Machine Learning on Kubernetes

Shruti Kavishwar
San Francisco Bay University
Guided by: Prof. Henry Chang

Agenda

- Introduction
- Setup Kubernetes Cluster
- Create necessary files
- Build Docker container and run it
- Access Web UI for Flask
- Kill the Docker container
- Conclusion

Introduction

Purpose:

- The goal of this project is to demonstrate the deployment of a machine learning model using Docker containers and Kubernetes clusters on Google Cloud Platform (GCP).

Key Components:

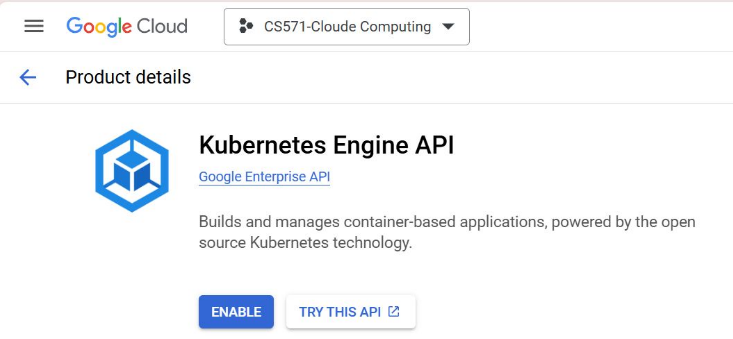
- **Machine Learning Model:**
  - A model trained to predict certain outcomes based on data inputs.
- **Docker:**
  - A platform for developing, shipping, and running applications inside containers, which provide a consistent environment for the application to run.
- **Kubernetes**:
  - An open-source platform for automating deployment, scaling, and operations of application containers across clusters of hosts.

Project Objectives:

- To set up a Kubernetes cluster on GCP.
- To containerize the machine learning application using Docker.
- To deploy and manage the application using Kubernetes.

Setup Kubernetes Cluster

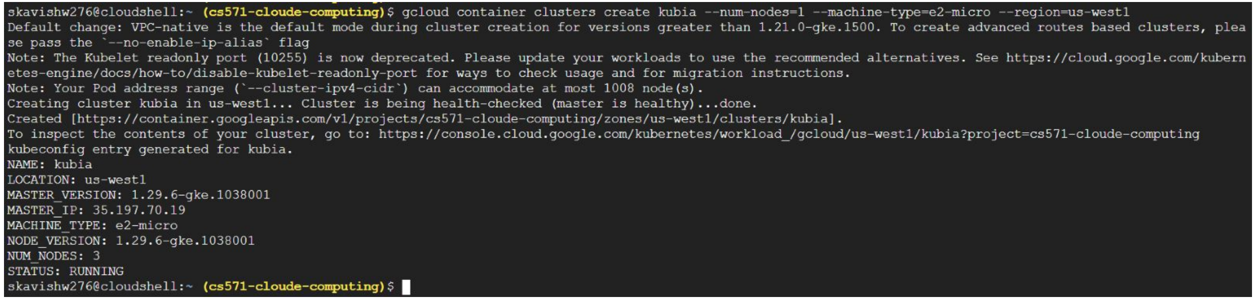
- Steps
  - Enable Kubernetes API in Google Cloud Platform
  - Create a 3 node GKE cluster using the GCP console and below command
    - $ gcloud container clusters create kubia --num-nodes=1 --machine-type=e2-micro --region=us-west1
  - Verify if the nodes are created
    - $ kubectl get nodes



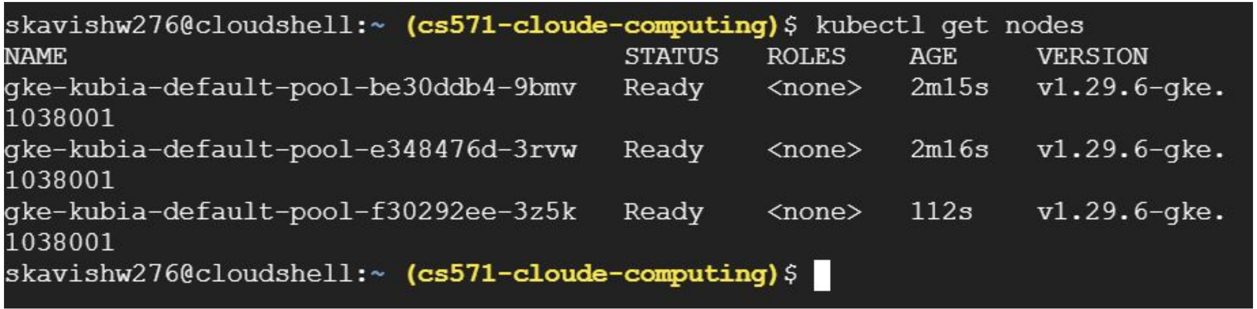
Kubernetes Engine API

Google Enterprise API

Builds and manages container-based applications, powered by the open source Kubernetes technology.



