL-49*) e (*FSVM-KERNEL-BEST*) - considerando menos classificadores, duas abordagens *Adaboost* (*BOOST-DEFAULT* e *BOOST-49*), *Bagging* (*BAGG-49*) e voto da maioria (*MV-49*). Nesse contexto, usando  $|C| = 49$  significa que todas as combinações possíveis de sete classificadores e sete descritores de imagem são consideradas no processo de fusão.

Os experimentos mostram que o *framework* proposto supera o melhor classificador utilizado e obtém resultados estatisticamente melhores do que às abordagens de fusão votação por maioria, *Bagging* e *Adaboost*, usando conjuntos de treinamento reduzidos. Além disso, é capaz de selecionar classificadores e também aprender, indiretamente, quais descritores são mais apropriados para a aplicação final.

Segundo os autores, para manter uma alta taxa de reconhecimento com o mínimo de esforço computacional, a estratégia proposta de seleção de classificadores explorando o uso de medidas de diversidade pode potencialmente melhorar a qualidade dos classificadores selecionados.

### 3.2.9 (SLUBAN; LAVRAC, 2015)

SLUBAN; LAVRAC (2015) investigam a relação entre a diversidade de *ensembles* heterogêneos de algoritmos de detecção de ruído e seu desempenho.

A diversidade dos *ensembles* foi determinada utilizando onze medidas de diversidade, sendo elas: estatística  $Q$ , coeficiente de correlação  $\rho$ , medida de discordância  $Dis$ , falha dupla  $df$ , entropia  $E$ , “Good” diversidade  $D_g$ , “Bad” diversidade  $D_b$ , concordância entre avaliadores  $k$ , ambiguidade  $A$ , *Kohavi-Wolpert*  $KW$ , *kappa*  $K_p$ .

Para avaliar a relação entre diversidade e desempenho de *ensemble* de detecção de ruído, foi calculado o coeficiente de correlação de *Spearman* (SPEARMAN, 1904) para todos os pares de uma medida de diversidade e uma medida de desempenho, separadamente para cada um dos dois esquemas de combinação (voto da maioria e votação por consenso).

Para realização dos experimentos foram utilizados 10 conjuntos de dados do *UCI*, com diferentes níveis de ruído de classe injetado aleatoriamente, e 10 algoritmos base.

Os experimentos mostraram que os resultados de detecção de ruído obtidos por voto da maioria não se correlacionam positivamente com os resultados das medidas de diversidade, pois todas as medidas de diversidade concordam que *ensembles* menos diversos

conduzem a um melhor desempenho na detecção de ruído.

Por outro lado, a experiência usando votação por consenso mostrou correlações significativamente maiores entre as medidas de diversidade e desempenho de detecção de ruído. Os resultados mostraram que conjuntos mais diversos tendem a alcançar uma maior precisão de detecção de ruído de classe, enquanto que conjuntos menos diversos tendem a alcançar um maior *recall* de detecção de ruído e maior pontuação  $F$ . Além disso, a medida de  $KW$  e a  $A$  estão fortemente correlacionadas com o *recall* e a medida  $F$ , e a medida de  $D_b$  com a precisão da detecção de ruído.

De acordo com os autores, a diversidade de um *ensemble* pode ser usada como orientação para se obter um bom desempenho apenas quando se utiliza o esquema de votação por consenso.

### 3.2.10 Análise Comparativa do uso de Medidas de Diversidade

A seguir é apresentada uma análise quanto ao uso de medidas de diversidade de acordo com os trabalhos estudados.

Os dois primeiros trabalhos exploram o uso das mesmas medidas de diversidade, onde o primeiro busca verificar a correlação entre os métodos de combinação, a correlação entre as medidas, e a correlação cruzada entre os métodos de combinação e as medidas (SHIPP; KUNCHEVA, 2002), e o segundo relaciona as medidas com a qualidade de um sistema de classificação (KUNCHEVA; WHITAKER, 2003). Ambos os trabalhos mostram que não existe uma relação bem definida entre a precisão do método de combinação e o valor da medida de diversidade, principalmente em problemas de reconhecimento de padrões da vida real. Em corroboração a essa ideia DYMITR; BOGDAN (2005) demonstra em seu trabalho a não apropriação do uso de medidas de diversidade como critérios de seleção de classificadores.

Os demais trabalhos abordam o uso de medidas de diversidade nos seguintes contextos: na geração de *ensembles* mais precisos, combinando a diversidade com a acurácia (OLIVEIRA, 2008); no aumento da precisão de classificação de um *ensemble* formado por vários classificadores *INN* (MUHAMMAD; JIM, 2010); na seleção de classificadores diversos e eficazes para formação de um *ensemble* final (MAKHTAR et al., 2012) e (FARIA et al., 2014); na orientação para se obter um bom desempenho na detecção de ruído utilizando o esquema de votação por consenso (SLUBAN; LAVRAC, 2015). Além disso, WHALEN; PANDEY (2013) salienta que a diversidade deve ser equilibrada com a precisão, pois o tipo errado de diversidade limita o desempenho do *ensemble*.

A Tabela 12 mostra uma série de medidas de diversidade, sinalizando as que foram utilizadas em cada trabalho e se foram consideradas apropriadas para a seleção de classificadores e/ou para a melhora da acurácia. Também é destacado a técnica de combinação dos classificadores. Os três primeiros trabalhos não associam a relação direta entre diversidade e acurácia do comitê. Os demais trabalhos apoiam a questão de pesquisa, mas com

objetivos muito específicos e utilizando poucas medidas de diversidade. A exceção é o trabalho de SLUBAN; LAVRAC (2015) que avalia um conjunto de onze medidas, sendo quatro delas não abordadas nesta dissertação. Como muitos trabalhos mostram resultados contraditórios em relação ao dilema diversidade-accurácia, e apenas um aborda o empilhamento como técnica de combinação, torna-se necessário um estudo mais amplo, para preencher esta lacuna na literatura científica. Esta é a principal motivação da proposta apresentada no capítulo seguinte.

Tabela 12: Análise do uso de medidas de diversidade de acordo com os trabalhos estudados.

Trabalho	$Q$	$Dis$	$df$	$KW$	$\rho$	$k$	$E$	$\theta$	$GD$	$CFD$	$FM$	$A$	$D_g$	$D_b$	$K_p$	Apropriada
(SHIPP; KUNCHEVA, 2002) voto da maioria, máximo, mínimo, média, produto, <i>Naive Bayes</i> , <i>Behavior-Knowledge Space</i> , método de <i>Wernicke</i> , modelos de decisão, e oráculo	•	•	•	•	•	•	•	•	•	•						
(KUNCHEVA; WHITAKER, 2003) voto da maioria, <i>Naive Bayes</i> , máximo, mínimo, média e produto dos métodos simples, <i>Behavior-Knowledge Space</i> , modelos de decisão, <i>Bagging</i> , e métodos fracos randômicos	•	•	•	•	•	•	•	•	•	•						
(DYMITR; BOGDAN, 2005) voto da maioria	•	•	•	•	•	•	•	•	•	•	•					
(OLIVEIRA, 2008) Soma e <i>Naive Bayes</i>	•				•	•										•
(MUHAMMAD; JIM, 2010) <i>DF-TS3-INN</i>		•					•									•
(MAKHTAR et al., 2012) <i>Bagging</i> , <i>Stacking</i> , <i>Naive Bayes</i> , e <i>Boosting</i>		•														•
(WHALEN; PANDEY, 2013) agregação simples, meta-aprendizagem, meta-aprendizagem baseada em <i>cluster</i> e seleção de <i>ensemble</i>	•															•
(FARIA et al., 2014) <i>SVM</i> ( <i>FSVM-KERNEL-49</i> e <i>FSVM-KERNEL-BEST</i> ), <i>Adaboost</i> ( <i>BOOST-DEFAULT</i> e <i>BOOST-49</i> ), <i>Bagging</i> ( <i>BAGG-49</i> ) e voto da maioria ( <i>MV-49</i> )	•	•	•		•	•										•
(SLUBAN; LAVRAC, 2015) voto da maioria e votação por consenso	•	•	•	•	•	•	•					•	•	•	•	•
Trab. Proposto <i>Stacking</i>	•	•	•	•	•	•	•									?
(?) questão de pesquisa																

## 4 ABORDAGEM PROPOSTA

Este capítulo apresenta em detalhes a abordagem proposta. A Seção 4.1 retoma o objetivo geral do trabalho e enfatiza os objetivos específicos, a questão de pesquisa e a principal contribuição científica. Na Seção 4.2 o método proposto para atingir os objetivos é detalhado.

### 4.1 Objetivo

O objetivo principal dessa dissertação é mostrar se a diversidade influencia na qualidade do empilhamento de classificadores supervisionados.

A abordagem proposta é baseada na seguinte questão: Quanto maior a diversidade entre os classificadores, maior será o ganho do empilhamento, quando comparado a precisão do melhor classificador base, independente do conjunto de dados utilizado.

Como contribuição científica demonstra-se, por meio de experimentos, a relação entre múltiplas medidas de diversidade e o ganho do empilhamento.

### 4.2 Método Proposto

Esta seção descreve o método proposto para analisar o impacto da diversidade no empilhamento de classificadores supervisionados, que é representado graficamente na Figura 3.

Para cada conjunto de dados analisado, diferentes algoritmos de aprendizagem são usados para treinar vários classificadores base. As predições retornadas por esses classificadores são avaliadas e utilizadas para calcular diversas medidas de diversidade. Essas medidas verificam se e como os classificadores concordam ou discordam sobre o rótulo da classe predita.

No nível 1, as predições dos classificadores base para cada instância original são usadas para compor um novo conjunto de dados que é submetido a outro algoritmo para treinar o meta-classificador. A predição final é determinada a partir da combinação do conhecimento aprendido pelos classificadores base.

O ganho do empilhamento  $G$  é calculado como mostrado pela Equação 8, e represen-

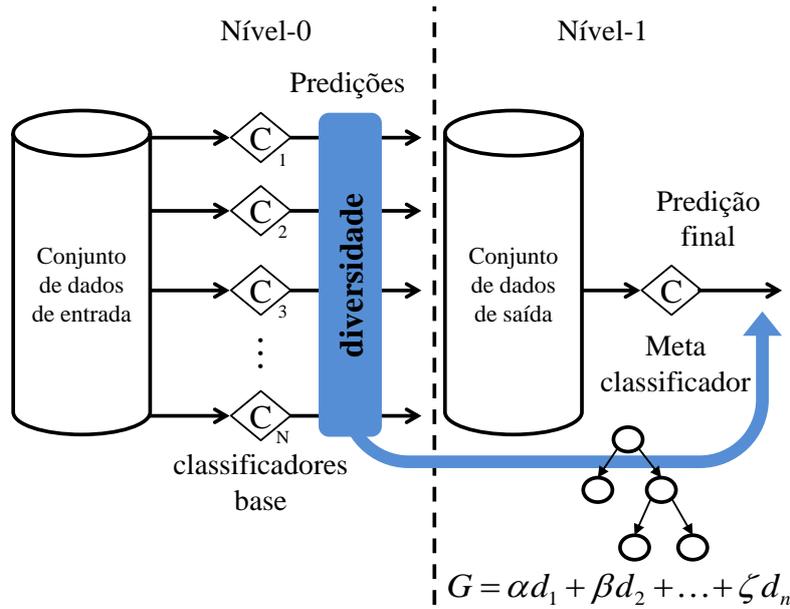


Figura 3: Metodologia do trabalho proposto.

tado como uma porcentagem, onde  $A_{MC}$  é a métrica de avaliação alcançada pelo melhor meta-classificador e  $A_{C_{melhor}}$  pelo melhor classificador base. Os valores negativos retornados para o ganho indicam perda de qualidade, ou seja, a acurácia do melhor empilhamento é pior do que a acurácia do melhor classificador base.

$$G = \left( \frac{A_{MC}}{A_{C_{melhor}}} \right) - 1 \quad (8)$$

Finalmente, a relação entre as medidas de diversidade e o ganho do empilhamento, calculado previamente para múltiplos conjuntos de dados, é induzida por meio de um modelo de regressão.

