Trabalho Final Data Mining

Proposta de trabalho para avaliação da Prof. Manoela Kohler

Aluno: Otávio Ciribelli Borges

email: otavio.ciribelli@gmail.com || tel: (21) 983163900

Introdução

A proposta de trabalho consiste em implementar uma rotina de classificação do tipo supervisionada em um sistema de medição de condições climáticas em dois ambientes domésticos. As medidas de temperatura e umidade foram tomadas em períodos compreendidos entre 2018 e 2021 e estão disponíveis em repositório do github no link https://github.com/ciribelli/autohome. O arquivo em específico fica em: https://github.com/ciribelli/autohome/tree/master/home/db.sqlite3.

O objetivo de classificação deste trabalho consiste em prever, a partir dos registros de temperatura e umidade, quando o aparelho de ar condicionado dos ambientes esteve ligado. Existe portanto duas classes de saída que são de natureza numérica, sendo '1' para o estado de ar condicionado ligado e '0' quando o aparelho está desligado.

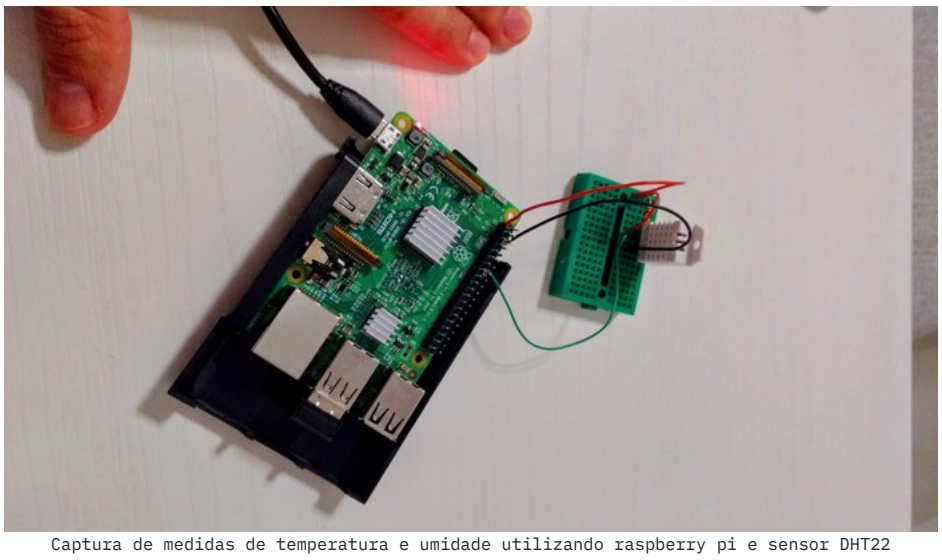
Em termos de propósito e abrangência mercadológica, esta rotina de classificação tem potencial contribuição com a simplificação de aplicações do tipo IoT ou IIoT, podendo atender com a geração de informações adicionais àquelas dos sensores de variáveis físicas.

Sobre as medições realizadas

A tomada das medições de temperatura e umidade foi feita utilizando um computador do tipo raspberry pi 3 com aplicação do sensor DHT22 (datasheet disponível em https://pdf1.alldatasheet.com/datasheet-pdf/view/1132459/ETC2/DHT22.html).

A rotina de captura está disponível no repositório github https://github.com/ciribelli/autohome/blob/master/motorHome.py.

O período de amostragem dos dados é de sessenta (60) segundos, tanto para temperatura quanto para umidade, e as informações são registradas na base de dados sqlite3 no arquivo db.sqlite3.



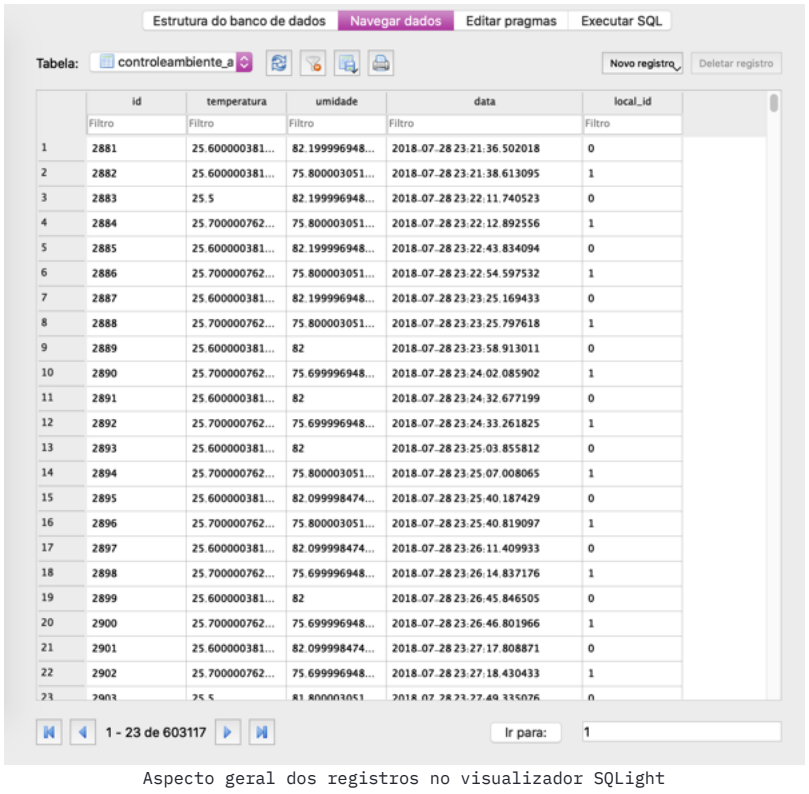
Captura de medidas de temperatura e umidade utilizando raspberry pi e sensor DHT22

