

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Rankeamento de espigas de acordo com a produtividade estimada

João Guilherme Ribeiro de Brito



Mestrado Integrado em Engenharia Informática e Computação

Orientador: João Pedro Mendes Moreira

Co-orientador: Pedro Manuel Mendes Moreira (ESAC)

20 de Fevereiro de 2014

Rankeamento de espigas de acordo com a produtividade estimada

João Guilherme Ribeiro de Brito

Mestrado Integrado em Engenharia Informática e Computação

Resumo

Nesta dissertação é apresentada uma comparação entre as abordagens Listwise (LW) e Pointwise (PW) para ordenamento de instâncias, usando a Regressão Linear Múltipla. É feita uma revisão teórica acerca destes métodos e das respectivas medidas de avaliação para cada uma destas abordagens. As experiências realizadas em cinco problemas, mostraram que em cada um destes a abordagem LW foi melhor ou similar a PW. No entanto considerando o desempenho geral das duas abordagens LW obteve melhores valores médios para a avaliação dos modelos criados que PW.

A utilização dos modelos de Regressão Linear, para fazer o ordenamento de instâncias deve-se à interpretabilidade dos modelos criados. Os resultados que foram obtidos, mostram que com a abordagem LW são selecionadas em média menos variáveis que nos modelos PW.

Abstract

In this dissertation is presented a comparison between Listwise (LW) and Pointwise (PW) approaches for instance ranking, using Multiple Linear Regression (MLR) Models. A theoretical review is performed of both approaches including the evaluation methods used for each one of these. Experiments done in five different problems, show that LW approach is better or similar than the PW approach, but in a overall look of the results, LW approach got better mean values for model evaluation than the PW.

Moreover this reseach goal, was to create intepretable Instance Ranking with interpretable models. The resuts obtained from the studied data, shown that LW models based on MLR methods are more interpretable in average, since the mean number of the selected variables is lower than in PW models.

Conteúdo

1	Introdução	1
1.1	Contexto/Enquadramento	1
1.2	Motivação e Objetivos	2
1.3	Estrutura da Dissertação	2
2	Revisão do Estado da Arte	3
2.1	Regressão Linear Múltipla	3
2.2	Pointwise	4
2.3	Listwise	4
2.4	Métodos de avaliação	5
2.4.1	Método dos Mínimos Quadrados	5
2.4.2	Normalized Discounted Cumulated Gain	6
3	Descrição das Experiências	9
3.1	Conjunto de Dados	9
3.2	Implementação	9
3.2.1	Visão Geral	9
3.2.2	Criação de Pastas	10
3.2.3	Seleção de Pastas	10
3.2.4	Criação de Modelos	11
3.2.5	Teste dos Modelos	13
4	Resultados	15
4.1	Introdução	15
4.2	Conjuntos de dados das espigas	16
4.3	Versão genérica	20
4.3.1	Conjuntos de dados	20
4.3.2	Resumo dos resultados	21
4.4	Conclusões	22
5	Conclusões e Trabalho Futuro	25
5.1	Satisfação dos Objetivos e Trabalho Futuro	25
	Referências	27
A	Tabelas de resultados para os casos genéricos	29
A.1	Wine white	29
A.2	Wine red	29
A.3	Parkinson (MOTOR)	30

CONTEÚDO

A.4	Parkinson (TOTAL)	30
A.5	Energy efficiency (Y1)	30
A.6	Energy efficiency (Y2)	31
A.7	LETOR	31
B	Gráficos CoV vs NDCG para os Conjuntos de dados	33
B.1	Espigas	33
B.1.1	LW DP	33
B.1.2	PW	35
B.2	Wine White	37
B.2.1	LW DP	37
B.2.2	PW	39
B.3	Wine Red	41
B.3.1	LW DP	41
B.3.2	PW	43
B.4	Parkinson (Total)	45
B.4.1	LW DP	45
B.4.2	PW	47
B.5	Parkinson (Motor)	49
B.5.1	LW DP	49
B.5.2	PW	51
B.6	Energy Efficiency(Y1)	53
B.6.1	LW DP	53
B.6.2	PW	55
B.7	Energy Efficiency(Y2)	57
B.7.1	LW DP	57
B.7.2	PW	59
B.8	LETOR	61
B.8.1	PW	61

Lista de Figuras

2.1	Representação do exemplo do artigo [CQL ⁺ 07]	5
3.1	Arquitetura Geral	10
3.2	Seleção de pastas	11
4.1	Gráfico de variação do CoV vs NDCG para a abordagem PW no conjunto das Espigas	18
4.2	Gráfico de variação do CoV vs NDCG para a abordagem "LW DP" no conjunto das Espigas	19
B.1	NDCG vs CoV, LW DP, 1ª execução	33
B.2	NDCG vs CoV, LW DP, 2ª execução	34
B.3	NDCG vs CoV, LW DP, 3ª execução	34
B.4	NDCG vs CoV, LW DP, 4ª execução	34
B.5	NDCG vs CoV, LW DP, 5ª execução	35
B.6	NDCG vs CoV, PW, 1ª execução	35
B.7	NDCG vs CoV, PW, 2ª execução	35
B.8	NDCG vs CoV, PW, 3ª execução	36
B.9	NDCG vs CoV, PW, 4ª execução	36
B.10	NDCG vs CoV, PW, 5ª execução	36
B.11	NDCG vs CoV, LW DP, 1ª execução	37
B.12	NDCG vs CoV, LW DP, 2ª execução	37
B.13	NDCG vs CoV, LW DP, 3ª execução	38
B.14	NDCG vs CoV, LW DP, 4ª execução	38
B.15	NDCG vs CoV, LW DP, 5ª execução	38
B.16	NDCG vs CoV, PW, 1ª execução	39
B.17	NDCG vs CoV, PW, 2ª execução	39
B.18	NDCG vs CoV, PW, 3ª execução	39
B.19	NDCG vs CoV, PW, 4ª execução	40
B.20	NDCG vs CoV, PW, 5ª execução	40
B.21	NDCG vs CoV, LW DP, 1ª execução	41
B.22	NDCG vs CoV, LW DP, 2ª execução	41
B.23	NDCG vs CoV, LW DP, 3ª execução	42
B.24	NDCG vs CoV, LW DP, 4ª execução	42
B.25	NDCG vs CoV, LW DP, 5ª execução	42
B.26	NDCG vs CoV, PW, 1ª execução	43
B.27	NDCG vs CoV, PW, 2ª execução	43
B.28	NDCG vs CoV, PW, 3ª execução	43

LISTA DE FIGURAS

B.29 NDCG vs CoV, PW, 4 ^a execução	44
B.30 NDCG vs CoV, PW, 5 ^a execução	44
B.31 NDCG vs CoV, LW DP, 1 ^a execução	45
B.32 NDCG vs CoV, LW DP, 2 ^a execução	45
B.33 NDCG vs CoV, LW DP, 3 ^a execução	46
B.34 NDCG vs CoV, LW DP, 4 ^a execução	46
B.35 NDCG vs CoV, LW DP, 5 ^a execução	46
B.36 NDCG vs CoV, PW, 1 ^a execução	47
B.37 NDCG vs CoV, PW, 2 ^a execução	47
B.38 NDCG vs CoV, PW, 3 ^a execução	47
B.39 NDCG vs CoV, PW, 4 ^a execução	48
B.40 NDCG vs CoV, PW, 5 ^a execução	48
B.41 NDCG vs CoV, LW DP, 1 ^a execução	49
B.42 NDCG vs CoV, LW DP, 2 ^a execução	49
B.43 NDCG vs CoV, LW DP, 3 ^a execução	50
B.44 NDCG vs CoV, LW DP, 4 ^a execução	50
B.45 NDCG vs CoV, LW DP, 5 ^a execução	50
B.46 NDCG vs CoV, PW, 1 ^a execução	51
B.47 NDCG vs CoV, PW, 2 ^a execução	51
B.48 NDCG vs CoV, PW, 3 ^a execução	51
B.49 NDCG vs CoV, PW, 4 ^a execução	52
B.50 NDCG vs CoV, PW, 5 ^a execução	52
B.51 NDCG vs CoV, LW DP, 1 ^a execução	53
B.52 NDCG vs CoV, LW DP, 2 ^a execução	53
B.53 NDCG vs CoV, LW DP, 3 ^a execução	54
B.54 NDCG vs CoV, LW DP, 4 ^a execução	54
B.55 NDCG vs CoV, LW DP, 5 ^a execução	54
B.56 NDCG vs CoV, PW, 1 ^a execução	55
B.57 NDCG vs CoV, PW, 2 ^a execução	55
B.58 NDCG vs CoV, PW, 3 ^a execução	55
B.59 NDCG vs CoV, PW, 4 ^a execução	56
B.60 NDCG vs CoV, PW, 5 ^a execução	56
B.61 NDCG vs CoV, LW DP, 1 ^a execução	57
B.62 NDCG vs CoV, LW DP, 2 ^a execução	57
B.63 NDCG vs CoV, LW DP, 3 ^a execução	58
B.64 NDCG vs CoV, LW DP, 4 ^a execução	58
B.65 NDCG vs CoV, LW DP, 5 ^a execução	58
B.66 NDCG vs CoV, PW, 1 ^a execução	59
B.67 NDCG vs CoV, PW, 2 ^a execução	59
B.68 NDCG vs CoV, PW, 3 ^a execução	59
B.69 NDCG vs CoV, PW, 4 ^a execução	60
B.70 NDCG vs CoV, PW, 5 ^a execução	60
B.71 NDCG vs CoV, PW, Pasta 1	61
B.72 NDCG vs CoV, PW, Pasta 2	61
B.73 NDCG vs CoV, PW, Pasta 3	62
B.74 NDCG vs CoV, PW, Pasta 4	62
B.75 NDCG vs CoV, PW, Pasta 5	63

Lista de Tabelas

4.1	Resultados para conjunto dos dados das espigas de milho	16
4.2	Número médio de variáveis selecionadas por abordagem	16
4.3	Descrição dos conjuntos de dados	20
4.4	Resumo de resultados	21
4.5	Número médio de variáveis selecionadas cada abordagem nos conjuntos de dados	21
4.6	Matriz de correlação dos dados para <i>Energy Efficiency</i>	23
A.1	Resultados para conjunto dos dados das Wine white	29
A.2	Resultados para conjunto dos dados das Wine white	29
A.3	Resultados para conjunto dos dados das Parkinson (MOTOR)	30
A.4	Resultados para conjunto dos dados das Parkinson (TOTAL)	30
A.5	Resultados para conjunto dos dados das Energy efficiency (Y1)	30
A.6	Resultados para conjunto dos dados das Energy efficiency (Y2)	31
A.7	Resultados para conjunto dos dados das LETOR	31

LISTA DE TABELAS

Lista de Algoritmos

1 Criação de modelos 12

LISTA DE ALGORITMOS

Abreviaturas e Símbolos

OI	Ordenamento de instâncias (tradução do termo inglês "Instance Ranking")
CoV	Coefficiente de Variação
NDCG	<i>Normalized Discounted Cumulated Gain</i>
LW	Listwise
PW	Pointwise
VC	Validação Cruzada
RN	Redes Neurais
MMQ	Método dos Mínimos Quadrados

Capítulo 1

Introdução

Neste primeiro capítulo vai ser feita uma introdução ao tema desta dissertação. Aqui vão ser apresentados, o contexto em que surgiu este tema, os objectivos e motivações do problema e também a estrutura em que se dividiu a dissertação.

1.1 Contexto/Enquadramento

O tema desta dissertação surge no seguimento de um estudo feito previamente que tem como base um concurso existente na região do Vale do Sousa, distrito do Porto, para a melhor espiga de milho. O objetivo deste é avaliar as espigas entregues pelos agricultores para concurso e escolher a melhor espiga de milho com base numa fórmula empírica. Esta utiliza características da espiga que bibliograficamente estão correlacionadas com a produção. Contudo não existia uma relação direta entre as espigas colhidas pelos agricultores e a produção. Deste modo o projeto FCT (PTDC/AGR-AAM/70845/2006) « Analysis of the breeding potential of fasciation traits in Portuguese maize landraces » permitiu gerar dados das espigas e produção.

Não só em Portugal como também em outros países existem ou existiram concursos semelhantes. Nos Estados Unidos os concursos e exibições das melhores espigas tornaram-se, muito populares nos inícios do século XX, a par do desenvolvimento ¹ científico das linhas puras, ambos elementos foram cruciais para o aumento de produção. Estes concursos têm como objetivo eleger o agricultor que tenha a melhor espiga de milho e a melhor espiga de milho é aquela que esteja mais correlacionada com a produção. Desta forma ao elegermos o melhor produtor estamos a encorajar outros produtores a seguirem o esquema de seleção subjacente à fórmula (características e sua ponderação), contribuindo indiretamente para o aumento de produção. Esta dissertação tem como objetivo desenvolver um método de avaliação interpretável que seja útil neste tipo de concursos.

¹<http://www.ncga.com/for-farmers/national-corn-yield-contest>

A maximização produção de cereais deve ser critério não só na área dos concursos, como também em todas as outras onde a maximização da produção seja relevante. Nos últimos anos tem-se observado, em parte devido à utilização dos cereais na produção de combustíveis², a um aumento do preço dos cereais e consequentemente ao aumento dos preços dos produtos alimentares dependentes destes. A utilização de cereais na produção de combustíveis está relacionada com diversos fatores. Exemplos destes são: as consequências ambientais que advêm da exploração dos combustíveis fósseis; o crescimento da procura de combustíveis nos próximos anos; e a rentabilidade económica dos combustíveis. [HNT⁺06].

1.2 Motivação e Objetivos

O conceito de concurso pressupõe que os candidatos depois de avaliados sejam ordenados consoante um critério de avaliação. Na área de *machine learning* este conceito é conhecido como Ordenamento de Instâncias (tradução do inglês *Instance Ranking*). Este consiste em estabelecer uma ordem aos objetos consoante as características destes. Este método tem como objetivo criar um modelo preditivo baseado em dados de treino, para posteriormente prever a ordem de novas instâncias. Como vai ser explicado mais a frente (capítulo 2) existem diversas abordagens possíveis para criar os modelos de previsão. Este método é utilizado em diversas áreas, como por exemplo ordenação de resultados em pesquisas na WEB, ações de bolsa de valores, sistemas de recomendação, etc.

Neste problema existe a necessidade de justificar a ordenação dos objetos (espigas de milho) no concurso. Tendo este ponto como restrição o espaço de técnicas disponíveis fica reduzido, sendo que a regressão linear múltipla é uma opção viável, uma vez que o resultado obtido é facilmente interpretável [HTF09]. Vários estudos sobre a abordagem LW feitos anteriormente, sugerem a utilização de Redes Neurais e Máquinas com Vetores de Suporte [XLW⁺08, CQL⁺07], mas tal como dito anteriormente estas soluções não são viáveis no âmbito dos concursos, uma vez que a solução não é interpretável.

1.3 Estrutura da Dissertação

A dissertação está dividida em mais quatro capítulos para além da introdução. No capítulo 2 é feita a revisão do estado da arte onde são apresentados exemplos das metodologias e abordagens aplicadas neste estudo. De seguida no capítulo 3 é feita a descrição do trabalho desenvolvido com a especificação dos algoritmos desenvolvidos e adaptação das abordagens escolhidas. No capítulo 4 são apresentados os resultados obtidos para as experiências descritas no capítulo 3. Por fim as conclusões do trabalho desenvolvido e propostas para trabalho futuro são apresentados no capítulo 5.

² <http://www.ncga.com/upload/files/documents/pdf/foodfueljune2010.pdf>

Capítulo 2

Revisão do Estado da Arte

De acordo com a motivação do problema, neste capítulo é feita a introdução teórica aos métodos adotados para cumprimento dos objetivos propostos anteriormente na secção 1.2.

Aqui é feita a apresentação teórica da Regressão Linear Múltipla e das abordagens a implementar (Listwise e Pointwise). Por fim são descritos os métodos de avaliação correspondentes a cada uma destas abordagens.

2.1 Regressão Linear Múltipla

Uma regressão linear simples [HKP11] é definida como o ajuste de uma reta a um conjunto de dados. Deste vai resultar um modelo linear representado por uma equação de reta (equação 2.1). Neste modelo a relação entre os dados vai ser representada por duas variáveis: (1) dependente Y correspondendo aos valores de resposta do modelo; e (2) independente X , valor do parâmetro.

$$Y = \alpha + \beta X + \varepsilon \quad (2.1)$$

Os valores de α e β correspondem respetivamente aos valores de: intersecção da reta com o eixo dos xx , isto é, quando $X = 0$; e o declive da reta. O parâmetro ε é um erro aleatório proveniente da aproximação da reta aos dados de entrada. Os valores de α e β não são valores conhecidos, pelo que têm de ser estimados $\hat{\alpha}$ e $\hat{\beta}$ respetivamente. Um procedimento comum para estimar estes valores é utilizando o Método dos Mínimos Quadrados (fórmulas 2.2 e 2.3).

$$\beta = \frac{\sum_{i=1}^s (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^s (x_i - \bar{x})^2} \quad (2.2)$$

$$\alpha = \bar{y} - \beta \bar{x} \quad (2.3)$$

Considerando os dados de entrada representados por $\{(x_1, y_1), (x_2, y_2), \dots, (x_s, y_s)\}$ na forma de pontos e sendo $\bar{x} = \text{media}(x_s)$ e $\bar{y} = \text{media}(y_s)$, este método estima os coeficientes da equação (α

e β) minimizando o erro global, sendo este a soma da distância quadrática entre os pontos antes definidos e a reta.

Para a regressão linear múltipla pode ser feita uma analogia com a regressão linear simples. Aqui em vez de uma variável independente vão existir tantas quantos os parâmetros seleccionados dos dados. A regressão linear múltipla (equação 2.4) apenas difere para a simples na parte direita, em que esta apresenta várias variáveis independentes (X_n sendo n o número de variáveis) e os correspondentes estimadores (para cada variável X_n existe um estimador β).

$$Y = \alpha + \beta_1 X_1 + \dots + \beta_n X_n + \varepsilon \quad (2.4)$$

O cálculo dos estimadores $\beta_i (i = 1, \dots, n)$ para a regressão linear múltipla, é feito usando o mesmo método apresentado anteriormente nas equações 2.2 e 2.3.

2.2 Pointwise

Consideremos como exemplo um conjunto de n instâncias de um tipo genérico, representado pelo vetor $E = \{e_1, e_2, \dots, e_n\}$. Em cada uma das instâncias de E vai existir conjunto de m parâmetros associado (tal como definido para a regressão linear múltipla em 2.1), podendo este ser representado por uma matriz n por m , em que cada linha corresponde ao conjunto de parâmetros de cada elemento do vetor E .

$$FE = \begin{bmatrix} fe_1^1 & fe_1^2 & \dots & fe_1^m \\ fe_2^1 & fe_2^2 & \dots & fe_2^m \\ \vdots & \vdots & \ddots & \vdots \\ fe_n^1 & fe_n^2 & \dots & fe_n^m \end{bmatrix}$$

Para fazer a ordenação das instâncias, este processo vai prever os valores da variável dependente com base num modelo de previsão. As previsões obtidas vão ser posteriormente ordenadas de modo a que se obtenha um vetor resultado $R(E) = e_1, e_2, \dots, e_n$ em que $e_1 < e_2 < \dots < e_n$.

Uma vez que ordenação das instâncias depende dos valores previstos pelo modelo preditivo, as medidas de avaliação do erro para a criação deste, têm também como base métodos baseados em distâncias como vai ser explicado mais a frente em 2.4. Ao ser construído o modelo preditivo, o objetivo desta abordagem é a minimização do erro de previsão para a variável dependente. Assim sendo quanto menor for o erro previsto, mais precisa vai ser a ordenação das instâncias.

2.3 Listwise

Tal como introduzido na secção 1.2, em [CQL⁺07] é defendido que esta abordagem apresenta melhores resultados face a *Pairwise* (considerada como uma evolução do método Pointwise), uma vez que considera a previsão para uma lista de objetos em vez de pares de objetos. É apresentado um caso exemplo explicativo do funcionamento da abordagem LW, em que este é baseado

na obtenção de documentos para uma pesquisa. Na figura 2.1 é feita uma representação deste exemplo.

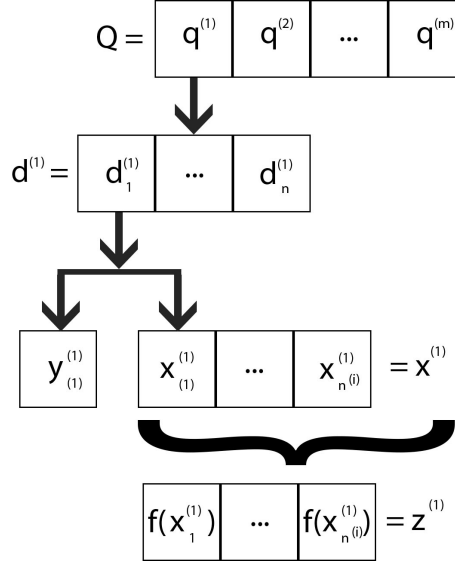


Figura 2.1: Representação do exemplo do artigo [CQL⁺07]

Considerando um conjunto de m pesquisas $Q = \{q^{(1)}, q^{(2)}, \dots, q^{(m)}\}$, cada uma das pesquisas tem associado um conjunto de n documentos que estão relacionadas com esta, definido como $d_{(n)}^{(m)}$, representando m e n , o identificador da pesquisa e o número de documentos relacionados com esta. Para o exemplo na figura 2.1, $d^{(1)}$ representa o conjunto de n documentos para a pesquisa $q^{(1)}$. Para cada um dos documentos de uma pesquisa está associado um vetor representativo das suas características (representado na figura pelo vetor $x^{(1)}$) e também a pontuação relativa à relevância do documento para a pesquisa (representada por $y_{(1)}^{(1)}$ para o caso exemplo). Para treino do modelo preditivo, é criada uma função f para fazer a avaliação das características de cada documento, resultando desta num vetor de pontuações, representado na figura por $z^{(1)}$ (vetor de pontuações para o conjunto das características de $d^{(1)}$, para a pesquisa $q^{(1)}$).