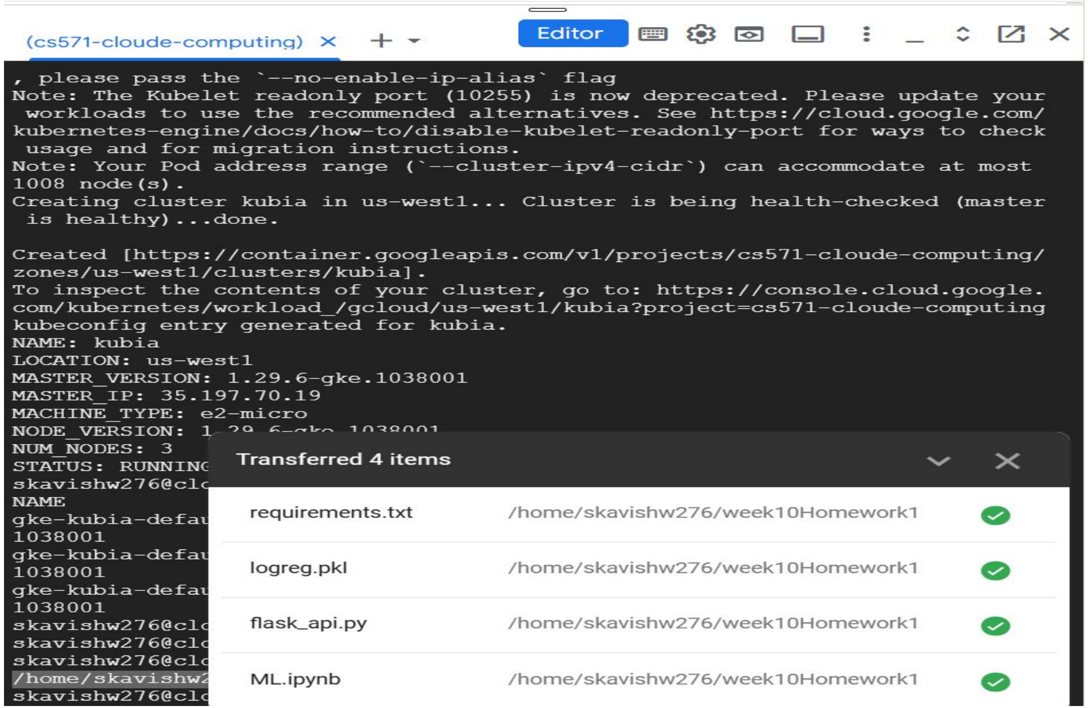
```
skavishw276@cloudshell:~ (cs571-cloude-computing)$ gcloud container clusters create kubia --num-nodes=1 --machine-type=e2-micro --region=us-west1
Default change: VPC-native is the default mode during cluster creation for versions greater than 1.21.0-gke.1500. To create advanced routes based clusters, plea
se pass the `--no-enable-ip-alias` flag
Note: The Kubelet readonly port (10255) is now deprecated. Please update your workloads to use the recommended alternatives. See https://cloud.google.com/kubern
etes-engine/docs/how-to/disable-kubelet-readonly-port for ways to check usage and for migration instructions.
Note: Your Pod address range (`--cluster-ipv4-cidr`) can accommodate at most 1008 node(s).
Creating cluster kubia in us-west1... Cluster is being health-checked (master is healthy)...done.
Created [https://container.googleapis.com/v1/projects/cs571-cloude-computing/zones/us-west1/clusters/kubia].
To inspect the contents of your cluster, go to: https://console.cloud.google.com/kubernetes/workload_/gcloud/us-west1/kubia?project=cs571-cloude-computing
kubeconfig entry generated for kubia.
NAME: kubia
LOCATION: us-west1
MASTER_VERSION: 1.29.6-gke.1038001
MASTER_IP: 35.197.70.19
MACHINE_TYPE: e2-micro
NODE_VERSION: 1.29.6-gke.1038001
NUM_NODES: 3
STATUS: RUNNING
skavishw276@cloudshell:~ (cs571-cloude-computing)$
```



```
skavishw276@cloudshell:~ (cs571-cloude-computing)$ kubectl get nodes
NAME                                   STATUS   ROLES    AGE     VERSION
gke-kubia-default-pool-be30ddb4-9bmv   Ready    <none>   2m15s   v1.29.6-gke.
1038001
gke-kubia-default-pool-e348476d-3rvw   Ready    <none>   2m16s   v1.29.6-gke.
1038001
gke-kubia-default-pool-f30292ee-3z5k   Ready    <none>   112s    v1.29.6-gke.
1038001
skavishw276@cloudshell:~ (cs571-cloude-computing)$
```

Create necessary files

- Steps
    - Create a directory to work in. (eg: week10Homework1) on GCP console.
        - flask_api.py
        - requirements.txt
        - logreg.pkl
        - ML.ipynb
        - Dockerfile

