UNIVERSITY OF NORTH CAROLINA
CHARLOTTE

Deliverable 3: Final Deliverable.

For ITCS 6110 :

Big Data Analytics for Competitive Advantage.

Under Guidance of:

Dr. Pamela Thompson

Submitted by:
(Final Project 2 Group)

Pallavi Shirodkar	801200666 pshirod1@uncc.edu
Krishna Vamsee Vangipurapu	801205116 kvangipu@uncc.edu
Prasanna Kumar Boddupalli	801202443 pboddupa@uncc.edu
Shashank Reddy Godala	801203673 sgodala@uncc.edu
Pratyusha Tummuri	801254220 ltummuri@uncc.edu

Collaborative Development Platform:

https://github.com/pallavi1505/DEA.git

(Access has been given to the TA. Invite has been sent to “pphalle@uncc.edu”)

Understanding Business Needs

"Digital Earth Africa transforms how earth observations can be applied to address some of the current pressing challenges coming in the path of Developing Africa."

Africa is a vast continent with a diverse environment, including various types of biodiversity and homes with one of the fastest growing populations in the world. The population explosion is an additional strain on the natural resources in Africa, is also putting the environment under constant threat from climate change, environment degradation, natural calamities causing severe issues like food security and access to safe water- not only to the humans but also to the flora and fauna.

As a result, a new technological spirit is taking resolving these issues into hands but is hampered by the lack of actionable data. There are many types of data needed which are almost decision ready, for example rate of desertification, illegal mining activities locations, rate of poaching, total revenue from over agriculture, deforestation hotspots.

"Digital Earth Africa" will address this gap between innovative technologies and needed for actionable, structure, veritable data by using analysis ready data produced by Satellites by leveraging The Landsat series of Earth Observation satellites, jointly led by USGS and NASA, which have been continuously acquiring images of the Earth's land surface since 1972.

Using this data, decision makers will be able to identify the common trends and patterns occurring in the given issue and thus identify the magnitude of the issue in the suture, thus combining Descriptive, Prescriptive, Predictive Data Analytics with DE Africa. Like we had earlier seen a use case with "Big Basket using AWS" getting insights about which products to keep in the shelf for a particular time, which stock to increase/ decrease, what marketing strategy to use - on parallel lines, DE will benefit the businesses at ground level as well- starting from Farmers - who will not only rely on the knowledge passed on to them through generations but will also rely on the current trends and insights in agriculture, current diseases among crops, new products in agriculture which are leading to a better yield thus increasing profitability. Growth in small businesses is an equally important factor when considering a nation's economic growth - mere industrial and technological development is never sufficient.

Leveraging the achievement and lessons from Digital Earth Australia and Africa Regional Data Cube - Digital Earth Africa will make it easy to access critical information to shape the positive growth spanning across the entire continent. In conclusion, keeping Digital Earth as a model, it will help not only Africa but also underdeveloped, developing nations in choosing better paths, making better decisions and handpicking options benefiting the national priorities long term sustainability and economic growth.

Source:

Digital Earth Africa Landsat Collection 2 Level 2 - Registry of Open Data on AWS

Reading References:

Data Catalog — Digital Earth Africa 2021 documentation

Platform | Digital Earth Africa

Project Objectives:

1. As a part of this project, we are performing water body(of Africa) analysis using the "Water Observations from Space" dataset (WOfS). By water body analysis, we will try to understand and analyze water resources. This dataset is a collection of images from satellites, in which each pixel is classified based on a single measurement called "water". Each pixel is either classified as dry pixel (water=0) or water pixel (water=128).

2. As a part of our descriptive analysis, once the pixels are extracted, we will calculate the area per pixel and then the total area of water pixels. By this analysis over timestamp of each pixel recorded, we can track how the water body area changes over time. By plotting a time series plot, we can identify the dates/months where there was more or less water within the area of interest.