2.4 Métodos de avaliação

Nesta secção são apresentadas medidas de avaliação de erro aplicáveis aos métodos acima descritos.

2.4.1 Método dos Mínimos Quadrados

Este método de avaliação, também conhecido como Método dos Quadrados Ordinários, tem como função o cálculo do erro na construção de um modelo preditivo. O erro é dado consoante a

distância entre o valor previsto e o valor real de uma dada instância, isto é, aquando da criação do modelo de regressão linear com base no conjunto de treino, como referido acima em 2.1.

$$SEQ = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - Y_i)^2 \quad (2.5)$$

Com o modelo criado, apresentado na forma de equação de reta (equação 2.4), é feita a previsão para um conjunto de dados e com base na equação 2.5 acima apresentada. Com o conjunto dos valores previstos ($\hat{Y}$) e os valores reais do conjunto de dados (Y) é calculado o erro de previsão. Este método é frequentemente usado para ajuste e cálculo do erro dos modelos em regressões lineares. Este método vai ser utilizado na abordagem PW para cálculo do erro de previsão dos modelos como se verá mais à frente na secção 3.

2.4.2 Normalized Discounted Cumulated Gain

Em [JK02], são definidos os conceitos relativos aos ganhos como:

$$CG[i] = \begin{cases} G[1], & \text{if } i = 1 \\ CG[i-1] + G[i], & \text{Caso contrário} \end{cases} \quad (2.6)$$

$$DCG[i] = \begin{cases} CG[i], & \text{if } i < b \\ CG[i-1] + G[i]/\log_b i, & \text{if } i \geq b \end{cases} \quad (2.7)$$

$$NDCG[i] = \frac{DCG[i]}{DCG_I[i]} \quad (2.8)$$

Nas fórmulas 2.6, 2.7 e 2.8 são definidos respetivamente os ganhos: cumulativo; cumulativo descontado; e cumulativo descontado normalizado. Na formula 2.8 é apresentado um termo denominado DCG_I , este representa o vetor de ganho ideal, sendo que os valores deverão estar ordenados por ordem decrescente.

Se considerarmos um vetor $V = \{5, 3, 4, 1, 6, 2\}$, o logaritmo com base $b = 2$ e i o número de instâncias no vetor V , por aplicação dos conceitos definidos anteriormente, vamos obter os seguintes resultados:

1. $CG(V) = \{5, 8, 12, 13, 19, 21\}$
2. $CG_I(V) = \{6, 11, 15, 18, 20, 21\}$
3. $DCG(V) = \{5, 8, 10.52, 11.02, 13.60, 14.38\}$
4. $DCG_I(V) = \{6, 11, 13.52, 15.02, 15.88, 16.27\}$
5. $NDCG(V) = \{0.83, 0.72, 0.77, 0.73, 0.85, 0.88\}$

Na utilização deste método para avaliação de um modelo, podemos considerar o valor médio de NDCG (sendo este $media(NDCG) = 0.80$ para a demonstração feita).

Revisão do Estado da Arte

Tomando como exemplo o caso apresentado anteriormente na abordagem LW (capítulo 2.3), considerando que temos um conjunto de documentos a apresentar, consoante uma pesquisa, a ordenação destes vai ser o conjunto de dados (V) para cálculo da eficácia de previsão para este problema (definido como pontuação, Y , na figura 2.1).

Capítulo 3

Descrição das Experiências

Neste capítulo é mostrada de forma esquematizada a implementação das abordagens e metodologias acima especificadas no capítulo 2. São também apresentadas considerações a ter em atenção relativamente ao conjunto de dados utilizado.

3.1 Conjunto de Dados

O conjunto de dados que foi fornecido ¹ contém 481 instâncias de espigas de milho caracterizadas por 43 atributos. No conjunto dos atributos existentes, alguns são relativos ao terreno em que estas foram plantadas. Apesar destes dados puderem dar resultados interessantes a estudar, estes não vão ser considerados uma vez que no âmbito de um concurso estes parâmetros não são possíveis de controlar, assim sendo, só 31 destes atributos vão ser usados.

3.2 Implementação

Nesta secção é apresentada a arquitetura geral do algoritmo desenvolvido (secção 3.2.1) para a aplicação dos conceitos estudados. Esta arquitetura é estruturada por módulos, em que cada um destes representa uma passo na preparação dos dados e criação dos modelos de previsão. Cada um dos módulos é descrito nas secções 3.2.2, 3.2.3, 3.2.4 e 3.2.5.

3.2.1 Visão Geral

Na figura 3.1 são apresentados os quatro passos que são necessários para a implementação do algoritmo. Uma vez que o conjunto dos dados é relativamente pequeno, é utilizada validação cruzada dos dados (VC). A vertente usada deste método, é conhecida como "*k-fold Cross-Validation*" (*k-fold CV*). Neste processo é dividido o conjunto de dados em k pastas. Após esta di-

¹ESAC responsible for the project FCT (PTDC/AGR-AAM/70845/2006) (a three- year project with the beginning on 1st of June of 2007) « Analysis of the breeding potential of fasciation traits in Portuguese maize landraces »

Descrição das Experiências

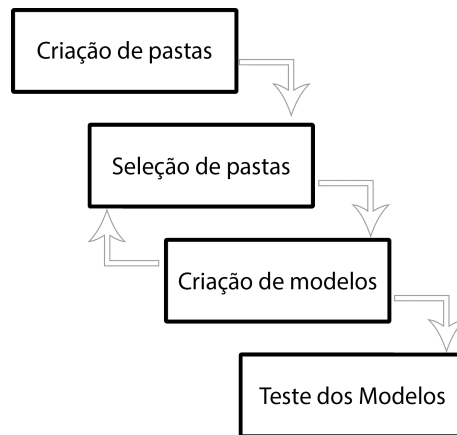


Figura 3.1: Arquitetura Geral

visão, pastas diferentes são utilizadas para criação, validação e teste de modelos [HTF09, HKP11]. O objetivo deste método é possibilitar a criação de vários modelos e testar qual o que se ajusta melhor ao conjunto dos dados.

Nas secções seguintes vão ser descritas as considerações tomadas para aplicação deste método e também o processo de criação de modelos para as abordagens LW e PW.

3.2.2 Criação de Pastas

Na divisão do conjunto de dados, são criadas dez pastas ($k = 10$) para treino, validação e teste dos modelos a criar. No processo VC aquando da divisão dos dados, é garantido que todas as pastas ficam com o mesmo número de elementos. No caso em que não é possível fazer a divisão dos dados pelas pastas de maneira a ficarem todas com o mesmo número de instâncias, isto é, se no resto da divisão por k não for 0, são retirados aleatoriamente as instâncias necessárias para equilibrar as pastas.

3.2.3 Seleção de Pastas

A cada iteração do algoritmo ($it \in [1, k]$, sendo k o número de pastas criadas), é feita a seleção de pastas para treino, validação e teste.

A subdivisão do conjunto de dados nestas três partes, está relacionado com a aplicação do método de VC. Este tem como objetivo avaliar a prestação do algoritmo num conjunto de dados. Com esta subdivisão, para cada pasta de teste vamos obter uma avaliação do modelo criado com base nas pastas de treino e validação. No final do processo de VC com o resumo de todas as avaliações feitas (uma para cada pasta de treino), podemos analisar o desempenho do algoritmo no conjunto de dados testado.

É considerada nesta fase uma iteração como, este processo de escolha das pastas para a criação dos modelos preditivos. Neste processo é garantido que as pastas escolhidas treino, validação e

Descrição das Experiências

teste, são as mesmas para cada abordagem, garantido uma melhor comparação entre as abordagens em comparação..

Como demonstrado na figura 3.2, abaixo apresentada, o número da iteração (it) vai ditar a pasta de teste a ser selecionada. Após esta ser selecionada, é escolhida aleatoriamente a pasta de validação das $k - 1$ restantes. Finalmente as $k - 2$ pastas restantes ficam para treino do modelo.

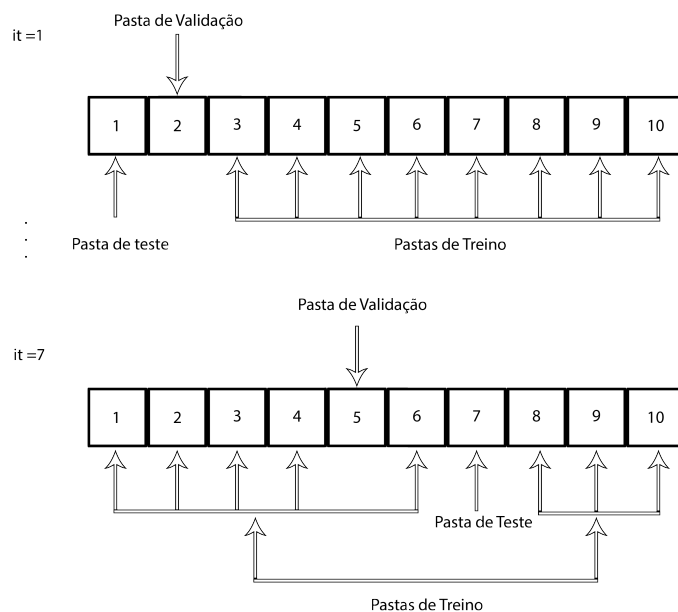


Figura 3.2: Seleção de pastas

3.2.4 Criação de Modelos

Na fase de criação dos modelos para ambas as abordagens definidas, o processo é muito semelhante. Para cada uma das k iterações do algoritmo são construídos dois modelos, um para a abordagem LW e outro para PW. A principal diferença na criação de cada um dos modelos é a medida de avaliação. Esta é aplicada no processo de seleção de variáveis.

No algoritmo desenvolvido, o modelo de seleção de variáveis adotado foi o modelo incremental (conhecido como "*Forward-selection*", [HTF09, HKP11]). No algoritmo 1 apresenta-se uma possível solução em pseudo-código visando englobar os passos comuns à criação de modelos para as abordagens LW e PW.

Tal como indica o nome do processo, a seleção de variáveis é feita incrementalmente. Partindo de um conjunto inicial de variáveis, na primeira iteração deste processo, são testadas todas as variáveis (*modelList* e *errorList*, representam o conjunto de modelos e erros para estes), é escolhido o modelo com a variável que maximize/minimize a função de avaliação consoante o objetivo desta. Na iteração seguinte as restantes variáveis são testadas com a variável já selecionada, mais uma vez é selecionado o modelo que cumpra o objetivo da função de avaliação.

Algoritmo 1 Criação de modelos

```

Input:  $varList, trainData, validationData$ 
 $selectedVariables \leftarrow \{\}$ 
loop
     $modelList \leftarrow buildModels(selectedVariables,$ 
 $varList, trainData)$ 
     $errorList \leftarrow errors(modelList,$ 
 $validationData)$ 
     $selectedModel \leftarrow bestModel(errorList, modelList)$ 
     $error \leftarrow errorList[selectedModel]$ 
    if  $isBetter(error, bestModelError)$  then
         $bestModel \leftarrow selectedModel$ 
         $selectedVariables \leftarrow modelVariables(bestModel)$ 
         $bestError \leftarrow error$ 
    else
        break
    end if
end loop

```

Este método é repetido até que, o conjunto de variáveis não contenha mais variáveis para além das selecionadas ou então das existentes nenhuma consegue otimizar o valor da função de avaliação.

3.2.4.1 LW

Na abordagem LW a função de avaliação a usar é o Ganho Cumulativo Descontado Normalizado (NDCG) já definida na secção 2.4.2. Nesta abordagem esta função é utilizada na validação do modelo no processo de seleção de variáveis e também no teste do modelo. Tanto no processo de validação do modelo, como no de teste, em caso de empate de instâncias na lista de previsão dos resultados, é considerada a ordem real dos dados. Considerando as instâncias previstas $P(I) = \{i_1, i_2, i_3, i_4, i_5\}$ e supondo que os valores de i_2 e i_3 são iguais, mas a ordenação real é $R(I) = \{i_1, i_3, i_2, i_4, i_5\}$ é considerada que a ordenação está certa, isto para evitar penalizações indevidas, uma vez que não é conhecido o critério de desempate da linguagem de programação usada.

3.2.4.2 PW

Para a tradicional abordagem PW, é usado o Método dos Mínimos Quadrados (MMQ) também definido na secção 2.4.1, como função de avaliação para validação do modelo. Uma vez que o objetivo desta dissertação é verificar os ganhos da abordagem LW relativamente a PW, a função de avaliação para teste do modelo usada foi mais uma vez Ganho Cumulativo Descontado Normalizado (NDCG), uma vez que as medidas de avaliação não são comparáveis entre si.

3.2.5 Teste dos Modelos

Como já referido anteriormente nas secções 3.2.4.1 e 3.2.4.2, para ambas as abordagens é calculado o valor de NDCG para todas as k pastas de teste. Estes valores vão permitir a comparação de resultados entre as duas abordagens, LW e PW. A avaliação dos modelos é feita através da média ponderada dos valores de NDCG das k pastas de teste de cada abordagem.

Descrição das Experiências

Capítulo 4

Resultados

Neste capítulo são apresentados os resultados obtidos para as metodologias abordadas consoante a arquitetura definida acima nos capítulos 2 e 3. Para além do conjunto de dados das espigas de milho, foi também aplicado o algoritmo para outros conjuntos de dados. Algumas considerações adicionais vão ser consideradas para casos excecionais em alguns dos conjuntos de dados a seguir apresentados.

4.1 Introdução

Para cada conjunto de dados o algoritmo foi executado cinco vezes, isto é, em cada uma das cinco corridas do algoritmo são geradas novas pastas e consequentemente criados novos modelos preditivos.

Para além deste ponto, foi também feita uma transformação no conjunto de dados, substituindo os valores da variável dependente, pelos valores de ordem ($O = \{n, \dots, 1\}$, sendo n o número de instâncias da pasta de treino, validação ou teste). Esta transformação faz com que a noção de distância entre instâncias do conjunto de dados mude, mas ao mesmo tempo dá-nos uma nova perspectiva sobre os resultados. Considerando por exemplo um conjunto de cinco instâncias, com os seguintes valores para a variável dependente, $I = \{55.2, 23.5, 12, 0.5, 6\}$, os valores que vão ser considerados para a variável dependente vão ser $I' = \{5, 4, 3, 1, 2\}$, sendo que a ordem é feita considerando que o valor mais alto é o que é mais relevante.

Esta diferenciação foi feita só para a abordagem LW. Os resultados desta variante do algoritmo está representada na coluna "LW rank", presente nas tabelas de resultados apresentadas nas próximas secções. Na coluna "LW DP", são apresentados os resultados para os valores reais para a variável dependente, isto é, sem transformação dos dados.

Para a abordagem PW, o erro que é apresentado na coluna "erro de treino", corresponde ao coeficiente de variância. O motivo da apresentação deste em vez do erro calculado pelo Método dos Mínimos Quadrados (definido no capítulo 2, equação 2.5), é a interpretação de erro, uma vez

Resultados

que o erro apresentado pelo MMQ é relativo ao conjunto dos dados, sendo apresentado em valor absoluto. No caso do Coeficiente de Variação dado pela equação 4.2.

$$AVG = \frac{1}{n} \sum_{i=1}^n (X_i) \quad (4.1)$$

$$CoV = \frac{\sqrt{SEQ}}{AVG} \quad (4.2)$$

Na continuação deste capítulo vão ser apresentados gráficos explicativos do comportamento dos métodos de avaliação para cada abordagem, para cada um dos conjuntos de dados estudados. Nestes gráficos os valores representados nos pontos azuis e verdes, são respetivamente os valores de CoV e NDCG na fase de seleção de variáveis.

4.2 Conjuntos de dados das espigas

Nas tabelas 4.1 e 4.2, são apresentados respetivamente, os resultados para cada execução do algoritmo (5 iterações) e média do número de variáveis selecionadas (5 iterações * 10 modelos devido à VC com 10 pastas) para cada abordagem.

Tabela 4.1: Resultados para conjunto dos dados das espigas de milho

	LW DP		LW rank		Pointwise	
	ndcg treino	ndcg teste	ndcg treino	ndcg teste	CoV de treino	ndcg teste
run 1	0.975	0.953	0.965	0.956	0.108	0.957
run 2	0.985	0.933	0.968	0.960	0.090	0.967
run 3	0.974	0.940	0.959	0.960	0.089	0.964
run 4	0.977	0.952	0.956	0.950	0.096	0.954
run 5	0.980	0.950	0.958	0.941	0.093	0.957
Médias		0.946		0.953		0.960

Tabela 4.2: Número médio de variáveis selecionadas por abordagem

LW DP	LW rank	Pointwise
6	6.96	12.86

Na figura 4.1 e também mais abaixo no apêndice B, são apresentados gráficos correspondendo às variações dos valores de NDCG e CoV, na fase de seleção de variáveis. Os valores apresentados nos gráficos são interpolados, uma vez que os valores para CoV e NDCG podem ser muito distantes, não sendo normalizados a comparação no gráfico pode ser difícil de fazer. Tal como referido em 4.1, a verde estão representados os valores de NDCG e a azul os valores de CoV. Estas figuras têm por objetivo perceber como se comporta a medida de avaliação usada pela abordagem LW quando a abordagem é PW (figura 4.1 e no anexo B.1.2) e, como se comporta a medida de avaliação usada pela abordagem PW quando a abordagem é LW (figura 4.2 e no anexo B.1.1). Nestes gráficos são apresentados dois eixos. Cada um destes tem a escala respetiva à métrica usada para a

Resultados

avaliação de cada modelo (CoV e NDCG). Foram feitas comparações só entre as abordagens LW DP e PW.