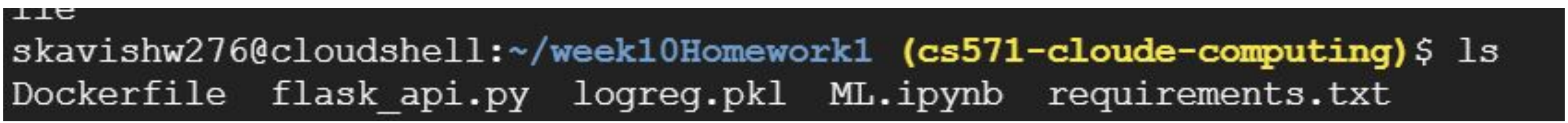


```
, please pass the `--no-enable-ip-alias` flag
Note: The Kubelet readonly port (10255) is now deprecated. Please update your
 workloads to use the recommended alternatives. See https://cloud.google.com/
kubernetes-engine/docs/how-to/disable-kubelet-readonly-port for ways to check
 usage and for migration instructions.
Note: Your Pod address range (`--cluster-ipv4-cidr`) can accommodate at most
1008 node(s).
Creating cluster kubia in us-west1... Cluster is being health-checked (master
 is healthy)...done.

Created [https://container.googleapis.com/v1/projects/cs571-cloude-computing/
zones/us-west1/clusters/kubia].
To inspect the contents of your cluster, go to: https://console.cloud.google.
com/kubernetes/workload_/gcloud/us-west1/kubia?project=cs571-cloude-computing
kubeconfig entry generated for kubia.
NAME: kubia
LOCATION: us-west1
MASTER_VERSION: 1.29.6-gke.1038001
MASTER_IP: 35.197.70.19
MACHINE_TYPE: e2-micro
NODE_VERSION: 1.29.6-gke.1038001
NUM_NODES: 3
STATUS: RUNNING
skavishw276@clo
NAME
gke-kubia-defau
1038001
gke-kubia-defau
1038001
gke-kubia-defau
1038001
skavishw276@clo
skavishw276@clo
skavishw276@clo
/home/skavishw2
skavishw276@clo
```

Transferred 4 items

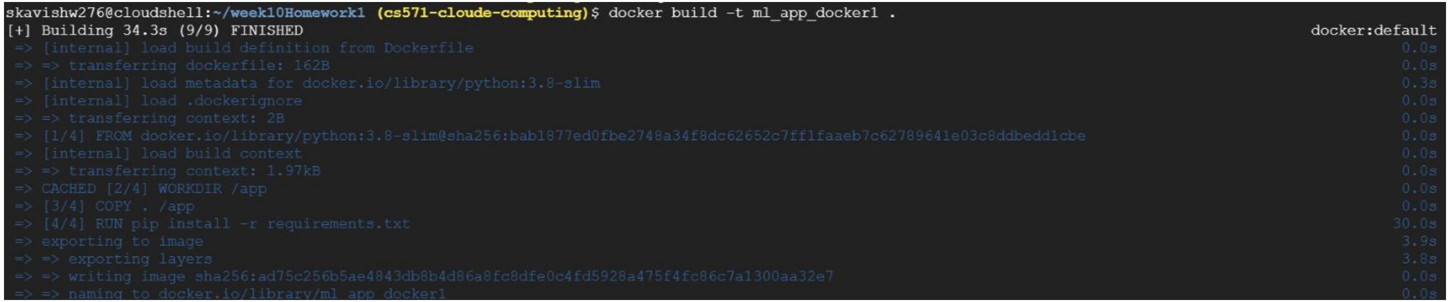
requirements.txt	/home/skavishw276/week10Homework1	✓
logreg.pkl	/home/skavishw276/week10Homework1	✓
flask_api.py	/home/skavishw276/week10Homework1	✓
ML.ipynb	/home/skavishw276/week10Homework1	✓



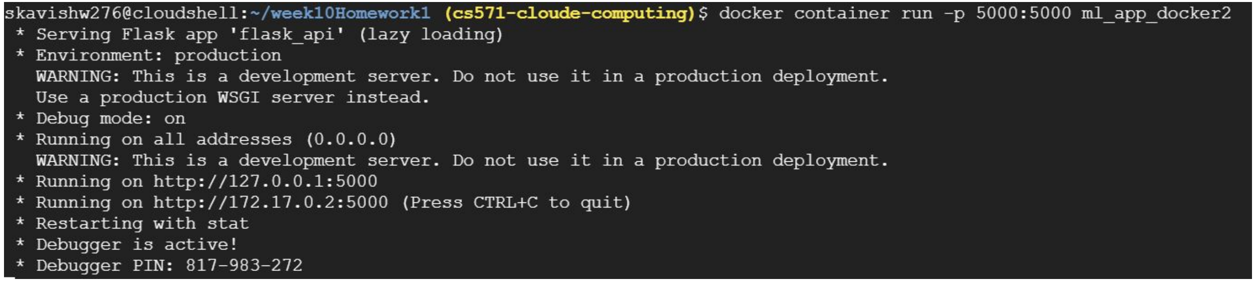
```
skavishw276@cloudshell:~/week10Homework1 (cs571-cloude-computing)$ ls
Dockerfile  flask_api.py  logreg.pkl  ML.ipynb  requirements.txt
```

