INSTITUTO FEDERAL
Brasília
Campus Brasília

TECNOLOGIA EM SISTEMAS PARA INTERNET

Vitor Rodrigues Ferreira
Nínive Helen Horácio da Silva

RELATÓRIO DE PRÁTICA INTEGRADA DE CIÊNCIA DE DADOS E APRENDIZADO DE MÁQUINA

Brasília - DF

03/02/2022

Sumário

1. Objetivos

O desenvolvimento da sprint 2 teve como objetivo o armazenamento dos dados coletados e preparados da Sprint 1, gerando um csv dos dados para serem armazenados no MongoDB, Após o armazenamento desses dados o objetivo era utilizar a plataforma Azure ML, uma plataforma da Microsoft para machine learning (aprendizado de máquina). Utilizando os dados do csv gerados na sprint 1 para criar um modelo preditivo a respeito dos dados da coluna movimento, movimentos que são esses gerados no cotidiano como: escovar dentes, correr, andar.

2. Descrição do problema

Ao final da Sprint 1, tendo completado a etapa de preparação, obtivemos um arquivo CSV (Comma Separated Values) contendo as medidas registradas pelo acelerômetro para os voluntários, em cada instante de tempo.

Entretanto, a forma como o dado foi armazenado - arquivo de texto simples - não permitiria uma fácil manipulação, dificultaria a implementação de controles de segurança e também de formas de acesso.

O armazenamento em um banco de dados seria uma solução para os problemas relacionados à segurança e facilidade na manipulação. Estando o banco de dados em nuvem, as questões de qualidade e facilidade no acesso também seriam solucionadas. Considerando esses fatores e a estrutura dos dados trabalhados, foi criado um banco no MongoDB Atlas (ferramenta para armazenamento em instâncias do MongoDB em nuvem) e utilizadas as bibliotecas PyMongo, para estruturação e armazenamento dos dados, do CSV para o banco.

Finalmente, na etapa de análise, o propósito era de obter mais informações e de criar modelos preditivos, para conseguir estimativas futuras com base nos dados já coletados e formatados.

A ferramenta de Machine Learning da Microsoft, a Azure ML, permite a montagem de uma pipeline de ações sobre os dados, que se assemelha à criação de um diagrama.

Para realizar a análise então, em linhas gerais, os procedimentos adotados para a predição, utilizando o Azure ML, passaram pela separação dos dados aleatoriamente em dois grupos (*Split Data*) de diferentes proporções, em que um destes grupos serviria como modelo de treinamento (*Train Model*) com a aplicação de uma regressão estatística (*Multiclass Logistic Regression*) sobre a variável de interesse na predição - nesse caso, o tipo do movimento - e então o segundo grupo serviria como modelo para a avaliação da qualidade das predições (*Score Model*) que poderiam ser feitas a partir do modelo de treinamento.

Montada essa estrutura, a análise consistiu na variação das proporções de tamanho entre grupo de teste e grupo de avaliação, e na observação dos diferentes resultados (*Evaluate Model*) obtidos a partir disso.

3. Desenvolvimento

Na sprint 2 foi utilizado em cada etapa a plataforma online google colab, uma plataforma que facilita para analisar dados de maneira gratuita e prática.

Nessa etapa da sprint 2 teve a etapa de armazenamento e a etapa de machine learning(aprendizado de máquina). Na etapa de armazenamento foi utilizado o mongodb, um banco de dados não relacional utilizado na nuvem. Foi criado o banco de dados é feito a conexão utilizando o python, após o banco de dados ter sido criado e conectado, foi utilizado o csv gerado na sprint 1 para guardar os dados no banco de dados, a importação dos dados foi utilizado linhas de códigos para importar o csv para o banco.

Já na etapa de machine learning(aprendizado de máquina), foi utilizado uma plataforma online da Microsoft o Azure ML, uma plataforma para criar modelos de machine learning de maneira mais prática e fácil . Foi criado uma conta gratuita no Azure ML e então foi feito a importação do csv medidas criado na sprint 1, após essa importação foi criado um modelo preditivo dos movimentos realizados no cotidiano,selecionando a coluna movimento.