#### 4.2.1 Classificadores

Os vetores de características de cada conjunto de treinamento são compostos por todos os atributos e pelo rótulo da classe. Os algoritmos citados a seguir são usados no nível 0 do método de empilhamento. Esta escolha foi motivada principalmente porque os algoritmos são bastante heterogêneos, uma vez que se baseiam em particularidades distintas:

- *MLP* (HAYKIN, 2007) - rede neural artificial, baseado em função;
- *SMO* (PLATT, 1999a) - variação do *SVM* (BOSER; GUYON; VAPNIK, 1992), baseado em função;
- *NB* (JOHN; LANGLEY, 1995) - baseado no teorema de *Bayes*;
- *RIPPER* (COHEN, 1995) - baseado em regras;
- *C4.5* (QUINLAN, 1993) - baseado em árvores de decisão;

- *RF* (BREIMAN, 2001) - baseado em um conjunto de árvores de decisão.

O meta-classificador é treinado utilizando qualquer algoritmo de classificação combinando o conhecimento aprendido pelos classificadores base, sendo finalmente usado para gerar uma predição final.

Os atributos que formam o conjunto de treinamento para a aprendizagem do meta-classificador variam de acordo com os algoritmos dos classificadores base. Para o *NB*, a predição corresponde à probabilidade a posteriori de um registro pertencer a mesma classe. Já para os algoritmos *RIPPER*, *C4.5* e *RF*, a predição corresponde à precisão da regra ou do nó que classificou cada amostra. Nos algoritmos baseados em função, a predição é mapeada diretamente para o rótulo da classe. Independentemente dos classificadores base, o último campo é o mesmo rótulo da classe original.

#### 4.2.2 Diversidade

As medidas de diversidade citadas a seguir são usadas para calcular a diversidade das predições retornadas pelos classificadores base que compõem o empilhamento. Algumas dessas trabalham operam sobre um par e outras sobre um conjunto de  $N$  classificadores.

- $df$  (GIACINTO; ROLI, 2001) - par;
- $Dis$  (HO, 1998) - par;
- $Q$  (AFIFI; AZEN, 2014) - par;
- $\rho$  (SNEATH; SOKAL, 1973) - par;
- $k$  (DIETTERICH, 2000) - conjunto;
- $KW$  (KOHAVI; WOLPERT et al., 1996) - conjunto;
- $E$  (CUNNINGHAM; CARNEY, 2000) - conjunto.

#### 4.2.3 Analisando o impacto da diversidade

O impacto da diversidade no empilhamento pode ser analisado observando-se a relação entre os valores das medidas de diversidade e o ganho do empilhamento para múltiplos conjuntos de dados.

Esta relação é deduzida usando modelos de regressão linear ou árvores de regressão, tendo como atributo alvo o ganho do empilhamento  $G$ . O vetor de características é composto pelas medidas de diversidade  $d_1, d_2, \dots, d_n$  citadas na seção anterior previamente calculadas. Os modelos de regressão mostram o quanto cada medida se relaciona com o ganho de empilhamento.

Esta proposta pode ser vista como uma estratégia genérica para relacionar diversidade de classificadores e qualidade do empilhamento, onde múltiplas medidas de diversidade

ou classificadores podem ser adicionados ou substituídos por outros mais efetivos ou eficientes em cenários específicos. Por exemplo, a falha dupla  $df$  pode ser substituída pela falha da maioria  $FM$ . Neste caso, apenas os modelos de regressão precisam ser treinados novamente. Caso um classificador base seja adicionado, também haverá necessidade de calcular todas as medidas de diversidade para este classificador e treinar os meta-classificadores novamente.

## 5 AVALIAÇÃO EXPERIMENTAL

Este capítulo descreve os experimentos realizados com a intenção de avaliar o método proposto para analisar o impacto da diversidade no empilhamento de classificadores supervisionados.

Primeiramente na Seção 5.1 são apresentadas as bases de dados usadas para a realização dos experimentos. Na Seção 5.2 são mostradas as configurações dos experimentos, e em seguida os resultados dos mesmos 5.3, onde é aplicado o teste-t, visando avaliar sua significância estatística. Por fim, são apresentados os modelos de regressão, que relacionam os valores das medidas de diversidade e o ganho do empilhamento.

### 5.1 Bases de Dados

Foram utilizados 54 conjuntos de dados de classificação extraídas do *Machine Learning Repository (UCI)*<sup>2</sup> (Tabela 13), dois conjuntos de dados sintéticos disponibilizados pela *University of Eastern Finland*<sup>3</sup> e um conjunto de dados extraído de (TOMASINI et al., 2016) (Tabela 14).

Para a escolha das bases do *UCI* diferentes características foram levadas em consideração. A Tabela 13 mostra para cada conjunto de dados as seguintes informações: nome, quantidade de instâncias (#I), atributos (#A) e rótulos de classe (#CL), tipos de dados (TD), área do conhecimento, ano de depósito no repositório e número de citações na literatura científica, de acordo com o *UCI*.

As bases de dados escolhidas do *UCI* abrangem várias áreas do conhecimento: *social, financial, life, game, computer, Physical* e *business*. Além disso, têm objetivos diversos. Quanto as características dos atributos, os mesmos podem ser do tipo inteiro (I), real (R) e categórico (C). O número de instâncias varia de 187 a 12960, e o número de atributos e rótulo de classe varia de 4 a 216 e de 2 a 48, respectivamente.

---

<sup>2</sup><http://archive.ics.uci.edu/ml>

<sup>3</sup><http://cs.joensuu.fi/sipu/datasets>

Tabela 13: Conjunto de dados do *UCI* usados na avaliação experimental.

#	Dataset	#I	#A	#CL	TD	Área	Ano	Citações
1	Abalone	4177	8	28	I,R,C	Life	1995	31
2	Annealing	898	38	6	I,R,C	Physical		6
3	Audiology (Std.)	226	69	24	C	Life	1992	20
4	Balance Scale	625	4	3	C	Social	1994	13
5	Banknote Authent.	1372	4	2	R	Computer	2013	
6	Blood Transf. Serv. Center	748	4	2	R	Business	2008	1
7	Breast Cancer Wisconsin	699	10	2	I	Life	1992	42
8	Car Evaluation	1728	6	4	C	Business	1997	18
9	Chess (K-R vs. K-P)	3196	36	2	C	Game	1989	29
10	Chronic Kidney Disease	400	24	2	R	Life	2015	
11	Congressional Voting Rec.	435	16	2	C	Social	1987	21
12	Connect. Bench (S,M vs. R)	208	60	2	R	Physical		54
13	Connect. Bench (VR-DD)	990	13	11	R	Physical		5
14	Contrac. Method Choice	1473	9	3	I,C	Life	1997	4
15	Credit Approval	690	15	2	I,R,C	Financial		6
16	Dermatology	366	34	6	I,C	Life	1998	9
17	Diabetic Retinopat. Debrec.	1151	19	2	I,R	Life	2014	
18	Dresses Attribute Sales	501	12	2	I,R,C	Computer	2014	
19	Ecoli	366	7	8	R	Life	1996	13
20	Forest type mapping	325	27	4	I,R,C	Life	2015	1
21	Glass Identification	214	9	7	R	Physical	1987	53
22	Hill-Valley	606	100	2	R	Physical	2008	
23	ILPD - Indian Liver Patient	583	10	2	I,R	Life	2012	2
24	Ionosphere	351	34	2	I,R	Physical	1989	56
25	Leaf	340	15	30	R	Computer	2014	
26	Low Resolution Spectrometer	531	102	48	I,R	Physical	1988	1
27	Mammographic Mass	961	5	2	I	Life	2007	1
28	Molecular Bio. (S-junction)	3190	60	3	C	Life	1992	34
29	Multiple Features	2000	216	10	I,R	Computer		7
30	Nursery	12,960	8	5	C	Social	1997	14
31	Opt. Recog. Handwrit. Dig.	5620	64	10	I	Computer	1998	8
32	Page Blocks Classification	5473	10	5	I,R	Computer	1995	5
33	Pen-based Recog. Handwrit.	10992	16	10	I	Computer	1998	20
34	Phishing Websites	11055	30	2	I	Security	2015	3
35	Primary Tumor	339	17	22	C	Life	1988	11
36	QSAR biodegradation	1055	41	2	I,R	Life	2013	1
37	Qualitative Bankruptcy	250	6	2	C	Computer	2014	
38	Seismic-Bumps	2584	18	2	R	Physical	2013	
39	Solar Flare	323	12	6	C	Physical		7
40	Soybean (Large)	683	35	19	C	Life	1988	51
41	Spambase	4601	57	2	I,R	Computer	1999	4
42	SPECT Heart	187	22	2	C	Life	2001	7
43	Statlog (Vehicle Silh.)	846	18	4	I	Business		26
44	Thoracic Surgery	470	16	2	I,R	Life	2013	1
45	Thyroid Disease (Hypothy.)	3772	29	4	C,R	Life	1987	28
46	Thyroid Disease (Sick)	3772	29	2	C,R	Life	1987	28
47	Tic-Tac-Toe Endgame	958	9	2	C	Game	1991	22
48	Turkiye Student Eval.	5819	32	3	I	Life	2013	
49	Vertebral Column	310	6	2	R	Life	2011	3
50	Waveform Database Gen. (V2)	5000	40	3	R	Physical	1988	36
51	Wholesale customers	440	7	2	I	Business	2014	2
52	Wilt	4839	5	2	C,R	Life	2014	1
53	Wine Quality	4898	11	7	R	Business	2009	1
54	Yeast	1484	8	10	R	Life	1996	15

I = Inteiro, R = Real, C = Categórico

Sob essas bases, foi aplicado um conjunto de operações de pré-processamento, visando padronizar o conteúdo de forma a deixá-las aptas para execução dos algoritmos na ferramenta de mineração de dados *Weka*<sup>4</sup> (WITTEN; FRANK, 2011). As principais operações foram a remoção de espaços duplos entre as instâncias, nomeação de atributos, padronização do delimitador de separação (,) e, a mudança do tipo de dados de numérico para nominal.

Após o pré-processamento, essas bases utilizadas para treinar modelos de classificação heterogêneos, isto é, utilizando os diferentes algoritmos descritos anteriormente (Subseção 4.2.1). As predições dos classificadores base são empilhadas compondo um conjunto de treinamento de nível 1 no qual o modelo de classificação final é aprendido.

Em relação aos dados sintéticos, foram escolhidos conjuntos de dados bidimensionais com características espaciais bem distintas, que permitem a interpretação dos resultados e do comportamento dos algoritmos por inspeção visual. O conjunto de dados *Spiral* (CHANG; YEUNG, 2008) é composto por 3 espirais concêntricas. Já *R15* (VEENMAN; REINDERS; BACKER, 2002) é composto por 15 distribuições gaussianas semelhantes. Nesse caso, o mesmo foi modificado para duas classes e incluído pouco ruído. *D13* (TOMASINI et al., 2016) são duas gaussianas parametrizadas aleatoriamente e com bastante ruído. A Figura 4 apresenta os referidos conjuntos de dados.

