Deutsche Gesellschaft
für Internationale
Zusammenarbeit (GIZ) GmbH

Open Machine Learning for Earth Observation (ML4EO) in Rwanda: Developing and Implementing an Application-Oriented Training Program for Young Professionals

Final Exam

27th August 2023

DTC, Kigali

Name: _____

Email: _____

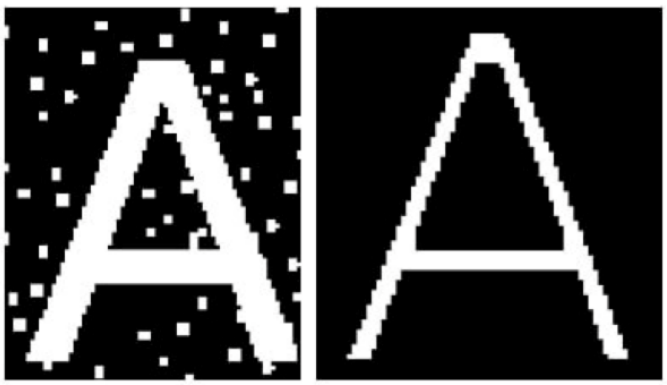
INITIALS: _____

1. Which of the following is not a step in the data cleaning process?
    a. Defining and determining error types
    b. Searching and identifying error instances
    c. Data collection
    d. Documenting error instances and error types

2. In which situation is it suitable to use the "Leave missing values in the dataset" approach for handling missing data?
    a. When the missing values occur systematically and follow a clear pattern.
    b. When using advanced model-based imputation techniques.
    c. When working with decision tree machine learning algorithms.
    d. When the missing data amount is small and occurs randomly.

3. Which of the following is not an EDA (Exploratory Data Analysis) method for single numerical features?
    a. Histogram
    b. Correlation
    c. Box plots
    d. Mean, median, mode

4. What image processing operation was applied to the image on the left to output the image on the right in the diagram shown below?



    a. Dilation
    b. Contrast Stretching
    c. Color segmentation
    d. Erosion

5. From what platform did we fetch the datasets in most of the exercises in this course?
    a. Sentinel Hub
    b. Google Earth Engine
    c. Amazon Web Services
    d. Google Drive

INITIALS: _____

6. A (An) _____ is a time series of images.
    a. RGB image
    b. Image collection
    c. Map
    d. Band

7. Which of the following methods (functions) was used to define the spatial boundaries of a Region-of-Interest (ROI) from an Image collection?
    a. select
    b. map
    c. filterBounds
    d. getMapId

8. To download data from Google Earth Engine (GEE) to Google Drive for training from Google Colaboratory (Google Colab), which of the following steps might be used?
    a. Directly download from GEE to your personal Google Drive
    b. Directly download from GEE to a Google Drive for service account
    c. Directly download from GEE to your personal Google Drive and transfer to a Google Drive for service account
    d. Directly download from GEE to a Google Drive for service account and transfer to a Google Drive for personal account

9. In some of the exercises in this training we can use GEE to train classifiers on our feature collection or export the feature collection to Google Drive in certain formats (e.g CSV) for use with canonical machine learning packages. Which of the following is NOT a disadvantage of training on GEE?
    a. GEE limits training set size
    b. GEE limits the number of bands that can be present in a sample for inference
    c. GEE limits the number of times you can train a classifier
    d. GEE limits model size
    e. GEE has a smaller selection of models

10. Transfer learning, used in machine learning, is the reuse of a pre-trained model on a new problem. Which of the following is not a reason for the use of transfer learning for EO data?
    a. The abundance of task-specific labels for EO applications
    b. Substantial reduction in compute gained by avoiding training a base model on a large image dataset
    c. Gains in model performance over training from scratch.
    d. Lack of enough data for training in an area of interest

INITIALS: _____

11. An EO dataset that contains fields around Kigali from the rainy season taken from Sentinel-2 was used to train a model for vegetation cover regression. The trained model was then used to predict vegetation cover on images taken by Sentinel-2 during the dry season for the same area. The model had a very bad performance. Which of the following is the most likely cause of the poor performance?
    a. Distribution shift resulting from temporal variability of the land cover
    b. Decrease in resolution of the Sentinel-2 sensors during the dry season.
    c. Obstruction of the land by clouds during the dry season
    d. Different distribution of vegetation in the are due to geographic variation
12. Which of the following hyperparameter tuning method where you define a subset of possible values for each hyperparameter and evaluate all possible combinations in the subset.
    a. Grid search
    b. Random search
    c. Bayesian optimization
    d. Genetic Algorithms