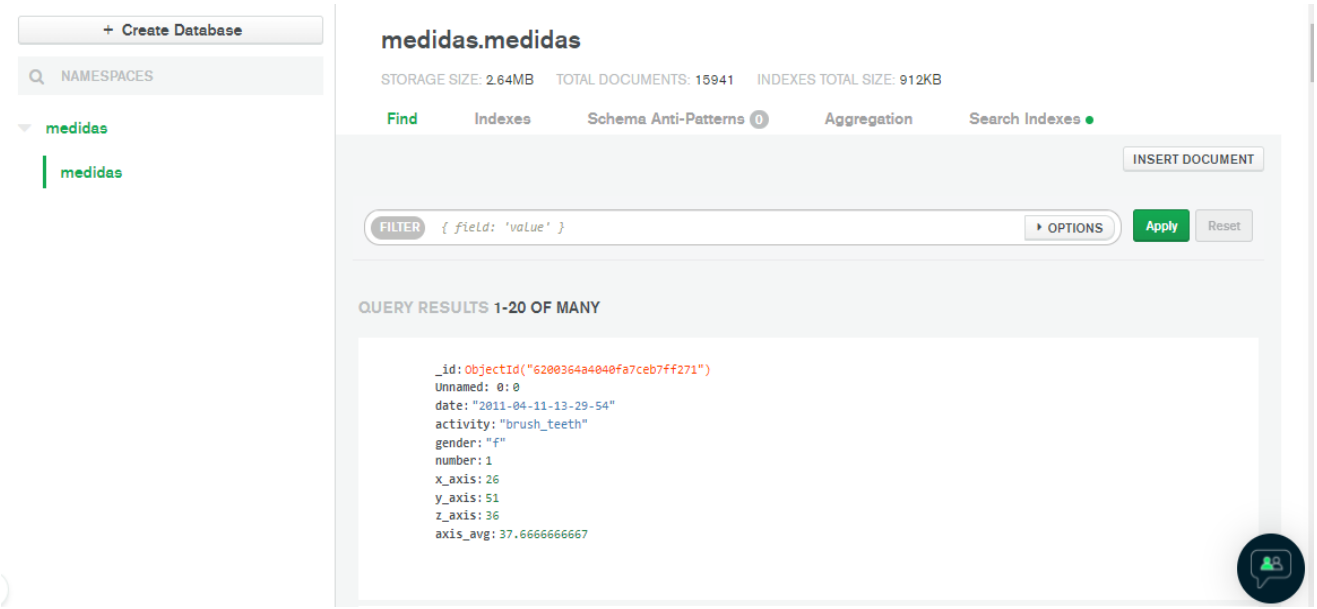
Então foi utilizado o modelo Logistic Regression, um modelo que prever o crescimento de determinados dados, esse modelo foi utilizado para prever a porcentagem de que um movimento acontece e quantos porcentos ele acerta na hora de prever o movimento.