Resultados

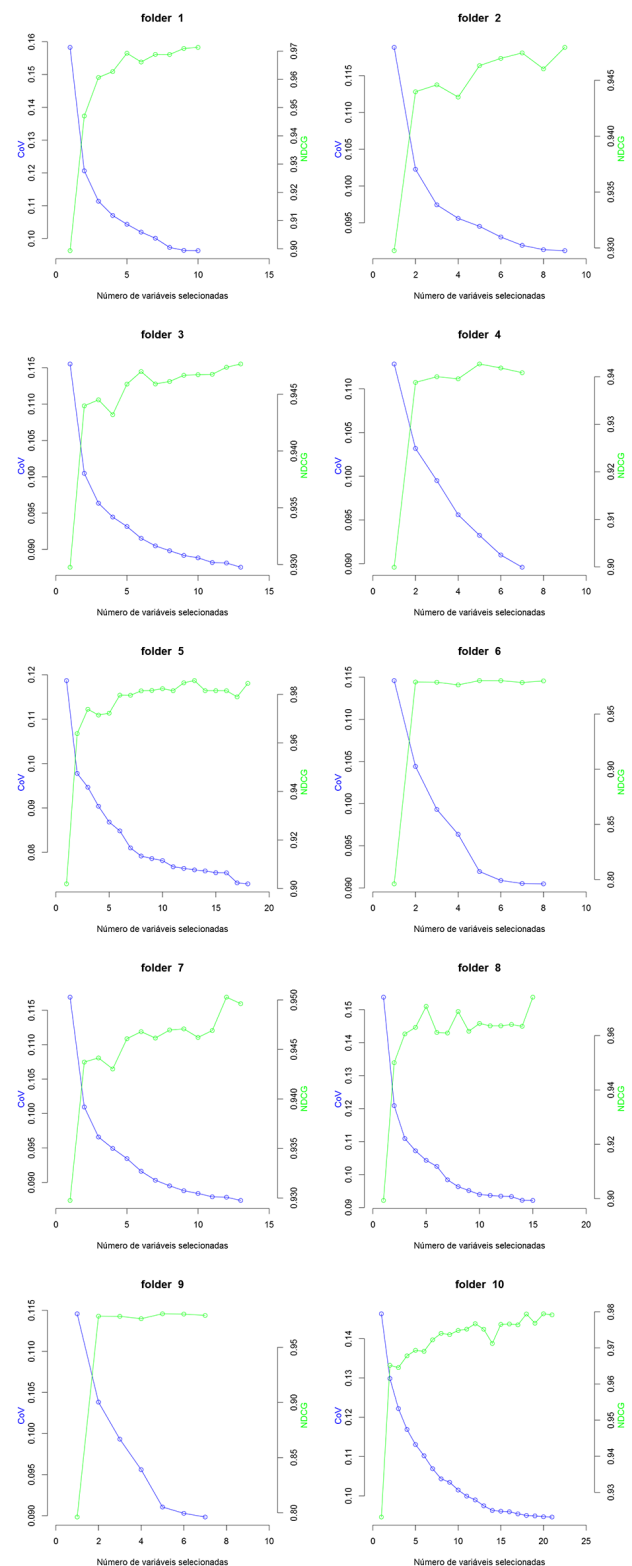


Figura 4.1: Gráfico de variação do CoV vs NDCG para a abordagem PW no conjunto das Espigas

Resultados

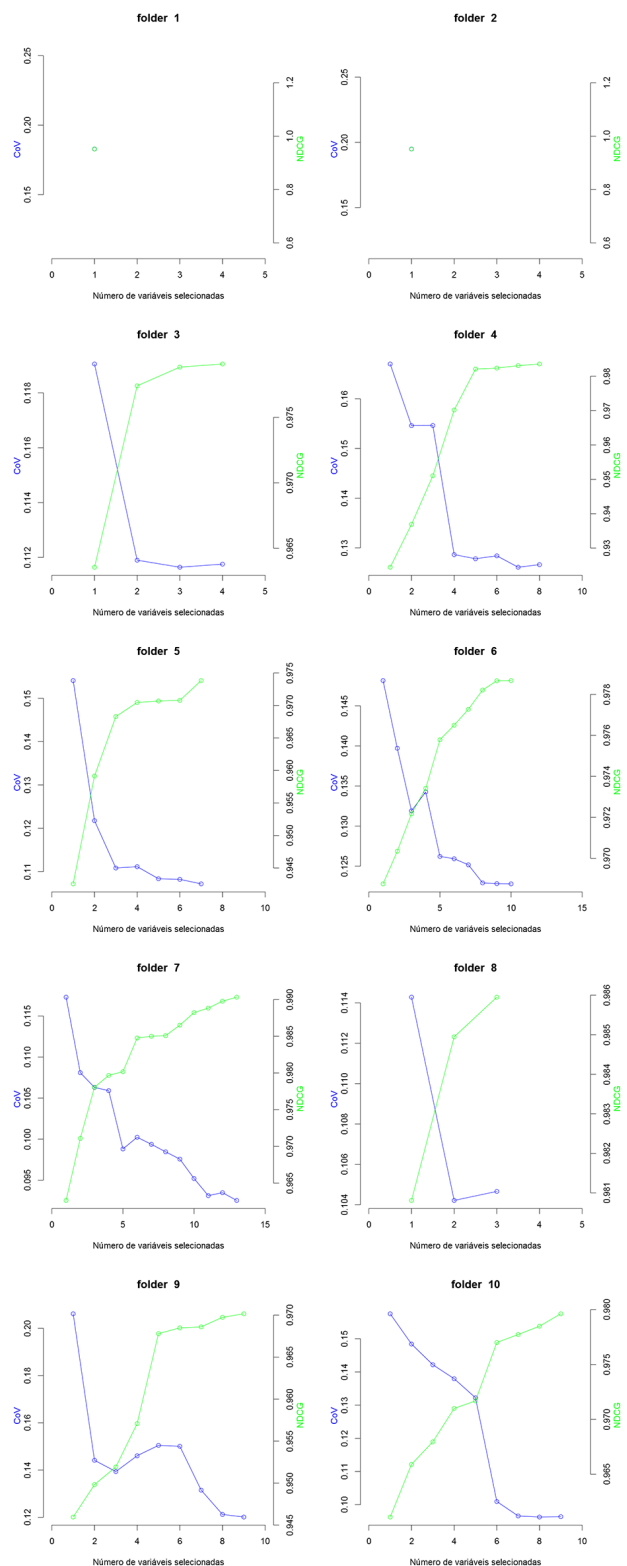


Figura 4.2: Gráfico de variação do CoV vs NDCG para a abordagem "LW DP" no conjunto das Espigas

4.3 Versão genérica

Para a aplicação do algoritmo num âmbito diferente do concurso das espigas, foram recolhidos alguns conjuntos de dados, passíveis de serem testados com uma regressão linear múltipla.

4.3.1 Conjuntos de dados

As características dos conjuntos de dados obtidos estão resumidamente descritos na tabela 4.3, abaixo apresentada.

Tabela 4.3: Descrição dos conjuntos de dados

Nome	Número de instâncias	Número de atributos	Fonte dos dados
Energy Efficiency	768	8	[BL13]
Wine Quality	4898	12	[BL13]
Parkinson Telemonitoring	5875	26	[BL13]
Microsoft Learning to Rank (LETOR)	30 000+	136	¹

Nos conjuntos de dados "*Energy Efficiency*" e "*Parkinson Telemonitoring*", existem duas variáveis objetivo a prever. No entanto, a previsão de cada uma das variáveis objetivo é independente. Assim sendo, quando o algoritmo é executado para prever uma das variáveis, a outra é desprezada, não sendo esta utilizada na execução do algoritmo.

Quanto ao conjunto de dados "*Wine Quality*", está subdividido em duas partes, uma para cada tipo de vinho (tinto e branco). Para estes não houve qualquer consideração adicional a tomar. Para ambos os conjuntos foi executado o algoritmo da mesma maneira que para o conjunto das espigas.

Tal como introduzido antes no capítulo 1, na secção 1.2, uma das áreas em que é muito importante OI, é na ordenação dos resultados de pesquisas na WEB. O conjunto de dados *Microsoft Learning to Rank (LETOR)* tem com base a avaliação dos resultados de uma pesquisa, baseando-se na relevância de um documento para uma pesquisa.

Na tabela 4.3 é referido que este conjunto de dados tem mais de 30 000 instâncias e com 136 atributos caracterizantes de cada instância, o que faz com que este conjunto de dados seja muito grande. Em R² a paralelização de funções ou processos não é executada por defeito. Apesar de já existirem alguns módulos para R que permitem a paralelização para alguns tipos de funções (nomeadamente para VC), estes não eram possíveis de aplicar neste caso. Para cada uma das abordagens, no processo criação de modelos são usadas as mesmas pastas, isto é, a pasta 1 em LW é igual à pasta 1 em PW, pasta 2 em LW é igual à pasta 2 em PW e assim sucessivamente.

A tentativa de computação dos resultados utilizando só um núcleo do processador demorou muito tempo e não chegou a concluir (passadas 24 horas de execução). Para contornar este problema, foi utilizado uma parte dos dados em vez da totalidade do conjunto. Este conjunto de

¹Este conjunto de dados foi obtido da página web microsoft através do endereço: <http://research.microsoft.com/en-us/projects/mslr/>

²Linguagem de programação para computação estatística. Mais informação disponível no endereço <http://www.r-project.org/>

Resultados

dados está previamente dividido 5 grupos de dados e cada um destes em pastas de treino, validação e teste.

Os resultados que são apresentados abaixo na tabela 4.4 para o conjunto de dados LETOR, foram obtidos com 1/100 parte dos dados para cada um dos 5 conjuntos predefinidos, selecionando-se respectivamente 1/100 parte dos dados em cada pasta de treino, teste e validação.

4.3.2 Resumo dos resultados

Na tabela 4.4 são apresentados os valores médios para cada abordagem, em cada corrida do algoritmo, para os conjuntos de dados apresentados anteriormente em 4.3.1. Tal como para o conjunto das espigas, apresentado anteriormente em 4.2, são apresentadas tabelas e gráficos de variação dos métodos de avaliação correspondentes a cada um dos conjuntos de dados. São também apresentadas tabelas com o número médio de variáveis escolhidas por cada abordagem. Os detalhes para estes conjuntos de dados anteriormente apresentados, estão presentes no anexo A.

Tabela 4.4: Resumo de resultados

	LW DP		LW rank		Pointwise	
	NDCG treino	NDCG teste	NDCG treino	NDCG teste	CoV treino	NDCG teste
Espigas	0.979	0.946	0.961	0.953	0.095	0.960
Wine (white)	0.937	0.928	0.839	0.929	0.128	0.932
Wine (red)	0.956	0.943	0.873	0.943	0.114	0.946
LETOR	0.925	0.925	0.955	0.916	0.024	0.652
Parkinson (motor)	0.852	0.804	0.797	0.796	0.338	0.793
Parkinson (total)	0.828	0.776	0.763	0.778	0.316	0.777
Energy	1.000	1.000	1.000	1.000	0.114	0.974
Efficiency (Y1)						
Energy	1.000	1.000	1.000	1.000	0.119	0.952
Efficiency (Y2)						
Médias		0.911		0.909		0.861

Tabela 4.5: Número médio de variáveis selecionadas cada abordagem nos conjuntos de dados

	LW DP	LW RANK	PW
Espigas	6	6.96	12.86
Wine White	5.36	4.48	7.86
Wine Red	5.78	5.76	7.14
Parkinson (TOTAL)	5.4	5.3	7.16
Parkinson (MOTOR)	5.92	4.76	8.88
Energy Efficiency (Y1)	1	1	4.92
Energy Efficiency (Y2)	1	1	5.22
LETOR	1	1	53.6

4.4 Conclusões

Para o conjunto das Espigas de Milho, apesar de os resultados para a abordagem PW serem muito parecidas com as outras duas em relação às médias de teste para os modelos treinados, é clara a diferença entre o número de variáveis selecionadas por cada uma das abordagens. Como mostrado na tabela 4.2, a abordagem PW seleciona quase o dobro das variáveis dos modelos LW DP e LW Rank.

Ao analisarmos os gráficos presentes na secção 4.2 ou no anexo A (aqui contendo todas as tabelas de todas as execuções do algoritmo), podemos observar ao comparar as abordagens LW DP e PW, que em geral existe uma variação semelhante nos erros, isto é, se observarmos na abordagem LW DP o comportamento do CoV, este na maioria dos casos é decrescente mesmo sendo a função desta abordagem a otimização dos valores de NDCG.

O facto de os modelos gerados pelas abordagens LW terem menos variáveis selecionadas é um ponto positivo em termos dos concursos das espigas, uma vez que torna o resultado muito mais fácil de interpretar. Comparando os resultados só no conjunto de dados das espigas, temos uma perda média em LW RANK de 0.0073% e LW DP de 0.0146% face a PW. Mas ao mesmo tempo estes escolhem em média respetivamente menos 45.88% e 53.34% variáveis que a abordagem PW. Em 4.3 são apresentados exemplos de modelos para uma corrida do algoritmo. Aqui foi utilizado o conjunto total dos dados, mas sem validação cruzada. Este foi dividido em pastas de treino, validação e teste, com respetivamente 50%, 30%, 20% dos dados. Este exemplo serve para mostrar a diferença de variáveis selecionadas em cada uma das abordagens.

$$\begin{cases}
 \text{LW DP Yield} = 0.4900 * L + 2.168 * DE2 - 3.544 * S + 1.782 * KD \\
 \text{LW RANK Yield} = 8.630 * EW15 - 8.996 * KW15 - 1506.8290 * CW_EW2 \\
 - 133.628 * S - 18.740 * DE4 + 0.316 * Alt1E - 0.0112 * KN \\
 \text{PW Yield} = 0.0186 * EW15 - 6.767 * KD + 8.549e - 05 * Stand \\
 + 0.02007 * SW15 + 0.0024 * Alt + 0.1201 * Fa - 1.0163 * DC4 + 0.4261 * DE3 \\
 + 0.431 * R - 1.944 * CW_EW + 0.029 * MOIST + 0.0061 * KN + 0.0017 * Alt1E
 \end{cases} \quad (4.3)$$

Para os subconjuntos de Wine, o comportamento das funções é muito semelhante ao do conjunto das espigas. As variações na abordagem LW Rank em relação aos valores do CoV são muito semelhantes. A grande diferença está no número médio de variáveis selecionadas. Neste são escolhidas em média um número bastante próximo de variáveis, no entanto este conjunto de dados só tem 12 atributos, o que representa uma grande parte das variáveis selecionadas.

Na tabela 4.6 pode-se observar a correlação entre as variáveis dependentes e as independentes, para o conjunto de dados "Energy Efficiency". Foi usado o método de *Spearman* para calcular a correlação entre as variáveis. Este calcula a correlação entre as variáveis baseando-se na distância entre os dados em vez da sua relação linear (método de *Pearson*). Neste caso particular as variáveis a prever têm uma correlação muito forte com o atributo "X4" e "X5". Nas abordagens LW DP e

Resultados

Tabela 4.6: Matriz de correlação dos dados para *Energy Efficiency*

	X1	X2	X3	X4	X5	X6	X7	X8
Y1	0.622	-0.622	0.471	-0.804	0.861	-0.004	0.322	0.068
Y2	0.651	-0.651	0.415	-0.803	0.864	0.017	0.288	0.046

LW Rank a única variável que é escolhida é o atributo "X5". Como consequência para estas abordagens a média de variáveis escolhidas é 1 (tabela 4.5) e por isso o algoritmo ser tão bem sucedido na ordenação das instâncias deste conjunto de dados. Apesar do algoritmo PW também obter bons resultados, os modelos criados têm uma média de variáveis escolhidas de 4.92 e 5.22 respetivamente para as variáveis Y1 e Y2 deste conjunto de dados.

Para o conjunto de dados LETOR, os resultados apresentados para a abordagem LW são bastante melhores que para PW. Aqui os ganhos em ambas as vertentes do algoritmo criadas, DP e RANK, para LW têm ganhos de 0.273 e 0.303 pontos (tabela 4.4) em relação a PW. Apesar do conjunto de dados ter bastantes atributos caraterísticos para cada pesquisa feita, em LW a média de variáveis escolhida é 1, enquanto para a abordagem PW é 53.6. Em relação à interpretabilidade dos resultados, o modelo LW é claramente mais interpretável, no entanto os valores para NDCG comparados aos outros conjuntos de dados são bastante mais baixos para a avaliação dos modelos.

Uma vez que em ambas as vertentes de LW é unicamente escolhida uma variável, os gráficos comparativos de CoV com NDCG para esta abordagem não são relevantes. No entanto os gráficos relativos à comparação do mesmo para a abordagem PW, é possível observar o comportamento do NDCG aquando da seleção de variáveis (nesta abordagem feita pela SEQ). Nos gráficos presentes na secção B.8.1, para as pastas 1, 2 e 3, as primeiras variáveis selecionadas baixam bastante o erro de previsão do modelo, no entanto, também penalizam em muito a previsão de ordenação das instâncias. Note-se que nestes 3 casos, os valores máximo e mínimo para NDCG, é atingido logo na escolha das primeiras variáveis. Nos gráficos em que não estão presentes as linhas correspondentes a CoV, deve-se a sobreposição das mesmas.

Resultados

Capítulo 5

Conclusões e Trabalho Futuro

5.1 Satisfação dos Objetivos e Trabalho Futuro

Dos problemas observados uma possível melhoria seria, a paralelização do algoritmo. O tempo demorado para processamento da informação para grandes conjuntos de dados, pode ser em muito reduzido com esta alteração.

Uma ponto inovador a desenvolver, é a elaboração de um método de avaliação híbrido, que aquando da criação dos modelos, em vez de usar unicamente uma medida de avaliação do erro, utilize uma combinação destas.

Apesar dos resultados obtidos serem indicadores de bons resultados, é importante compará-los com outros métodos de avaliação ou modelos preditivos. No caso dos modelos interpretáveis podia ser adaptado o algoritmo para MARS. Para o caso de comparação de resultados com modelos não interpretáveis podiam ser utilizadas Redes-Neuronais como proposto em [CQL⁺07, XLW⁺08], utilizando também a abordagem *Pairwise*.

Conclusões e Trabalho Futuro

Referências

- [BL13] K. Bache e M. Lichman. UCI machine learning repository, 2013. URL: <http://archive.ics.uci.edu/ml>.
- [CQL⁺07] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai e Hang Li. Learning to rank: From pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning, ICML '07*, pages 129–136, New York, NY, USA, 2007. ACM. URL: <http://doi.acm.org/10.1145/1273496.1273513>, doi:10.1145/1273496.1273513.
- [HKP11] Jiawei Han, Micheline Kamber e Jian Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011.
- [HNT⁺06] Jason Hill, Erik Nelson, David Tilman, Stephen Polasky e Douglas Tiffany. Environmental, economic, and energetic costs and benefits of biodiesel and ethanol biofuels. *Proceedings of the National Academy of Sciences*, 103(30):11206–11210, 2006.
- [HTF09] Trevor Hastie, Robert Tibshirani e Jerome Friedman. *The Elements of Statistical Learning*. New York: Springer 2nd edition, 2009.
- [JK02] Kalervo Järvelin e Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.
- [XLW⁺08] Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang e Hang Li. Listwise approach to learning to rank: theory and algorithm. In *Proceedings of the 25th international conference on Machine learning*, pages 1192–1199. ACM, 2008.

REFERÊNCIAS

Anexo A

Tabelas de resultados para os casos genéricos

A.1 Wine white

Tabela A.1: Resultados para conjunto dos dados das Wine white

	LW DP		LW rank		Pointwise	
	ndcg treino	ndcg teste	ndcg treino	ndcg teste	erro de treino	ndcg teste
run 1	0.939	0.929	0.825	0.926	0.128	0.932
run 2	0.936	0.925	0.846	0.927	0.128	0.934
run 3	0.937	0.928	0.841	0.929	0.127	0.933
run 4	0.937	0.929	0.840	0.932	0.127	0.929
run 5	0.935	0.927	0.845	0.929	0.130	0.934
Médias		0.928		0.929		0.932

A.2 Wine red

Tabela A.2: Resultados para conjunto dos dados das Wine white

	LW DP		LW rank		Pointwise	
	ndcg treino	ndcg teste	ndcg treino	ndcg teste	erro de treino	ndcg teste
run 1	0.960	0.946	0.866	0.945	0.112	0.946
run 2	0.956	0.945	0.881	0.944	0.115	0.945
run 3	0.951	0.940	0.857	0.947	0.113	0.945
run 4	0.958	0.945	0.880	0.941	0.113	0.946
run 5	0.953	0.940	0.880	0.940	0.115	0.947
Médias		0.943		0.943		0.946

A.3 Parkinson (MOTOR)

Tabela A.3: Resultados para conjunto dos dados das Parkinson (MOTOR)

	LW DP		LW rank		Pointwise	
	ndcg treino	ndcg teste	ndcg treino	ndcg teste	erro de treino	ndcg teste
run 1	0.845	0.825	0.789	0.791	0.351	0.787
run 2	0.874	0.787	0.802	0.826	0.340	0.790
run 3	0.835	0.816	0.789	0.756	0.344	0.779
run 4	0.852	0.794	0.802	0.820	0.336	0.804
run 5	0.856	0.797	0.801	0.786	0.320	0.803
Médias		0.804		0.796		0.793

A.4 Parkinson (TOTAL)

Tabela A.4: Resultados para conjunto dos dados das Parkinson (TOTAL)

	LW DP		LW rank		Pointwise	
	ndcg treino	ndcg teste	ndcg treino	ndcg teste	erro de treino	ndcg teste
run 1	0.839	0.781	0.768	0.794	0.304	0.774
run 2	0.813	0.777	0.763	0.771	0.304	0.785
run 3	0.820	0.774	0.768	0.776	0.331	0.783
run 4	0.836	0.794	0.768	0.791	0.321	0.767
run 5	0.834	0.757	0.751	0.759	0.318	0.776
Médias		0.776		0.778		0.777

A.5 Energy efficiency (Y1)

Tabela A.5: Resultados para conjunto dos dados das Energy efficiency (Y1)

	LW DP		LW rank		Pointwise	
	ndcg treino	ndcg teste	ndcg treino	ndcg teste	erro de treino	ndcg teste
run 1	1.000	1.000	1.000	1.000	0.117	0.976
run 2	1.000	1.000	1.000	1.000	0.113	0.975
run 3	1.000	1.000	1.000	1.000	0.112	0.975
run 4	1.000	1.000	1.000	1.000	0.115	0.973
run 5	1.000	1.000	1.000	1.000	0.112	0.973
Médias		1.000		1.000		0.974

A.6 Energy efficiency (Y2)

Tabela A.6: Resultados para conjunto dos dados das Energy efficiency (Y2)

	LW DP		LW rank		Pointwise	
	ndcg treino	ndcg teste	ndcg treino	ndcg teste	erro de treino	ndcg teste
run 1	1.000	1.000	1.000	1.000	0.120	0.945
run 2	1.000	1.000	1.000	1.000	0.114	0.953
run 3	1.000	1.000	1.000	1.000	0.121	0.956
run 4	1.000	1.000	1.000	1.000	0.124	0.954
run 5	1.000	1.000	1.000	1.000	0.116	0.950
Médias		1.000		1.000		0.952

A.7 LETOR

Tabela A.7: Resultados para conjunto dos dados das LETOR

	LW DP		LW rank		Pointwise	
	ndcg treino	ndcg teste	ndcg treino	ndcg teste	erro de treino	ndcg teste
fold 1	0.906	0.923	0.945	0.923	0.024	0.649
fold 2	0.951	0.952	0.971	0.952	0.024	0.649
fold 3	0.929	0.908	0.953	0.908	0.023	0.671
fold 4	0.919	0.924	0.956	0.882	0.024	0.652
fold 5	0.921	0.916	0.948	0.916	0.023	0.641
Médias		0.925		0.916		0.652