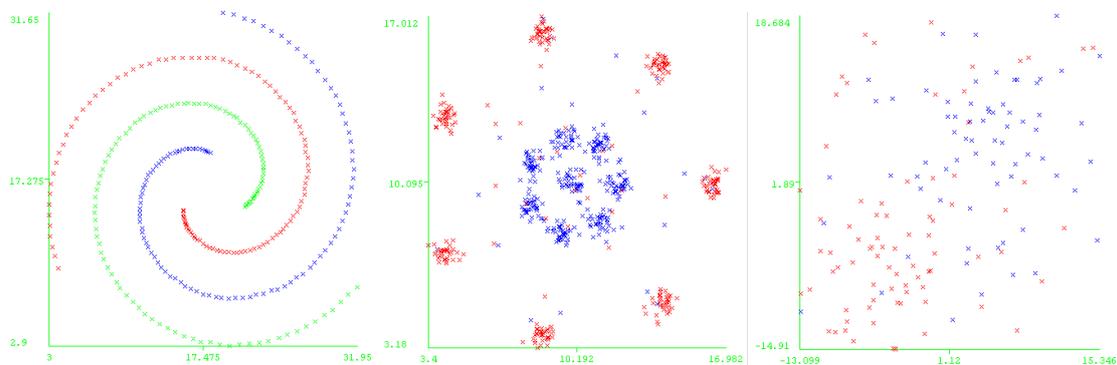


Figura 4: Base de dados *Spiral*, *R15* e *D13*

A Tabela 14 mostra para cada conjunto de dados as seguintes informações: nome, quantidade de instâncias (#I), atributos (#A) e rótulos de classe (#CL), tipos de dados (TD), e ano de disponibilização.

Tabela 14: Conjunto de dados sintéticos usados na avaliação experimental.

	Dataset	#I	#A	#CL	Tipos de Dados	Ano
1	<i>Spiral</i>	312	2	3	R	2008
2	<i>R15</i>	599	2	2	R	2002
3	<i>D13</i>	149	2	2	R	2016

<sup>4</sup><http://www.cs.waikato.ac.nz/ml/weka>

Quanto as características dos atributos, os mesmos são do tipo real (R). O número de instâncias varia de 150 a 600, o número de atributos é 2 para as 3 bases, e o rótulo de classe é 2 para as bases R15 e D13, e 3 para a Spiral.

Esses conjuntos de dados também são utilizados para treinar modelos de classificação heterogêneos, conforme mencionado anteriormente para os dados do *UCI*.

## 5.2 Configuração dos Experimentos

O método detalhado na Seção 4.2 foi exercitado utilizando cada algoritmo citado para treinar os classificadores base e também o meta-classificador. Os classificadores base e o empilhamento foram avaliados a partir da acurácia (TAN; STEINBACH; KUMAR, 2005), que estima a qualidade da classificação, ou seja, avalia a capacidade de predição do modelo.

Os experimentos foram realizados em um computador pessoal usando a ferramenta *Weka*. Para as bases do *UCI* os algoritmos foram executados com os valores padrão desta ferramenta. Já para as bases sintéticas, foram realizadas duas categorias de experimentos: primeiramente os algoritmos foram executados com seus valores padrão, em seguida foram parametrizados visando melhorar a avaliação da classificação. Para todos os experimentos foram utilizadas 10 partições na validação cruzada.

Usando as bases sintéticas, foram escolhidos, para cada algoritmo, um ou mais parâmetros para serem testados, sendo eles:

- *MLP* - número de nós da camada oculta (*hidden layers*) = 4 e número de épocas (*training time*) = 1000;
- *SMO* - *Kernel* = *Puk* e usando um parâmetro  $\gamma$  *Kernel* = *RBF*;
- *NB* - utilizar um estimador *kernel* para atributos numéricos ao invés de uma distribuição normal (*useKernelEstimator* = *True*);
- *JRip*<sup>5</sup> - quantidade de dados utilizados na poda (*folds*) = 2 e 4;
- *J48*<sup>6</sup> - fator de confiança utilizado para poda da árvore (*confidence factor*) = variando de 0,30 até 0,50, com incremento de 0,05 e número mínimo de instâncias por folha (*minNumObj*) = variando de 1 até 10 (incremento 1);
- *RF* - semente utilizada para randomização dos dados (*seed*) = 30.

A Tabela 15 mostra a base de dados e o melhor parâmetro para cada algoritmo, ou seja, aquele que proporcionou ao algoritmo melhor acurácia, sendo esse o critério de seleção. Para a obtenção desses valores foram realizados testes com todos os possíveis valores

<sup>5</sup>Implementação do algoritmo *RIPPER* na ferramenta de mineração de dados *Weka*

<sup>6</sup>Implementação do algoritmo *C4.5* na ferramenta de mineração de dados *Weka*

que cada parâmetro pode assumir, para cada uma das bases. Os demais parâmetros, não apresentados nessa tabela, continuaram o padrão da ferramenta *Weka*.

Tabela 15: Configuração de parâmetros para cada algoritmo.

	<i>J48</i>	<i>RF</i>	<i>NB</i>	<i>MLP</i>	<i>SMO</i>	<i>JRip</i>
Spiral	<i>minNumObj</i> = 1	<i>seed</i> = 30	<i>useKernelEstimator</i> = True	<i>hidden layers</i> = 4	<i>Kernel</i> = Puk	<i>folds</i> = 2
R15	<i>confidence factor</i> = 0,30	<i>seed</i> = 30	<i>useKernelEstimator</i> = True	<i>hidden layers</i> = 4	<i>Kernel</i> = Puk	<i>folds</i> = 2
D13	<i>minNumObj</i> = 7	<i>seed</i> = 30	<i>useKernelEstimator</i> = True	<i>training time</i> = 1000	<i>Kernel</i> = Puk	<i>folds</i> = 2

Essas configurações de parâmetros foram utilizadas tanto no nível 0 quanto no nível 1 do empilhamento.

Após obter-se as predições dos classificadores base, foram calculados os valores das medidas de diversidade citadas na Seção 2.3, usando uma aplicação desenvolvida em *Python*.

### 5.3 Diversidade e Resultados do Empilhamento

Os resultados experimentais são resumidos nas Tabelas 16, 17 e 18 que mostram para cada base de dados as seguintes informações: os valores calculados para as medidas de diversidade falha dupla ( $df$ ), medida de discordância ( $Dis$ ), estatística  $Q$ , coeficiente de correlação ( $\rho$ ), concordância entre avaliadores ( $k$ ), variância Kohavi-Wolpert ( $KW$ ) e entropia ( $E$ ); o algoritmo utilizado para aprender o melhor classificador base ( $L_0$ ) e sua acurácia ( $A_{L_0}$ ) em porcentagem; o algoritmo utilizado para aprender o melhor meta-classificador ( $L_1$ ) e sua acurácia ( $A_{L_1}$ ) em porcentagem; e o ganho do empilhamento ( $G$ ), usado para classificar os resultados, também em porcentagem. Os valores de  $df$ ,  $Dis$ ,  $Q$  e  $\rho$  são médias dos valores calculados para cada par de classificadores base. Já os valores de ( $k$ ), ( $KW$ ) e ( $E$ ) representam a diversidade entre um conjunto de  $N$  classificadores.

Além disso, estas tabelas apresentam  $Q'$ ,  $\rho'$ ,  $k'$  e  $KW'$  que são as medidas originais de diversidade padronizadas em uma distribuição de valores no intervalo fechado entre  $[0, 1]$ , assim como  $df$ ,  $Dis$  e  $E$ .

A Tabela 16 apresenta os resultados sobre as bases de dados que atingiram os piores e melhores valores de  $G$ . Já na Tabela 17 são mostrados os resultados quando o ganho de empilhamento varia entre -1 e 1%. Essas duas tabelas são referentes as bases do *UCI*. Já a Tabela 18 mostra os resultados das bases sintéticas, com as duas categorias de experimentos realizadas.

Observando a Tabela 16, nota-se que o empilhamento funcionou bem apenas para 8 dos 54 conjuntos de dados, onde o ganho variou de 1,2 a 5,1% (linhas 1-8). O melhor ganho de empilhamento foi alcançado com o conjunto de dados Balance Scale, em um resultado preciso (90,7%) que é difícil de melhorar. O algoritmo mais frequente que obteve a melhor acurácia para o nível 0 foi o *MLP* variando  $26,6 \leq A_{L_0} \leq 90,7$ , seguido por *RF* com  $84,8 \leq A_{L_0} \leq 92,9$ . No nível 1, os melhores meta classificadores foram treinados com *SMO* ( $26,9 \leq A_{L_1} \leq 94,9$ ) e *RF* ( $83,7 \leq A_{L_1} \leq 95,4$ ).

No entanto, o empilhamento diminuiu a qualidade de classificação para alguns conjuntos de dados (linhas 9-17) atingindo no pior caso  $G = -9,4\%$ . Os algoritmos mais frequentes com pior acurácia foram *RF* ( $L_0$ ) e *SMO* ( $L_1$ ). Para alguns conjuntos de dados, mais de um classificador usado no nível 1 retornou o mesmo resultado. Por exemplo, *SMO* e *MLP* atingem valores iguais ( $A_{L_1} = 78,8\%$ ) para o conjunto de dados Leaf (linha 9).

Tabela 16: Medidas de diversidade e os melhores e piores resultados do empilhamento para as bases do UCI.

Base de Dados	$df \downarrow$	$Dis \uparrow$	$Q \downarrow$	$Q' \downarrow$	$\rho \downarrow$	$\rho' \downarrow$	$k \downarrow$	$k' \downarrow$	$KW \uparrow$	$KW' \uparrow$	$E \uparrow$	$L_0$	$A_{L_0}$	$L_1$	$A_{L_1}$	G
1 Balance Scale	0,78	0,13	0,87	0,93	0,50	0,75	0,48	0,74	0,06	0,11	0,17	MLP	90,7	RF	95,4	5,1
2 Connect. Bench (S,M vs. R)	<b>0,62</b>	<b>0,27</b>	<b>0,58</b>	<b>0,79</b>	<b>0,28</b>	<b>0,64</b>	<b>0,26</b>	<b>0,63</b>	<b>0,11</b>	<b>0,23</b>	<b>0,35</b>	MLP	82,2	JRip	85,6	4,1
3 Statlog (Vehicle Silh.)	<b>0,55</b>	<b>0,30</b>	<b>0,64</b>	<b>0,82</b>	<b>0,32</b>	<b>0,66</b>	<b>0,28</b>	<b>0,64</b>	<b>0,13</b>	<b>0,25</b>	<b>0,37</b>	MLP	81,7	RF	83,7	2,5
4 Diabetic Retinopat. Debrec.	<b>0,49</b>	<b>0,32</b>	<b>0,56</b>	<b>0,78</b>	<b>0,29</b>	<b>0,65</b>	<b>0,29</b>	<b>0,64</b>	<b>0,13</b>	<b>0,27</b>	<b>0,41</b>	MLP	72,0	SMO	73,8	2,4
5 Ionosphere	0,83	0,12	0,83	0,92	0,41	0,70	0,38	0,69	0,05	0,10	0,13	RF	92,9	SMO	94,9	2,2
6 Contrac. Method Choice	<b>0,38</b>	<b>0,29</b>	0,71	0,85	0,42	0,71	0,42	0,71	<b>0,12</b>	<b>0,24</b>	<b>0,36</b>	MLP	54,2	SMO	55,3	2,0
7 Vertebral Column	0,72	0,19	0,74	0,87	0,39	0,69	0,38	0,69	0,08	0,16	0,23	RF	84,8	RF	86,1	1,5
8 Abalone	<b>0,09</b>	<b>0,28</b>	<b>0,47</b>	<b>0,74</b>	<b>0,21</b>	<b>0,60</b>	<b>0,21</b>	<b>0,60</b>	<b>0,12</b>	<b>0,24</b>	<b>0,36</b>	MLP	26,6	SMO	26,9	1,2
9 Leaf	<b>0,52</b>	<b>0,30</b>	0,70	0,85	0,37	0,68	0,33	0,67	<b>0,12</b>	<b>0,25</b>	<b>0,37</b>	MLP	79,7	SMO*	78,8	-1,1
10 Glass Identification	<b>0,49</b>	<b>0,32</b>	<b>0,61</b>	<b>0,80</b>	<b>0,33</b>	<b>0,66</b>	<b>0,30</b>	0,65	<b>0,13</b>	<b>0,27</b>	<b>0,40</b>	RF	79,9	RF	79,0	-1,2
11 Credit Approval	0,77	0,14	0,85	0,93	0,50	0,75	0,48	0,74	0,06	0,12	0,17	RF	86,7	NB	85,7	-1,2
12 Ecoli	0,79	0,11	0,92	0,96	0,57	0,79	0,57	0,78	0,05	0,09	0,14	RF	87,2	RF	86,0	-1,4
13 Solar Flare	<b>0,62</b>	0,15	0,91	0,96	0,64	0,82	0,64	0,82	0,06	0,13	0,18	J48	72,1	SMO	70,9	-1,7
14 Dresses Attribute Sales	<b>0,46</b>	<b>0,26</b>	0,77	0,88	0,47	0,73	0,47	0,73	<b>0,11</b>	<b>0,22</b>	<b>0,32</b>	JRip	63,0	NB	60,2	-4,4
15 Audiology (Std.)	0,72	0,13	0,94	0,97	0,64	0,82	0,63	0,81	0,05	0,10	0,16	MLP	83,2	SMO	79,2	-4,8
16 Low Resolution Spectrometer	<b>0,18</b>	<b>0,36</b>	<b>0,47</b>	<b>0,73</b>	<b>0,25</b>	<b>0,62</b>	<b>0,23</b>	<b>0,61</b>	<b>0,15</b>	<b>0,30</b>	<b>0,46</b>	RF	54,0	SMO	51,0	-5,6
17 Primary Tumor	<b>0,33</b>	0,19	0,89	0,95	0,61	0,80	0,60	0,80	0,08	0,16	<b>0,25</b>	NB	50,1	RF	45,4	-9,4
Média (todas as 54 bases de dados)	0,74	0,15	0,76	0,88	0,41	0,71	0,39	0,69	0,06	0,13	0,19					