Technologies and Datasets used:

- Sandbox alias: ls8_sr
- Data type: Surface reflectance
- Data timespan: March 2013 – present
- Available regions: Entire African continent
- Spatial resolution: 30 x 30 m (size of one pixel)
- Spectral resolution: 7 spectral bands — Coastal/Aerosol, Blue, Green, Red, Near-Infrared, Short Wavelength Infrared 1, Short Wavelength Infrared 2

Deciding the Use case:

While deciding the use case we went through the data sets available in WOfS to study the surface water patterns in Africa. We came across the case of Floods in Niamey, Nigeria in September 2020.
To confirm the event we have constructed the case through a data set available in WOfS.

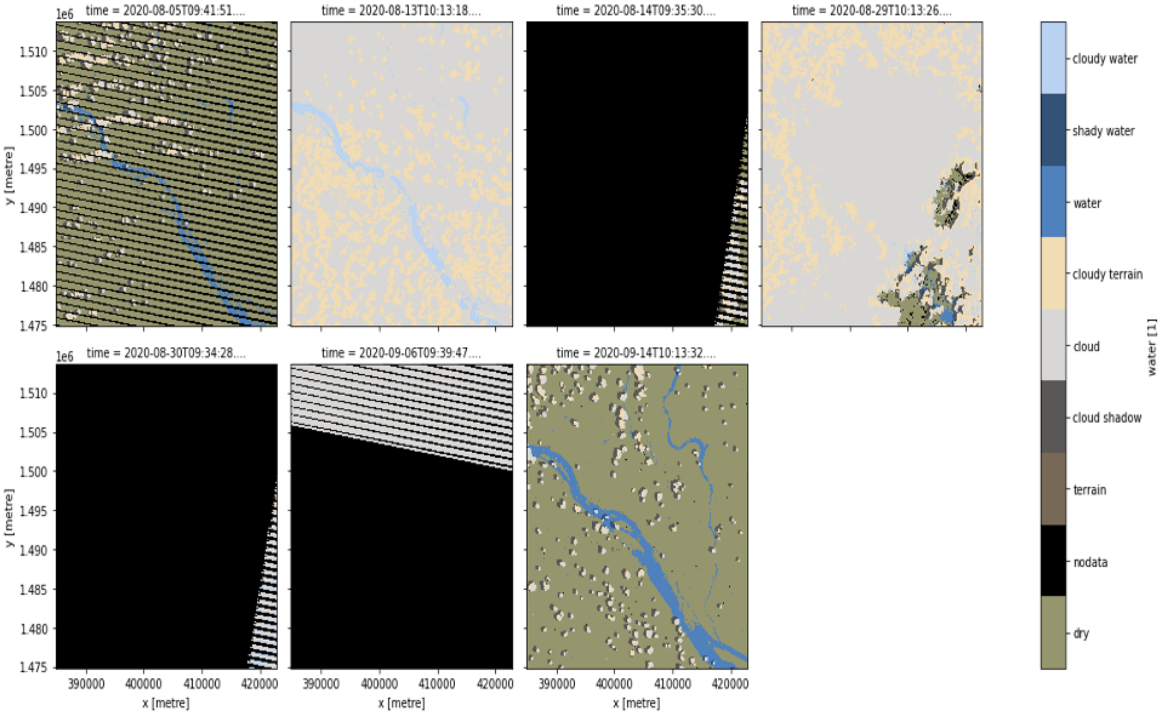

time = 2020-08-05T09:41:51....
time = 2020-08-13T10:13:18....
time = 2020-08-14T09:35:30....
time = 2020-08-29T10:13:26....
time = 2020-08-30T09:34:28....
time = 2020-09-06T09:39:47....
time = 2020-09-14T10:13:32....
1e6
1.510
1.505
1.500
1.495
1.490
1.485
1.480
1.475
y [metre]
390000
400000
410000
420000
x [metre]
cloudy water
shady water
water
cloudy terrain
cloud
cloud shadow
terrain
nodata
dry
water [1]

Figure 1 - Data available through satellites through August 05th - September 14t

Also, after considering the visuals of surface water bodies on August 05th, 2020 - Fig(2) and September 14th, 2020 - Fig(3) we can clearly conclude the flood data through the Satellite Observations are accurate.