Tabelas de resultados para os casos genéricos

Anexo B

Gráficos CoV vs NDCG para os Conjuntos de dados

B.1 Espigas

B.1.1 LW DP

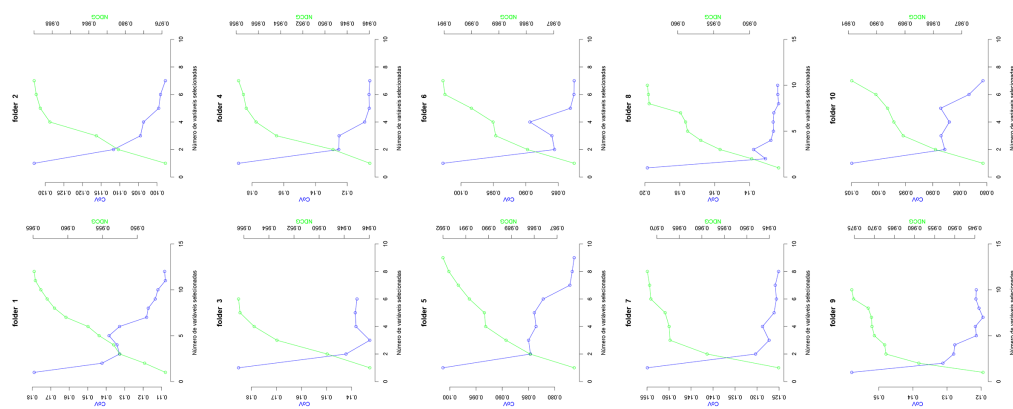


Figura B.1: NDCG vs CoV, LW DP, 1ª execução

Gráficos CoV vs NDCG para os Conjuntos de dados

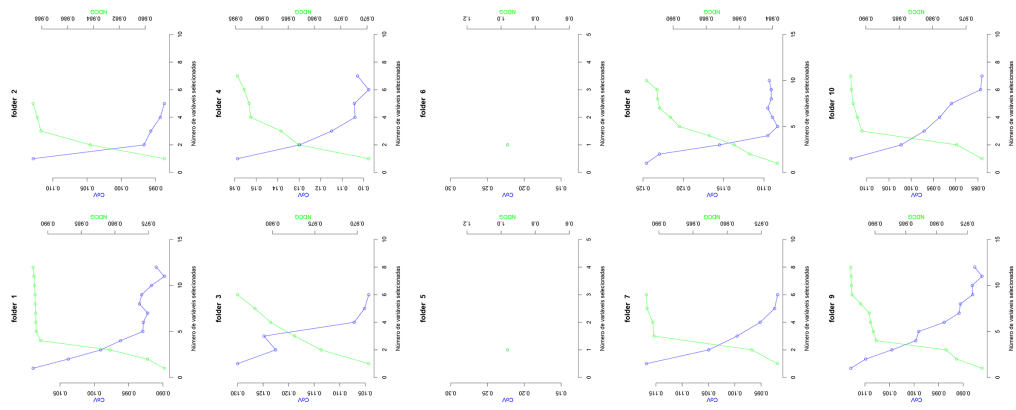


Figura B.2: NDCG vs CoV, LW DP, 2ª execução

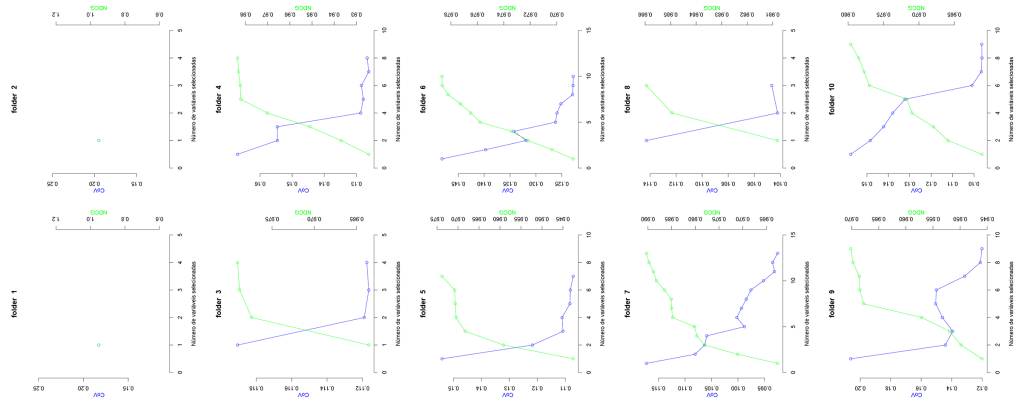


Figura B.3: NDCG vs CoV, LW DP, 3ª execução

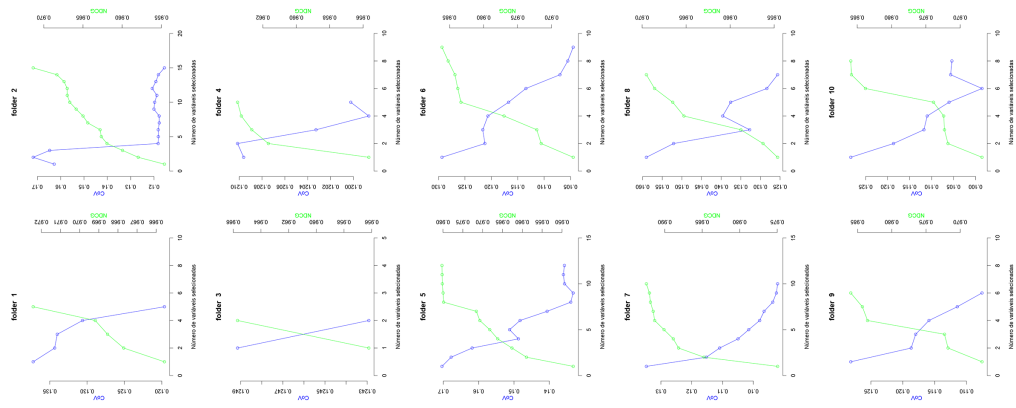


Figura B.4: NDCG vs CoV, LW DP, 4ª execução

Gráficos CoV vs NDCG para os Conjuntos de dados

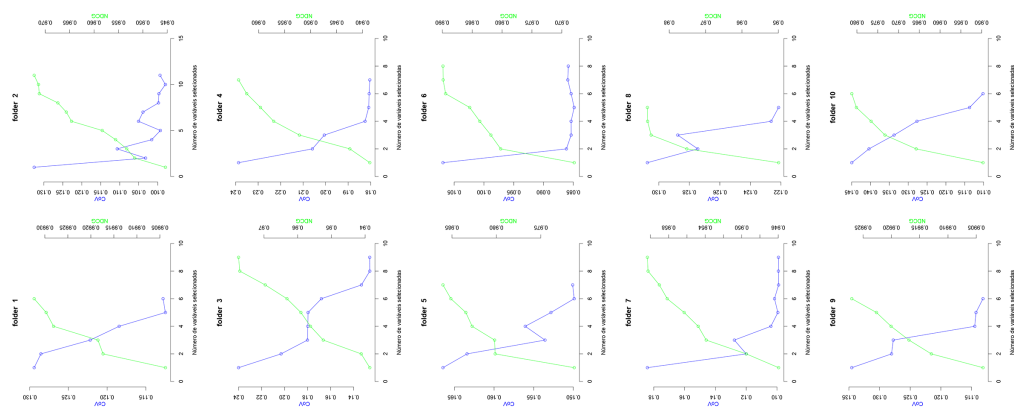


Figura B.5: NDCG vs CoV, LW DP, 5ª execução

B.1.2 PW

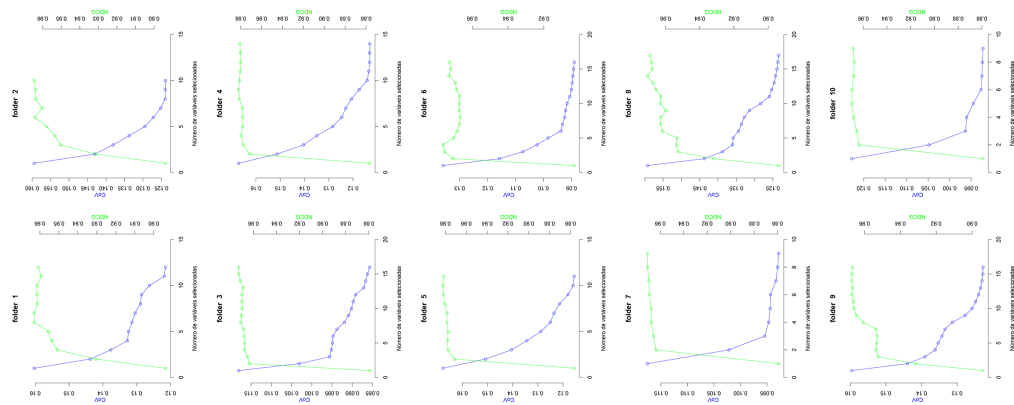


Figura B.6: NDCG vs CoV, PW, 1ª execução

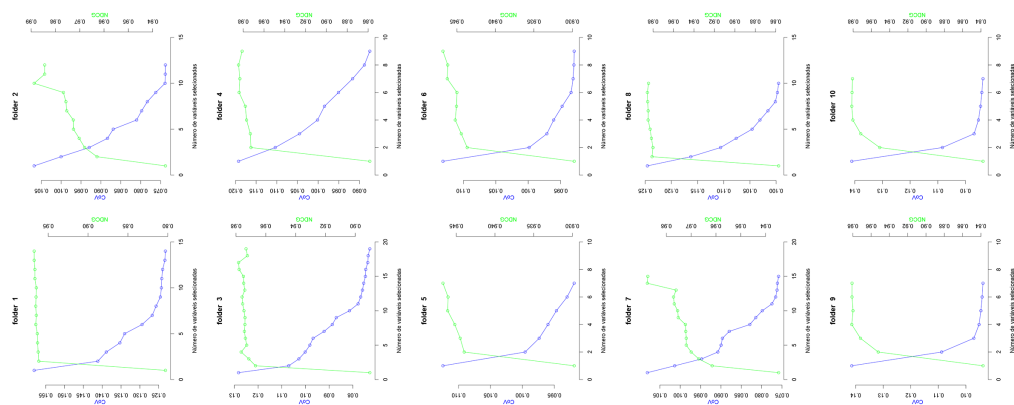


Figura B.7: NDCG vs CoV, PW, 2ª execução

Gráficos CoV vs NDCG para os Conjuntos de dados

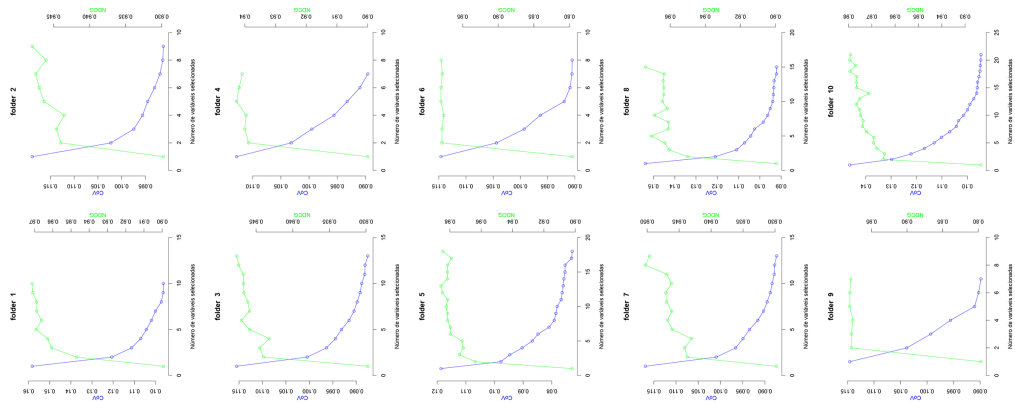


Figura B.8: NDCG vs CoV, PW, 3ª execução

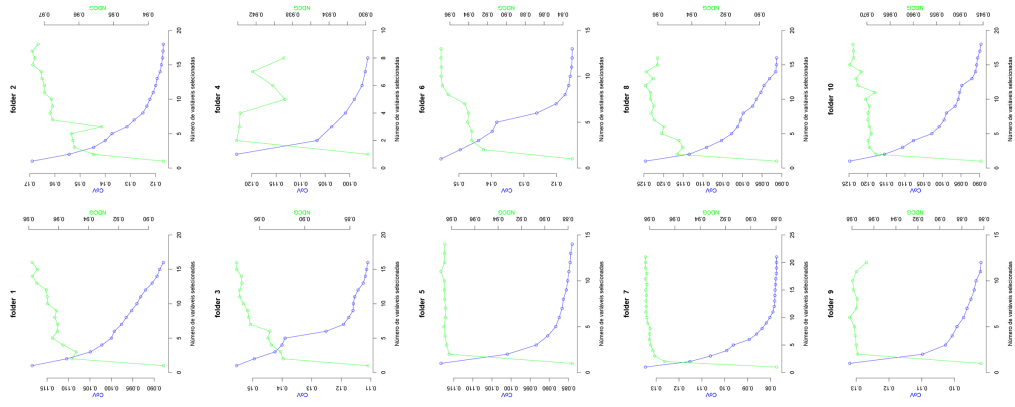


Figura B.9: NDCG vs CoV, PW, 4ª execução

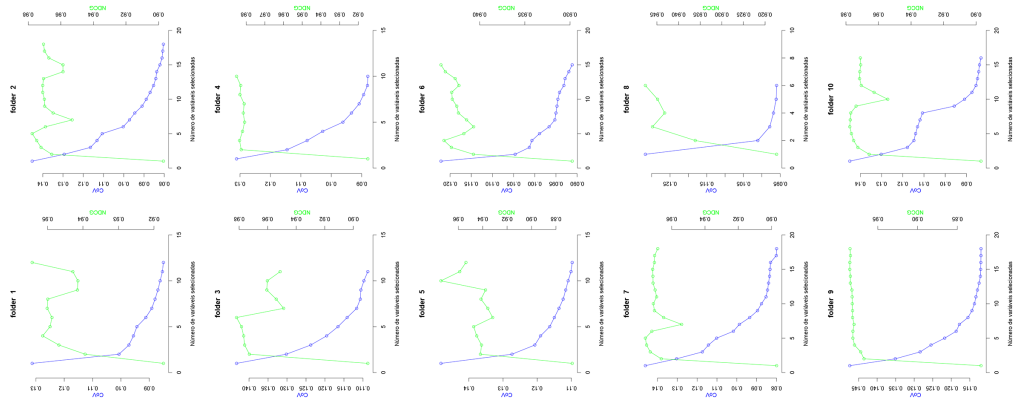


Figura B.10: NDCG vs CoV, PW, 5ª execução

B.2 Wine White

B.2.1 LW DP

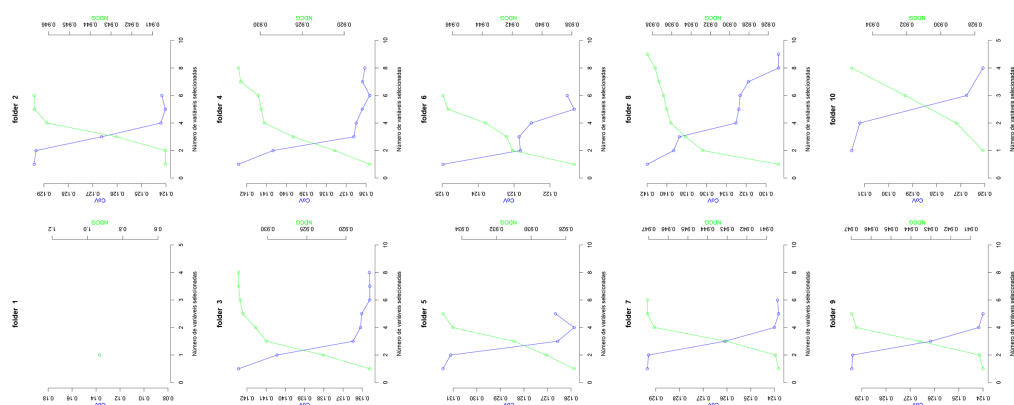


Figura B.11: NDCG vs CoV, LW DP, 1ª execução

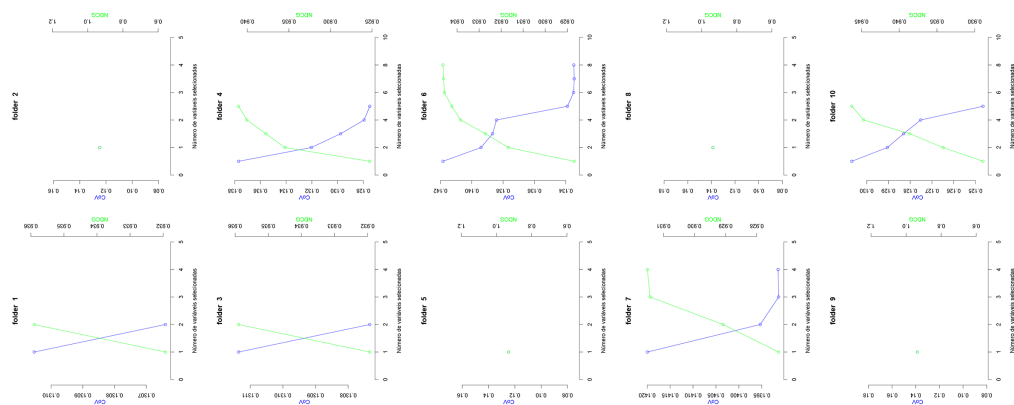


Figura B.12: NDCG vs CoV, LW DP, 2ª execução

Gráficos CoV vs NDCG para os Conjuntos de dados

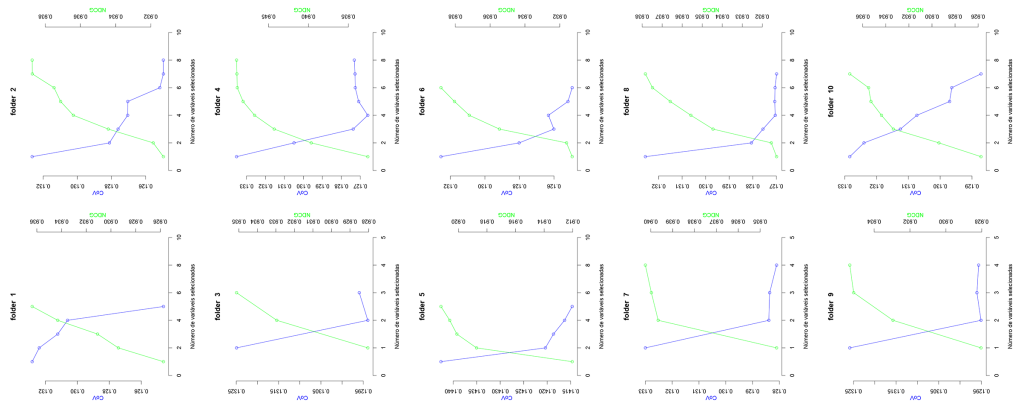


Figura B.13: NDCG vs CoV, LW DP, 3ª execução

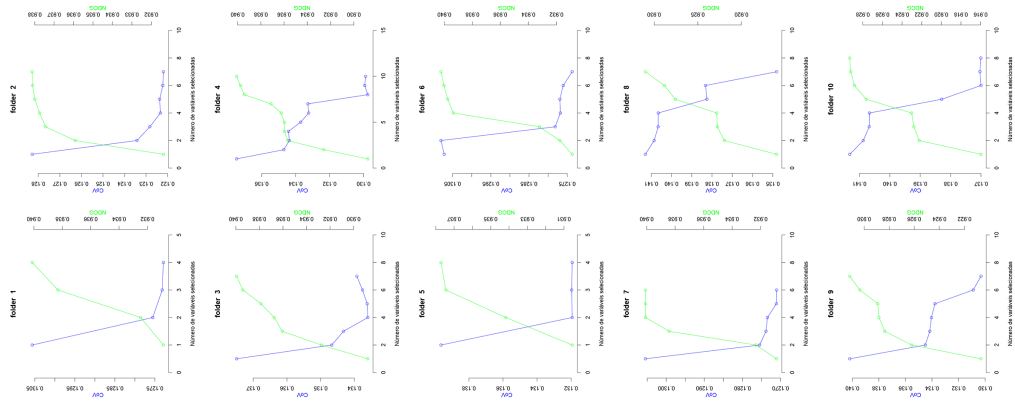


Figura B.14: NDCG vs CoV, LW DP, 4ª execução

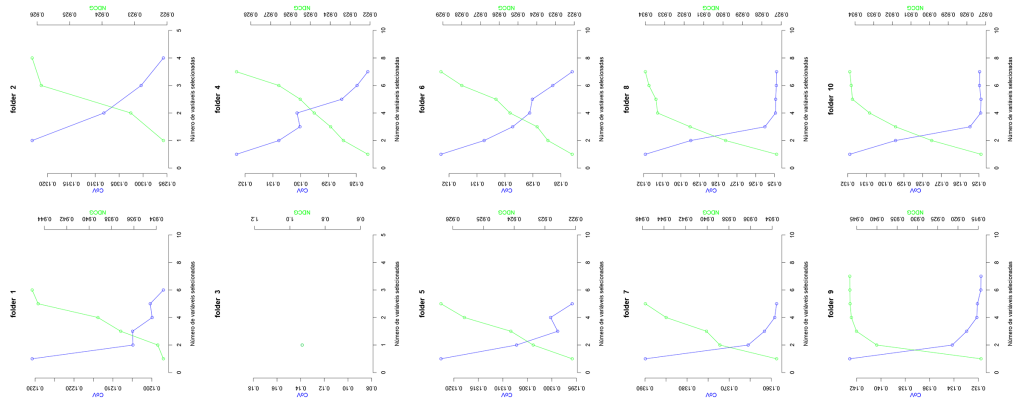