3.1 Código implementado

Armazenamento

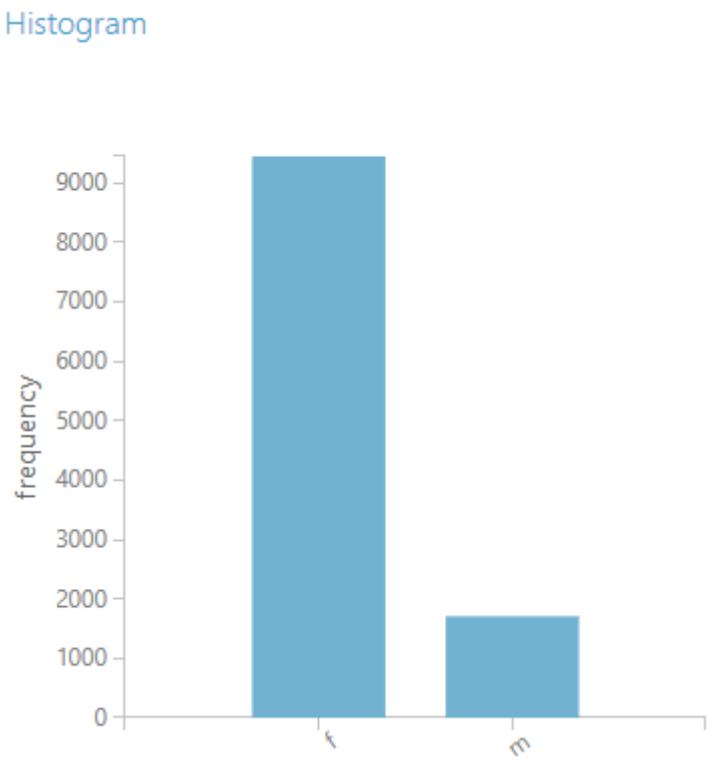
```
# Instalando bibliotecas necessárias
!pip install pymongo
# Instalando bibliotecas necessárias
!pip install dnspython
# Conectando ao banco de dados
import pymongo
myclient =
pymongo.MongoClient("mongodb://g8-vn:dados123456@cluster0-shard-00-00.o1
nvr.mongodb.net:27017,cluster0-shard-00-01.o1nvr.mongodb.net:27017,clust
```

```
er0-shard-00-02.o1nvr.mongodb.net:27017/medidas?ssl=true&replicaSet=atla
s-cbkvj0-shard-0&authSource=admin&retryWrites=true&w=majority")

      # Verificando se o banco existe e se a conexão funcionou.
      dblist = myclient.list_database_names()
      if "medidas" in dblist:
        print("Banco existente.")
      else:
        print("Banco não existente")

      # Código para carregar o CSV no banco de dados.
      import pandas as pd
      from pymongo import MongoClient
      import json

      #Url do banco de dados criado
      db_url                                                          =
"mongodb://g8-vn:dados123456@cluster0-shard-00-00.o1nvr.mongodb.net:2701
7,cluster0-shard-00-01.o1nvr.mongodb.net:27017,cluster0-shard-00-02.o1nv
r.mongodb.net:27017/medidas?ssl=true&replicaSet=atlas-cbkvj0-shard-0&aut
hSource=admin&retryWrites=true&w=majority"
      #Link do csv criado na sprint 1
      csv_url                                                         =
"https://raw.githubusercontent.com/infocbra/pratica-integrada-cd-e-am-20
21-2-g8-vn/master/Sprint1/medidas_preparacao.csv?token=GHSAT0AAAAABRCD7
GPKR5F6LP3JQHCX3GEYQEPGOA"
      #Função para conectar o banco
      def mongoimport(csv_path, db_name, coll_name, db_url):
          client = MongoClient(db_url)
          db = client[db_name]
          coll = db[coll_name]
          data = pd.read_csv(csv_path)
          payload = json.loads(data.to_json(orient='records'))
          coll.delete_many({})
          coll.insert_many(payload)
          return coll.count_documents({})
      #Codigo para enviar csv para o banco.
      mongoimport(csv_url, "medidas", "medidas", db_url)
```