\* MLP atingem resultados iguais

Considera-se bons valores de diversidade aqueles que foram maiores ou menores do que 0,5 desvio padrão em relação à média para todos os 54 conjuntos de dados, levando em consideração se a medida é diretamente ou inversamente proporcional à diversidade entre os classificadores. Estes valores estão em negrito. Uma análise geral deles indica que há mais diversidade nos experimentos em que houve ganho no empilhamento (linhas 1-8) do que naqueles em que houve perda de qualidade (linhas 9-17).

O conjunto de dados Abalone (linha 8) teve o melhor valor de falha dupla  $df$  devido à baixa precisão apresentada pelos classificadores base ( $A_{L_0} = 26,6\%$ ). Muitos deles falham juntos porque se trata de um problema de multi-classificação envolvendo 28 rótulos de classe distintas. Para os conjuntos de dados Connect. Bench (S,M vs. R) (linha 2), Statlog (Vehicle Silh.) (linha 3) e Diabetic Retinopat. Debrec. (linha 4) os valores da medida  $df$  foram menos significativos, comparados ao Abalone, mas bons em relação a média. Esses valores de  $df$  estão relacionados a melhor acurácia apresentada pelos classificadores base, e ao menor número de classes para esses conjuntos. Observamos que para estes conjuntos de dados, todas as medidas de diversidade retornam bons valores, colaborando com a questão de pesquisa de que quanto maior a diversidade, maior a qualidade do empilhamento.

No entanto, a experiência envolvendo o conjunto de dados Low Resolution Spectrometer (linha 16) revelou o comportamento oposto onde o ganho do empilhamento foi negativo ( $G = -5,6\%$ ), ou seja, a qualidade da classificação diminuiu consideravelmente, mesmo com valores bons para todas as medidas de diversidade. Estes valores são retornados porque existem 531 instâncias distribuídas em 48 classes, tornando ainda mais difícil o acordo de muitos classificadores. O conjunto de dados Balance Scale (linha 1) é outro contraexemplo em que não havia diversidade entre os classificadores, entretanto o empilhamento atingiu o melhor  $G$  entre todas as experiências realizadas.

A Tabela 17 mostra que para 37 conjuntos de dados os resultados do ganho de empilhamento varia entre -1 e 1%. Os valores positivos do  $G$  variam de 0,1 a 0,8 % (linhas 1-26), e os valores negativos de -0,8 a -0,2 % (linhas 34-37). Além disso, 7 bases (linhas 27-33) apresentaram valor nulo para o  $G$ .

Uma análise geral das medidas de diversidade mostra que apenas 4 desses 37 conjuntos de dados apresentaram bons valores para todas as medidas de diversidade originais. Para as demais bases a diversidade foi considerada inexistente ou baixa.

Analisando esses 4 conjuntos de dados, notamos que para Connect. Bench (VR-DD) (linha 16) e Wine Quality (linha 26) o valor do  $G$  foi positivo, sendo 0,2 % e 0,1 %, respectivamente. Já as outras 2 bases apresentaram valor negativo para o  $G$ , sendo -0,2 % para ILPD - Indian Liver Patient (linha 34) e -0,8 % para Hill-Valley (linha 37). Esses valores vão contra a questão de pesquisa, pois apesar da alta diversidade o  $G$  foi próximo de nulo e negativo. Em corroboração a isso, percebe-se que nas bases que apresentam os melhores valores para o  $G$  (linhas 1-6) não existe diversidade.

A Tabela 18 mostra os resultados sobre as bases sintéticas usando a configuração padrão do *Weka* (linhas 1-3) e a configuração manual de parâmetros para alcançar a melhor acurácia (linhas 4-6).

Observando primeiramente os resultados obtidos com a configuração padrão, verifica-se que as bases R15 (linha 1) e Spiral (linha 2) apresentam valores positivos para o  $G$ , sendo 0,7 % e 0,6 %, respectivamente. R15 obteve o melhor ganho do empilhamento sob um resultado já muito preciso (93,8%), atingido pelos classificadores base *J48* e *NB*. Os experimentos realizados sobre a base D13 (linha 3) apresentaram resultado negativo para o ganho ( $G = -0,8$  %). Essa perda de qualidade no empilhamento pode estar associada a grande presença de ruídos na referida base. A Figura 5 destaca os erros de classificação associados ao meta-classificador com o símbolo  $\square$ .

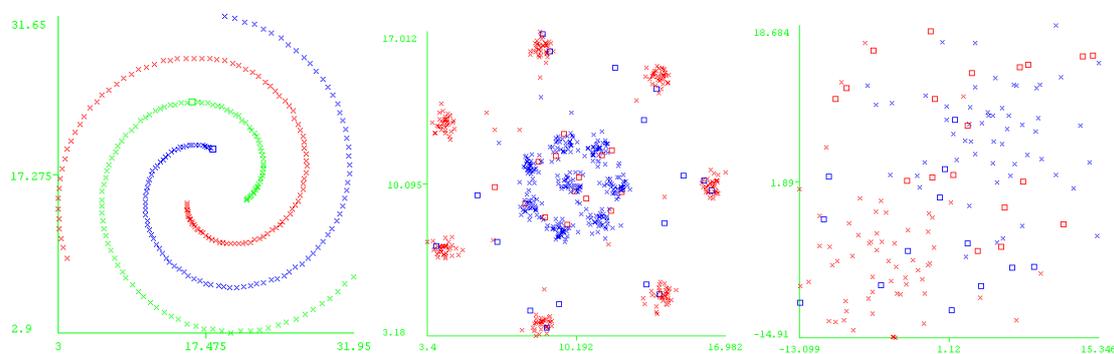


Figura 5: Erros de classificação do empilhamento usando configuração padrão dos algoritmos

No entanto, configurando manualmente os algoritmos, com os melhores parâmetros, verificou-se perda no ganho do empilhamento para as três bases de dados, com valores negativos para  $G$  variando entre -3,4 % e -0,3 %. Com isso, pode-se inferir que ao realizar essas configurações a técnica de empilhamento não é considerada a mais apropriada, sendo que o uso da mesma é indicada quando não sabe-se qual o algoritmo mais adequado a ser usado, e também para evitar esses procedimentos repetitivos de configurações de parâmetros.

Em relação a diversidade, utilizou-se os mesmos limiares definidos na Tabela 16. A base Spiral apresentou alta diversidade nas duas configurações. No entanto, usando a configuração padrão, o melhor valor de  $G$  foi alcançado por R15, em que apenas duas das sete medidas de diversidade retornaram valores satisfatórios. Esse resultado vai contra a questão de pesquisa pois baixa diversidade foi associada ao melhor empilhamento.

Tabela 17: Medidas de diversidade e o resultado do empilhamento variando entre -1 e 1 para as bases do *UCI*.

