ACTIVE LEARNING: UMA ABORDAGEM INTELIGENTE PARA AMOSTRAGEM DE DADOS

1 INTRODUÇÃO

No campo de aprendizado de máquina há aplicações que exigem quantidades massivas de dados de treinamento para alcançar a performance desejada. No entanto, os dados anotados[1] são difíceis e caros de obter, devido ao tempo necessário para realizar as anotações. Geralmente, a quantidade de dados é muito grande para rotulá-los manualmente e torna-se bastante desafiador para as equipes treinarem bons modelos supervisionados[2](KONYUSHKOVA; RAPHAEL; FUA, 2017).

Uma alternativa para minimizar o custo da anotação, é aplicar o método chamado *active learning* (aprendizado ativo, em português), cujo objetivo é obter o máximo de desempenho selecionando os datapoints (amostras) mais úteis de um conjunto de dados não rotulado (REN et al., 2020). Por outro lado, há também o *passive learning* (aprendizado passivo), que seleciona um *datapoint* aleatório de um conjunto de dados não rotulado e, portanto, não incorpora uma técnica sofisticada para treinar um modelo ou classificador (TONG, 2001).

Visto isso, este trabalho apresenta uma abordagem comparativa ao aplicar um classificador linear sobre o mesmo conjunto de dados sintético empregando o *active learning* e a amostragem aleatória, e desta forma destacar a superioridade do primeiro método no tempo de treinamento.

2 ACTIVE LEARNING

Dado um modelo de aprendizado de máquina e um conjunto de dados não rotulados, o objetivo do *active learning* é selecionar quais dados devem ser anotados para aprender o modelo o mais rápido possível. Na prática, isso significa que, em vez de pedir a especialistas para anotar todos os dados, seleciona-se de forma iterativa e adaptativa quais pontos de dados devem ser anotados em seguida (KONYUSHKOVA; RAPHAEL; FUA, 2017).

A seguir é e apresentado o fluxo geral das múltiplas estratégias de *active learning*:

1. O algoritmo começa a partir de um conjunto de dados não rotulados. Dos quais um número muito pequeno deles é rotulado.
2. O modelo é treinado com a subamostra rotulada.
3. Um procedimento de seleção é usado para definir o próximo *datapoint* que será anotado na iteração seguinte. Dessa forma, o algoritmo seleciona os pontos de dados mais informativos para os quais pode ter incerteza na predição.
4. Este *datapoin*t recém-escolhido é rotulado por especialistas humanos (também chamados de *oracle*) e adicionado ao conjunto de dados rotulados.
5. O modelo é novamente treinado no conjunto de dados atualizado e seu desempenho é calculado.

[1] Anotação de dados é o processo de rotulagem de dados em formatos como imagens, vídeo, áudio ou texto. Os dados rotulados são utilizados no aprendizado de máquina supervisionado.
[2] O aprendizado supervisionado ocorre quando o modelo ou classificador é treinado a partir de resultados conhecidos.

6. Este processo é repetido até que o desempenho esperado seja alcançado ou a rotulagem se esgote. Dessa forma, o algoritmo atinge seu desempenho máximo com o mínimo de dados e rapidamente.

2.1 Amostragem por incerteza (Uncertainty sampling)

A seleção de exemplos de treinamento geralmente é feita de maneira iterativa, ou seja, o algoritmo alterna entre retreinamento e seleção de novos exemplos. Em cada iteração, a utilidade de um *datapoint* candidato é estimada em termos de pontuação de utilidade (*score priosisation*), e aquele com a pontuação mais alta é selecionado. Nesse sentido, a noção de utilidade normalmente se refere à redução da incerteza. Na amostragem de incerteza (*uncertainty sampling*) que está entre as abordagens mais populares, a utilidade é quantificada em termos de incerteza preditiva, ou seja, o *active learning* seleciona os *datapoints* para as quais sua previsão atual é maximamente incerta. As previsões, bem como as medidas utilizadas para quantificar o grau de incerteza, como a entropia, são quase exclusivamente de natureza probabilística. Tais abordagens de fato provaram ser bem-sucedidas em muitas aplicações (NGUYEN; SHAKER; HÜLLERMEIER, 2022).