Figura B.15: NDCG vs CoV, LW DP, 5ª execução

Gráficos CoV vs NDCG para os Conjuntos de dados

B.2.2 PW

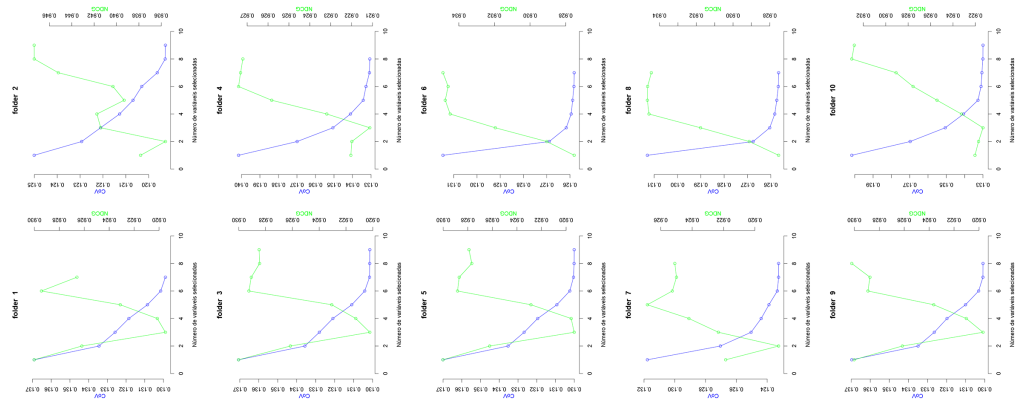


Figura B.16: NDCG vs CoV, PW, 1ª execução

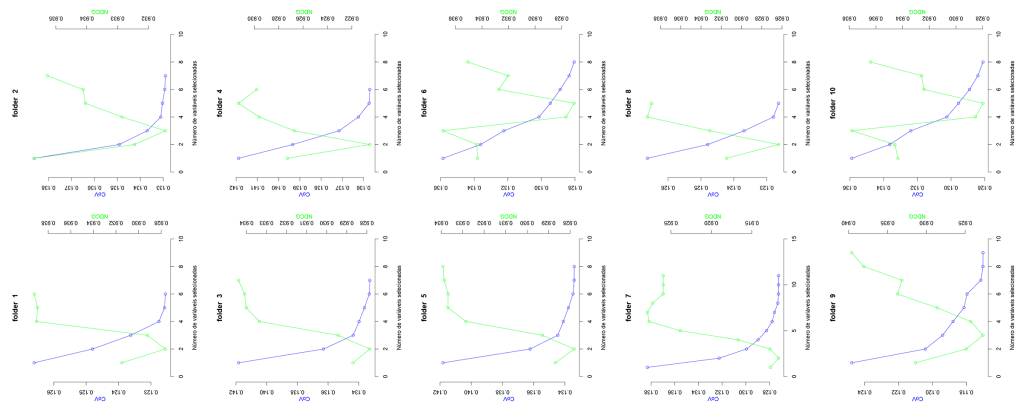


Figura B.17: NDCG vs CoV, PW, 2ª execução

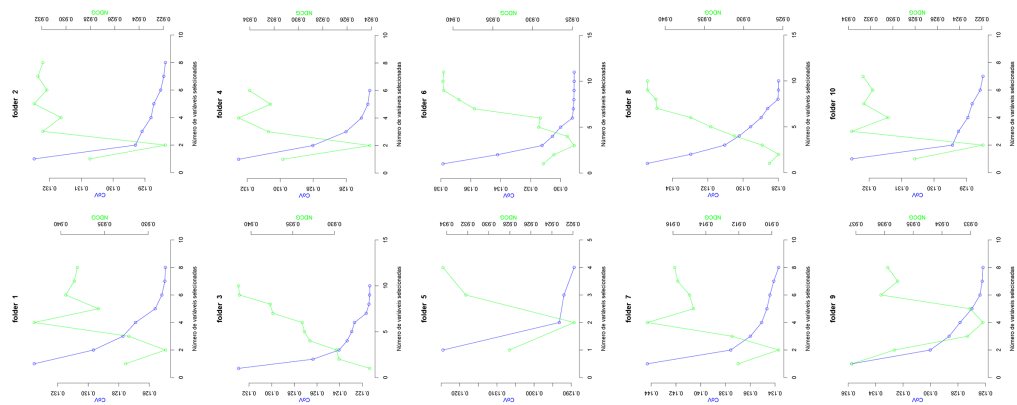


Figura B.18: NDCG vs CoV, PW, 3ª execução

Gráficos CoV vs NDCG para os Conjuntos de dados

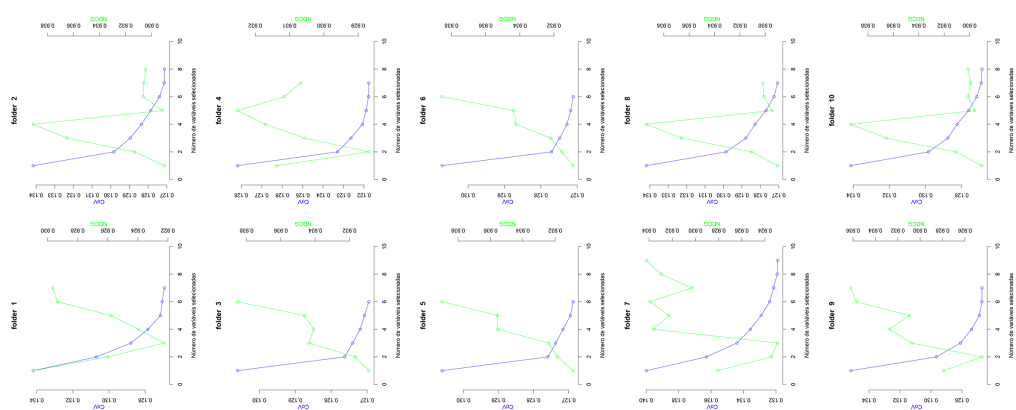


Figura B.19: NDCG vs CoV, PW, 4ª execução

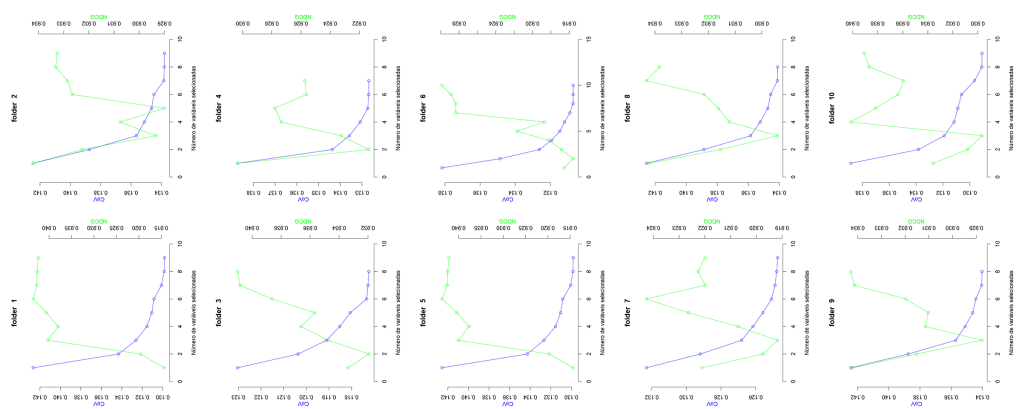


Figura B.20: NDCG vs CoV, PW, 5ª execução

B.3 Wine Red

B.3.1 LW DP

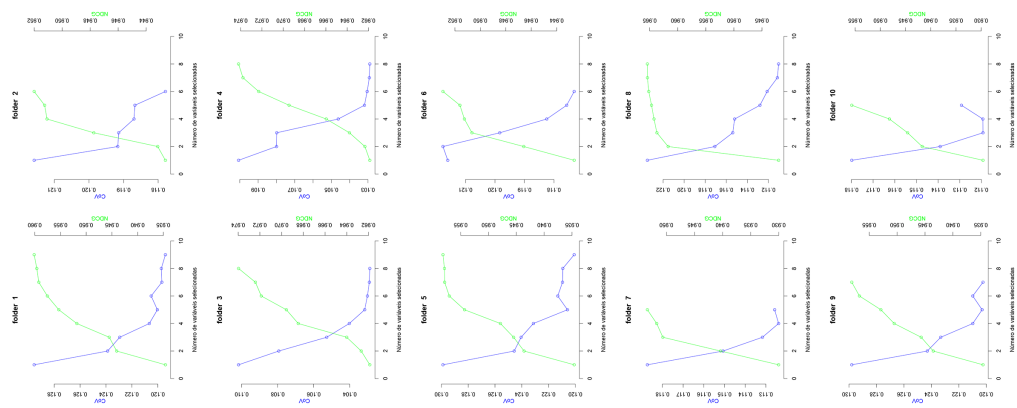


Figura B.21: NDCG vs CoV, LW DP, 1ª execução

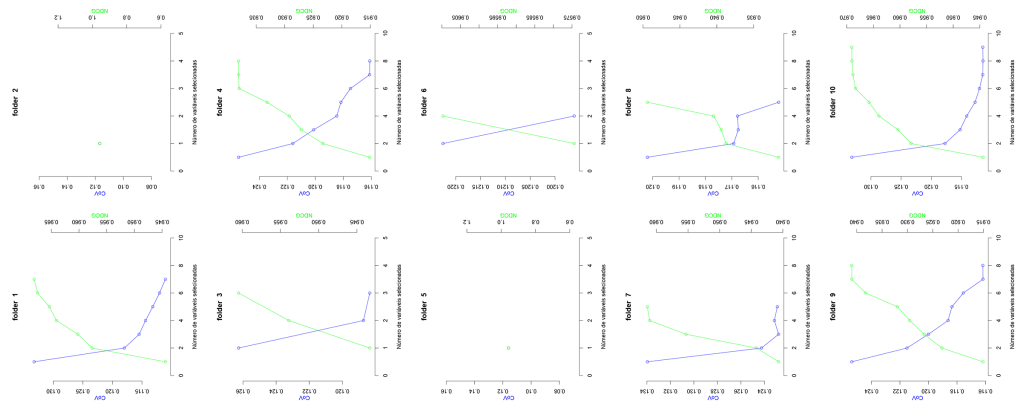


Figura B.22: NDCG vs CoV, LW DP, 2ª execução

Gráficos CoV vs NDCG para os Conjuntos de dados

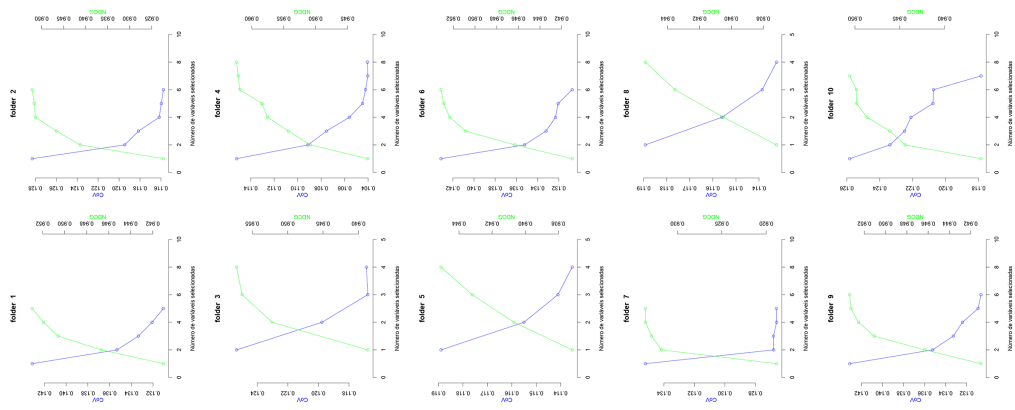


Figura B.23: NDCG vs CoV, LW DP, 3ª execução

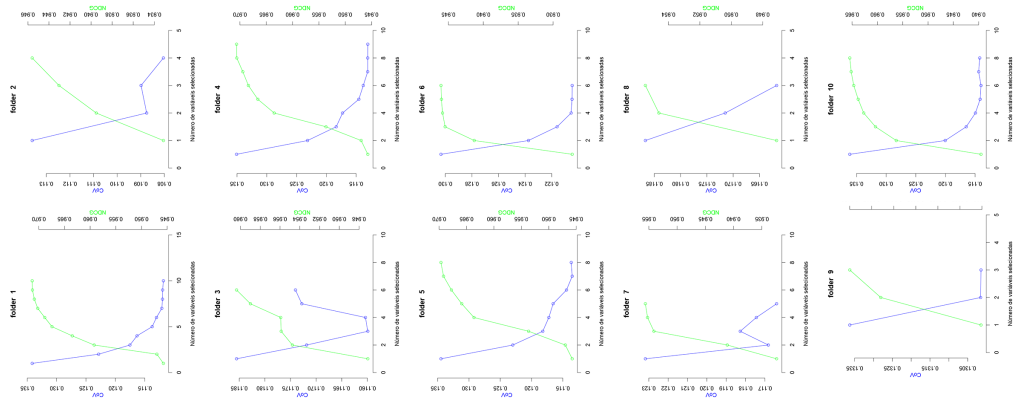


Figura B.24: NDCG vs CoV, LW DP, 4ª execução

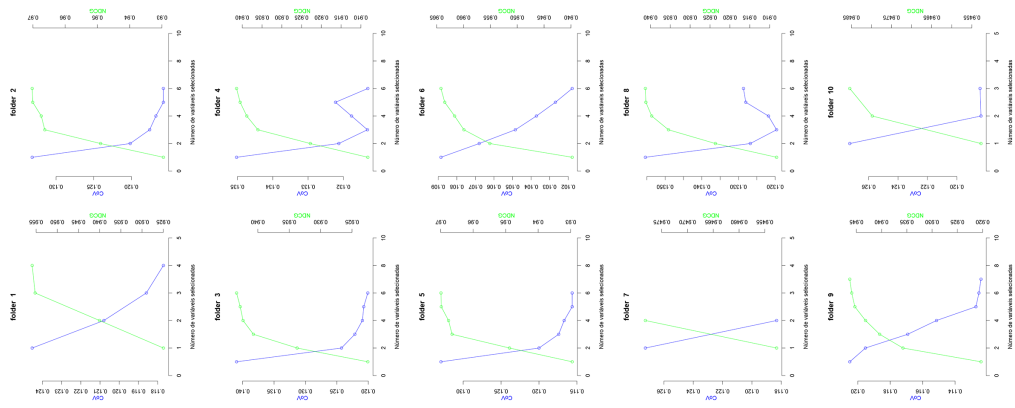


Figura B.25: NDCG vs CoV, LW DP, 5ª execução

Gráficos CoV vs NDCG para os Conjuntos de dados

B.3.2 PW

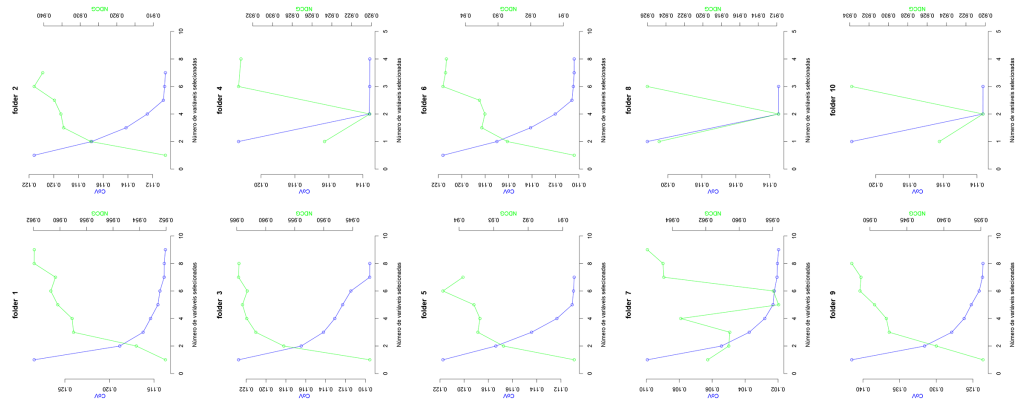


Figura B.26: NDCG vs CoV, PW, 1ª execução

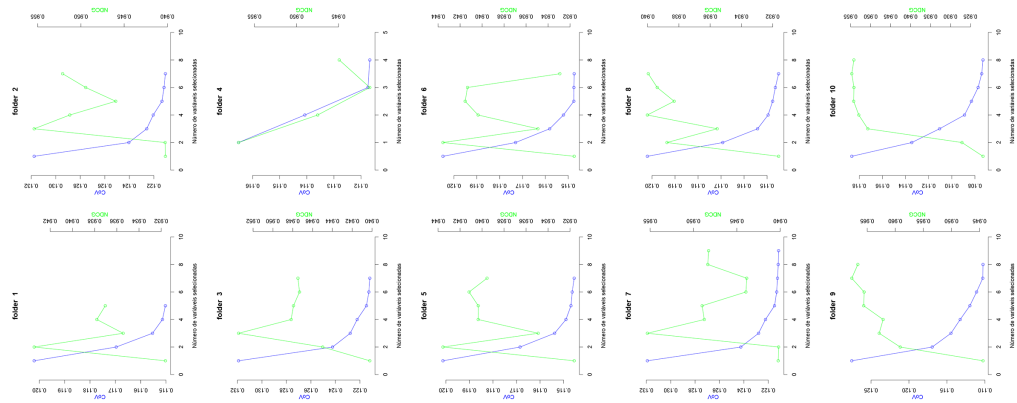


Figura B.27: NDCG vs CoV, PW, 2ª execução

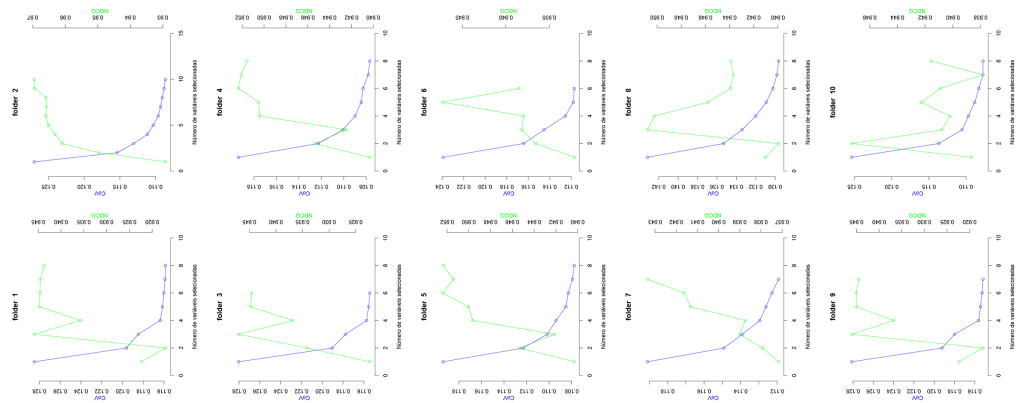


Figura B.28: NDCG vs CoV, PW, 3ª execução

Gráficos CoV vs NDCG para os Conjuntos de dados

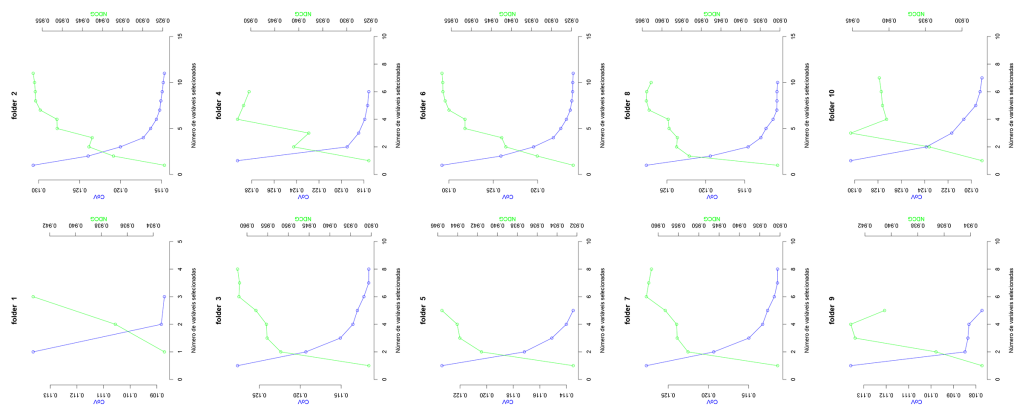


Figura B.29: NDCG vs CoV, PW, 4ª execução

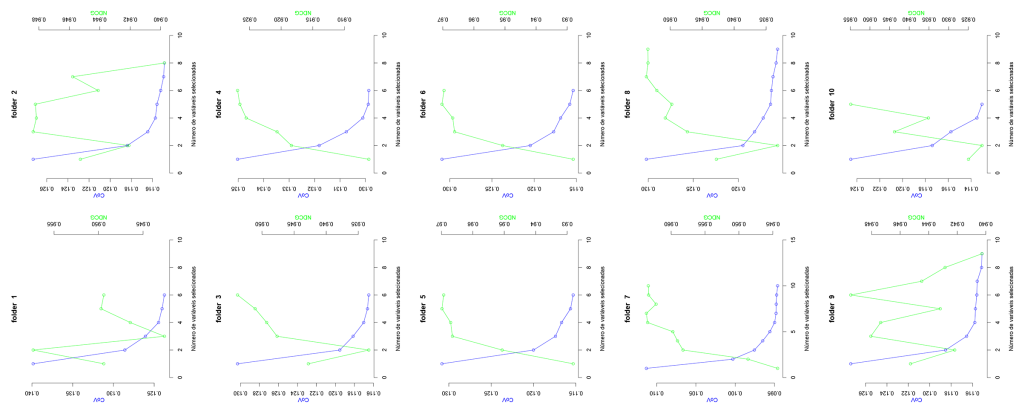


Figura B.30: NDCG vs CoV, PW, 5ª execução

B.4 Parkinson (Total)