Build Docker container and Run it

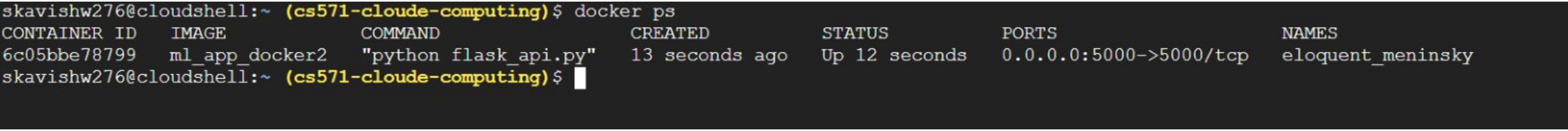
- Steps
  - Build docker image using cli
    - $ sudo docker build –t ml_app_docker1 .
  - Run the docker image
    - $ docker run –t 5000:5000 ml_app_docker1
  - Verify if the container is running
    - $ docker ps



```
skavishw276@cloudshell:~/week10Homework1 (cs571-cloude-computing)$ docker build -t ml_app_docker1 .
[+] Building 34.3s (9/9) FINISHED                                                                                    docker:default
 => [internal] load build definition from Dockerfile                                                                          0.0s
 => => transferring dockerfile: 162B                                                                                          0.0s
 => [internal] load metadata for docker.io/library/python:3.8-slim                                                            0.3s
 => [internal] load .dockerignore                                                                                             0.0s
 => => transferring context: 2B                                                                                               0.0s
 => [1/4] FROM docker.io/library/python:3.8-slim@sha256:bab1877ed0fbe2748a34f8dc62652c7ff1faaeb7c62789641e03c8ddbeddlcbe      0.0s
 => [internal] load build context                                                                                             0.0s
 => => transferring context: 1.97kB                                                                                           0.0s
 => CACHED [2/4] WORKDIR /app                                                                                                 0.0s
 => [3/4] COPY . /app                                                                                                         0.0s
 => [4/4] RUN pip install -r requirements.txt                                                                                30.0s
 => exporting to image                                                                                                        3.9s
 => => exporting layers                                                                                                       3.8s
 => => writing image sha256:ad75c256b5ae4843db8b4d86a8fc8dfe0c4fd5928a475f4fc86c7a1300aa32e7                                  0.0s
 => => naming to docker.io/library/ml_app_docker1                                                                             0.0s
```



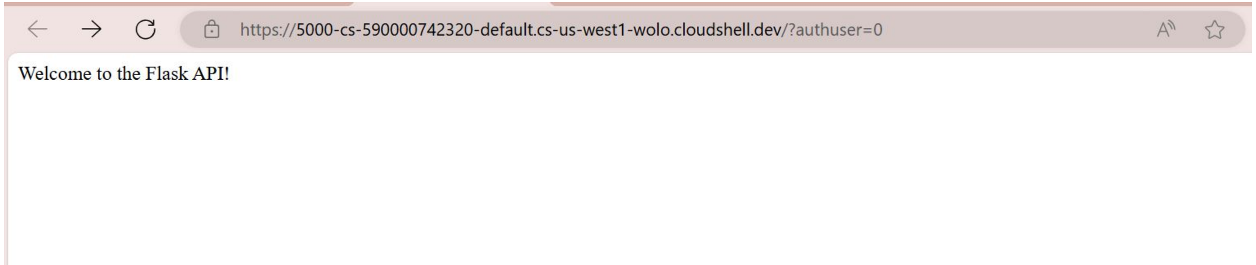
```
skavishw276@cloudshell:~/week10Homework1 (cs571-cloude-computing)$ docker container run -p 5000:5000 ml_app_docker2
 * Serving Flask app 'flask_api' (lazy loading)
 * Environment: production
   WARNING: This is a development server. Do not use it in a production deployment.
   Use a production WSGI server instead.
 * Debug mode: on
 * Running on all addresses (0.0.0.0)
   WARNING: This is a development server. Do not use it in a production deployment.
 * Running on http://127.0.0.1:5000
 * Running on http://172.17.0.2:5000 (Press CTRL+C to quit)
 * Restarting with stat
 * Debugger is active!
 * Debugger PIN: 817-983-272
```



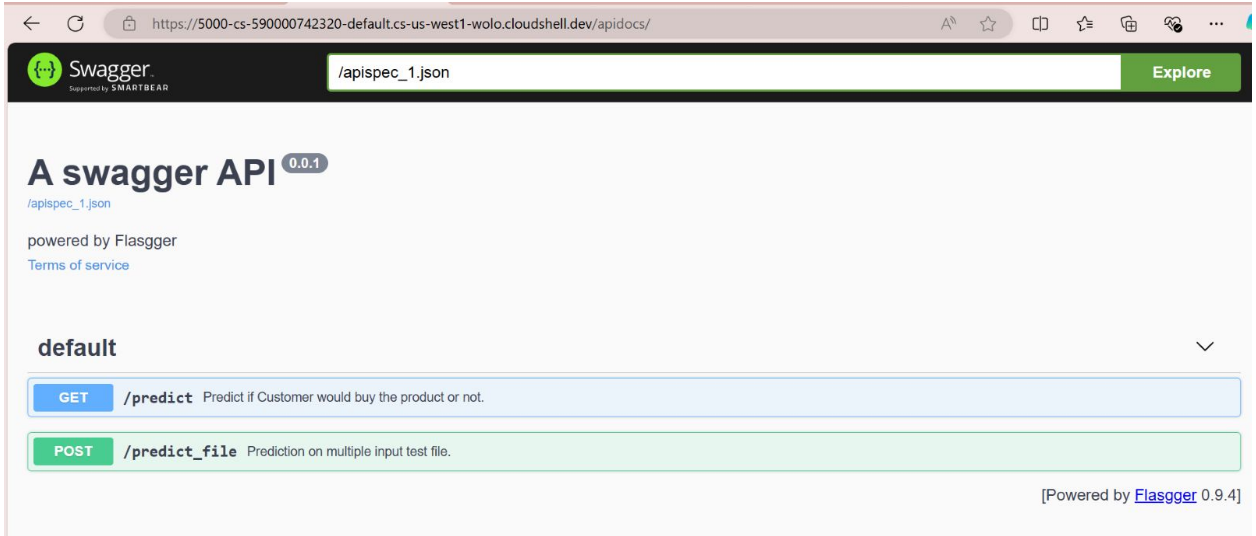
```
skavishw276@cloudshell:~ (cs571-cloude-computing)$ docker ps
CONTAINER ID   IMAGE            COMMAND                  CREATED          STATUS          PORTS                    NAMES
6c05bbe78799   ml_app_docker2   "python flask_api.py"    13 seconds ago   Up 12 seconds   0.0.0.0:5000->5000/tcp   eloquent_meninsky
skavishw276@cloudshell:~ (cs571-cloude-computing)$ 
```

