DataTalks ML Zoomcamp Week 7 - Midterm

For the midterm project, I have chosen to apply my knowledge using the data about Metabolic Syndrome from Kaggle. The information about dataset from Kaggle is as follows:
"This dataset contains information on individuals with metabolic syndrome, a complex medical condition associated with a cluster of risk factors for cardiovascular diseases and type 2 diabetes. The data includes demographic, clinical, and laboratory measurements, as well as the presence or absence of metabolic syndrome.

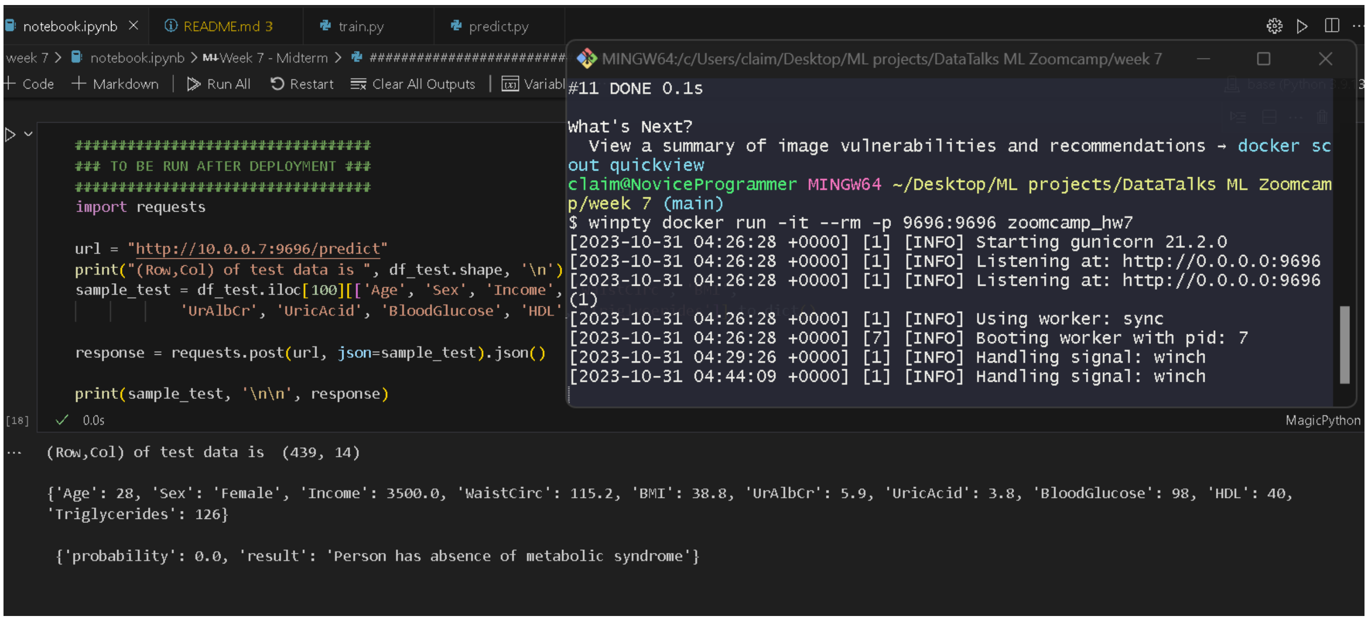
Column Descriptors:

- seqn: Sequential identification number.
- Age: Age of the individual.
- Sex: Gender of the individual (e.g., Male, Female).
- Marital: Marital status of the individual.
- Income: Income level or income-related information.
- Race: Ethnic or racial background of the individual.
- WaistCirc: Waist circumference measurement.
- BMI: Body Mass Index, a measure of body composition.
- Albuminuria: Measurement related to albumin in urine.
- UrAlbCr: Urinary albumin-to-creatinine ratio.
- UricAcid: Uric acid levels in the blood.
- BloodGlucose: Blood glucose levels, an indicator of diabetes risk.
- HDL: High-Density Lipoprotein cholesterol levels (the "good" cholesterol).
- Triglycerides: Triglyceride levels in the blood.
- MetabolicSyndrome: Binary variable indicating the presence (1) or absence (0) of metabolic syndrome."

Using the concepts of Machine Learning, we clean the data and apply a model to figure out what factors contribute to Metabolic Syndrome.

Instructions:

- Fork the repo.
- run `pip install pipenv` to ensure you can create virtual environment.
- run `pipenv shell` and then `pipenv install` to install the dependecnies.
- run `pipenv run python train.py` to re-create the needed models.
- assuming you have all the files (check Dockerfile for reference), run `docker build -t random_image_name .` Note that random_image_name can be any name you want to give.
- Now run `winpty docker run -it --rm -p 9696:9696 zoomcamp_hw7` to start the web service (you may not need winpty but my device needs it.)



```
####################################
### TO BE RUN AFTER DEPLOYMENT ###
####################################
import requests

url = "http://10.0.0.7:9696/predict"
print("(Row,Col) of test data is ", df_test.shape, '\n')
sample_test = df_test.iloc[100][['Age', 'Sex', 'Income',
            'UrAlbCr', 'UricAcid', 'BloodGlucose', 'HDL'

response = requests.post(url, json=sample_test).json()

print(sample_test, '\n\n', response)
```

```
(Row,Col) of test data is  (439, 14)

{'Age': 28, 'Sex': 'Female', 'Income': 3500.0, 'WaistCirc': 115.2, 'BMI': 38.8, 'UrAlbCr': 5.9, 'UricAcid': 3.8, 'BloodGlucose': 98, 'HDL': 40,
'Triglycerides': 126}

 {'probability': 0.0, 'result': 'Person has absence of metabolic syndrome'}
```

```
#11 DONE 0.1s

What's Next?
  View a summary of image vulnerabilities and recommendations → docker sc
out quickview
claim@NoviceProgrammer MINGW64 ~/Desktop/ML projects/DataTalks ML Zoomcam
p/week 7 (main)
$ winpty docker run -it --rm -p 9696:9696 zoomcamp_hw7
[2023-10-31 04:26:28 +0000] [1] [INFO] Starting gunicorn 21.2.0
[2023-10-31 04:26:28 +0000] [1] [INFO] Listening at: http://0.0.0.0:9696
[2023-10-31 04:26:28 +0000] [1] [INFO] Listening at: http://0.0.0.0:9696
(1)
[2023-10-31 04:26:28 +0000] [1] [INFO] Using worker: sync
[2023-10-31 04:26:28 +0000] [7] [INFO] Booting worker with pid: 7
[2023-10-31 04:29:26 +0000] [1] [INFO] Handling signal: winch
[2023-10-31 04:44:09 +0000] [1] [INFO] Handling signal: winch
```

In the image abovem it can be seen as the model is deployed, we can interact with it to send test data and get results.

Feel free to share your feedback and suggesttions!