Atividade prática 1 **Aprendizagem de Máquina**
Franklin Giovani Bastidas Cuya Universidade Federal do Rio grande do Sul
Email: fgbcuya@inf.ufgrs.br Instituto de Informática
ID: 322632 Instructor: Mariana Recamonde Mendoza

Problem 2.
Utilize os conjuntos de dados de treinamento e teste fornecidos no diretório "Dados Normalizados" para avaliar a aplicação do KNN com diferentes valores de k. Como o nome do arquivo indica, estes dados foram previamente normalizados, tal que todos os atributos variam no mesmo intervalo ($[0, 1]$). Os dados se referem à classificação de tumores de mama em maligno (1) ou benigno (0), de acordo com a coluna target. Os atributos (29) descrevem características dos núcleos celulares presentes em uma imagem digitalizada do material coletado na biópsia pelo método fine needle aspirate (FNA).

- Sua implementação deve permitir informar/variar o valor do hiperparâmetro k para cada execução do algoritmo
- Utilize como medida de distância padrão a Distância Euclidiana. Se desejar, pode permitir que esta medida seja configurada, por exemplo, que seja usada a Distância de Manhattan.
- Assuma que sua implementação se destina à análise de dados quantitativos, isto é, não é necessário se preocupar em implementar suporte ao tratamento de dados qualitativos.

Solution: The KNN algorithm was developed in Python using the Jupyter tool, this section uses the normalized data from the classification of breast tumors. The Euclidean and Manhattan equations are used to determine the distance between instances, then the number of errors and the hit rate are determined according to the hyperparameter k (number of k data to analyze), these are shown in the table 1. It should be taken into account that the set of 455 data were for training and 114 data for the test.

K	Euclidian		Manhattan	
	N of Errors	Hit Rate (%)	N of Errors	Hit Rate (%)
1	9	92.105	10	91.228
3	5	95.614	5	95.614
5	6	94.736	5	95.614
7	6	94.736	6	94.736

Table 1: Hit rate and error count of normalized data

Problem 3.
Refaça o item 2, agora para o conjunto de dados no arquivo "Dados NaoNormalizados".

- Observe o intervalo em que varia cada atributo no dado de treinamento: há uma grande diferença entre os valores máximo e mínimo de cada atributo?
- Analise e comente se houveram ou não diferenças em relação à taxa de acerto do algoritmo treinado e testado com dados normalizados. A normalização impactou? De que forma: melhorando ou piorando o desempenho? Esta tendência foi observada para todos os valores de k?

Solution:
In this section the KNN is determined with the non-normalized data and, as in the previous numeral, the distance between instances is also calculated with the equations of Euclid and Manhattan, then the number of errors and the hit rate are determined, these values are shown in the table 2. It can be seen that unlike the Normalized values, these tend to obtain less favorable results, reaching hit rates of up to

86%, compared to the Normalized data that obtained 91% in their hit rate plus estimated low. It should be taken into account that the set of 455 data were for training and 114 data for the test. Looking at all the estimated K values, it can be seen that the number of errors increased, resulting in a worse hit rate.

K	Euclidian		Manhattan	
	N of Errors	Hit Rate (%)	N of Errors	Hit Rate (%)
1	14	87.719	11	90.350
3	15	86.842	14	87.719
5	12	89.473	11	90.350
7	13	88.596	13	88.596

Table 2: Hit rate and error count of Non-normalized data

The interval of each attribute of the training data is calculated, the 30 intervals are not represented in this report but according to the list of intervals, table 3 highlights the maximum, minimum value and the number of the attribute that belongs to the normalized and non-normalized data.

Also in the analysis it was possible to detect that the "Normalidos" data set is not properly normalized, discovering that the attribute (column) 27 has a value of 375.

	Non-Normalized Data		Normalized Data	
	Attribute	Interval	Attribute	Interval
Max	21	9960.7	27	375
Min	15	0.029	16	0.782

Table 3: Interval analysis between instances