Access the Web UI

- Steps
    - Click on the web preview on the console and change the port to 5000 is necessary and click ‘Preview on Port 5000’
    - Add ‘/apidocs’ to the URL that opens up
    - There are two Tabs: GET and POST
    - GET is for single value prediction and POST is for multiple value prediction
    - In GET add the values and execute
    - In POST add the test_data.csv and execute


https://5000-cs-590000742320-default.cs-us-west1-wolo.cloudshell.dev/?authuser=0

Welcome to the Flask API!


https://5000-cs-590000742320-default.cs-us-west1-wolo.cloudshell.dev/apidocs/
Swagger
Supported by SMARTBEAR
/apispec_1.json
Explore

A swagger API 0.0.1

/apispec_1.json

powered by Flasgger

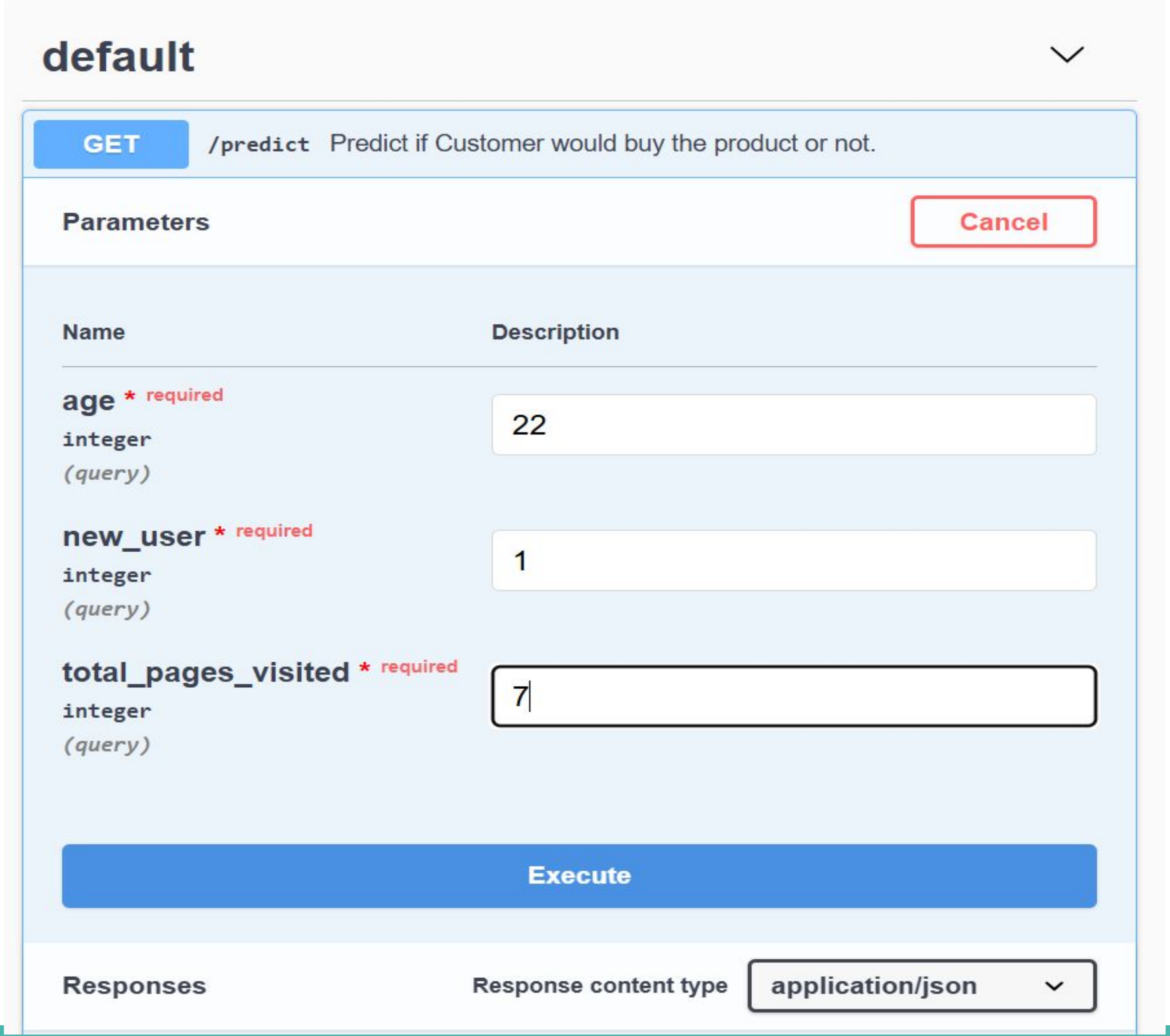
Terms of service

default

GET /predict Predict if Customer would buy the product or not.

POST /predict_file Prediction on multiple input test file.

