

Universidade Federal de Pernambuco – UFPE

Programa de Pós-graduação em Ciência da Computação

Soluções de Mineração de Dados

IN1166

Prof: Leandro Maciel

Aluno: Irvin Soares Bezerra

Entrega do primeiro relatório sobre a variação de parâmetros dos modelos (tunning) visando encontrar as melhores configurações para a base de dados. Apresentar a estratégia de variação de parâmetros e justificativa para ela. Entregar relatório e scripts.

### Tunning para dos modelos para a base Dry Bean

#### Introdução

Todas as medições de acurácia e desvio padrão foram realizadas utilizando a técnica de K-fold cross validation com o valor de 10.

#### 1 – KNN

Foram testados diferentes valores para K, começando com o valor de 5 até o valor ótimo para K de 15, utilizando este valor atingiu-se a acurácia de 91.85%, utilizando o cross validation com k=10. Valores maiores para K resultaram em uma acurácia menor. A partir do valor de K igual a 8 houve uma discreta melhoria na acurácia, mas atingindo um desvio padrão menor, sendo este desvio igual a 0.0109. A matriz de correlação apresenta 7 variáveis com uma forte relação com a variável resposta, sendo este o valor de vizinho com maior relevância para predizer a classe resposta.

#### 2 – Decision Tree.

Para a árvore de decisão, os dados de treinamento não foram normalizados. Visto que este modelo é robusto para dados em diferentes escalas. Foram testados diferentes valores para a profundidade da árvore. Os testes se iniciaram utilizando uma ramificação igual a 1, sendo testados os seguintes valores: 5, 8, 10 e 13. A partir de 8 ramificações foi possível chegar a uma acurácia de 89.89% O valor ideal foi atingido utilizando 13 ramificações, atingindo a acurácia de 88.80% e o menor valor para o desvio padrão de 0.0081.

#### 3 – Random Forest.

Para a floresta de árvores, foi adotado o valor de 13 ramificações para cada árvore da floresta, visto que foi o valor de ótimo encontrado durante a experimentação de hiperparâmetros da árvore de decisão. Para aumentar a acurácia da floresta, foi experimentado diferentes valores para as quantidades de árvores, começando com o valor de 20, e incrementado até o valor de 100. A partir de 70 árvores foi obtido uma acurácia de 91.10%, para

outros incrementos, houve uma discreta melhoria para a acurácia. O valor de ótimo foi obtido com 100 árvores, apresentando uma acurácia de 91.36% e um desvio padrão de 0.0074.

#### 4 – Rede Neural MLP

Para a rede neural de múltiplas camadas perceptron, foram alterados os seguintes hiperparâmetros, densidade de neurônios e a quantidade de epochs que calculam a saída e os pesos do conjunto de treinamento. Foram experimentados valores de 20 até 90 para a densidade, e de 100 até 500 para os epochs. Houve uma grande variação nos resultados de acurácia, sendo de 35.38% para densidade 20 e 100 epochs. O valor de ótimo foi encontrado com a densidade 90 e 500 epochs, apresentando uma acurácia de 70.90% e um desvio padrão de 0.0218. Como teste, foi possível obter uma acurácia maior, utilizando grandes valores de densidade e de epochs, mas sendo necessário um elevado custo computacional.

#### 5 – Ensemble de Redes Neurais

Após encontrar os valores de ótimo para a rede neural MLP, foram criados 5, 8 e 10 modelos para a avaliação do Ensemble. Foi treinado um meta-classificador utilizando como modelo os modelos obtidos a partir da execução solo do algoritmo MLP. Apesar de variar a quantidade de redes para encontrar o melhor valor, todas as execuções apresentaram os mesmos resultados, sendo 92.30% de acurácia e 0.0057 de desvio padrão. Dessa forma foi adotado o valor de 5 modelos como ótimo para a criação do Ensemble, visto que mais modelos elevam o custo computacional. O ensemble apresentou um desempenho superior a execução solo do algoritmo que foi de 70.90% e um desvio padrão de 0.0218.

#### 6 – Ensemble heterogêneo

Após encontrar os valores de ótimo para os quatro modelos avaliados, foi possível criar um ensemble com diferentes modelos. A execução do ensemble apresentou a seguinte acurácia 92.01% e um desvio padrão de 0.0075. Obtendo um melhor desempenho que todos os modelos solo.