B.4.1 LW DP

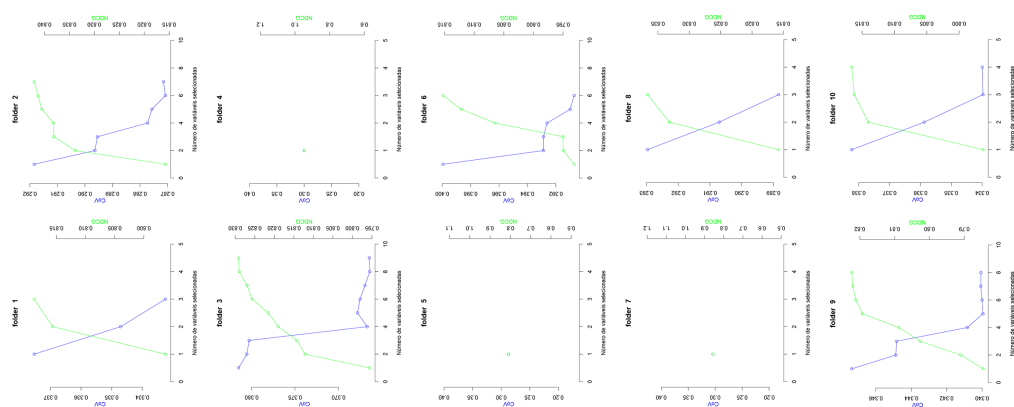


Figura B.31: NDCG vs CoV, LW DP, 1ª execução

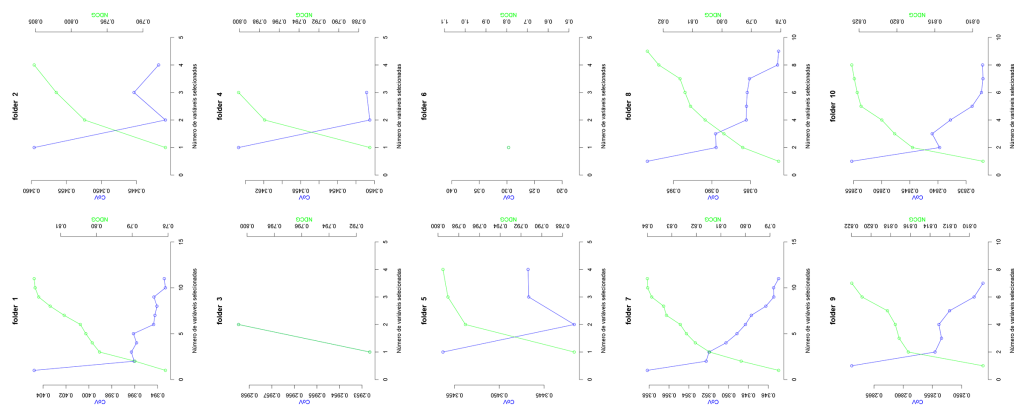


Figura B.32: NDCG vs CoV, LW DP, 2ª execução

Gráficos CoV vs NDCG para os Conjuntos de dados

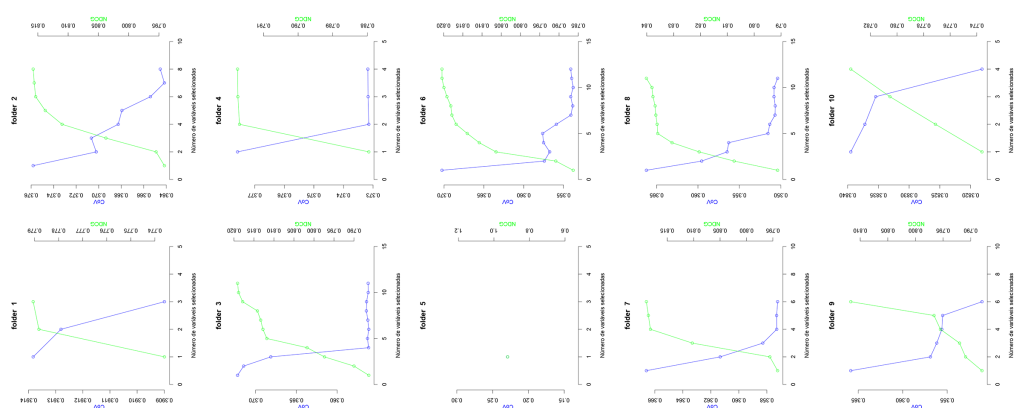


Figura B.33: NDCG vs CoV, LW DP, 3ª execução

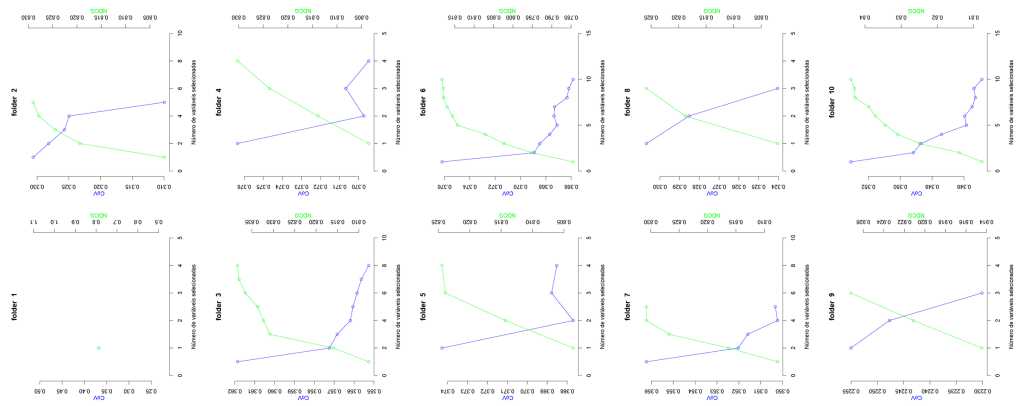


Figura B.34: NDCG vs CoV, LW DP, 4ª execução

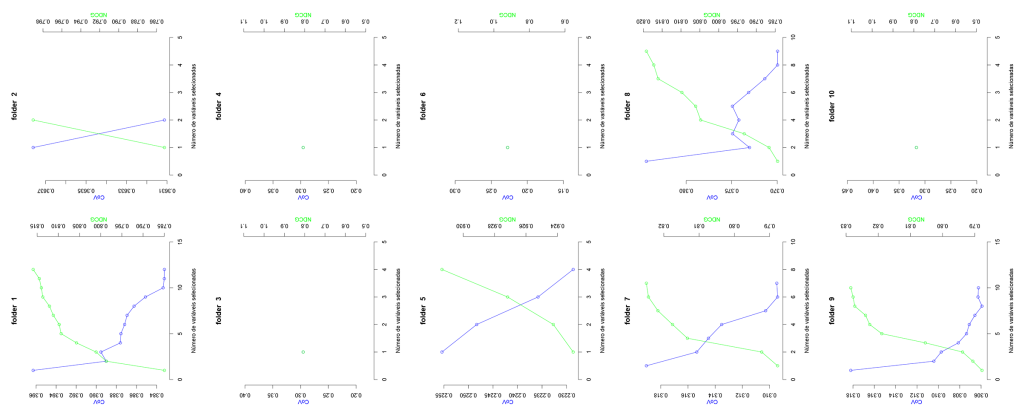


Figura B.35: NDCG vs CoV, LW DP, 5ª execução

Gráficos CoV vs NDCG para os Conjuntos de dados

B.4.2 PW

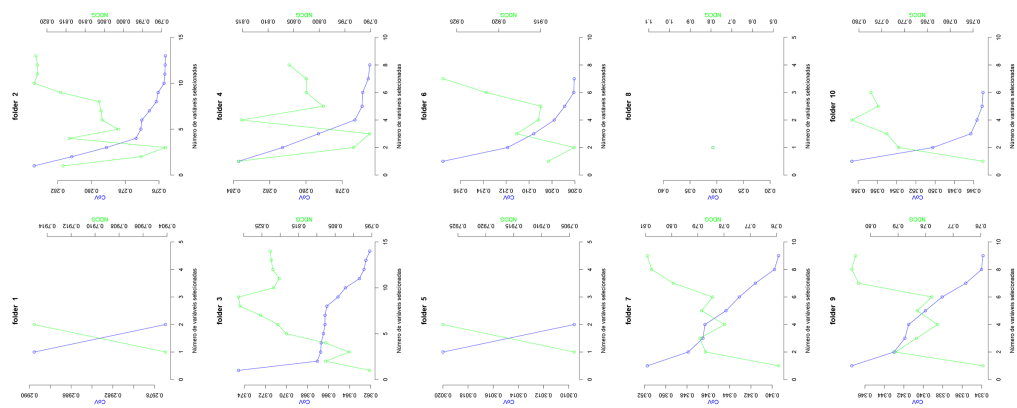


Figura B.36: NDCG vs CoV, PW, 1ª execução

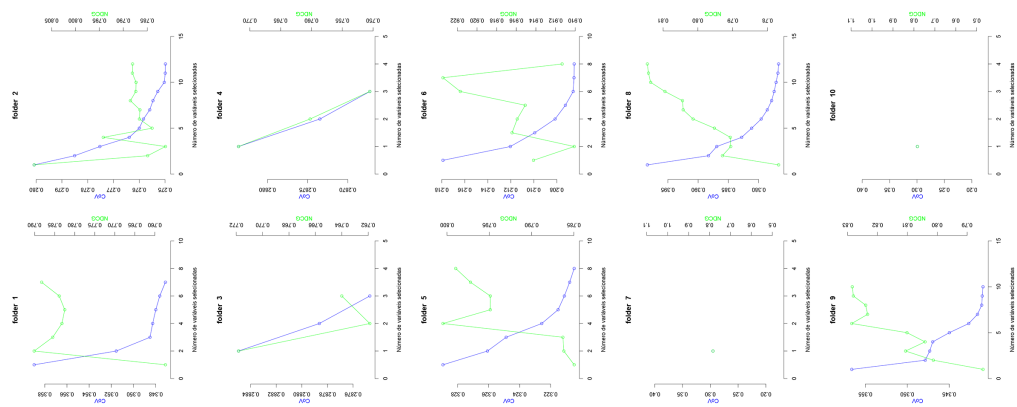


Figura B.37: NDCG vs CoV, PW, 2ª execução

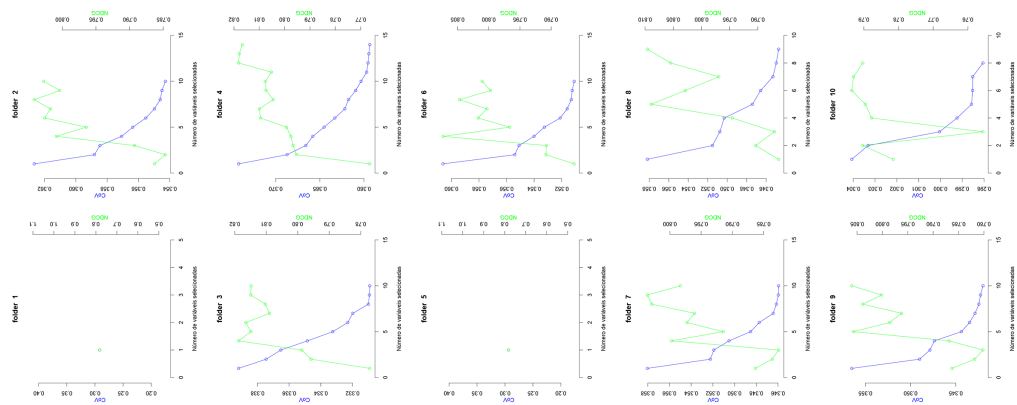


Figura B.38: NDCG vs CoV, PW, 3ª execução

Gráficos CoV vs NDCG para os Conjuntos de dados

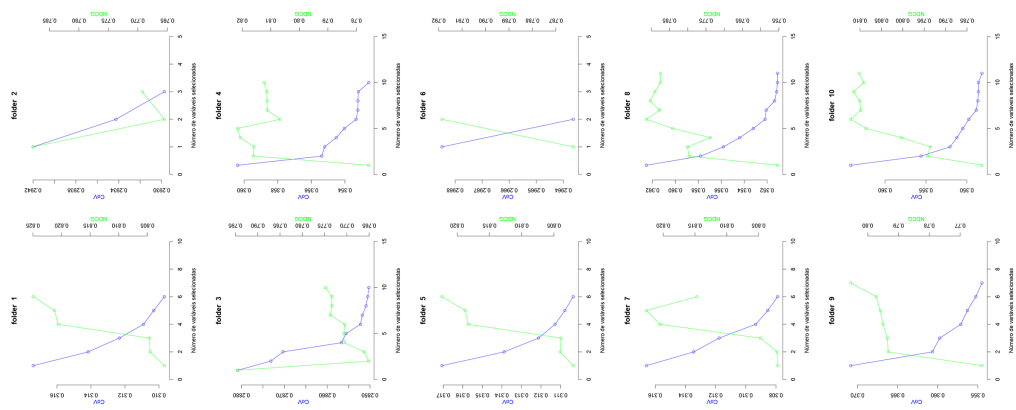


Figura B.39: NDCG vs CoV, PW, 4ª execução

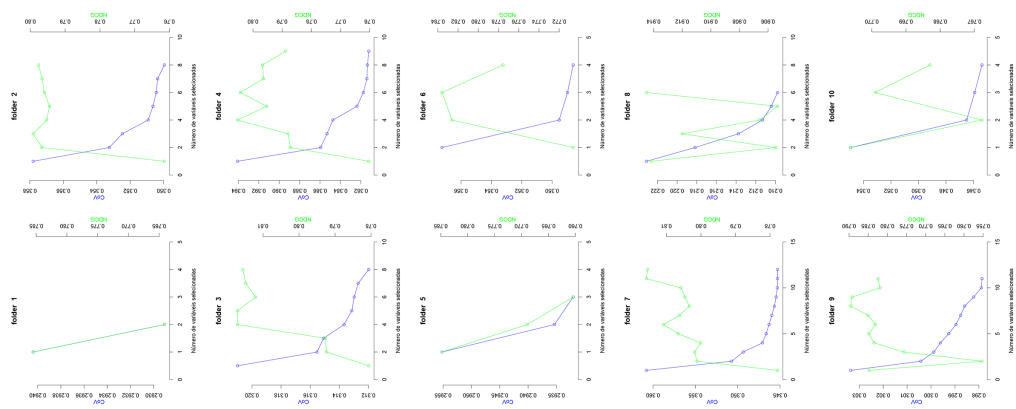


Figura B.40: NDCG vs CoV, PW, 5ª execução

B.5 Parkinson (Motor)