[Powered by Flasgger 0.9.4]



default

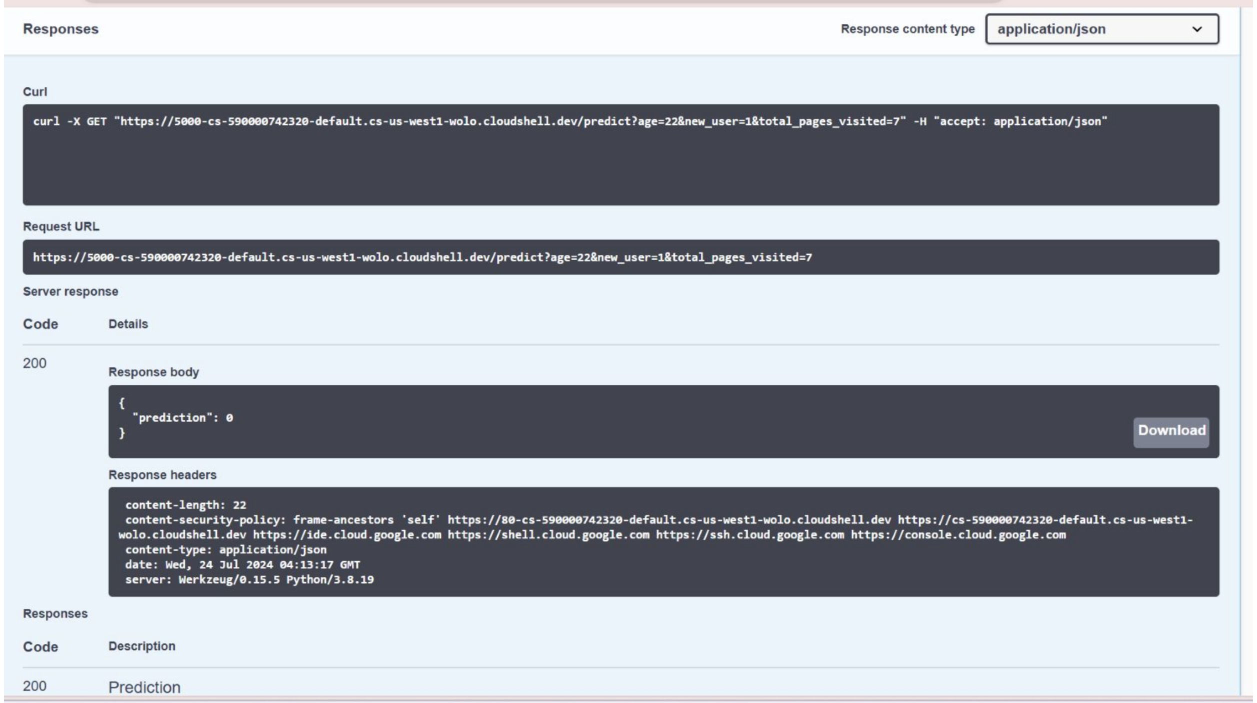
GET /predict Predict if Customer would buy the product or not.

Parameters

Name	Description
age * required integer *(query)*	22
new_user * required integer *(query)*	1
total_pages_visited * required integer *(query)*	7

Responses

Response content type: application/json



Curl

```
curl -X GET "https://5000-cs-590000742320-default.cs-us-west1-wolo.cloudshell.dev/predict?age=22&new_user=1&total_pages_visited=7" -H "accept: application/json"
```

Request URL

```
https://5000-cs-590000742320-default.cs-us-west1-wolo.cloudshell.dev/predict?age=22&new_user=1&total_pages_visited=7
```

Server response

Code	Details
200	Response body

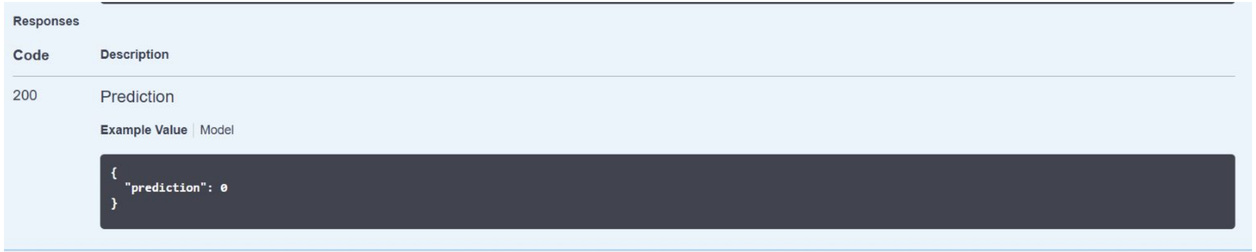
```
{
  "prediction": 0
}
```

Response headers

```
 content-length: 22
 content-security-policy: frame-ancestors 'self' https://80-cs-590000742320-default.cs-us-west1-wolo.cloudshell.dev https://cs-590000742320-default.cs-us-west1-
wolo.cloudshell.dev https://ide.cloud.google.com https://shell.cloud.google.com https://ssh.cloud.google.com https://console.cloud.google.com
 content-type: application/json
 date: Wed, 24 Jul 2024 04:13:17 GMT
 server: Werkzeug/0.15.5 Python/3.8.19
```