A entropia é uma medida de incerteza popular amplamente utilizada em estudos anteriores sobre o *active learning* (CHEN et al., 2006; JINGBO; HOVY, 2007; TANG; LUO; ROUKOS, 2002). Como leitura complementar, o exemplo apresentado em (SAJIL C.K., [s.d.]), faz uma breve introdução do uso da entropia para medidas de incerteza. Esse método é descrito pela Equação 1:

$$S_E = argmax\left(-\sum_i P(\hat{y}_i|x)\, logP(\hat{y}_i|x)\right) \quad (1)$$

Onde $P(\hat{y}_i|x)$ é a probabilidade *a posteriori* de um *datapoint* pertencer às classes existentes [0,1], e S_E é o argumento máximo extraído do somatório.

Para clarificar o uso da entralpia na pontuação de utilidade, é apresentado um exemplo na Tabela 1 que simula um conjunto de quatro dados (X1, X2, X3 e X4) e três classes (Classe 1, Classe 2 e Classe 3) com a probabilidade de os dados pertencerem às classes presentes. Note que ao calcular S_E para cada linha da tabela, o *datapoint* X3 é o que apresenta maior pontuação, indicando que ele será selecionado pelo algoritmo e inserido aos dados de treinamento para que o modelo possa reduzir sua incerteza.

Tabela 1 – Exemplo de aplicação da amostragem de incerteza usando entralpia

Data	Classe 1	Classe 2	Classe 3
X1	0.9	0.07	0.03
X2	0.87	0.03	0.1
X3	0.2	0.5	0.3
X4	0.00001	0.01	0.99
X1 = -0.9*log(0.9)-0.07*log(0.07)-0.03*log(0.03) = 0.386			
X2: -0.87*log(0.87)-0.03*log(0.03)-0.1*log(0.1) = 0.457			
X3: -0.2*log(0.2)-0.5*log(0.5)-0.3*log(0.3) = **1.03**			
X4: -0*log(0.00001)-0.01*log(0.01)-0.99*log(0.99) = 0.056			

Fonte:(SOLAGUREN-BEASCOA, 2020).

3 ALGORITMO IMPLEMENTADO

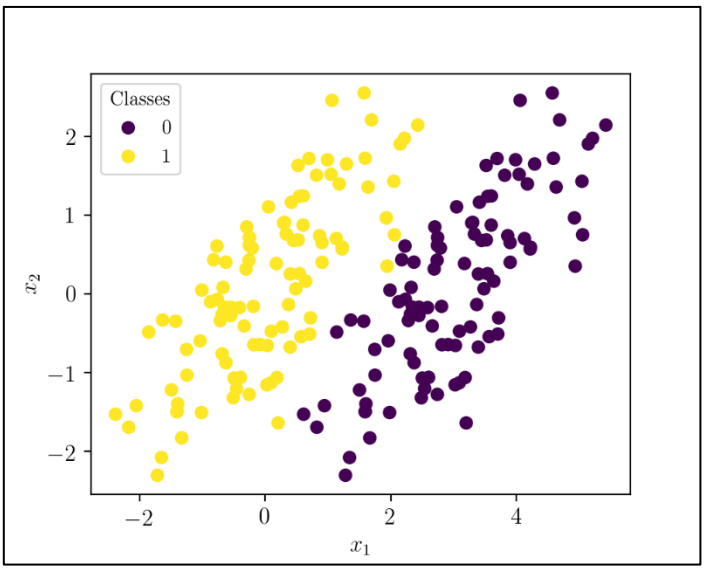
O algoritmo foi desenvolvido sob o seguinte cenário: foi gerado um conjunto de dados sintético de 200 *datapoints*, com duas classes: Classe A e Classe B. A Figura 1 destaca os dados em que os pontos amarelos denotam os *datapoints* de Classe A e os *datapoints* violetas representam a Classe B. Note que a *label* para os dados é binária, assumindo valores 0 (Classe B) ou 1 (Classe A).

Foi empregado um classificador binário para prever a qual das duas classes (categorias) um *datapoint* pertence. A entrada do algoritmo de classificação (Regressão Logística) é um conjunto de exemplos rotulados, em que cada rótulo é um número inteiro 0 ou 1. No processo de treinamento do classificador, os dados foram divididos da seguinte maneira: 70% treino (140 *datapoints*) e 30% teste (60 *datapoints*).

Ao empregar o *active learning* com amostragem de incerteza, o classificador inicia escolhendo aleatoriamente 10 datapoints do conjunto de dados de treinamento. Em seguida, a cada iteração, a amostragem de incerteza seleciona um novo *datapoint* entre os 130 candidatos, cujo anotação é desconhecida, i.e, apesar de todos os dados deste conjunto ser anotado, durante o processo de amostragem, esta informação é ignorada. Após a seleção, o *datapoint* é inserido ao conjunto de dados anotados[3]. O modelo é treinado no novo conjunto rotulado e a precisão é calculada em relação ao conjunto de dados de validação.