B.5.1 LW DP

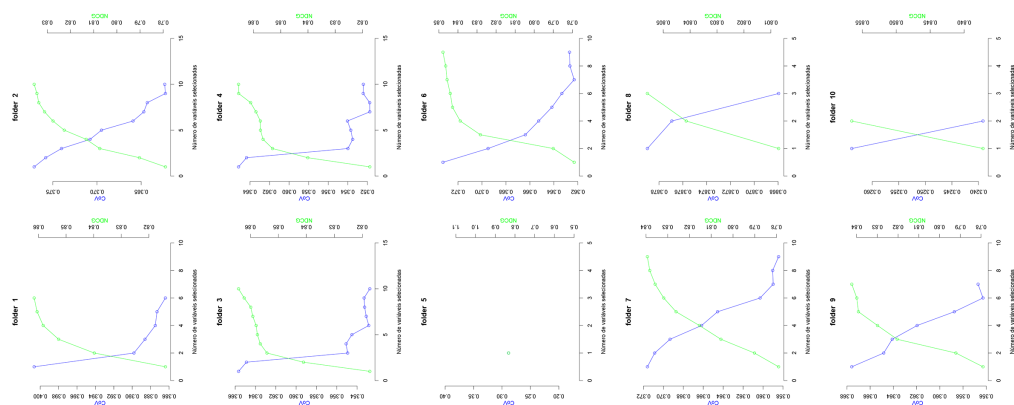


Figura B.41: NDCG vs CoV, LW DP, 1ª execução

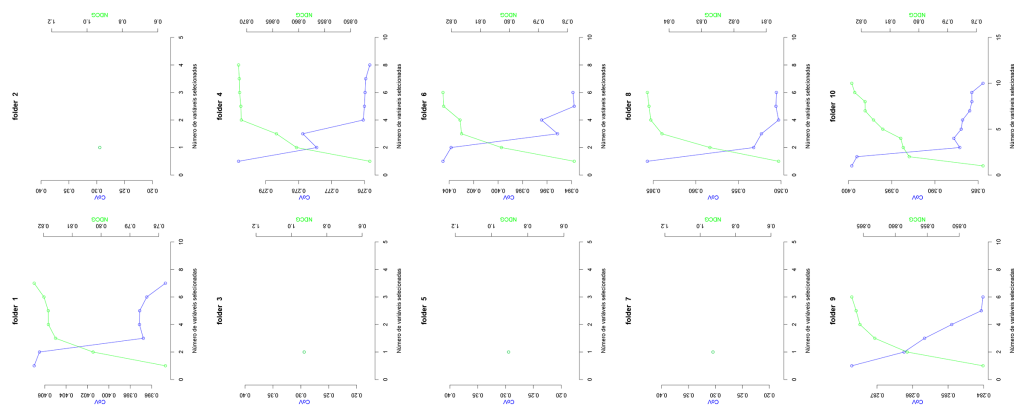


Figura B.42: NDCG vs CoV, LW DP, 2ª execução

Gráficos CoV vs NDCG para os Conjuntos de dados

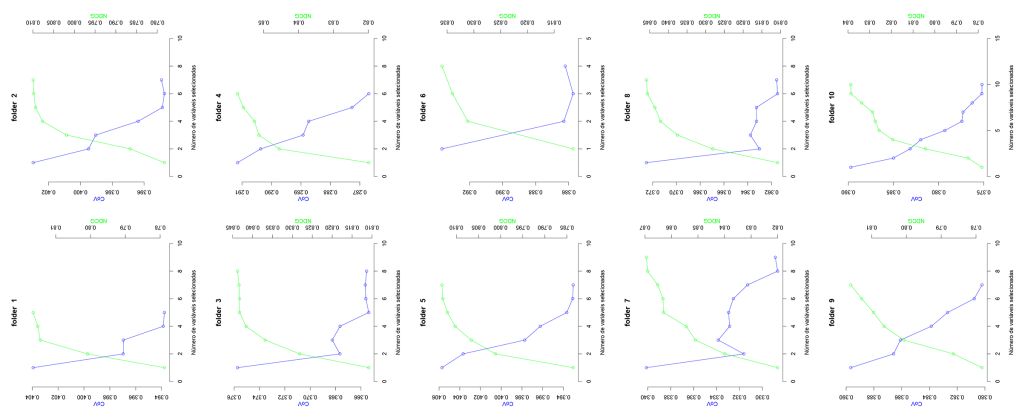


Figura B.43: NDCG vs CoV, LW DP, 3ª execução

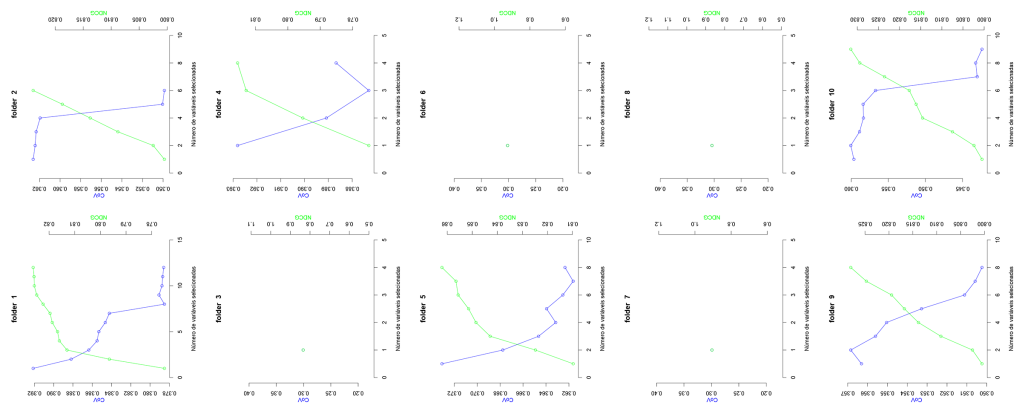


Figura B.44: NDCG vs CoV, LW DP, 4ª execução

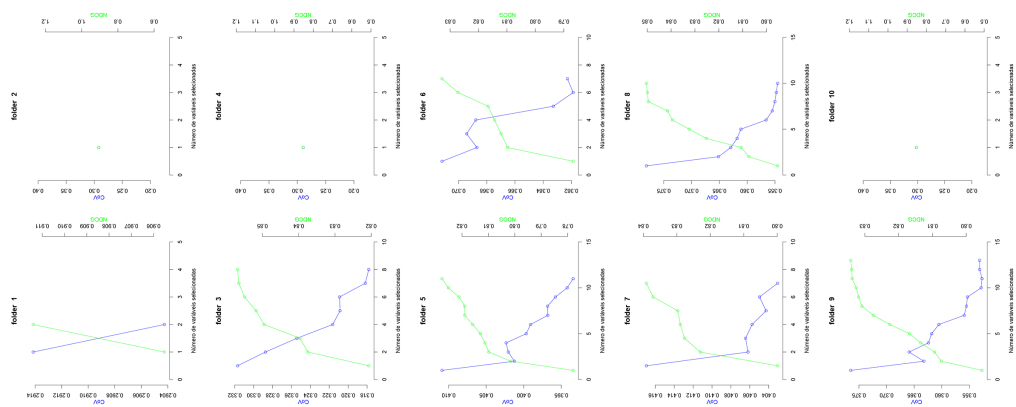


Figura B.45: NDCG vs CoV, LW DP, 5ª execução

Gráficos CoV vs NDCG para os Conjuntos de dados

B.5.2 PW

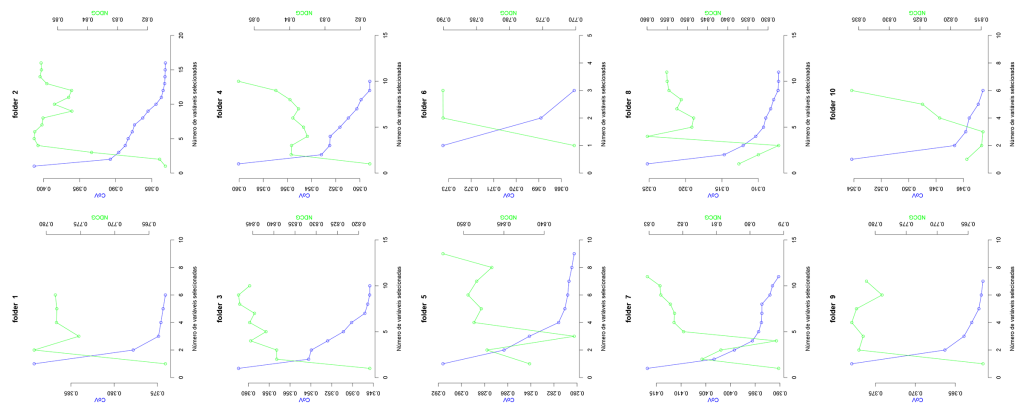


Figura B.46: NDCG vs CoV, PW, 1ª execução

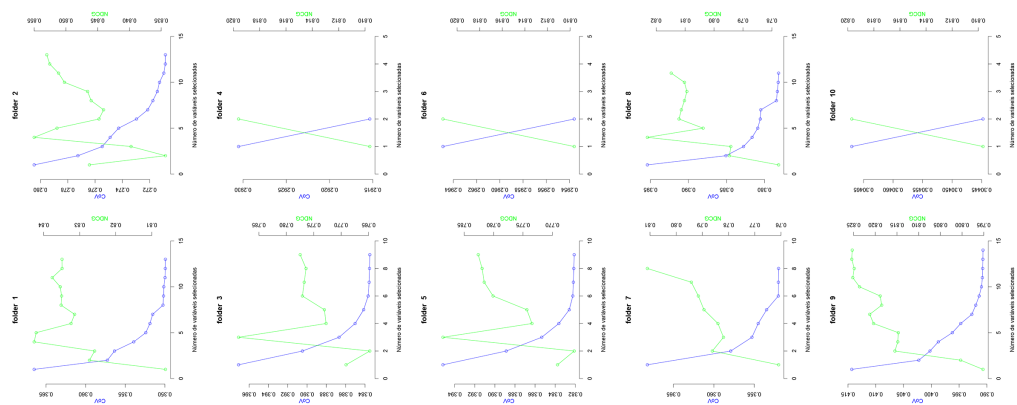


Figura B.47: NDCG vs CoV, PW, 2ª execução

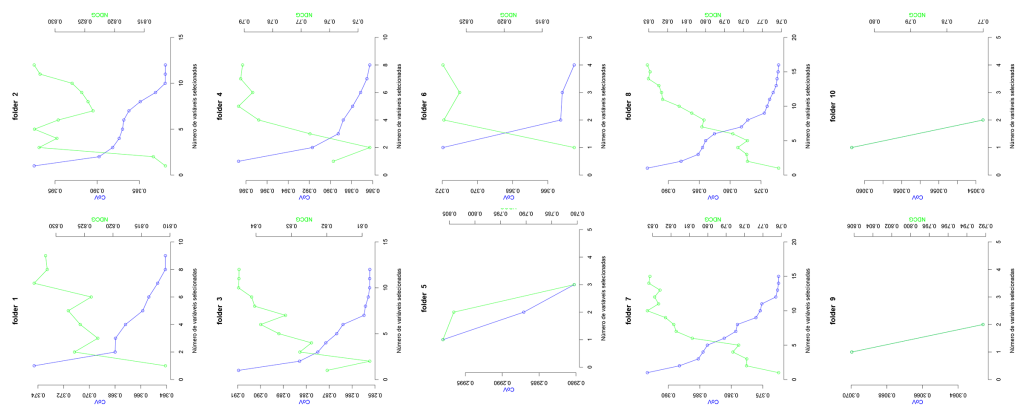


Figura B.48: NDCG vs CoV, PW, 3ª execução

Gráficos CoV vs NDCG para os Conjuntos de dados

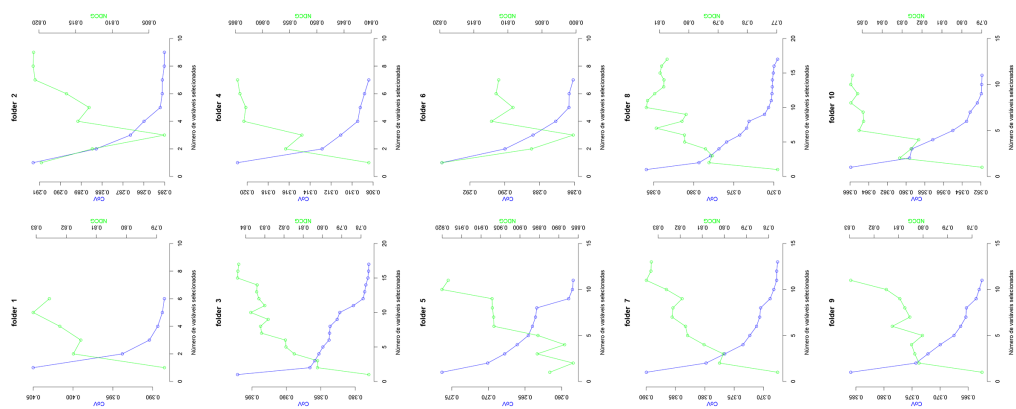


Figura B.49: NDCG vs CoV, PW, 4ª execução

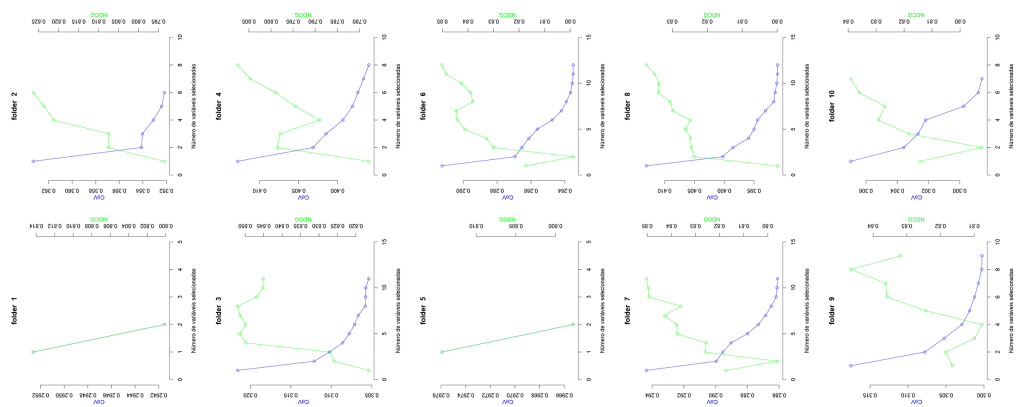


Figura B.50: NDCG vs CoV, PW, 5ª execução

B.6 Energy Efficiency(Y1)

B.6.1 LW DP

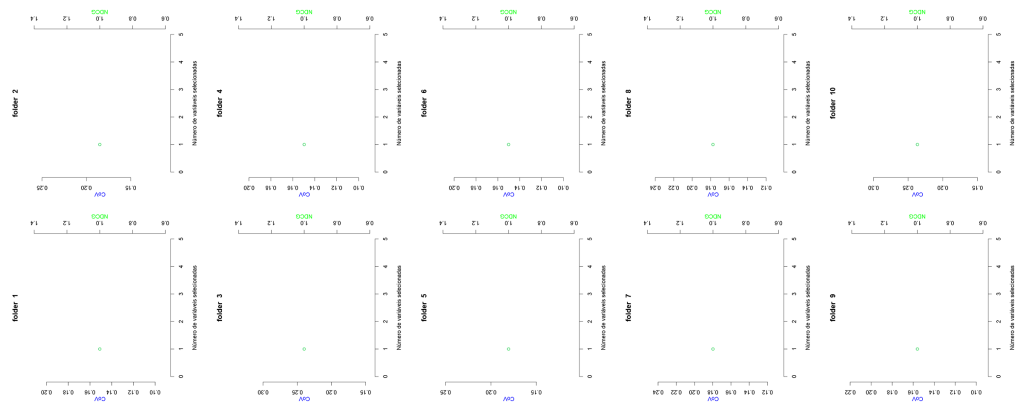


Figura B.51: NDCG vs CoV, LW DP, 1ª execução

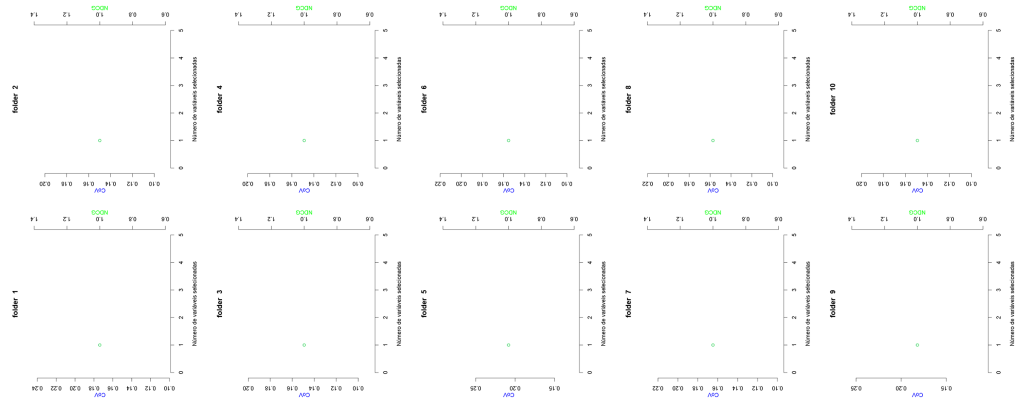


Figura B.52: NDCG vs CoV, LW DP, 2ª execução

Gráficos CoV vs NDCG para os Conjuntos de dados

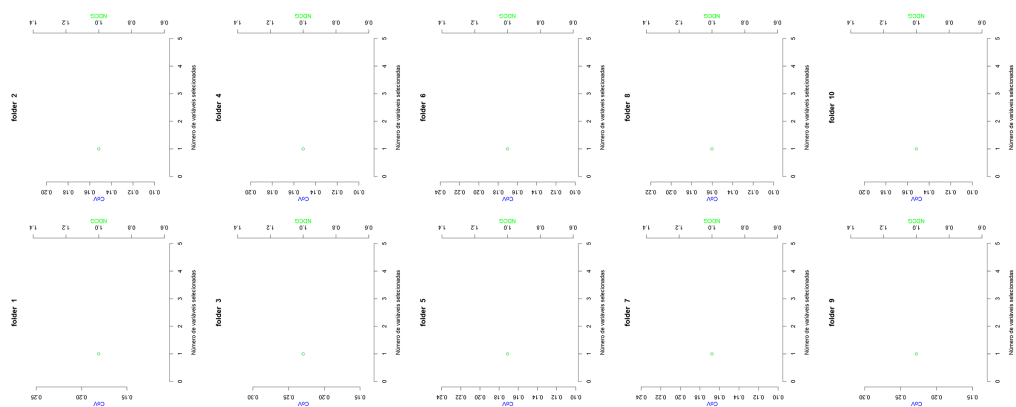


Figura B.53: NDCG vs CoV, LW DP, 3ª execução

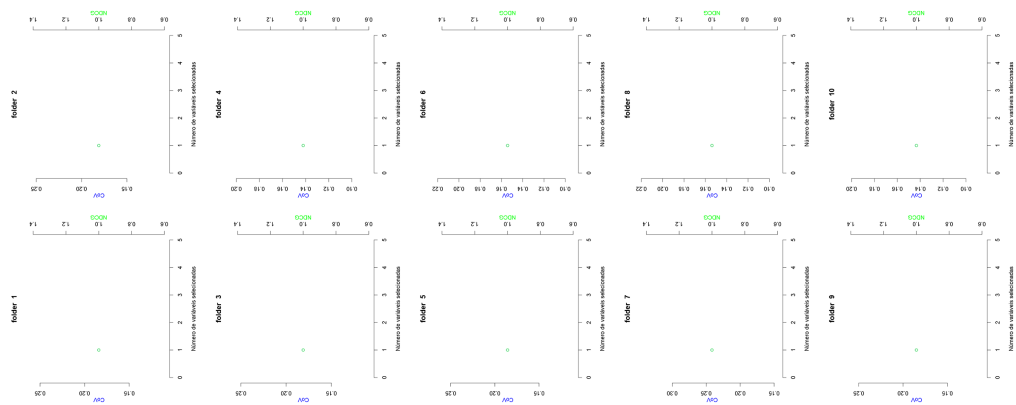


Figura B.54: NDCG vs CoV, LW DP, 4ª execução

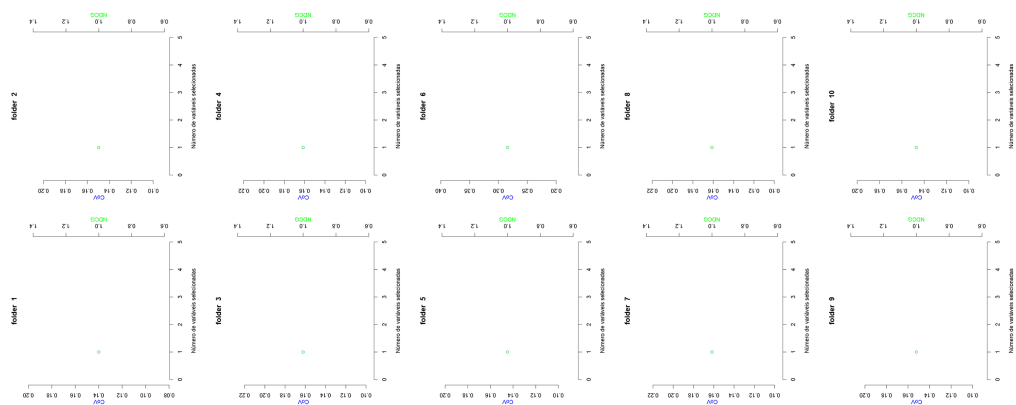


Figura B.55: NDCG vs CoV, LW DP, 5ª execução

Gráficos CoV vs NDCG para os Conjuntos de dados

B.6.2 PW

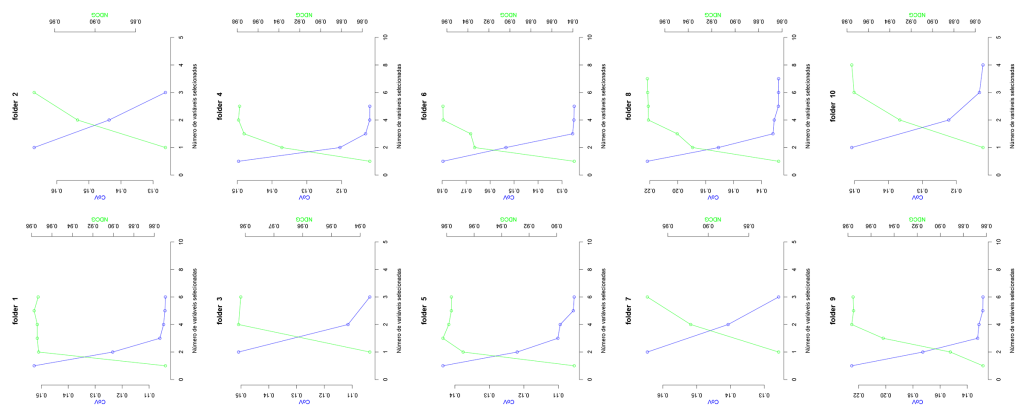


Figura B.56: NDCG vs CoV, PW, 1ª execução

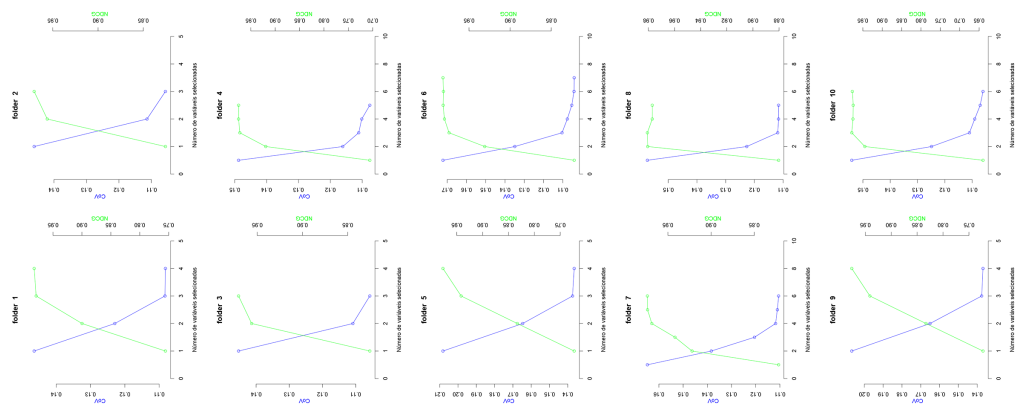


Figura B.57: NDCG vs CoV, PW, 2ª execução

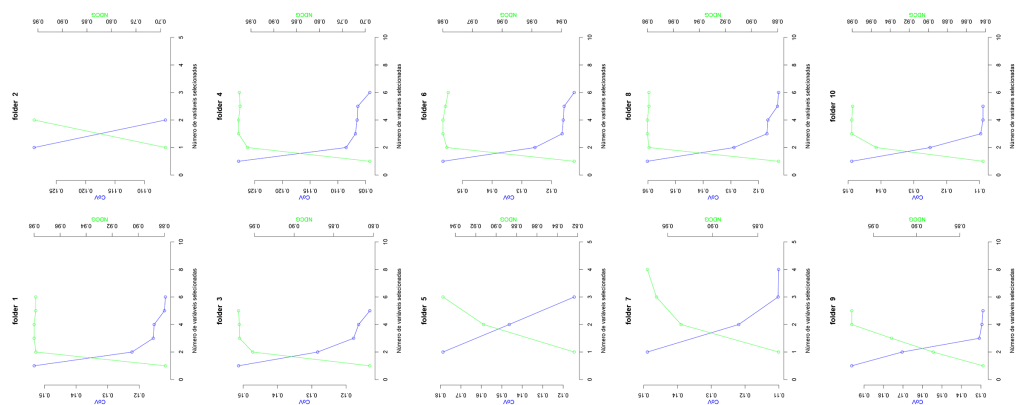


Figura B.58: NDCG vs CoV, PW, 3ª execução

Gráficos CoV vs NDCG para os Conjuntos de dados

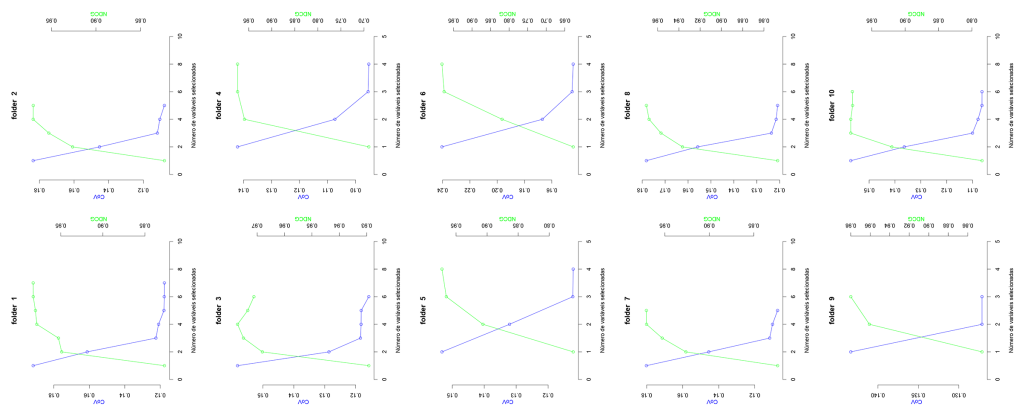


Figura B.59: NDCG vs CoV, PW, 4ª execução

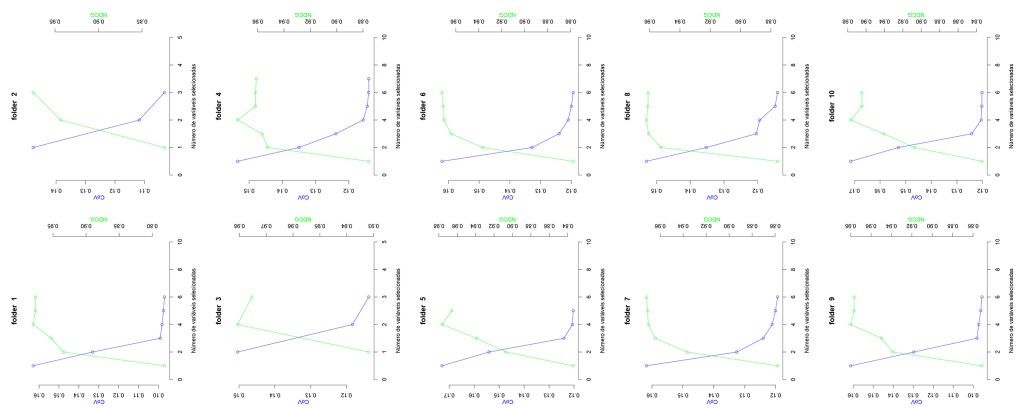


Figura B.60: NDCG vs CoV, PW, 5ª execução

B.7 Energy Efficiency(Y2)