Base de Dados	$df \downarrow$	$Dis \uparrow$	$Q \downarrow$	$Q' \downarrow$	$\rho \downarrow$	$\rho' \downarrow$	$k \downarrow$	$k' \downarrow$	$KW \uparrow$	$KW' \uparrow$	$E \uparrow$	$L_0$	$A_{L_0}$	$L_1$	$A_{L_1}$	G
1 Molecular Bio. (Splice)	0,92	0,05	0,94	0,97	0,46	0,73	0,45	0,72	0,02	0,04	0,06	MLP	96,0	SMO	96,7	0,8
2 Congressional Voting Rec.	0,93	0,04	0,98	0,99	0,63	0,81	0,57	0,78	0,02	0,03	0,05	J48*	96,3	RF	97,0	0,7
3 Forest type mapping	0,78	0,11	0,92	0,96	0,59	0,80	0,59	0,79	0,05	0,09	0,13	SMO	86,5	RF	87,1	0,7
4 Dermatology	0,91	0,07	0,92	0,96	0,44	0,72	0,37	0,69	0,03	0,06	0,08	NB	97,3	SMO	97,8	0,6
5 Mammographic Mass	0,76	0,11	0,93	0,97	0,63	0,81	0,62	0,81	0,05	0,09	0,14	JRip	83,1	NB	83,7	0,6
6 Soybean (Large)	0,90	0,06	0,95	0,98	0,56	0,78	0,56	0,78	0,02	0,05	0,07	SMO	93,9	RF	94,4	0,6
7 Opt. Recog. of Handwrit. Dig.	0,91	0,08	0,92	0,96	<b>0,36</b>	<b>0,68</b>	<b>0,26</b>	<b>0,63</b>	0,03	0,06	0,09	SMO	98,3	RF	98,9	0,5
8 Blood Transf. Serv. Center	0,70	0,14	0,92	0,96	0,63	0,81	0,62	0,81	0,06	0,11	0,18	JRip	78,5	JRip	78,9	0,5
9 Tic-Tac-Toe Endgame	0,83	0,15	0,78	0,89	<b>0,30</b>	<b>0,65</b>	0,92	0,96	0,06	0,12	0,17	SMO	98,3	RF	98,9	0,5
10 QSAR biodegradation	0,75	0,16	0,81	0,90	0,45	0,72	0,43	0,71	0,07	0,13	0,20	RF	86,9	NB	87,3	0,4
11 Pen-Based Recog. Handwrit.	0,91	0,07	0,88	0,94	<b>0,27</b>	<b>0,63</b>	<b>0,21</b>	<b>0,60</b>	0,03	0,06	0,09	RF	99,1	RF	99,4	0,3
12 Wilt	0,93	0,06	0,93	0,96	0,39	0,70	<b>0,27</b>	<b>0,63</b>	0,02	0,05	0,06	RF	98,4	SMO	98,6	0,3
13 Yeast	<b>0,48</b>	<b>0,21</b>	0,86	0,93	0,56	0,78	0,56	0,78	0,09	<b>0,18</b>	<b>0,26</b>	RF	61,9	SMO	62,1	0,3
14 Multiple Features	0,89	0,08	0,92	0,96	0,40	0,70	<b>0,31</b>	0,66	0,04	0,07	0,10	SMO	97,6	RF	97,7	0,2
15 Nursery	0,93	0,06	0,88	0,94	<b>0,23</b>	<b>0,62</b>	<b>0,21</b>	<b>0,61</b>	0,03	0,05	0,07	MLP	99,7	MLP	100,0	0,2
16 Connect. Bench (VR-DD)	<b>0,65</b>	<b>0,28</b>	<b>0,55</b>	<b>0,78</b>	<b>0,17</b>	<b>0,58</b>	<b>0,12</b>	<b>0,56</b>	<b>0,12</b>	<b>0,24</b>	<b>0,35</b>	RF	98,3	RF	98,5	0,2
17 Thoracic Surgery	0,78	0,08	0,96	0,98	0,73	0,86	0,71	0,85	0,03	0,07	0,09	SMO	84,9	JRip**	85,1	0,2
18 Breast Cancer Wisconsin	0,94	0,03	0,98	0,99	0,61	0,80	0,60	0,80	0,01	0,03	0,04	RF	96,9	SMO	97,0	0,1
19 Banknote Authent.	0,93	0,06	<b>0,24</b>	<b>0,62</b>	<b>0,14</b>	<b>0,57</b>	<b>0,08</b>	<b>0,54</b>	0,03	0,05	0,07	MLP	99,9	SMO*	100,0	0,1
20 Car Evaluation	0,87	0,11	0,83	0,92	<b>0,28</b>	<b>0,64</b>	<b>0,25</b>	<b>0,63</b>	0,05	0,09	0,13	MLP	99,5	SMO	99,7	0,1
21 Chess (KR vs. K-P)	0,94	0,06	0,91	0,95	<b>0,28</b>	<b>0,64</b>	<b>0,12</b>	<b>0,56</b>	0,02	0,05	0,06	J48	99,4	RF	99,6	0,1
22 Phishing Websites	0,93	0,04	0,96	0,98	0,52	0,76	0,50	0,75	0,02	0,04	0,06	RF	97,3	SMO	97,3	0,1
23 Spambase	0,84	0,13	0,77	0,88	<b>0,34</b>	0,67	<b>0,28</b>	<b>0,64</b>	0,05	0,11	0,15	RF	95,5	SMO	95,6	0,1
24 Turkiye Student Eval.	0,70	<b>0,26</b>	0,72	0,86	<b>0,29</b>	<b>0,64</b>	<b>0,06</b>	<b>0,53</b>	<b>0,11</b>	<b>0,22</b>	<b>0,31</b>	J48	98,1	RF	98,2	0,1
25 Thyroid Disease (Hypothy.)	0,95	0,04	0,94	0,97	<b>0,34</b>	0,67	<b>0,30</b>	0,65	0,02	0,04	0,06	J48	99,6	RF	99,6	0,1
26 Wine Quality	<b>0,39</b>	<b>0,34</b>	<b>0,57</b>	<b>0,79</b>	<b>0,33</b>	<b>0,66</b>	<b>0,31</b>	0,66	<b>0,14</b>	<b>0,28</b>	<b>0,43</b>	RF	70,3	SMO	70,4	0,1
27 Annealing	0,94	0,06	0,91	0,96	<b>0,32</b>	<b>0,66</b>	<b>0,15</b>	<b>0,57</b>	0,02	0,05	0,06	RF	99,6	JRip****	99,6	0,0
28 Chronic Kidney Disease	0,97	0,03	<b>-0,29</b>	<b>0,36</b>	<b>0,04</b>	<b>0,52</b>	<b>0,05</b>	<b>0,52</b>	0,01	0,03	0,04	RF	100,0	NB*****	100,0	0,0
29 Qualitative Bankruptcy	0,99	0,01	<b>0,20</b>	<b>0,60</b>	0,37	0,68	<b>0,30</b>	0,65	0,01	0,01	0,01	SMO*	99,6	NB*****	99,6	0,0
30 Seismic-bumps	0,90	0,05	0,96	0,98	0,74	0,87	0,69	0,84	0,02	0,04	0,05	SMO	93,4	SMO	93,4	0,0
31 Thyroid Disease (Sick)	0,94	0,05	0,91	0,96	<b>0,32</b>	<b>0,66</b>	<b>0,25</b>	<b>0,62</b>	0,02	0,04	0,06	J48	98,8	J48*	98,8	0,0
32 Waveform Gen. (V 2)	0,73	0,18	0,81	0,90	0,43	0,72	0,41	0,71	0,07	0,15	0,21	SMO	86,7	SMO	86,7	0,0
33 Page Blocks Classification	0,93	0,05	0,95	0,98	0,49	0,74	0,42	0,71	0,02	0,04	0,06	RF	97,5	RF	97,6	-0,1
34 ILPD - Indian Liver Patient	<b>0,51</b>	<b>0,33</b>	<b>0,42</b>	<b>0,71</b>	<b>0,28</b>	<b>0,64</b>	<b>0,26</b>	<b>0,63</b>	<b>0,14</b>	<b>0,27</b>	<b>0,41</b>	SMO	71,4	SMO	71,2	-0,2
35 SPECT Heart	0,84	0,10	0,90	0,95	0,53	0,76	0,51	0,75	0,04	0,08	0,12	J48*	92,0	SMO	91,4	-0,6
36 Wholesale customers	0,87	0,08	0,93	0,96	0,57	0,78	0,55	0,78	0,03	0,06	0,10	RF	92,7	NB***	92,0	-0,7
37 Hill-Valley	<b>0,31</b>	<b>0,47</b>	<b>0,08</b>	<b>0,54</b>	<b>0,06</b>	<b>0,53</b>	<b>0,06</b>	<b>0,53</b>	<b>0,19</b>	<b>0,39</b>	<b>0,62</b>	SMO	61,1	MLP	60,6	-0,8
Média (todas as 54 bases)	0,74	0,15	0,76	0,88	0,41	0,71	0,39	0,69	0,06	0,13	0,19					

\* RF atingem resultados iguais

\*\* SMO e J48 atingem resultados iguais

\*\*\* J48 atingem resultados iguais

\*\*\*\* J48 e RF atingem resultados iguais

\*\*\*\*\* SMO e MLP atingem resultados iguais

Tabela 18: Medidas de diversidade e o resultado do empilhamento para as bases de dados sintéticas.

	Base de Dados	Parâmetros	$df \downarrow$	$Dis \uparrow$	$Q \downarrow$	$Q' \downarrow$	$\rho \downarrow$	$\rho' \downarrow$	$k \downarrow$	$k' \downarrow$	$KW \uparrow$	$KW' \uparrow$	$E \uparrow$	$L_0$	$A_{L_0}$	$L_1$	$A_{L_1}$	G
1	R15		0,76	0,19	<b>0,66</b>	<b>0,83</b>	0,48	0,74	<b>0,23</b>	<b>0,61</b>	0,08	0,16	0,21	NB*	93,8	SMO	94,5	0,7
2	Spiral	Padrão	<b>0,44</b>	<b>0,43</b>	<b>0,52</b>	<b>0,76</b>	<b>0,21</b>	<b>0,60</b>	<b>0,05</b>	<b>0,53</b>	<b>0,18</b>	<b>0,36</b>	<b>0,63</b>	RF	98,7	MLP	99,3	0,6
3	D13		0,71	0,12	0,94	0,97	0,67	0,84	0,67	0,83	0,05	0,10	0,15	JRip	79,2	MLP	78,5	-0,8
4	Spiral'	Configuração manual	0,75	<b>0,23</b>	<b>0,47</b>	<b>0,73</b>	<b>0,11</b>	<b>0,56</b>	<b>0,02</b>	<b>0,51</b>	<b>0,10</b>	<b>0,19</b>	<b>0,26</b>	RF	99,4	SMO**	99	-0,3
5	R15'		0,93	0,02	1	1	0,87	0,94	0,87	0,94	0,01	0,01	0,02	J48	94	J48***	93,7	-0,4
6	D13'		0,72	0,10	0,95	0,98	0,71	0,85	0,70	0,85	0,04	0,09	0,13	JRip*	79,2	SMO	76,5	-3,4

\* J48 atingem resultados iguais

\*\* MLP atingem resultados iguais

\*\*\*NB e JRip atingem resultados iguais

### 5.3.1 Teste Estatístico

O teste-t (*Student's t-test*) avalia se as médias de duas distribuições normais de valores são estatisticamente diferentes, apresentando bons resultados mesmo quando as distribuições não são perfeitamente normais (STUDENT, 1908).

Devido ao fato da maioria dos resultados do ganho ser muito próximo de zero foi aplicado o teste-t para verificar se o ganho do empilhamento é de fato estatisticamente significativo. Para isso, foram comparadas a acurácia do melhor meta classificador ( $A_{L_1}$ ) com a acurácia do melhor classificador base ( $A_{L_0}$ ), para as 54 bases do *UCI* e os três conjuntos de dados sintéticos. Para esses últimos foram verificados os valores obtidos pelos algoritmos usando a configuração padrão do *weka* e também a parametrização manual. Nesse contexto, o nível de confiança ( $\alpha$ ) utilizado foi 0,05, indicando diferença significativa quando o resultado for inferior a esse. Caso contrário, não há diferença estatística.

A Tabela 19 apresenta o resultado da aplicação do teste-t para as bases do *UCI*. Nela é mostrado o conjunto de dados, o valor do teste, e o ganho do empilhamento  $G$ , em porcentagem. Apenas com 18 dos 54 conjuntos de dados analisados apresentou-se um valor estatisticamente significativo para  $G$ . Para os outros conjuntos de dados o ganho de empilhamento foi considerado nulo, ou seja, um empate técnico entre os níveis 0 e 1.

Observando a Tabela 19 percebe-se que a maioria dos conjuntos de dados apresentam um valor significativo positivo para o ganho (linhas 1-13), e as demais um valor negativo (linhas 14-18).

Tabela 19: Teste estatístico aplicado ao ganho do empilhamento para as bases do *UCI*

#	Conjunto de Dados	teste-t	$G$
1	<b>Connect. Bench (S,M vs. R)</b>	0,0078	4,1
2	Ionosphere	0,0080	2,2
3	Vertebral Column	0,0453	1,5
4	<b>Abalone</b>	0,0002	1,2
5	Mammographic Mass	0,0253	0,6
6	Soybean	0,0454	0,6
7	Tic-Tac-Toe Endgame	0,0253	0,5
8	QSAR biodegradation	0,0454	0,4
9	Wilt	0,0005	0,3
10	Spambase	0,0143	0,1
11	Chess (KR vs. K-P)	0,0455	0,1
12	Turkiye Student Eval.	0,0253	0,1
13	Phishing Websites	0,0027	0,1
14	Credit Approval	0,0081	-1,2
15	Ecoli	0,0453	-1,4
16	Solar Flare	0,0453	-1,7
17	Dresses Attribute Sales	0,0002	-4,4
18	Audiology	0,0025	-4,8

Comparando os conjuntos de dados da Tabela 19, com aqueles que apresentaram os

melhores e piores valores para  $G$  (Tabela 16), verificamos que dos conjuntos onde  $G$  é estatisticamente significativo com valores maiores que 1 (linhas 1-4), dois desses (Connect. Bench (S,M vs. R) e Abalone) possuem bons valores para todas as medidas de diversidade, ou seja, são considerados diversos. Esses conjuntos de dados estão representados em negrito na Tabela 19.