Responses

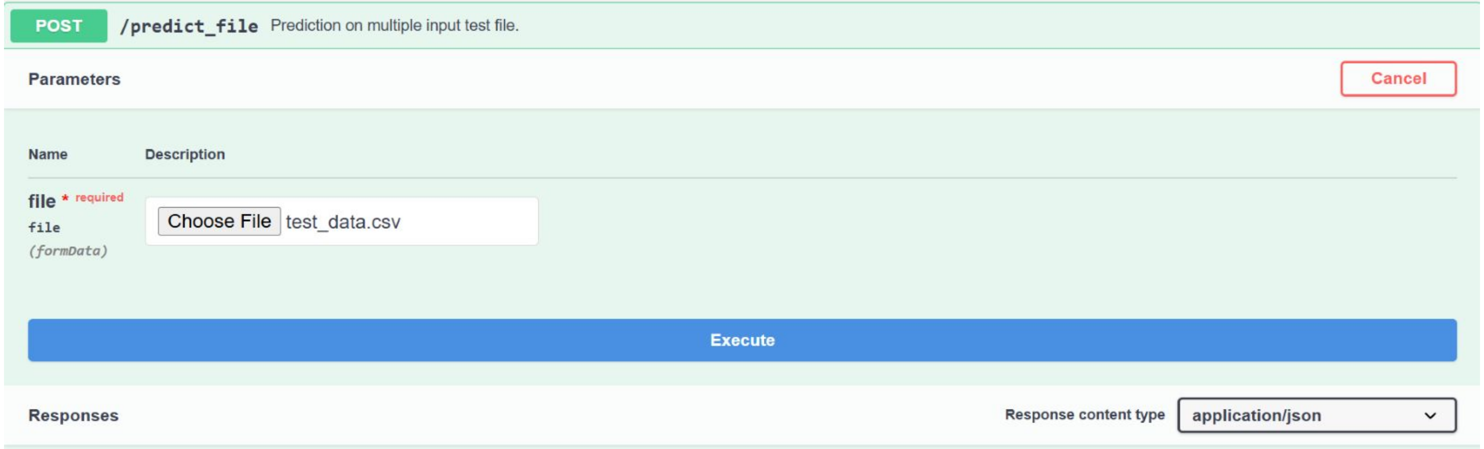
Code	Description
200	Prediction



Responses

Code	Description
200	Prediction

```
{
  "prediction": 0
}
```

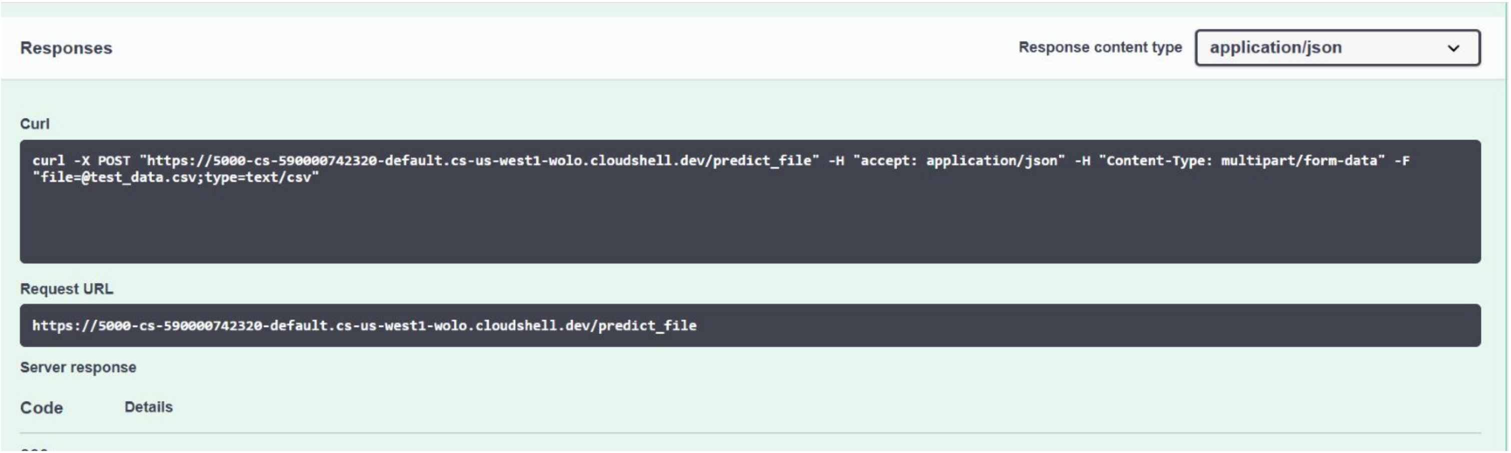


/predict_file Prediction on multiple input test file.

file * required
file
(formData)