Sobre o formato dos dados brutos

O formato dos dados brutos está apresentado abaixo por meio de uma captura de tela do software 'DB Browser for SQLite'.

Algumas informações relevantes do banco de dados:

- são 603.117 registros para dois ambientes diferentes
- a coluna 'local_id' indica a origem do local das medições (são dois ambientes diferentes)
- a coluna data indica o timestamp do instante em que os registros são capturados do sensor
- temperaturas são registradas em graus celsius
- umidades são registradas em valores percentuais



	id	temperatura	umidade	data	local_id
1	2881	25.600000381...	82.199996948...	2018-07-28 23:21:36.502018	0
2	2882	25.600000381...	75.800003051...	2018-07-28 23:21:38.613095	1
3	2883	25.5	82.199996948...	2018-07-28 23:22:11.740523	0
4	2884	25.700000762...	75.800003051...	2018-07-28 23:22:12.892556	1
5	2885	25.600000381...	82.199996948...	2018-07-28 23:22:43.834094	0
6	2886	25.700000762...	75.800003051...	2018-07-28 23:22:54.597532	1
7	2887	25.600000381...	82.199996948...	2018-07-28 23:23:25.169433	0
8	2888	25.700000762...	75.800003051...	2018-07-28 23:23:25.797618	1
9	2889	25.600000381...	82	2018-07-28 23:23:58.913011	0
10	2890	25.700000762...	75.699996948...	2018-07-28 23:24:02.085902	1
11	2891	25.600000381...	82	2018-07-28 23:24:32.677199	0
12	2892	25.700000762...	75.699996948...	2018-07-28 23:24:33.261825	1
13	2893	25.600000381...	82	2018-07-28 23:25:03.855812	0
14	2894	25.700000762...	75.800003051...	2018-07-28 23:25:07.008065	1
15	2895	25.600000381...	82.099998474...	2018-07-28 23:25:40.187429	0
16	2896	25.700000762...	75.800003051...	2018-07-28 23:25:40.819097	1
17	2897	25.600000381...	82.099998474...	2018-07-28 23:26:11.409933	0
18	2898	25.700000762...	75.699996948...	2018-07-28 23:26:14.837176	1
19	2899	25.600000381...	82	2018-07-28 23:26:45.846505	0
20	2900	25.700000762...	75.699996948...	2018-07-28 23:26:46.801966	1
21	2901	25.600000381...	82.099998474...	2018-07-28 23:27:17.808871	0
22	2902	25.700000762...	75.699996948...	2018-07-28 23:27:18.430433	1
23	2903	25.5	81.800003051...	2018-07-28 23:27:49.335076	0

Aspecto geral dos registros no visualizador SQLight

Sobre a composição do dataset para o trabalho proposto

Para a composição do dataset deste trabalho são tomadas medidas indiretas a partir dos dados brutos apresentados. Essas medidas consistem do resultado de operações de janelas deslizantes de diferentes tamanhos aplicadas sobre os dados disponíveis.