Análise dos dados com o Azure ML

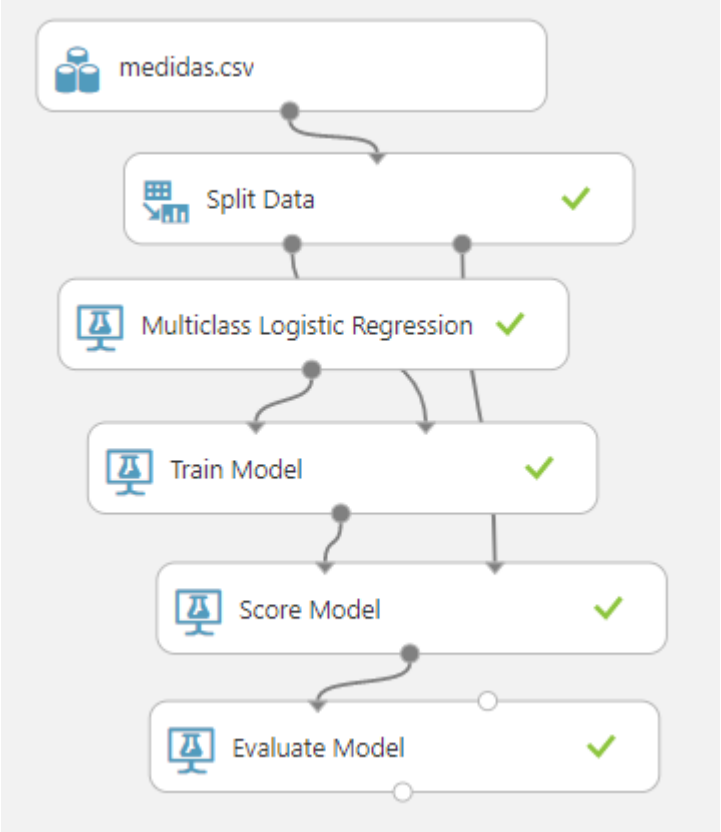
Gráfico de gênero gerado do medidas.csv.



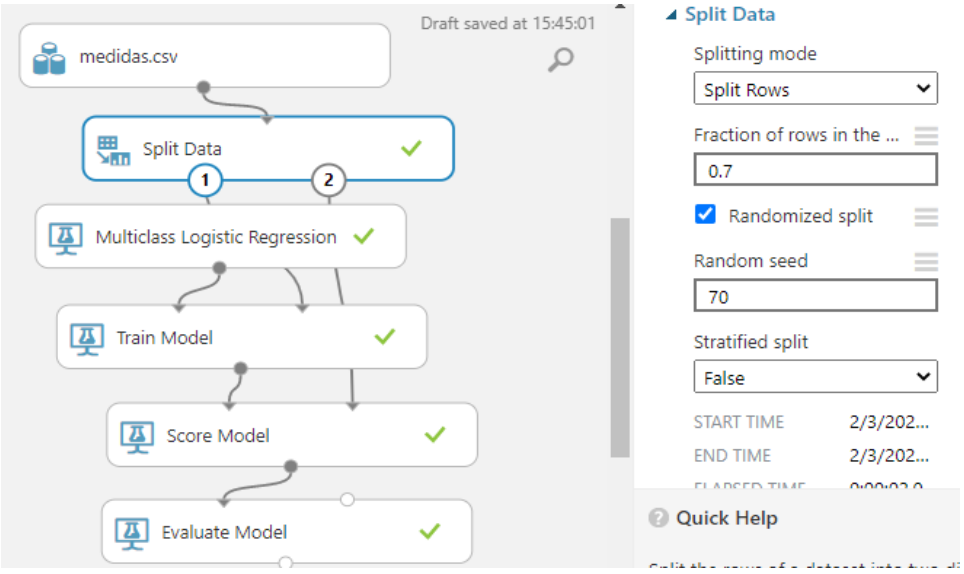
Machine Learning/Azure ML

Utilização da plataforma da microsoft Azure Ml para fazer a previsão dos dados da coluna movimento, prevendo os movimentos.