No entanto, nas demais bases com valor significativo positivo para o ganho, a diversidade é baixa. Para os conjuntos de dados Credit Approval, Ecoli, e Audiology, onde  $G$  é estatisticamente significativo com valores menores que -1 (linhas 14-18), não existe diversidade. Além disso, para os conjuntos de dados Solar Flare e Dresses Attribute Sales a diversidade é baixa.

A Tabela 20 mostra os dois conjuntos sintéticos que apresentaram valor significativo para o ganho do empilhamento. O conjunto R15 (linha 1), usando a configuração padrão para os algoritmos, obteve um valor significativo positivo para  $G$ . Já o conjunto D13' (linha 2), com os algoritmos parametrizados, apresentou um valor significativo negativo.

Tabela 20: Resultado do teste-t para as bases sintéticas

#	Conjunto de Dados	teste-t	$G$
1	R15	0,0054	0,7
2	D13'	0,0451	-3,4

Conforme observado na Tabela 18, o conjunto de dados Spiral apresentou alta diversidade usando configuração padrão e parametrizada. No entanto, de acordo com o teste-t o mesmo não apresentou valores estatisticamente significativos para o ganho do empilhamento. Esse fato pode ser justificado devido a alta acurácia apresentada pelo classificador base *Random Forest*, em ambos os casos, sendo esse um algoritmo eficiente.

Concluindo a análise, percebe-se que na maioria dos experimentos parece não haver uma relação significativa entre a diversidade dos classificadores e o ganho do empilhamento.

### 5.3.2 Interpretação gráfica da relação entre diversidade e qualidade do empilhamento

Para um melhor entendimento dos efeitos da diversidade nos experimentos realizados para cada conjunto de dados onde o ganho do empilhamento passou no teste-t, ou seja, onde a acurácia do empilhamento foi estatisticamente diferente do melhor resultado do classificador base, é mostrado graficamente a relação entre as medidas de diversidade e o ganho do empilhamento.

São considerados bons valores de diversidade aqueles que foram suficientemente maiores ou menores do que a média para todos os 54 conjuntos de dados. Nesse caso, foram consideradas apenas as medidas de diversidade originais e cada gráfico apresentado aborda uma dessas medidas.

Os valores de  $G$  são apresentados no eixo das abcissas, em porcentagem, e os valores da medida de diversidade no eixo das ordenadas. Além disso, os pontos destacados em cada gráfico (verde/vermelho) representam os conjuntos de dados que apoiam a questão de pesquisa, e os demais pontos (azul) representam os conjuntos de dados que rejeitam a mesma. Em relação a esses pontos, estão representados na cor verde aqueles que obtiveram alta diversidade e alto ganho (+), em vermelho os que apresentaram baixa diversidade e baixo ganho (-), e em azul os pontos com alta diversidade e baixo ganho ou baixa diversidade e alto ganho.

Analisando estas figuras nota-se que apenas alguns pontos apresentaram bons valores de diversidade. Muitos deles estão localizados perto do centro do eixo das abcissas, bem como a grande maioria dos valores de baixa diversidade.

As Figuras 6, 7 e 8 representam, respectivamente, a relação entre  $Dis$ ,  $KW$ ,  $E$  e o ganho do empilhamento. Observando os conjuntos de dados que apoiam a questão de pesquisa, percebe-se que os mesmos conjuntos destacaram-se para as três medidas de diversidade, sendo essas diretamente proporcionais a diversidade dos classificadores.

Os conjuntos de dados que apresentaram bons valores de diversidade e alto ganho são: Connect. Bench (S,M vs. R) ( $Dis = 0,27$ ;  $KW = 0,11$ ;  $E = 0,35$  e  $G = 4,1$  %) e Abalone ( $Dis = 0,28$ ;  $KW = 0,12$ ;  $E = 0,36$  e  $G = 1,2$  %). Já os conjuntos Credit Approval ( $Dis = 0,14$ ;  $KW = 0,06$ ;  $E = 0,17$  e  $G = -1,2$  %), Ecoli ( $Dis = 0,11$ ;  $KW = 0,05$ ;  $E = 0,14$  e  $G = -1,4$  %), Solar Flare ( $Dis = 0,15$ ;  $KW = 0,06$ ;  $E = 0,18$  e  $G = -1,7$  %) e Audiology ( $Dis = 0,13$ ;  $KW = 0,05$ ;  $E = 0,16$  e  $G = -4,8$  %) apresentaram baixa diversidade e baixo ganho.

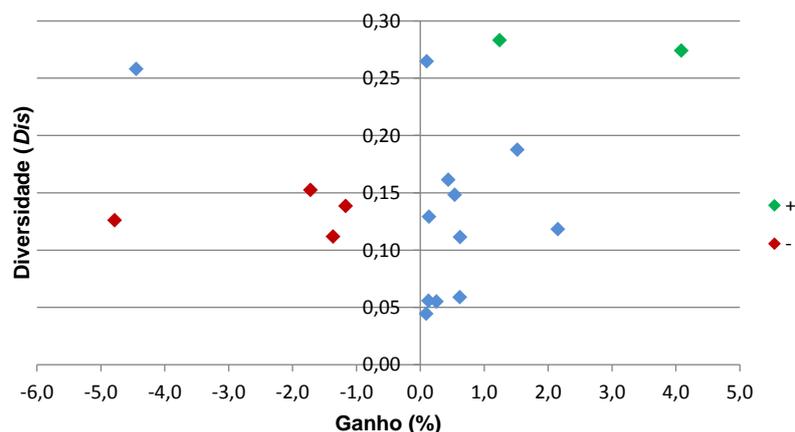


Figura 6: Relação entre medida de discordância e o ganho do empilhamento

As Figuras 9, 10 e 11 representam a relação entre  $Q$ ,  $\rho$ ,  $k$  e o ganho do empilhamento, respectivamente. Os mesmos conjuntos destacaram-se para as três medidas de diversidade, sendo essas inversamente proporcionais a diversidade dos classificadores.

Analisando essas figuras, nota-se que cinco dos seis os conjuntos de dados anteriormente citados continuam destacados apresentando boa relação entre diversidade e ganho

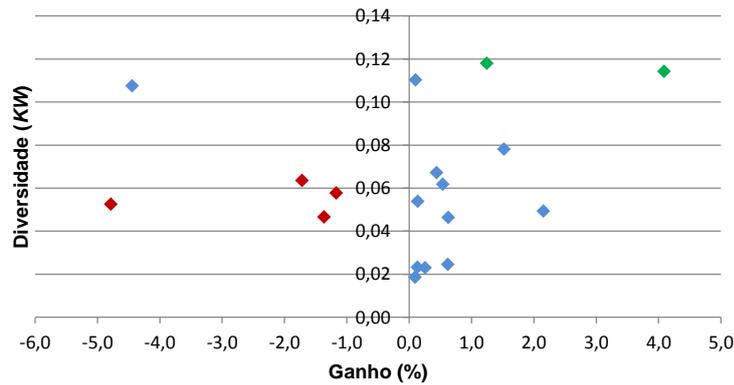


Figura 7: Relação entre Variância *Kohavi-Wolpert* e ganho do empilhamento

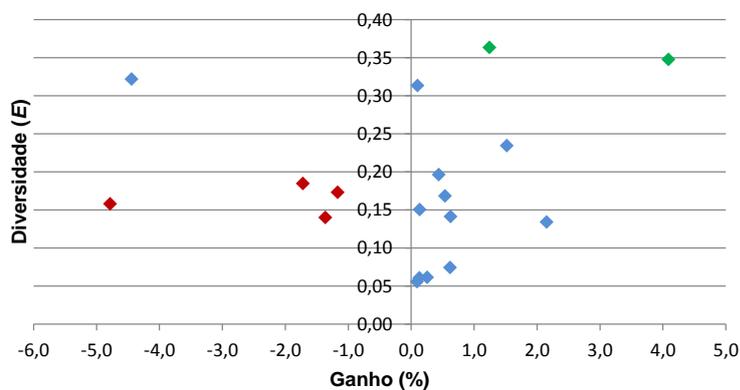


Figura 8: Relação entre entropia e ganho do empilhamento

do empilhamento: Abalone ( $Q = 0,47$ ;  $\rho = 0,21$ ;  $k = 0,21$  e  $G = 1,2$  %), Connect. Bench (S,M vs. R) ( $Q = 0,58$ ;  $\rho = 0,28$ ;  $k = 0,26$  e  $G = 4,1$  %), Credit Approval ( $Q = 0,85$ ;  $\rho = 0,50$ ;  $k = 0,48$  e  $G = -1,2$  %), Ecoli ( $Q = 0,92$ ;  $\rho = 0,57$ ;  $k = 0,57$  e  $G = -1,4$  %), Solar Flare ( $Q = 0,91$ ;  $\rho = 0,64$ ;  $k = 0,64$  e  $G = -1,7$  %) e Audiology ( $Q = 0,94$ ;  $\rho = 0,64$ ;  $k = 0,63$  e  $G = -4,8$  %). Adicionalmente, os valores calculados para essas medidas de diversidade também apoiaram a questão de pesquisa no conjunto de dados Dresses Attribute Sales ( $Q = 0,77$ ;  $\rho = 0,47$ ;  $k = 0,47$  e  $G = -4,4$  %).

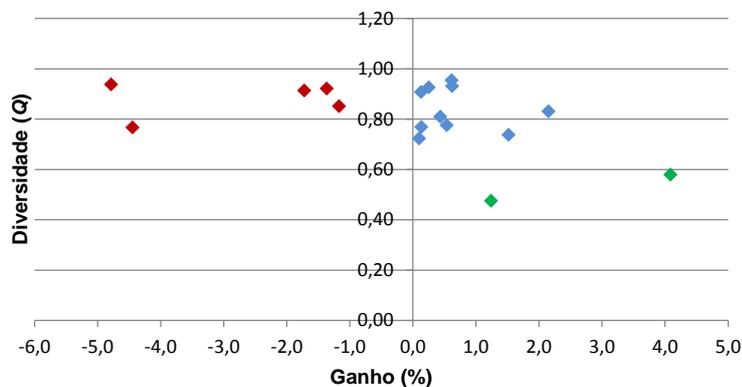


Figura 9: Relação entre estatística  $Q$  e o ganho do empilhamento

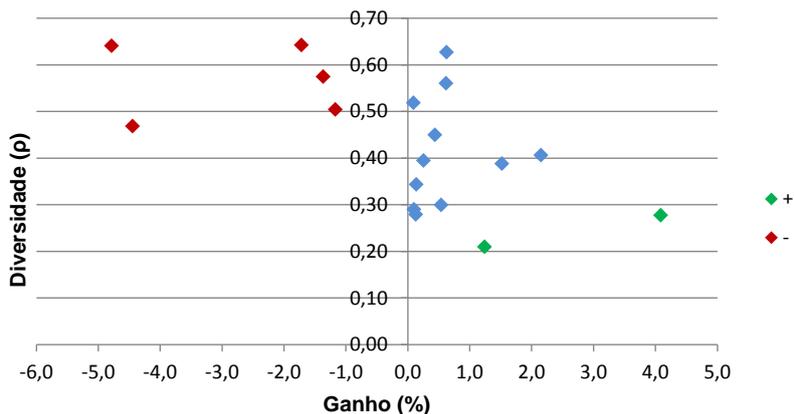


Figura 10: Relação entre coeficiente de correlação e ganho do empilhamento

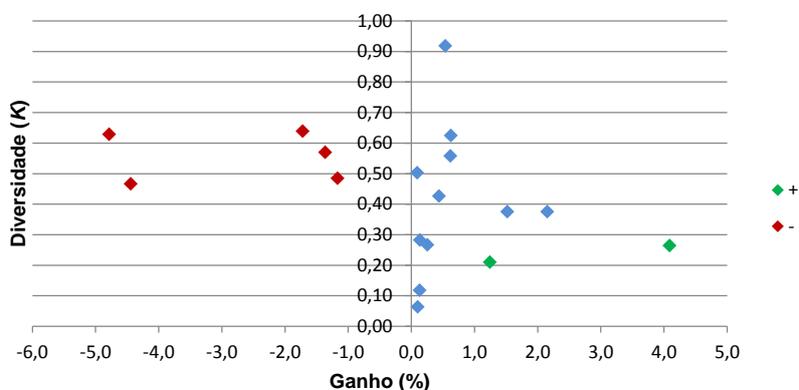


Figura 11: Relação entre concordância entre avaliadores e ganho do empilhamento

A relação entre  $df$  e o ganho do empilhamento é apresentada na Figura 12. Observando os conjuntos de dados que apoiam a questão de pesquisa, nota-se que destacaram-se apenas Abalone ( $df = 0,09$  e  $G = 1,2$  %), Connect. Bench (S,M vs. R) ( $df = 0,62$  e  $G = 4,1$  %), Credit Approval ( $df = 0,77$  e  $G = -1,2$  %), Ecoli ( $df = 0,79$  e  $G = -1,4$  %) e Audiology ( $df = 0,72$  e  $G = -4,8$  %). Os resultados dos demais conjuntos não mencionados rejeitam a questão de pesquisa.