De forma similar, o uso do *passive learning* escolhe aleatoriamente 10 *datapoints* do conjunto de dados de treinamento. Em seguida, a cada iteração, é selecionado um *datapoint* aleatório entre os 130 candidatos, cujo anotação é desconhecida. Após a seleção, o *datapoint* é inserido ao conjunto de dados anotados. O modelo é treinado no novo conjunto rotulado e a precisão é calculada em relação ao conjunto de dados de validação.

Figura 1 – Distribuição do conjunto de dados sintético



Fonte: Autor.

[3] Note que ao selecionar o *datapoint*, neste exemplo, não foi necessária a rotulação por parte humanos, pois após o processo de seleção dos datapoints, em que as anotações são ignoradas, o algoritmo insere este *datapoint* ao conjunto de treinamento reconsiderando a sua anotação. Esta metodologia foi utilizada para facilitar o desenvolvimento do algoritmo

Os algoritmos foram implementados em Python utilizando o Google Colab, por sua facilidade em permitir escrever e executar Python em um navegador. O código documentado está disponível para acesso no GitHub através do seguinte link: https://github.com/Lucas-Mantuan/ActiveLearningVSPassiveLearning.

3.1 Active learning

O algoritmo geral do *active learning* implementado é apresentado na Tabela *2* e baseado nos experimentos realizados por (SAJIL C.K., [s.d.]).

Tabela 2 – Algoritmo geral da aplicação do *active learning*

```
O conjunto de dados de treinamento contém 140 datapoints

Inicia-se o treinamento do classificador com 10 datapoints
rotulados e 130 datapoints não rotulados

for 100 iterações:
     Treina o classificador
     Calcula a acurácia do dataset rotulado com 10 amostras
     for i in range(130)
          Seleciona um datapoint com base na amostragem de
          incerteza (Entropia) do conjunto não rotulado
          Adiciona o datapoint selecionado ao conjunto de
          datapoints rotulados
          Atualiza o conjunto de datapoints não rotulados
          Calcula a acurácia do dataset rotulado
          Armazena em um vetor a acurácia calculada
Calcula a média das acurácias ao longo de 100 iterações
```

Fonte: Autor.

3.2 Passive learning

O algoritmo geral do *passive learning* implementado é apresentado na Tabela 3 e baseado nos experimentos realizados por (SAJIL C.K., [s.d.]).

Tabela 3 – Algoritmo geral da aplicação do *passive learning*

```
O conjunto de dados de treinamento contém 140 datapoints

Inicia-se o treinamento do classificador com 10 datapoints
rotulados e 130 datapoints não rotulados

for 100 iterações:
     Treina o classificador
     Calcula a acurácia do dataset rotulado com 10 amostras
     for i in range(130)
          Seleciona aleatoriamente um datapoint do conjunto não
          rotulado
          Atualiza o conjunto de datapoints não rotulados
```

```
            Adiciona o datapoint selecionado ao conjunto de
            datapoints rotulados
            Calcula a acurácia do dataset rotulado
            Armazena em um vetor a acurácia calculada
Calcula a média das acurácias ao longo de 100 iterações
```

Fonte: Autor.