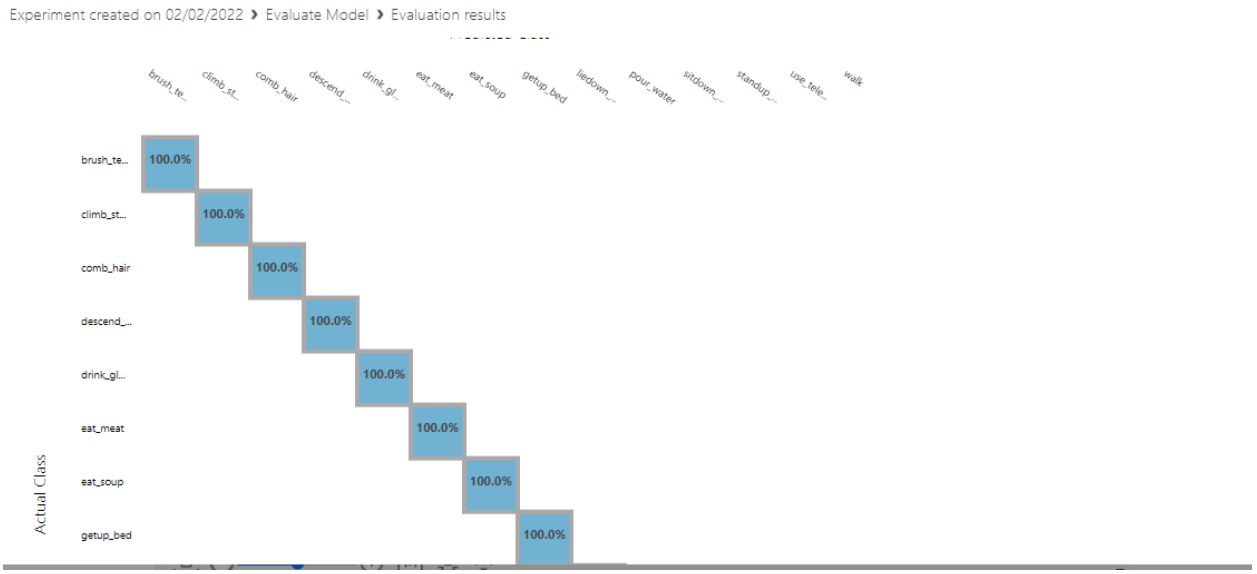
1. Criação do modelo preditivo no Azure ML.
2. Etapas: Adicionar o csv criado na sprint 1 chamado medidas.
3. adicionar o módulo split data e selecionar a coluna movimento.
4. adicionar o modelo de predição logistic regression.
5. Adicionar a etapa de treino, para separar os dados e utilizar eles para a previsão do modelo.
6. Adicionar etapa score model para a porcentagem dos dados.
7. adicionar modelo evaluate model para ver os resultados e a evolução do modelo preditivo é sua e verificar sua previsão.



Etapa de separar os dados/ split Data.



Resultados da previsão do modelo. O modelo teve como 100% a previsão de todos os movimentos. Com a opção Randomized split marcada para utilizar dados do dataset movimento de maneira aleatória na hora de escolher os dados



Com o modelo sem a opção Randomized split desmarcada foram geradas porcentagens diferentes para a previsão dos dados. Com 100% de acerto na previsão do movimento: Getup Bed/ Acordar, Comb Hair/ Pentear o cabelo , Walk/ Caminhar. E com a porcentagem menor de previsão o movimento: Eat Soup/ Tomar sopa.

4. Considerações finais

Foi interessante conhecer e ter este primeiro contato com a Azure ML e o MongoDB Atlas, ver o poder e as possibilidades que trazem.

Conseguimos concluir as etapas propostas nesta sprint sem maiores problemas. As bibliotecas e ferramentas foram de fácil utilização no geral, sendo a Azure ML a que exigiu uma pesquisa mais aprofundada, não pela montagem da estrutura em si, mas pelo entendimento da função de alguns dos componentes.

Referências

MICROSOFT. Docs Azure Machine Learning, c2022. Página inicial. Disponível em: <https://docs.microsoft.com/en-us/azure/machine-learning/>. Acesso em: 26 de janeiro de 2022.

MICROSOFT. Microsoft Machine Learning Studio (classic), c2022. Página inicial. Disponível em: <https://studio.azureml.net/>. Acesso em: 05 de fevereiro de 2022.

W3SCHOOLS. Python MongoDB, c2022. Página inicial. Disponível em: <https://www.w3schools.com/python/python_mongodb_getstarted.asp/>. Acesso em: 27 de janeiro de 2022.