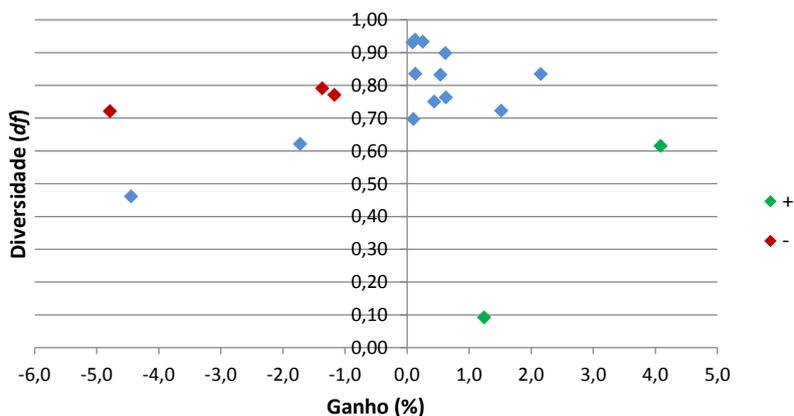


Figura 12: Relação entre falha dupla e o ganho do empilhamento

Observando os gráficos percebe-se que, para todas as medidas de diversidade, existe mais conjuntos que rejeitam a questão de pesquisa. Devido a isso, não consegue-se verificar quais medidas estão mais relacionadas com o ganho, o que induz a aplicação de modelos de regressão para tentar capturar essas relações implícitas e desconhecidas.

### 5.3.3 Modelos de Regressão

O impacto da diversidade no ganho do empilhamento foi analisado utilizando uma função de regressão linear e um modelo de árvore de regressão induzida pelo algoritmo *M5* (QUINLAN, 1992). Esses modelos foram treinados com os 17 conjuntos de dados presentes na Tabela 16, com todos os 54 conjuntos de dados, e também com os 18 conjuntos de dados que apresentaram significância estatística de acordo com o teste-t (Tabela 19). Além disso, foram realizadas duas categorias de treinamento para cada conjunto de dados, ou seja, usando todas as medidas de diversidade originais, ou substituindo as medidas originais ( $Q$ ,  $\rho$ ,  $k$  e  $KW$ ) pelos seus valores padronizados no intervalo fechado entre  $[0, 1]$ . Os modelos de regressão foram avaliados utilizando o coeficiente de correlação e a raiz do erro quadrático relativo (RRSE).

A equação 9 mostra o modelo linear treinado com todos os 54 conjuntos de dados. Para este modelo apenas  $df$  e  $KW$  tiveram impacto no ganho do empilhamento. Outras medidas de diversidade não foram usadas. A correlação e o RRSE foi de 0,4081 e 91,58%, respectivamente.

$$G = 0,0971 df + 0,3757 KW - 0,0957 \quad (9)$$

Para o conjunto de treinamento composto pelos 17 conjuntos de dados com os melhores e piores valores de  $G$ , o modelo de árvore apresentou melhor correlação. O número mínimo de instâncias permitidas em um nó folha no algoritmo *M5* variou de 2 a 4, porém o resultado foi a mesma árvore com apenas um nó contendo o modelo descrito pela Equação 10. Para este modelo,  $df$  continua tendo um impacto positivo sobre o ganho, mas a influência de  $\rho$  foi negativa. Outras medidas de diversidade foram irrelevantes na estimativa do ganho. A correlação com o ganho do empilhamento foi de 0,5243 e o RRSE foi de 79,67 %.

$$G = 0,1278 df - 0,2189 \rho + 0,0168 \quad (10)$$

A Equação 11 mostra o modelo linear gerado considerando apenas os 18 conjuntos de dados que passaram no teste-t. Nesse caso, a única medida que apresentou influência em relação ao ganho foi  $\rho$ , cuja correlação foi de 0,3532 e RRSE igual a 99,89 %.

$$G = -0,0847 \rho + 0,0362 \quad (11)$$

Observando as referidas equações, notamos que o uso das medidas de diversidade

padronizadas não se mostrou relevante para a geração dos modelos de regressão.

A Tabela 21 resume os melhores resultados comparando a avaliação da regressão linear e do modelo de árvore.

Tabela 21: Avaliação dos modelos de regressão.

Conjunto de Dados	Modelo	Correlação	RRSE (%)
17	M5	0.5243	79.67
54	linear	0.4081	91.58
18	linear	0.3532	99.89

Analisando a Tabela 21 percebe-se que a correlação entre as medidas de diversidade e o ganho do empilhamento foi relativamente fraca para todos os modelos induzidos. O valor do RRSE apresentou-se elevado.

Além disso, o melhor valor de correlação e RRSE foi obtido usando os 17 conjunto de dados, ou seja, aqueles que apresentaram os melhores e piores valores para o  $G$ . Esse fato pode ser justificado pela presença de mais diversidade nos experimentos em que houve ganho no empilhamento do que naqueles em que houve perda de qualidade. No entanto, o pior valor de correlação e RRSE foi apresentado justamente pelos conjuntos de dados selecionados pelo teste-t, ou seja, aqueles que possuem significância estatística em relação ao ganho do empilhamento.

## 6 CONCLUSÃO

Este trabalho teve por objetivo apresentar uma análise do impacto da diversidade de classificadores supervisionados na qualidade do empilhamento.

Os experimentos realizados mostraram baixa relação entre as medidas de diversidade estudadas e o ganho do empilhamento considerando 54 conjuntos de dados reais e 3 conjuntos de dados sintéticos.

Os modelos de regressão revelaram conexões entre algumas medidas e a qualidade do empilhamento.  $df$ ,  $KW$  e  $\rho$  estão relacionados com a acurácia da classificação final, mas valores baixos dos coeficientes de correlação e valores elevados de RRSE implicam uma relação fraca.

De acordo com os resultados encontrados na avaliação experimental, percebe-se que não existe uma relação significativa entre diversidade e ganho do empilhamento, fato esse que refuta a questão de pesquisa levantada.

Assim, como sugerido pela literatura para *bagging*, votação por maioria e outros *ensembles*, prever o ganho de acurácia em cima do melhor classificador individual usando medidas de diversidade é possivelmente uma estratégia inadequada.

Como produção científica resultante desta dissertação foram produzidos os seguintes artigos:

- (LANES; BORGES, 2016) – Uma análise do impacto da diversidade sobre o resultado do empilhamento de classificadores supervisionados. In: MOSTRA DE PRODUÇÃO UNIVERSITÁRIA DA FURG, 15., ENCONTRO DE PÓS-GRADUAÇÃO, 18., 2016, Rio Grande. **Anais...** Universidade Federal do Rio Grande, 2016. p.1–2. Este resumo estendido apresenta uma visão geral da abordagem proposta nesta dissertação de mestrado.
- (LANES et al., 2017) – Qualis B1 – An analysis of the impact of diversity on stacking supervised classifiers. In: INTERNATIONAL CONFERENCE ON ENTERPRISE INFORMATION SYSTEMS (ICEIS), 19., 2017. **Proceedings...** Springer International Publishing, 2017, to appear. p.1–8. Artigo aceito para publicação que apresenta uma análise de parte dos resultados, demonstrando a relação entre

as sete medidas de diversidade abordadas e a qualidade do empilhamento, usando conjuntos de dados reais de domínio público.

- (LANES; BORGES; GALANTE, 2017) – Qualis B1 – The effects of classifiers diversity on the accuracy of stacking. In: INTERNATIONAL CONFERENCE ON SOFTWARE ENGINEERING AND KNOWLEDGE ENGINEERING (SEKE), 29., 2017. Artigo submetido contendo os resultados finais da dissertação. A confirmação do aceite está prevista para 20 de abril de 2017.

Como trabalhos futuros, pretende-se usar a medida de avaliação *Micro-F1* e realizar experimentos com medidas de diversidade adicionais e com mais conjuntos de dados sintéticos, visando compreender melhor as relações entre distribuição de dados, fronteiras de decisão, diversidade de classificadores e qualidade de empilhamento.

## REFERÊNCIAS

AFIFI, A. A.; AZEN, S. P. **Statistical analysis: a computer oriented approach**. New York: Academic press, 2014.

AHA, D. W.; KIBLER, D.; ALBERT, M. K. Instance-based learning algorithms. **Machine Learning**, [S.l.], v.6, n.1, p.37–66, 1991.

ALI, S.; MAJID, A. Can-Evo-Ens. **Journal of Biomedical Informatics**, [S.l.], v.54, n.C, p.256–269, 2015.

BERNARDINI, F. C. **Combinação de classificadores simbólicos para melhorar o poder preditivo e descritivo de ensembles**. 2002. Tese (Doutorado em Ciência da Computação) — Universidade de São Paulo (USP). Instituto de Ciências Matemáticas e de Computação de São Carlos.

BORGES, E. N. **Um método para deduplicação de metadados bibliográficos baseado no empilhamento de classificadores**. 2013. Tese (Doutorado em Ciência da Computação) — Universidade Federal do Rio Grande do Sul (UFRGS). Instituto de Informática.

BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A training algorithm for optimal margin classifiers. In: ANNUAL WORKSHOP ON COMPUTATIONAL LEARNING THEORY, 1992. **Proceedings...** [S.l.: s.n.], 1992. p.144–152.

BREIMAN, L. Stacked regressions. **Machine Learning**, [S.l.], v.24, n.1, p.49–64, 1996.

BREIMAN, L. Random forests. **Machine learning**, [S.l.], v.45, n.1, p.5–32, 2001.

CHANG, H.; YEUNG, D.-Y. Robust path-based spectral clustering. **Pattern Recognition**, [S.l.], v.41, n.1, p.191 – 203, 2008.

CLEARY, J. G.; TRIGG, L. E. K\*: An instance-based learner using an entropic distance measure. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 1995. **Proceedings...** [S.l.: s.n.], 1995. p.108–114.

COHEN, W. W. Fast effective rule induction. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 1995. **Proceedings...** [S.l.: s.n.], 1995. p.115–123.

COST, S.; SALZBERG, S. A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features. **Machine Learning**, [S.l.], v.10, n.1, p.57–78, 1993.