As operações são realizadas utilizando a função *rolling* da biblioteca Pandas (https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.rolling.html) com janelas que têm a seguinte composição:

```
janela_0 = 15 # janela deslizante de 15 elementos
janela_1 = 30 # janela deslizante de 30 elementos
janela_2 = 60 # janela deslizante de 60 elementos
```

Código-exemplo de aplicação das janelas ao dado bruto:

```
l_janelas = [janela_0, janela_1, janela_2]

for l_j in l_janelas:

    # para temperaturas
    s_arcon_t = Y_t.rolling(l_j).std()              # desvio padrao
    v_arcon_t = Y_t.rolling(l_j).var()              # variância
    m_arcon_t = Y_t.rolling(l_j).mean()             # média
    min_arcon_t = Y_t.rolling(l_j).min()            # calc. auxiliar de mínimo
    max_arcon_t = Y_t.rolling(l_j).max()            # calc. auxiliar de máximo
    a_arcon_t = max_arcon_t - min_arcon_t           # amplitude (max - min)
```

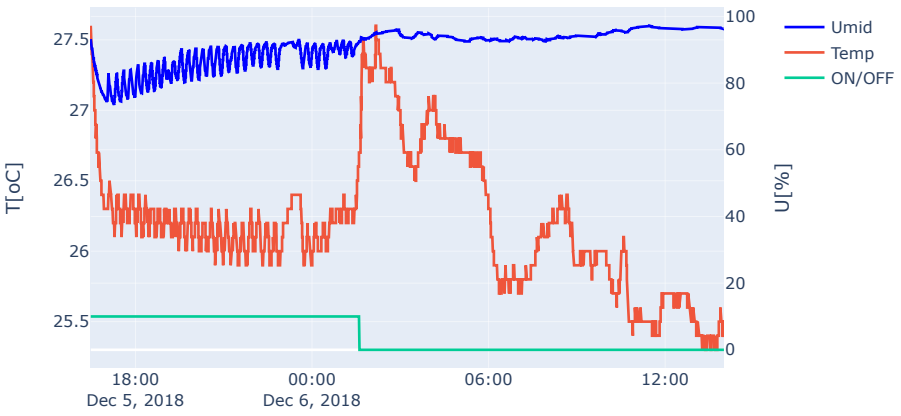
Dataset proposto no trabalho:

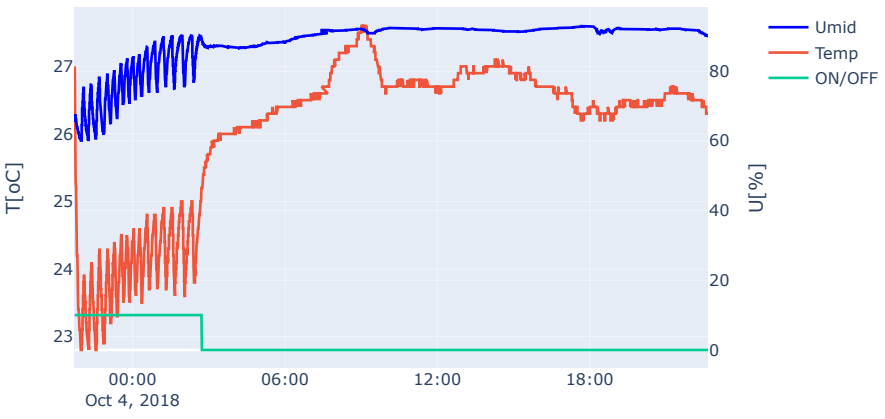
	tempo	temperatura	umidade	temp_desvpad_15	temp_variancia_15	temp_media_15	temp_amplitude_15	umid_desvpad_15	umid_variancia_15
324689	Dec 5, 2018 4:28 PM	27.6000	93.2000	0.0258	0.0007	27.5933	0.1000	0.0258	0.0007
324690	Dec 5, 2018 4:28 PM	27.5000	91.9000	0.0352	0.0012	27.5867	0.1000	0.0352	0.0012
324691	Dec 5, 2018 4:29 PM	27.6000	91.9000	0.0352	0.0012	27.5867	0.1000	0.0352	0.0012
324692	Dec 5, 2018 4:29 PM	27.5000	91	0.0352	0.0012	27.5867	0.1000	0.0352	0.0012
324693	Dec 5, 2018 4:30 PM	27.5000	91	0.0414	0.0017	27.5800	0.1000	0.0414	0.0017
324694	Dec 5, 2018 4:30 PM	27.4000	90.3000	0.0617	0.0038	27.5667	0.2000	0.0617	0.0038
324695	Dec 5, 2018 4:31 PM	27.4000	90.3000	0.0743	0.0055	27.5533	0.2000	0.0743	0.0055
324696	Dec 5, 2018 4:31 PM	27.3000	89.3000	0.0976	0.0095	27.5333	0.3000	0.0976	0.0095
324697	Dec 5, 2018 4:32 PM	27.3000	89.3000	0.1125	0.0127	27.5133	0.3000	0.1125	0.0127
324698	Dec 5, 2018 4:32 PM	27.3000	88.3000	0.1223	0.0150	27.4933	0.3000	0.1223	0.0150
324699	Dec 5, 2018 4:34 PM	27.2000	87.4000	0.1397	0.0195	27.4667	0.4000	0.1397	0.0195