Choose File test_data.csv

Response content type: application/json



Response content type: application/json

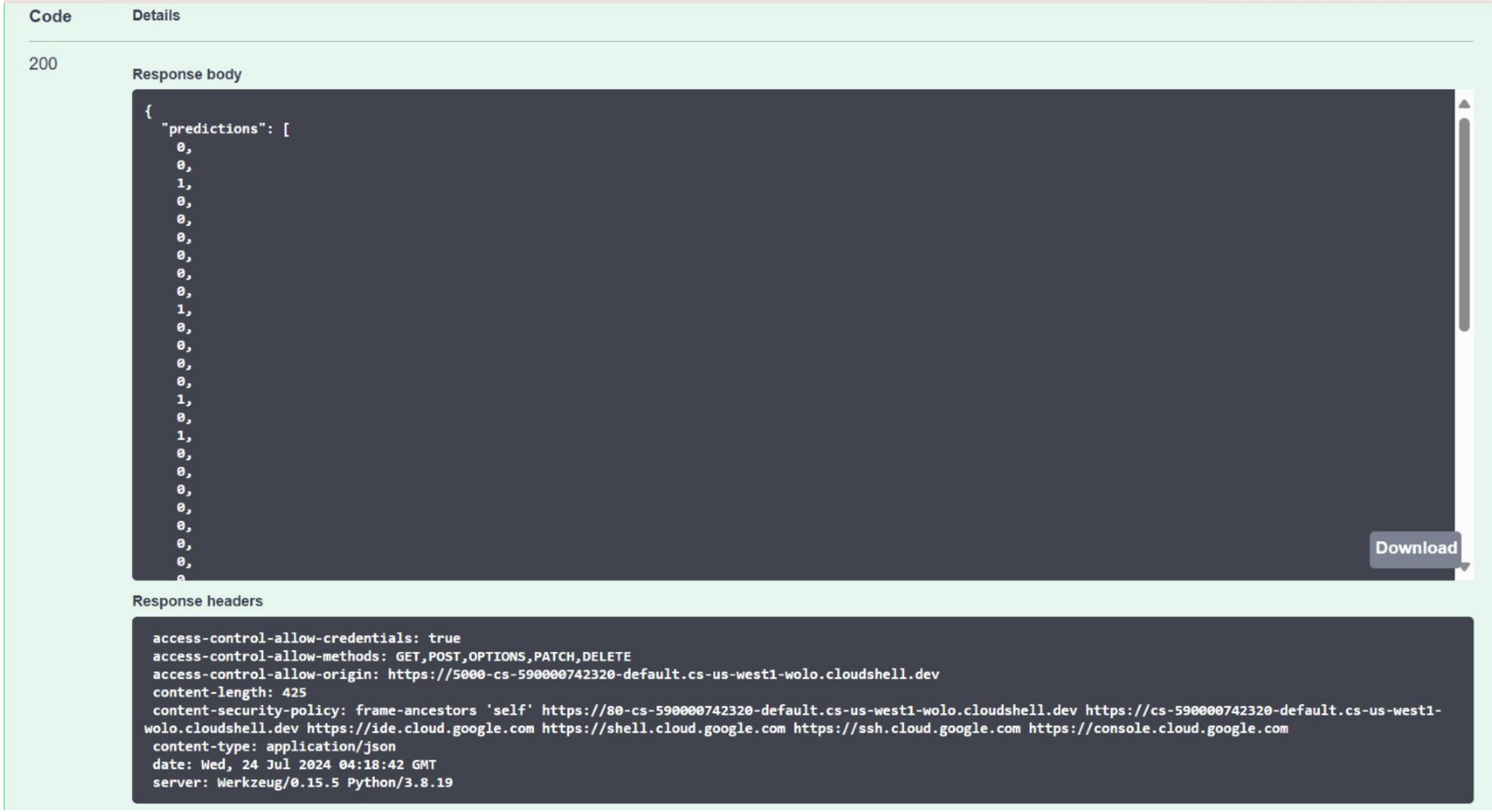
Curl

```
curl -X POST "https://5000-cs-590000742320-default.cs-us-west1-wolo.cloudshell.dev/predict_file" -H "accept: application/json" -H "Content-Type: multipart/form-data" -F
"file=@test_data.csv;type=text/csv"
```

Request URL

```
https://5000-cs-590000742320-default.cs-us-west1-wolo.cloudshell.dev/predict_file
```

Server response



200

Response body

```
{
  "predictions": [
    0,
    0,
    1,
    0,
    0,
    0,
    0,
    0,
    0,
    1,
    0,
    0,
    0,
    0,
    1,
    0,
    1,
    0,
    0,
    0,
    0,
    0,
    0,
    0,
```

Response headers

```
 access-control-allow-credentials: true
 access-control-allow-methods: GET,POST,OPTIONS,PATCH,DELETE
 access-control-allow-origin: https://5000-cs-590000742320-default.cs-us-west1-wolo.cloudshell.dev
 content-length: 425
 content-security-policy: frame-ancestors 'self' https://80-cs-590000742320-default.cs-us-west1-wolo.cloudshell.dev https://cs-590000742320-default.cs-us-west1-wolo.cloudshell.dev https://ide.cloud.google.com https://shell.cloud.google.com https://ssh.cloud.google.com https://console.cloud.google.com
 content-type: application/json
 date: Wed, 24 Jul 2024 04:18:42 GMT
 server: Werkzeug/0.15.5 Python/3.8.19
```

Kill the Docker Container

- Steps
  - Find out the docker container id
    - $ docker ps
  - Delete the container id that is found
    - $ docker kill <container_id>

Conclusion

Summary of the Project:

- The project successfully demonstrated the deployment of a machine learning model using Docker and Kubernetes on the Google Cloud Platform (GCP).
- Key steps included enabling the Kubernetes Engine API, creating a Kubernetes cluster, downloading necessary files, building and running a Docker container, and accessing the deployed application.

Future Work and Improvements:

- **Enhancing the Model:**
  - Further training and fine-tuning of the machine learning model for better accuracy.
- **Security and Monitoring:**
  - Implementing security best practices and monitoring solutions to ensure the application's stability and security.
- **Scaling the Application:**
  - Exploring auto-scaling features in Kubernetes to handle varying levels of traffic and workloads.

Github:
https://github.com/ShrutiK02/Cloud-Computing/tree/aa9fbe4e5b07fb7978ca8417f43327bd653307af/Kubernetes/Machine%20Learning