

Mãos na massa: Analisando os dados e criando um experimento

Chegou a hora de você executar o que foi visto na aula! Para isso, execute os passos listados abaixo.

Será utilizado o **RStudio** como interface gráfica, então todos os comandos aqui mostrados deverão ser executados no seu console.

1) Abra o **RStudio** e carregue o script **Exporta_Base_MongoDB.R**, que você pode baixar [aqui](https://s3.amazonaws.com/caelum-online-public/731-pipeline-big-data/05/arquivos/Exporta_Base_MongoDB.R) (https://s3.amazonaws.com/caelum-online-public/731-pipeline-big-data/05/arquivos/Exporta_Base_MongoDB.R). Caso a acentuação do conteúdo do script esteja errada, abra-o novamente, acessando o menu **File -> Reopen with Encoding...**, escolhendo **UTF-8** em seguida.

2) Especifique o diretório de trabalho, executando o comando `setwd` do script, por exemplo:

```
setwd("C:\\Users\\eduar\\OneDrive\\aBig Data\\aAlura\\R")
```

3) Caso você ainda não tenha feito, instale o *package* **mongolite**:

```
install.packages('mongolite')
```

Com ele instalado, coloque-o em memória:

```
library(mongolite)
```

4) Com o MongoDB já rodando na sua máquina, faça uma conexão com o mesmo:

```
m <- mongo("ovnis", url = "mongodb://localhost:27017/ovni")
```

5) Traga tudo o que está no MongoDB e salve em um *data frame*:

```
df_OVNI <- m$find ('{}')
```

6) Por fim, salve os dados em um CSV:

```
write.csv(rbind(df_OVNI), file = "OVNIs_Preparados.csv")
```

Criando um experimento

7) Acesse o **Azure Machine Learning** (<https://studio.azureml.net/>) e faça o login com a sua conta da Microsoft. Primeiramente, é preciso disponibilizar um *dataset*, então, no menu à esquerda, acesse **DATASET**, clique em **NEW**, no menu inferior, escolha **DATASET** e faça o upload do CSV que você acabou de gerar, o **OVNIs_Preparados.csv**.

8) Crie um experimento, acessando **EXPERIMENTS**, no menu à esquerda, clicando em **NEW**, no menu inferior, escolhendo **EXPERIMENT** e clicando em **Blank Experiment**.

9) Primeiramente, adicione o *dataset* salvo, pesquise por **ovni** e arraste o seu *dataset* para o experimento. Aproveite e modifique também o seu nome. Ao clicar com o botão direito do mouse no *dataset* e selecionar **Visualize**, você pode explorar um pouco os seus dados.

10) Adicione um **Split Data** ao experimento, conectando a porta inferior do *dataset* a ele. Nas suas propriedades, modifique a sua proporção para 70% (0.7) e em **Random seed**, coloque qualquer número inteiro.

11) Adicionado o **Split Data** ao experimento, acrescente o algoritmo **Multiclass Logistic Regression** e o **Train Model**, conectando a porta inferior da esquerda do **Split Data** à porta superior da direita do **Train Model**, e à sua porta superior da esquerda, conecte a porta inferior do **Multiclass Logistic Regression**.

12) Para escolher a coluna que o *train* tentará descobrir o seu conteúdo, acesse as suas propriedades e clique em **Launch column selector**. Em seguida, escolha a coluna **Shape**.

13) Para medir a eficiência do modelo, adicione um **Score Model**, conectando a sua porta superior da esquerda à porta inferior do **Train Model**, e à sua porta superior da direita, conecte a porta inferior da direita do **Split Data**.

14) Por último, para avaliar o modelo, adicione um **Evaluate Model**, conectando a sua porta superior da esquerda à porta inferior do **Score Model**.

15) Salve o experimento e execute-o, clicando em **RUN**. Finalizada a execução, clique com o botão direito do mouse em **Evaluate Model** e selecione **Visualize** para ver as métricas.

E se remover o tipo Light?

16) Adicione um **Apply SQL Transformation** ao experimento, conecte a sua porta superior da esquerda à porta inferior do *dataset*, removendo a sua conexão com o **Split Data**.

17) Nas propriedades do **Apply SQL Transformation**, modifique o seu SQL:

```
select * from t1
where shape <> 'Light';
```

Com isso, você irá selecionar os OVNI's que não tenho o tipo **Light**.

18) Por fim, conecte a porta inferior do **Apply SQL Transformation** à porta superior do **Split Data**. Salve o experimento e execute-o, clicando em **RUN**. Finalizada a execução, clique com o botão direito do mouse em **Evaluate Model** e selecione **Visualize** para ver as métricas.