



**Inteligência Artificial**  
Prof. Bruno M. Nogueira  
*Faculdade de Computação - UFMS*  
**Trabalho Prático II**

Neste segundo trabalho, o objetivo é comparar o desempenho de algoritmos de classificação em tarefas diversas. Este trabalho deverá ser realizado em grupos de no mínimo três e no máximo quatro integrantes. O procedimento experimental, as bases de dados e o procedimento de entrega do trabalho serão descritos nas seções a seguir.

## 1 Procedimento experimental

Deverão ser utilizados os algoritmos de classificação de **Árvore de Decisão**<sup>1</sup>, **Vizinhos Mais Próximos (KNN)**<sup>2</sup>, **Naïve Bayes**<sup>3</sup>, **Regressão Logística**<sup>4</sup> e **Redes Neurais MLP**<sup>5</sup>. Devem ser utilizados os algoritmos disponíveis na biblioteca Scikit-learn.

Neste trabalho, deve-se adotar o procedimento de validação cruzada para avaliação dos algoritmos de classificação. Deve-se utilizar o procedimento de **10-fold cross validation estratificado**<sup>6</sup>. Para cada fold, deve-se coletar a **acurácia** e a **f1\_score**<sup>8</sup>. Ao final, para cada algoritmo em cada dataset, deve-se fazer uma média e o desvio padrão dos valores obtidos das medidas de avaliação, utilizando os valores obtidos em cada fold.

Para todos os algoritmos testados e em todos os datasets, diferentes valores de parâmetros deverão ser testados (no mínimo 2 combinações para cada algoritmo em cada dataset). Deve-se reportar o desempenho de cada algoritmo com cada combinação de parâmetro. Recomenda-se utilizar dispositivos de calibração de parâmetros como o *GridSearch*<sup>9</sup>.

## 2 Conjuntos de dados

Neste trabalho, cada grupo deve escolher 10 conjuntos de dados de classificação dentre os vários conjuntos disponíveis no repositório da UCI<sup>10</sup>. A única restrição imposta é que estes datasets escolhidos devem ter, obrigatoriamente, mais do que 100 exemplos. Havendo dúvidas se o dataset pode ou não ser utilizado, consulte o professor. Também, caso o grupo queira utilizar bases de dados de outros repositórios ou de outras fontes, será necessário consultar o professor.

Em alguns casos, vocês terão que pré-processar as bases de dados antes de aplicar os algoritmos de aprendizado de máquina. Geralmente, os algoritmos utilizados esperam como entrada uma matriz do tipo atributo valor. Caberá a você gerar essas matrizes de maneira adequada, sempre ficando atento à qualidade dos dados. Existem, na literatura, diversos métodos de pré-processamento de dados que podem auxiliar nestas etapas. Inclusive, vários deles encontram-se implementados nas API's que utilizamos neste semestre.

---

<sup>1</sup><http://scikit-learn.org/stable/modules/tree.html>

<sup>2</sup><http://scikit-learn.org/stable/modules/neighbors.html>

<sup>3</sup>[http://scikit-learn.org/stable/modules/naive\\_bayes.html](http://scikit-learn.org/stable/modules/naive_bayes.html)

<sup>4</sup>[http://scikit-learn.org/stable/modules/generated/sklearn.linear\\_model.LogisticRegression.html](http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html)

<sup>5</sup>[http://scikit-learn.org/stable/modules/generated/sklearn.neural\\_network.MLPClassifier.html](http://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html)

<sup>6</sup>[http://scikit-learn.org/stable/modules/cross\\_validation.html](http://scikit-learn.org/stable/modules/cross_validation.html)

<sup>7</sup>[http://scikit-learn.org/stable/modules/generated/sklearn.model\\_selection.KFold.html](http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html)

<sup>8</sup>[https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1\\_score.html](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html)

<sup>9</sup>[http://scikit-learn.org/stable/modules/grid\\_search.html](http://scikit-learn.org/stable/modules/grid_search.html)

<sup>10</sup><https://archive.ics.uci.edu/ml/index.php>

### 3 Entrega

Vocês deverão entregar um conjunto de até 10 slides com a apresentação do trabalho. Nessa apresentação, vocês deverão apresentar:

- Uma breve descrição dos algoritmos de classificação utilizados;
- Uma explicação do procedimento experimental;
- Enumeração e descrição dos datasets utilizados, contendo: nome do dataset, número de exemplos, número de atributos, número de classes e número de exemplos da classe majoritária;
- Resultados obtidos;
- Discussão dos resultados;
- Considerações finais.

Os resultados obtidos deverão ser apresentados por meio de tabelas e/ou gráficos. Nestes, deverão ser mostrados, para todos os datasets, a média geral de cada dataset para a medida coletada. Os gráficos ficam a critério dos grupos e devem ser utilizados para ajudar a interpretar os resultados.

Um dos integrantes do grupo deverá gravar um vídeo apresentando os slides submetidos. O vídeo deverá ser lançado em uma plataforma de streaming (Youtube, preferencialmente) e o link deverá constar no último slide da apresentação enviada.

O prazo para entrega do trabalho será as **23:55** do dia **26/06**, impreterivelmente. Deverão ser entregues, além da apresentação, todos os scripts e referências para os conjuntos de dados, para fins de reprodução dos experimentos. Todos os scripts com os códigos das soluções devem ser devidamente comentados e identificados com os nomes de todos os integrantes do grupo. Uma única entrega por grupo deve ser feita, via AVA.

Entregas fora do prazo, ou feitas por outros meios, serão desconsideradas. Não será tolerado plágio de quaisquer fontes, mesmo que parcial. Quando detectado plágio, o trabalho terá nota zero.

Quaisquer dúvidas, não hesitem em procurar o professor.

Bom trabalho!