B.7.1 LW DP

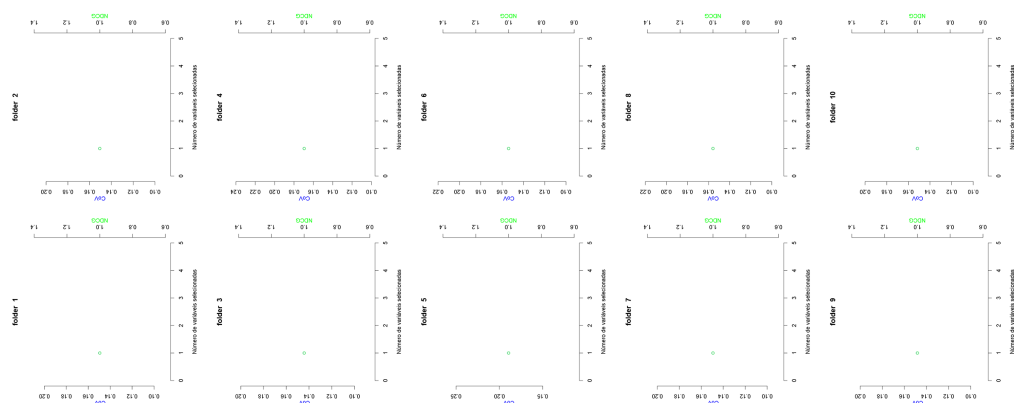


Figura B.61: NDCG vs CoV, LW DP, 1ª execução

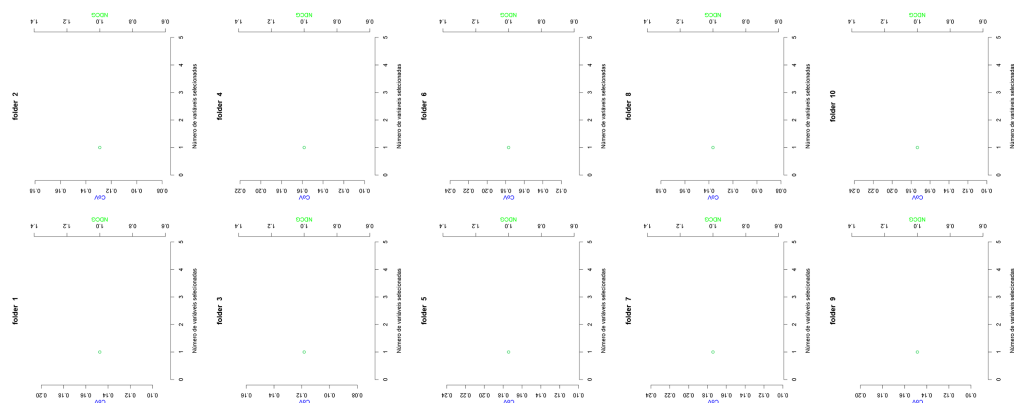


Figura B.62: NDCG vs CoV, LW DP, 2ª execução

Gráficos CoV vs NDCG para os Conjuntos de dados

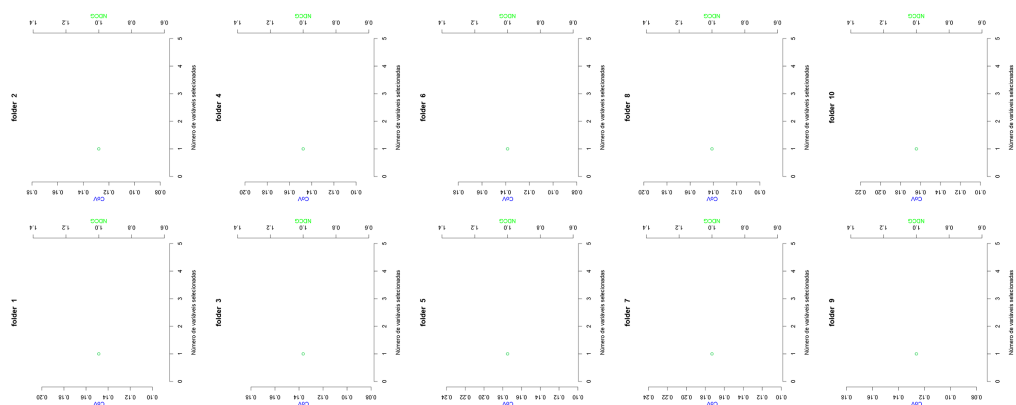


Figura B.63: NDCG vs CoV, LW DP, 3ª execução

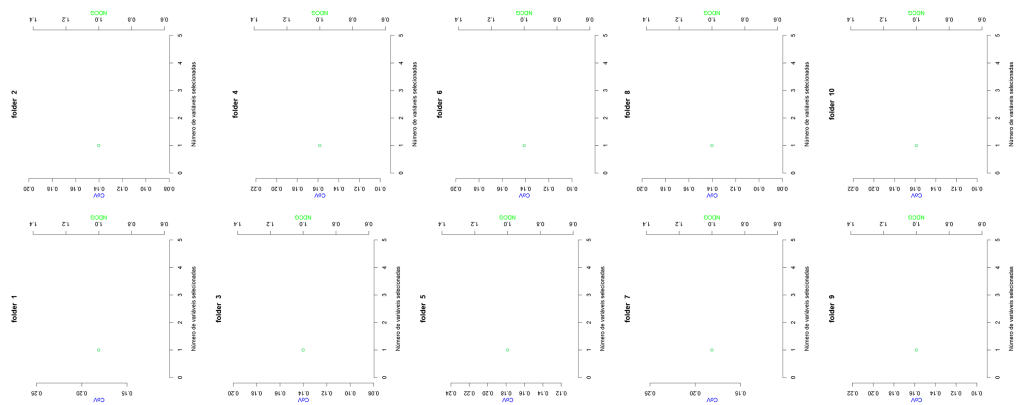


Figura B.64: NDCG vs CoV, LW DP, 4ª execução

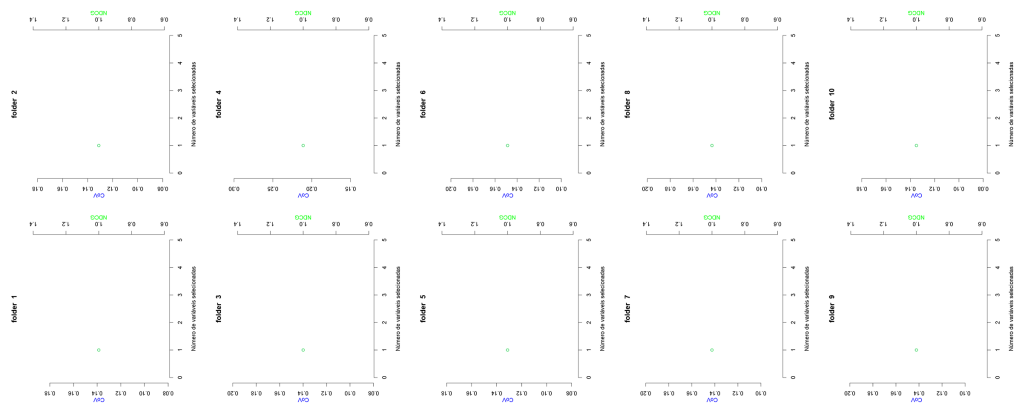


Figura B.65: NDCG vs CoV, LW DP, 5ª execução

Gráficos CoV vs NDCG para os Conjuntos de dados

B.7.2 PW

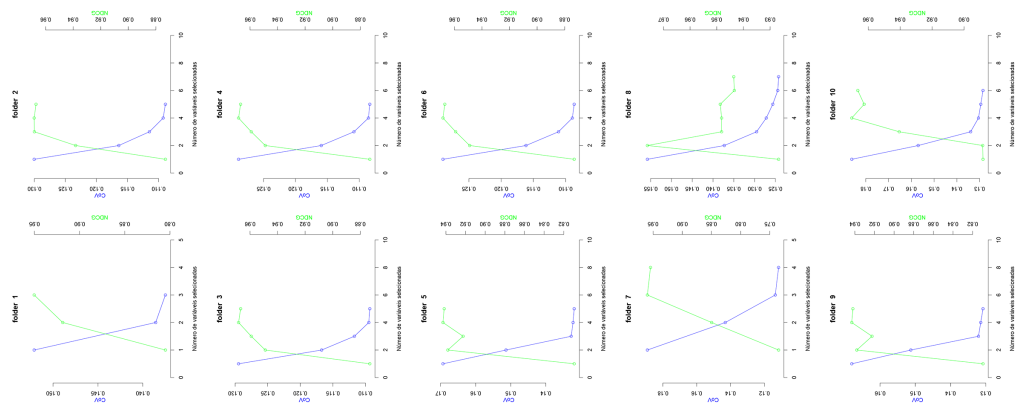


Figura B.66: NDCG vs CoV, PW, 1ª execução

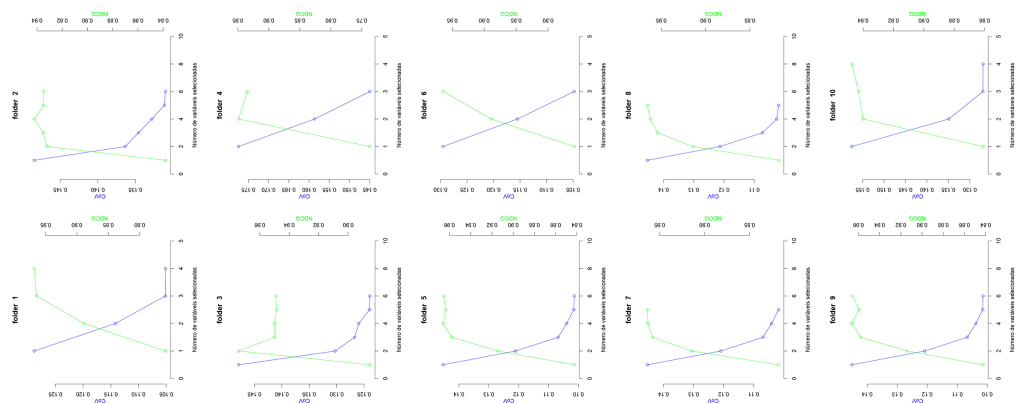


Figura B.67: NDCG vs CoV, PW, 2ª execução

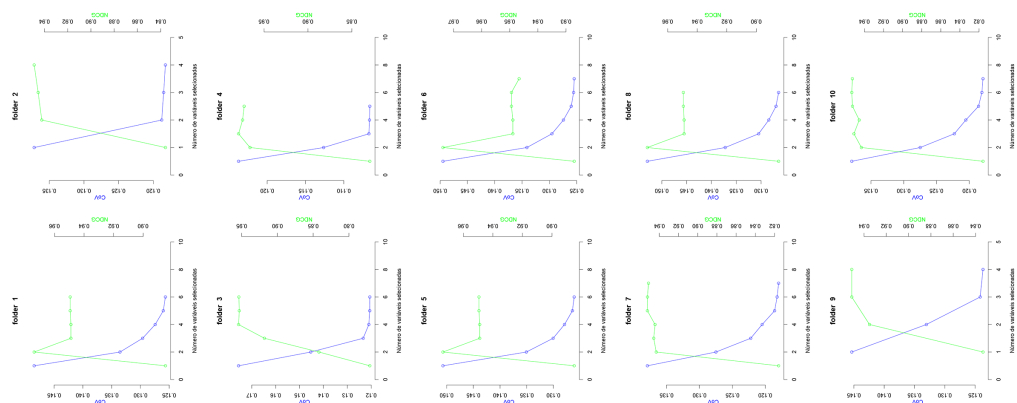


Figura B.68: NDCG vs CoV, PW, 3ª execução

Gráficos CoV vs NDCG para os Conjuntos de dados

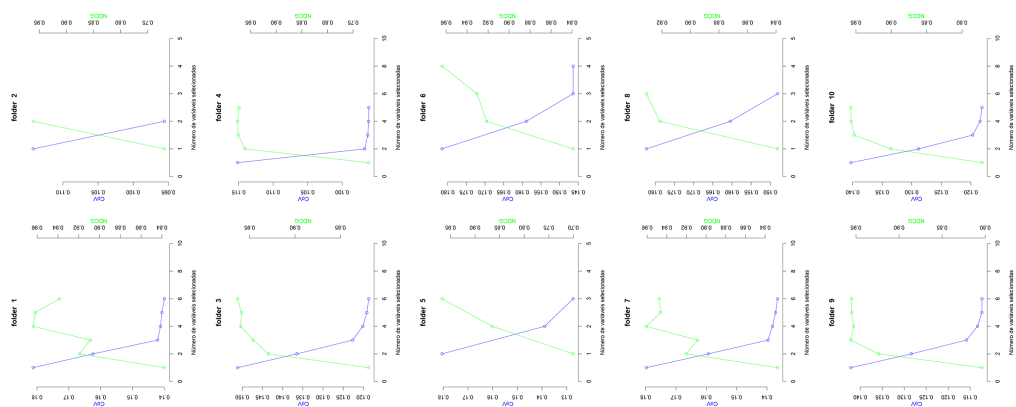


Figura B.69: NDCG vs CoV, PW, 4ª execução

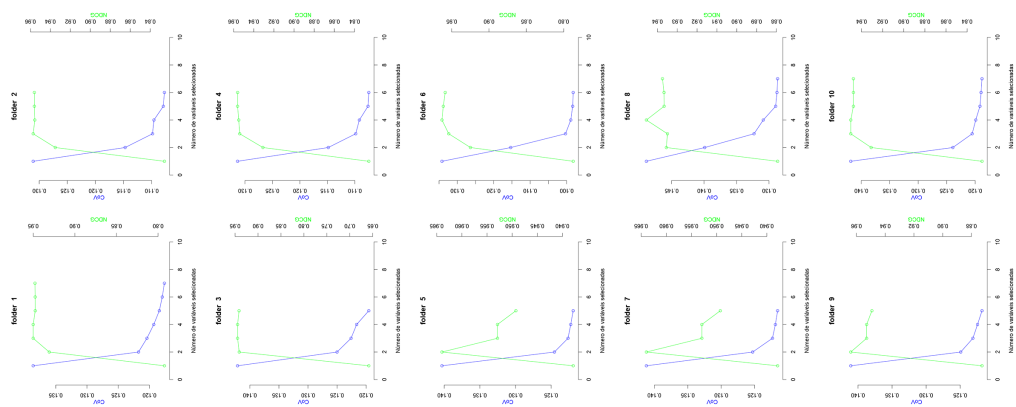


Figura B.70: NDCG vs CoV, PW, 5ª execução

B.8 LETOR

B.8.1 PW

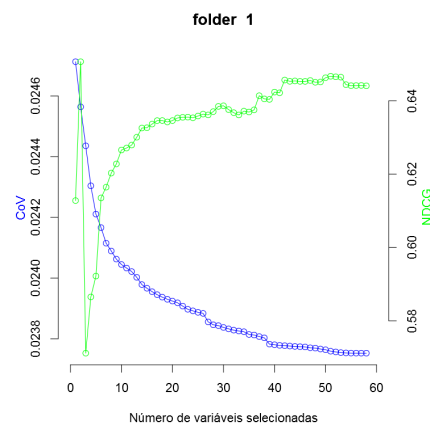


Figura B.71: NDCG vs CoV, PW, Pasta 1

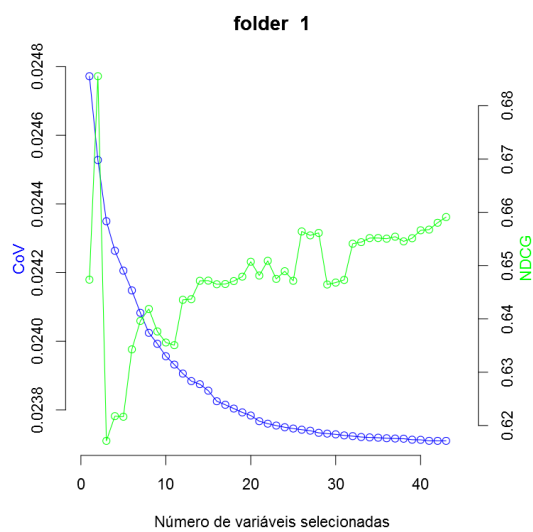


Figura B.72: NDCG vs CoV, PW, Pasta 2

Gráficos CoV vs NDCG para os Conjuntos de dados

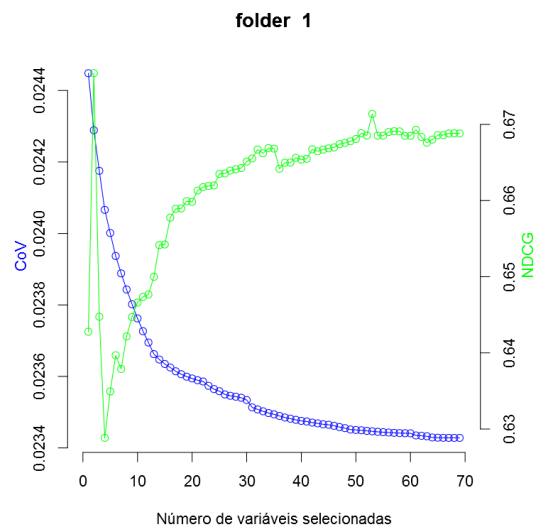


Figura B.73: NDCG vs CoV, PW, Pasta 3

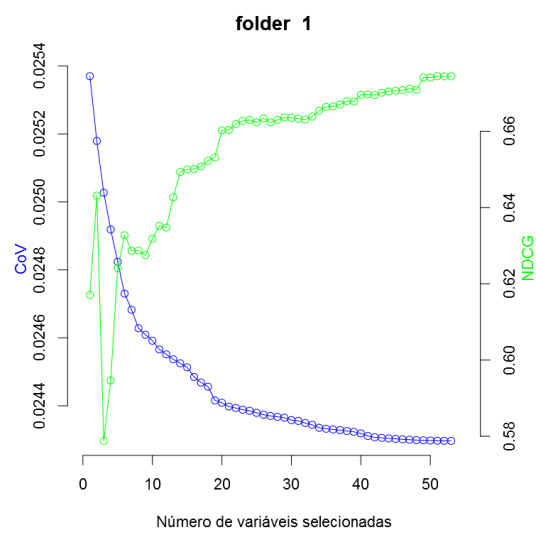


Figura B.74: NDCG vs CoV, PW, Pasta 4

Gráficos CoV vs NDCG para os Conjuntos de dados

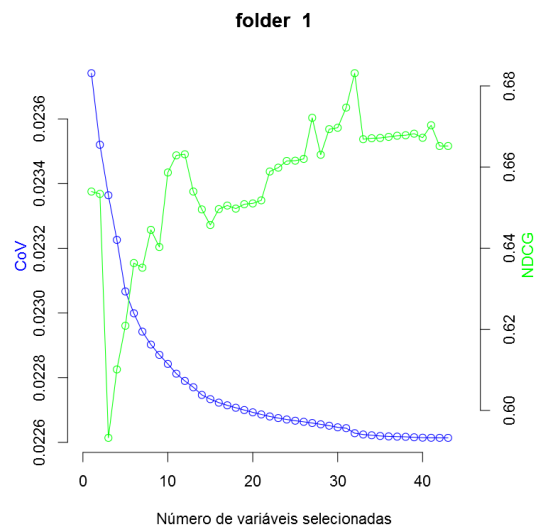


Figura B.75: NDCG vs CoV, PW, Pasta 5