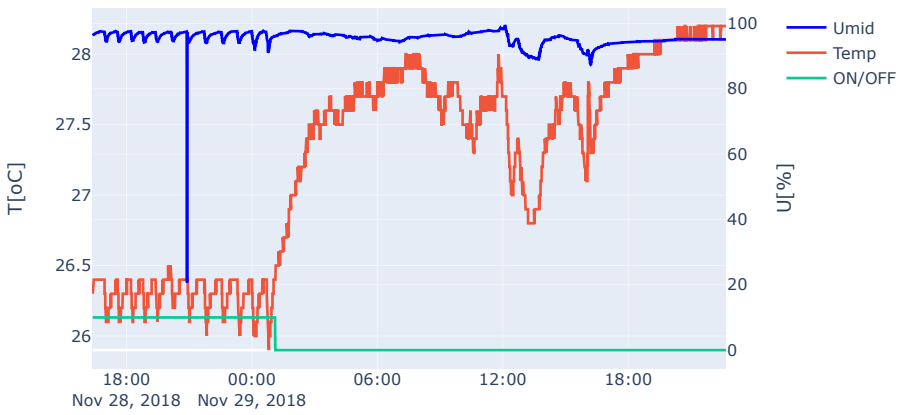
O número de colunas do dataset resultante é de: 27

Sobre a rotulagem dos dados

A sequência de gráficos apresentados abaixo mostra o aspecto das curvas de temperatura e umidade, bem como a sinalização do estado do ar condicionado, se ligado ou desligado. Essas rotulagens foram feitas manualmente com base na experiência e observação dos sinais frente aos comandos de liga e desliga dos aparelhos.







Sobre as implementações propostas no Trabalho da Disciplina

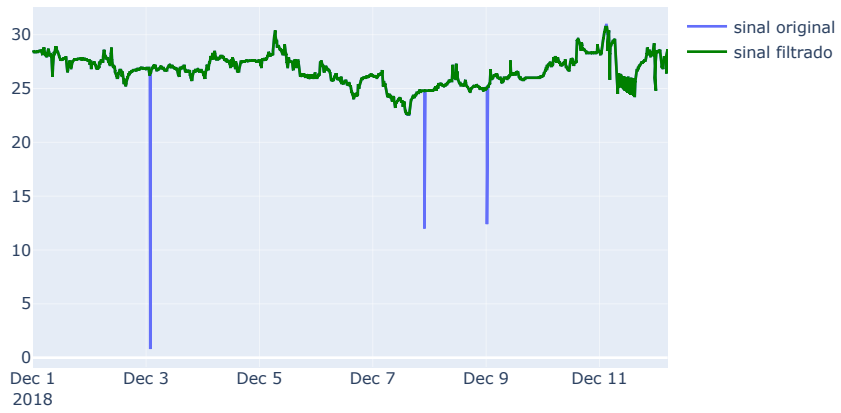
- Análise exploratória e descarte de atributos desnecessários

Esta etapa já foi iniciada por meio da organização dos dados, rotulagem e seleção de períodos de interesse. Também foram realizados testes rápidos com features não otimizadas para comprovar a hipótese de classificação por meio de *machine learning*. A proposição de diferentes tempos de janelamento (*rolling*) para as duas variáveis medidas também fez parte da análise exploratória dos dados.

Essa etapa contará ainda com uma interpretação exploratória para seleção/descarte de atributos utilizando software RapidMiner e bibliotecas python indicadas na discplina de DM (Seaborn et al).

- Missing values

O trabalho deverá compreender análise de termos faltantes e filtragem de dados. As etapas de aquisição, por vezes, envolvem perda do dado ou medição errática que deverão ser tratadas nessa etapa com uma abordagem de pré-processamento. Existem também casos de outliers e sensores que entraram e falha o degradação com o passar do tempo. No gráfico abaixo, é apresentado um caso de filtragem que será aplicada no trabalho.



- Balanceamento e normalização

Será realizada uma análise de balanceamento dos dados considerando as métricas de performance apresentadas na disciplina de DM (precisão, revocação, F1 score, matriz de confusão, et al). Também serão considerados testes de normalização para avaliar o impacto nos resultados do processo.

- Testes de Modelos ML e transformação dos dados

Por fim, será realizado uma sessão de testes compreensivos nos modelos de ML apresentados na disciplina de ML. Para tal, serão levadas em conta eventuais necessidades de transformações nos dados. Dentre os métodos de **classificação** a serem testados, pode-se citar SVM (e suas variações), *nearest neighbors*, regressão logística, árvores de decisão e Emsembles (Gradient Boosting, por exemplo).