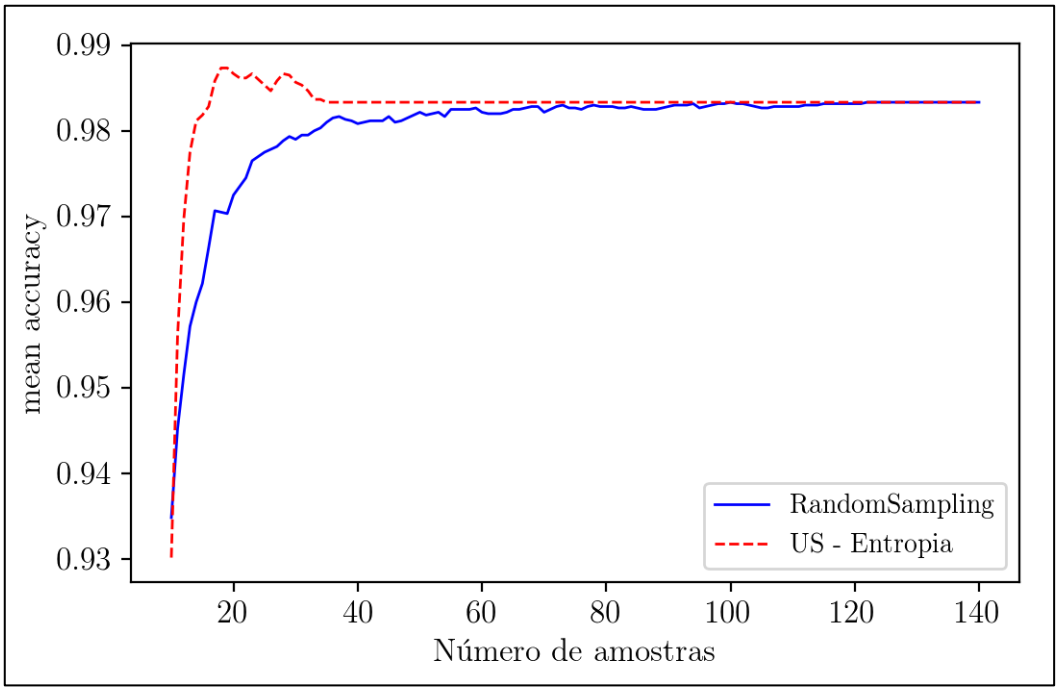
4 RESULTADOS

Abaixo são apresentados os resultados comparativos entre a acurácia média do classificador utilizando a amostragem de incerteza com cálculo de entropia e a amostragem aleatória. Como destacado anteriormente, o classificador é inicialmente treinado com 10 *datapoints* e a acurácia média inicial reflete o início das curvas destacadas em azul contínuo (amostram aleatória) e vermelho tracejado (amostragem de incerteza).

À medida que mais amostras são acrescentadas ao conjunto de dados rotulados e o classificador é retreinado, a acurácia tende a crescer. A estratégia de seleção de *datapoints* usada pelo amostragem de incerteza apresenta uma taxa de aprendizagem maior, sendo que com menos de 40 *datapoints*, a acurácia média da curva vermelha alcança uma amplitude de desempenho que só é alcançada pela curva azul a partir de 100 *datapoints*, nos quatro testes realizados, destacados na Figura 2, Figura 3, Figura 4 e Figura 5.

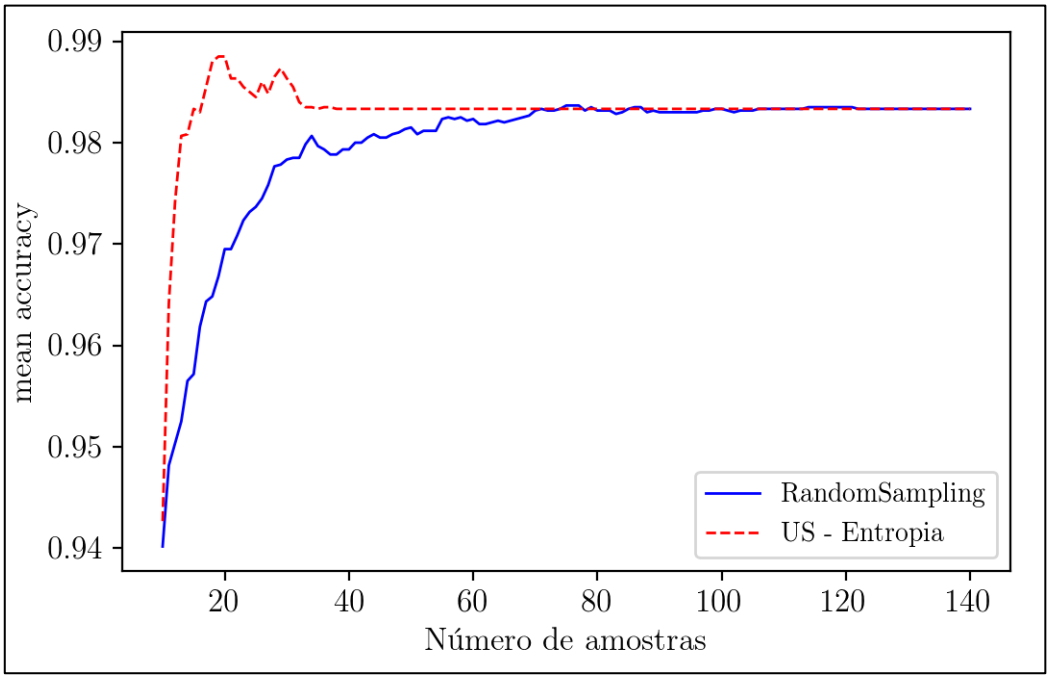
Em resumo, o algoritmo de *active learning* com amostragem de incerteza aprende muito mais rápido e atinge o desempenho máximo com um número mínimo de amostras. A amostragem aleatória e iterativa aprende à medida que recebe mais pontos de dados rotulados.

