



UNIVERSIDADE FEDERAL DO PIAUÍ  
CAMPUS SENADOR HELVÍDIO NUNES DE BARROS  
DISCIPLINA: SISTEMAS INTELIGENTES  
PROFESSORA: DEBORAH MAGALHÃES  
MONITOR: DAVI LUÍS

## **TRABALHO COMPUTACIONAL**

### **1. Descrição do Trabalho**

Nesta etapa, nós construímos um notebook para prever os valores de venda de casas de um região de Seattle, EUA. Desse modo, nos familiarizamos com a linguagem Python para análise de dados: executando comandos no jupyter notebook, carregando e transformando dados, calculando estatísticas dos dados e, finalmente, construindo um modelo de regressão. Com base nisso, **o objetivo deste trabalho é construir um modelo de regressão mais preciso para prever os preços das casas considerando mais características dos dados.**

#### **1.1. O que você eu vou precisar?**

- Ambiente configurado (seguir os slide SI-setup disponibilizados no Sigaa);
- Realizar o download dos dados (**home\_data.gl.zip**) disponível do repositório da disciplina.

#### **1.2. O que deve ser feito?**

Três passos deverão ser seguidos neste trabalho computacional:

1. **Seleção e estatística:** selecione apenas as casas com o zipcode = XXX e calcule o preço médio de vendas.
2. **Filtrar os dados:** usar filtros lógicos<sup>1</sup> para selecionar apenas as casas com o atributo 'sqft\_living' maior que 2000 sqft e menor que 4000 sqft. Calcule a fração das casas que estão nesse intervalo em relação ao todo.
3. **Construir um modelo de regressão:** considerando um número maior de características (caracteristicas\_avancadas), construa um novo modelo de regressão. Em seguida, calcule o RMSE (root mean squared error) nos dados de teste para o modelo usando o vetor

---

<sup>1</sup> <https://turi.com/products/create/docs/generated/graphlab.SFrame.html>

"caracteristicas" e o modelo usando o vetor "caracteristicas\_avancadas". Qual é a diferença no RMSE entre o modelo treinado com "caracteristicas" e "caracteristicas\_avancadas"?

```
caracteristicas = ['bedrooms', 'bathrooms', 'sqft_living', 'sqft_lot', 'floors',  
'zipcode']  
  
caracteristicas_avancadas = ['bedrooms', 'bathrooms', 'sqft_living', 'sqft_lot',  
'floors', 'zipcode', 'conditions', 'waterfront', 'grade', 'view', 'sqft_above',  
'sqft_basement' ]
```

**Nota 1:** ambos os modelos devem ser treinados com o conjunto de dados original e não o conjunto filtrado;

**Nota 2:** ao fazer a divisão treino/teste, certifique-se de passar o parâmetro seed = 0 para a função random\_split;

**Nota 3:** quando utilizar a função linear\_regression.create(), certifique-se de usar o parâmetro validation\_set = None.

## 2. Avaliação

Este trabalho entra na composição da nota referente a segunda avaliação parcial da disciplina e deverá ser entregue no dia **07/12**. A nota do trabalho é **individual** e assumirá o valor de **0-1.5**. Faça o Pull Request com apenas a pasta com seu nome contendo o notebook com extensão "ipynb". **Não** é necessário incluir os dados e, por favor, **não** exclua nenhum arquivo do repositório. **O prazo final para a entrega deste trabalho é 12/04/2017 (quinta-feira).**

**Atenção:** se identificada a cópia de código, a nota **zero** será atribuída aos envolvidos.