COVER, T.; HART, P. Nearest neighbor pattern classification. **IEEE transactions on information theory**, [S.l.], v.13, n.1, p.21–27, 1967.

CUNNINGHAM, P.; CARNEY, J. Diversity versus Quality in Classification Ensembles Based on Feature Selection. In: EUROPEAN CONFERENCE ON MACHINE LEARNING, 2000, Berlin, Heidelberg. **Anais...** Springer Berlin Heidelberg, 2000. p.109–116.

DIETTERICH, T. G. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. **Machine Learning**, [S.l.], v.40, n.2, p.139–157, 2000.

DYMITR, R.; BOGDAN, G. Classifier selection for majority voting. **Information Fusion**, [S.l.], v.6, n.1, p.63 – 81, 2005.

DZEROSKI, S.; ZENKO, B. Is Combining Classifiers with Stacking Better than Selecting the Best One? **Machine Learning**, [S.l.], v.54, n.3, p.255–273, 2004.

EBRAHIMPOUR, R.; SADEGHNEJAD, N.; AMIRI, A.; MOSHTAGH, A. Low resolution face recognition using combination of diverse classifiers. In: INTERNATIONAL CONFERENCE OF SOFT COMPUTING AND PATTERN RECOGNITION, 2010. **Proceedings...** [S.l.: s.n.], 2010. p.265–268.

FARIA, F. A.; SANTOS, J. A. dos; ROCHA, A.; S. TORRES, R. da. A framework for selection and fusion of pattern classifiers in multimedia recognition. **Pattern Recognition Letters**, [S.l.], v.39, p.52 – 64, 2014. Advances in Pattern Recognition and Computer Vision.

GARCÍA-GUTIÉRREZ, J.; MATEOS-GARCÍA, D.; RIQUELME-SANTOS, J. EVORSTACK: A label-dependent evolutive stacking on remote sensing data fusion. **Neurocomputing**, [S.l.], v.75, n.1, p.115–122, 2012.

GARCIA-GUTIERREZ, J.; MATEOS-GARCIA, D.; RIQUELME-SANTOS, J. C. A SVM and k-NN Restricted Stacking to Improve Land Use and Land Cover Classification. In: HYBRID ARTIFICIAL INTELLIGENCE SYSTEMS: 5TH INTERNATIONAL CONFERENCE, HAIS 2010, SAN SEBASTIÁN, SPAIN, JUNE 23-25, 2010. PROCEEDINGS, PART II, 2010, Berlin, Heidelberg. **Anais...** Springer Berlin Heidelberg, 2010. p.493–500.

- GIACINTO, G.; ROLI, F. Design of effective neural network ensembles for image classification purposes. **Image and Vision Computing**, [S.l.], v.19, n.9–10, p.699–707, 2001.
- HAYKIN, S. **Neural Networks: A Comprehensive Foundation**. Upper Saddle River, USA: Prentice-Hall, Inc., 2007.
- HECHT-NIELSEN, R. Theory of the backpropagation neural network. In: INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS, 1989. **Proceedings...** [S.l.: s.n.], 1989. p.593–605.
- HO, T. K. The random subspace method for constructing decision forests. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, [S.l.], v.20, n.8, p.832–844, 1998.
- JOHN, G. H.; LANGLEY, P. Estimating continuous distributions in Bayesian classifiers. In: CONFERENCE ON UNCERTAINTY IN ARTIFICIAL INTELLIGENCE, 1995. **Proceedings...** [S.l.: s.n.], 1995. p.338–345.
- JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis**. [S.l.]: Prentice hall Englewood Cliffs, 2002. v.5, n.8.
- KITTLER, J.; ALKOOT, F. M. Sum versus vote fusion in multiple classifier systems. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, [S.l.], v.25, n.1, p.110–115, 2003.
- KOHAVI, R.; WOLPERT, D. H. et al. Bias plus variance decomposition for zero-one loss functions. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 1996. **Anais...** [S.l.: s.n.], 1996. p.275–83.
- KOTSIANTIS, S. B.; PINTELAS, P. E. A hybrid decision support tool-using ensemble of classifiers. In: INTERNATIONAL CONFERENCE ON ENTERPRISE INFORMATION SYSTEMS, 2004. **Proceedings...** [S.l.: s.n.], 2004. p.448–453.
- KUNCHEVA, L. **Fuzzy classifier design**. [S.l.]: Springer Science & Business Media, 2000. v.49.
- KUNCHEVA, L. I.; WHITAKER, C. J. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. **Machine learning**, [S.l.], v.51, n.2, p.181–207, 2003.
- LANES, M. A.; BORGES, E. N. Uma análise do impacto da diversidade sobre o resultado do empilhamento de classificadores supervisionados. In: MOSTRA DE PRODUÇÃO UNIVERSITÁRIA DA FURG, 15., ENCONTRO DE PÓS-GRADUAÇÃO, 18., 2016, Rio Grande. **Anais...** Universidade Federal do Rio Grande, 2016. p.1–2.

LANES, M. A.; BORGES, E. N.; GALANTE, R. The effects of classifiers diversity on the accuracy of stacking. In: INTERNATIONAL CONFERENCE ON SOFTWARE ENGINEERING AND KNOWLEDGE ENGINEERING, 29., 2017. **Anais...** [S.l.: s.n.], 2017.

LANES, M. A.; SCHIAVO, P.; PEREIRA JR., S.; BORGES, E. N.; GALANTE, R. An analysis of the impact of diversity on stacking supervised classifiers. In: INTERNATIONAL CONFERENCE ON ENTERPRISE INFORMATION SYSTEMS, 19., 2017. **Proceedings...** Springer International Publishing: to appear, 2017.

LARIOS, N.; LIN, J.; ZHANG, M.; LYTLE, D.; MOLDENKE, A.; SHAPIRO, L.; DIETTERICH, T. Stacked spatial-pyramid kernel: An object-class recognition method to combine scores from random trees. In: IEEE WORKSHOP ON APPLICATIONS OF COMPUTER VISION, 2011. **Proceedings...** [S.l.: s.n.], 2011. p.329–335.

LAZEBNIK, S.; SCHMID, C.; PONCE, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2006. **Proceedings...** [S.l.: s.n.], 2006. v.2, p.2169–2178.

MAKHITAR, M.; YANG, L.; NEAGU, D.; RIDLEY, M. Optimisation of Classifier Ensemble for Predictive Toxicology Applications. In: INTERNATIONAL CONFERENCE ON COMPUTER MODELLING AND SIMULATION, 14., 2012. **Anais...** [S.l.: s.n.], 2012. p.236–241.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. et al. **Introduction to information retrieval**. [S.l.]: Cambridge university press Cambridge, 2008.

MERZ, C. J. Using correspondence analysis to combine classifiers. **Machine Learning**, [S.l.], v.36, n.1-2, p.33–58, 1999.

MUHAMMAD, A. T.; JIM, S. Creating diverse nearest-neighbour ensembles using simultaneous metaheuristic feature selection. **Pattern Recognition Letters**, [S.l.], v.31, n.11, p.1470–1480, 2010.

NESS, S. R.; THEOCHARIS, A.; TZANETAKIS, G.; MARTINS, L. G. Improving automatic music tag annotation using stacked generalization of probabilistic svm outputs. In: ACM INTERNATIONAL CONFERENCE ON MULTIMEDIA, 2009. **Proceedings...** [S.l.: s.n.], 2009. p.705–708.

OLIVEIRA, D. F. d. **Dilema da Diversidade-Acurácia**: Um estudo empírico no contexto de multiclassificadores. 2008. 91p. Dissertação (Mestrado em Ciência da Computação) — Dissertação (Mestrado em Ciência da Computação) - Universidade Federal do Rio Grande do Norte.

OPITZ, D.; MACLIN, R. Popular ensemble methods: An empirical study. **Journal of Artificial Intelligence Research**, [S.l.], v.11, p.169–198, 1999.

PARZEN, E. On Estimation of a Probability Density Function and Mode. **The Annals of Mathematical Statistics**, [S.l.], v.33, n.3, p.1065–1076, 1962.

PEPPOLONI, L.; SATLER, M.; LUCHETTI, E.; AVIZZANO, C. A.; TRIPICCHIO, P. Stacked generalization for scene analysis and object recognition. In: IEEE INTERNATIONAL CONFERENCE ON INTELLIGENT ENGINEERING SYSTEMS, 2014. **Proceedings...** [S.l.: s.n.], 2014. p.215–220.

PLATT, J. C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: ADVANCES IN LARGE MARGIN CLASSIFIERS, 1999. **Proceedings...** MIT Press, 1999.

PLATT, J. C. Using analytic QP and sparseness to speed training of support vector machines. In: CONFERENCE ON ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 1999. **Proceedings...** [S.l.: s.n.], 1999. p.557–563.

QUININO, R. C.; REIS, E. A.; SUYAMA, E.; BESSEGATO, L. F. **Uma abordagem alternativa para o ensino do método dos mínimos quadrados no nível médio e início do curso superior**. Relatório Técnico RTP-03/2013, Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas.

QUINLAN, J. R. Learning with continuous classes. In: AUSTRALIAN JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, 1992, Singapore. **Proceedings...** World Scientific, 1992. v.92, p.343–348.

QUINLAN, J. R. **C4.5: programs for machine learning**. San Francisco, USA: Morgan Kaufmann Publishers Inc., 1993.

SENI, G.; ELDER, J. F. Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions. **Synthesis Lectures on Data Mining and Knowledge Discovery**, [S.l.], v.2, n.1, p.1–126, 2010.

SHIPP, C. A.; KUNCHEVA, L. I. Relationships between combination methods and measures of diversity in combining classifiers. **Information Fusion**, [S.l.], v.3, n.2, p.135 – 148, 2002.

SLUBAN, B.; LAVRAC, N. Relating ensemble diversity and performance: A study in class noise detection. **Neurocomputing**, [S.l.], v.160, p.120 – 131, 2015.

SNEATH, P. H. A.; SOKAL, R. R. **Numerical taxonomy. The principles and practice of numerical classification**. San Francisco, USA: W.H. Freeman and Company, 1973.

SPEARMAN, C. The proof and measurement of association between two things. **The American journal of psychology**, [S.l.], v.15, n.1, p.72–101, 1904.

STUDENT. The probable error of a mean. **Biometrika**, [S.l.], p.1–25, 1908.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. [S.l.]: Addison-Wesley, 2005.

TING, K. M.; WITTEN, I. H. Issues in stacked generalization. **Journal of Artificial Intelligence Research**, [S.l.], v.10, p.271–289, 1999.

TOMASINI, C.; EMMENDORFER, L.; BORGES, E. N.; MACHADO, K. A Methodology for Selecting the Most Suitable Cluster Validation Internal Indices. In: ANNUAL ACM SYMPOSIUM ON APPLIED COMPUTING, 31., 2016, New York, NY, USA. **Proceedings...** ACM, 2016. p.901–903.

VEENMAN, C. J.; REINDERS, M. J. T.; BACKER, E. A maximum variance cluster algorithm. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, [S.l.], v.24, n.9, p.1273–1280, Sep 2002.

WHALEN, S.; PANDEY, G. A Comparative Analysis of Ensemble Classifiers: Case Studies in Genomics. In: IEEE 13TH INTERNATIONAL CONFERENCE ON DATA MINING, 2013., 2013. **Anais...** [S.l.: s.n.], 2013. p.807–816.

WITTEN, I. H.; FRANK, E. **Data Mining**: Practical machine learning tools and techniques. [S.l.]: Morgan Kaufmann, 2011.

WOLPERT, D. H. Stacked generalization. **Neural networks**, [S.l.], v.5, n.2, p.241–259, 1992.