Figura 2 – Comparação entre amostragem de incerteza e amostragem aleatória – TESTE 1



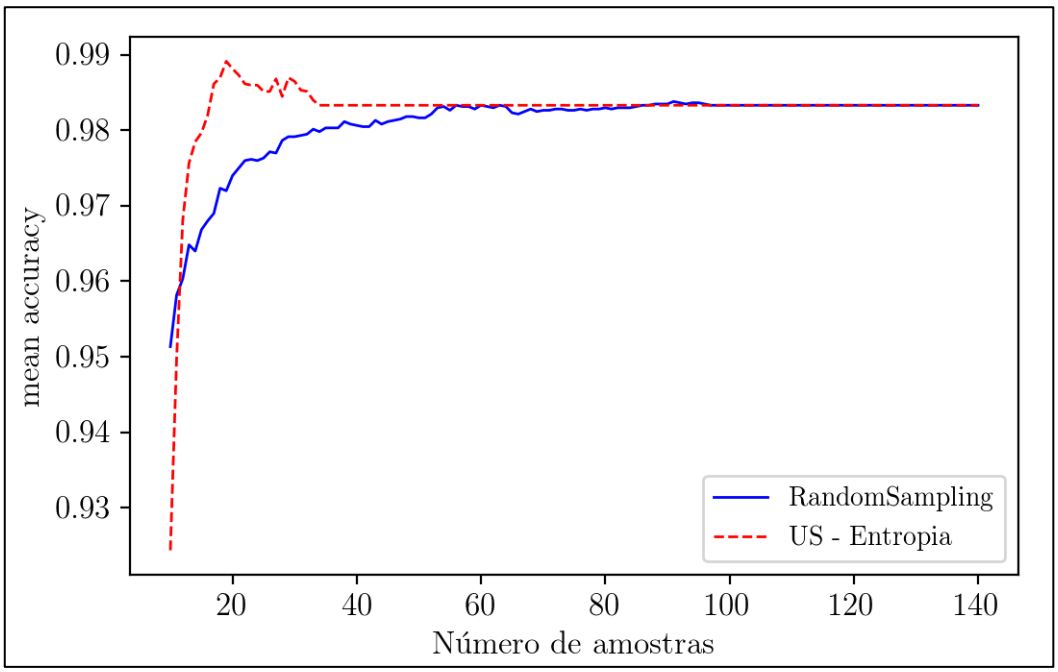
Fonte: Autor.

Figura 3 – Comparação entre amostragem de incerteza e amostragem aleatória – TESTE 2



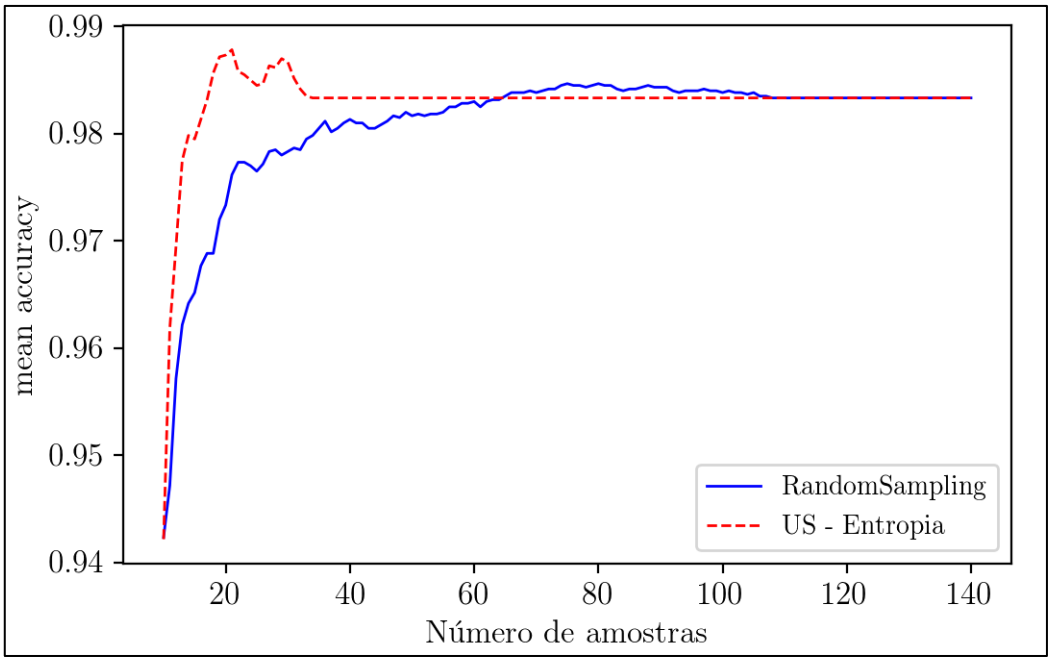
Fonte: Autor.