13. Which of the following tools can be used for versioning data and model artifacts in machine learning projects?
    a. Data Version Control
    b. Git Large File Storage
    c. Git
    d. Pytorch

14. Which of the following tools can be used to track experiments from multiple runs, compare runs, visualize metrics from runs automatically?
    a. Weights and Biases
    b. Matplotlib
    c. PyTorch
    d. Keras

15. What is the main difference between Continuous Integration/Continuous Deployment (CI/CD) for ML and CI/CD for traditional software development?
    a. Changes to code trigger CI/CD tests to be run for CI/CD for ML but not for traditional software development
    b. Only changes to hyperparameters trigger CI/CD tests to be run for CI/CD for ML while CI/CD tests for traditional software development are triggered by changes to code only.
    c. Changes to code, data, and hyperparameters trigger CI/CD tests to be run for CI/CD for ML while only changes to code trigger CI/CD tests for traditional software development
    d. Changes to code trigger CI/CD tests to be run for CI/CD for traditional software development but not for ML

INITIALS: _____

16. Which of the following deployment platforms is a Platform-as-a-Service?
    a. Heroku
    b. Amazon Web Services Elastic Cloud Compute
    c. Google Cloud Compute Engine
    d. Microsoft Azure

17. In the only Heroku deployment exercise we did as part of this course, for what purposes did we use Heroku for?
    a. Only to train the ML model deployed
    b. Only to serve the model (process requests, run inference using the model, and respond to requests)
    c. To both train the ML model deployed and serve it
    d. To train the ML model deployed, serve it, and re-train the model based on the performance metrics

18. Which of the following has the product development stages in the correct order?
    a. Proof of concept, Minimum Viable Product, Prototype, Production
    b. Minimum Viable Product, Prototype, Proof of Concept, Production
    c. Prototype, Proof of Concept, Minimum Viable Product, Production
    d. Minimum Viable Product, Proof of Concept, Prototype, Production

19. Which of the following is true about instance segmentation and semantic segmentation?
    a. Instance segmentation assigns segment id to detected groups of pixels while semantic segmentation assigns class labels in addition to detecting groups of pixels
    b. Semantic segmentation assigns segment id to detected groups of pixels while instance segmentation assigns class labels in addition to detecting groups of pixels
    c. The two are the same thing
    d. Semantic segmentation can be used with EO data while instance segmentation cannot be used with EO data

20. When using the Sentinel-2 surface reflectance dataset, and Sentinel-2 cloud probability dataset (In Exercise 5.3) to mask clouds and cloud shadows which of the following operations is not applied?
    a. Apply a threshold to the cloud probability band in S2 cloud probability band and use it to mask clouds
    b. Determine which black pixels could have resulted from cloud shadows by using the direction of the sun and nearby clouds
    c. Exclude water bodies using the scene classification band
    d. Remove nighttime data by using timestamps

INITIALS: _____

21. When training deep neural networks, different runs on the same dataset can result in two different models because?
    a. It is mathematically impossible to have two neural networks have the same weights even when trained on the same data
    b. The current implementations of neural networks are buggy
    c. The precision of the floating point numbers used for neural network training is small
    d. Neural networks are usually initialized with random weights

22. Which of the following is the true about EO datasets?
    a. Very high resolution datasets are usually not freely available
    b. It is not possible to use a higher resolution dataset at a lower resolution
    c. In computational resource-limited applications, running on higher resolution is recommended
    d. When there is no computational resource constraint, running on higher resolution dataset and lower resolution dataset always gives same quality results

23. In a few sentences describe one of the ways we generate samples for training, testing, from an Image Collection after filtering by date, filtering by region etc…

    _____
    _____
    _____
    _____
    _____
    _____
    _____
    _____

24. In one sentence explain what the following snippet of code does:

    ```
    img.normalizedDifference(['B8', 'B4'])
    ```

    where *img* is a GEE Image, ‘B8’ is the Near-infrared band (NIR), and ‘B4’ is the red band.

    _____
    _____
    _____
    _____
    _____
    _____
    _____
    _____

INITIALS: _____

25. In Exercise 6.1, we used the **Visible Infrared Imaging Radiometer Suite (VIIRS)** dataset. The **VIIRS** dataset contains average monthly radiance of nighttime for all locations in the world as shown in the image below.

We used this the radiance in this dataset as a feature to classify land as inhabited and uninhabited. Briefly explain how average monthly radiance can be an indicator of inhabited or uninhabited areas.

_____
_____
_____
_____
_____
_____
_____
_____