Figura 4 – Comparação entre amostragem de incerteza e amostragem aleatória – TESTE 3



Fonte: Autor.

Figura 5 – Comparação entre amostragem de incerteza e amostragem aleatória – TESTE 4



Fonte: Autor.

5 CONSIDERAÇÕES FINAIS

O poder da técnica de amostragem inteligente se mostrou superior mesmo aplicado ao cenário com um *dataset* de poucas amostras. Ao escalar para o domínio de *big data*, o *active learning*, é relevante por auxiliar as equipes em economia de tempo, menor custo computacional e financeiro, visto que o modelo é treinado mais rapidamente, exige menor recurso humano e ao atingir a acurácia desejada o processo de treinamento pode ser finalizado.

REFERÊNCIAS

CHEN, J. et al. **An Empirical Study of the Behavior of Active Learning for Word Sense Disambiguation**. Proceedings of the Human Language Technology Conference of the NAACL, Main Conference. **Anais**...2006. Disponível em: <https://aclanthology.org/N06-1016/>. Acesso em: 29 maio. 2022

JINGBO, Z.; HOVY, E. **Active Learning for Word Sense Disambiguation with Methods for Addressing the Class Imbalance Problem**. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). **Anais**...2007. Disponível em: <https://aclanthology.org/D07-1082/>. Acesso em: 29 maio. 2022

KONYUSHKOVA, K.; RAPHAEL, S.; FUA, P. Learning Active Learning from Data. **Advances in Neural Information Processing Systems**, v. 2017- December, p. 4226–4236, 9 mar. 2017.

NGUYEN, V. L.; SHAKER, M. H.; HÜLLERMEIER, E. How to measure uncertainty in uncertainty sampling for active learning. **Machine Learning**, v. 111, n. 1, p. 89–122, 1 jan. 2022.

REN, P. et al. A Survey of Deep Active Learning. **ACM Computing Surveys**, v. 54, n. 9, 30 ago. 2020.

SAJIL C.K. **Entropy Simplified - Intuitive Tutorials**. Disponível em: <https://intuitivetutorial.com/2021/05/18/entropy-simplified/>. Acesso em: 30 maio. 2022a.

SAJIL C.K. **Active Learning with Uncertainty Sampling from Scratch**. Disponível em: <https://intuitivetutorial.com/2021/05/15/active-learning-with-uncertainty-sampling-from-scratch/>. Acesso em: 30 maio. 2022b.

SOLAGUREN-BEASCOA, A. **Active Learning in Machine Learning**. Disponível em: <https://towardsdatascience.com/active-learning-in-machine-learning-525e61be16e5>. Acesso em: 29 maio. 2022.

TANG, M.; LUO, X.; ROUKOS, S. Active Learning for Statistical Natural Language Parsing. p. 120–127, 2002.

TONG, S. **Active Learning: Theory and Applications**. [s.l.] Standford University, 2001.

ACTIVE LEARNING: A SMART APPROACH TO DATA SAMPLING

Abstract: *Labeling the training data for a machine learning algorithm is a tedious and time-consuming, error-prone process. Furthermore, in some application domains, labeling each example can also be extremely costly financially. The active learning technique helps to deal with this problem by detecting and asking the user to label only the most informative examples in the domain, being a tool that contributes to the data team in the labeling process. Considering the power of smart sampling, this work presents a comparison of classifiers trained with the technique of active learning and passive learning. The results achieved reinforce the superiority and gain that active learning employs in a scenario where data annotation is a critical factor.*

Keywords: *active learning, passive learning, uncertainty sampling, random random sampling, machine learning*