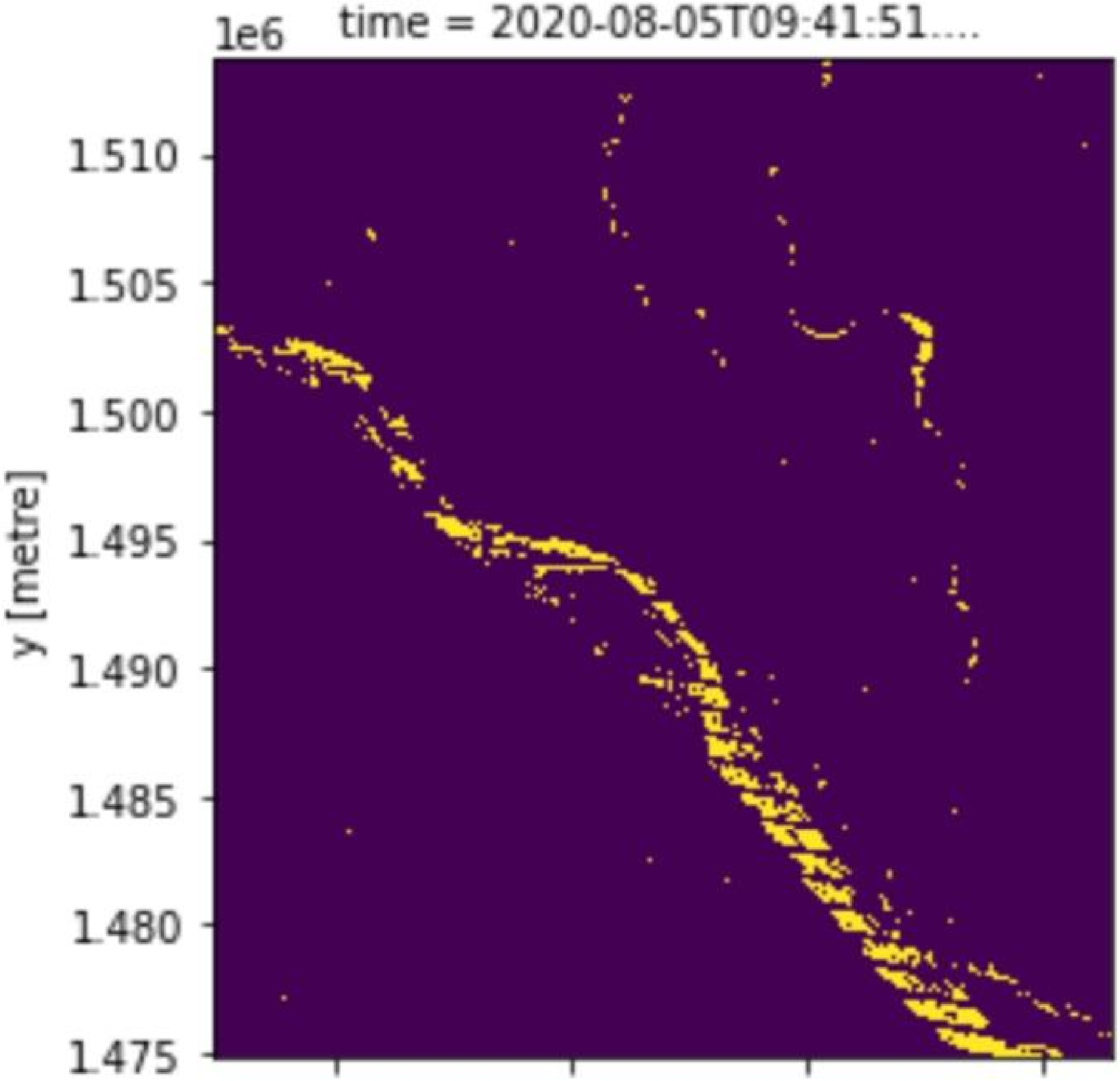


Figure 1. August 05th, 2020
The original water body of river Niger

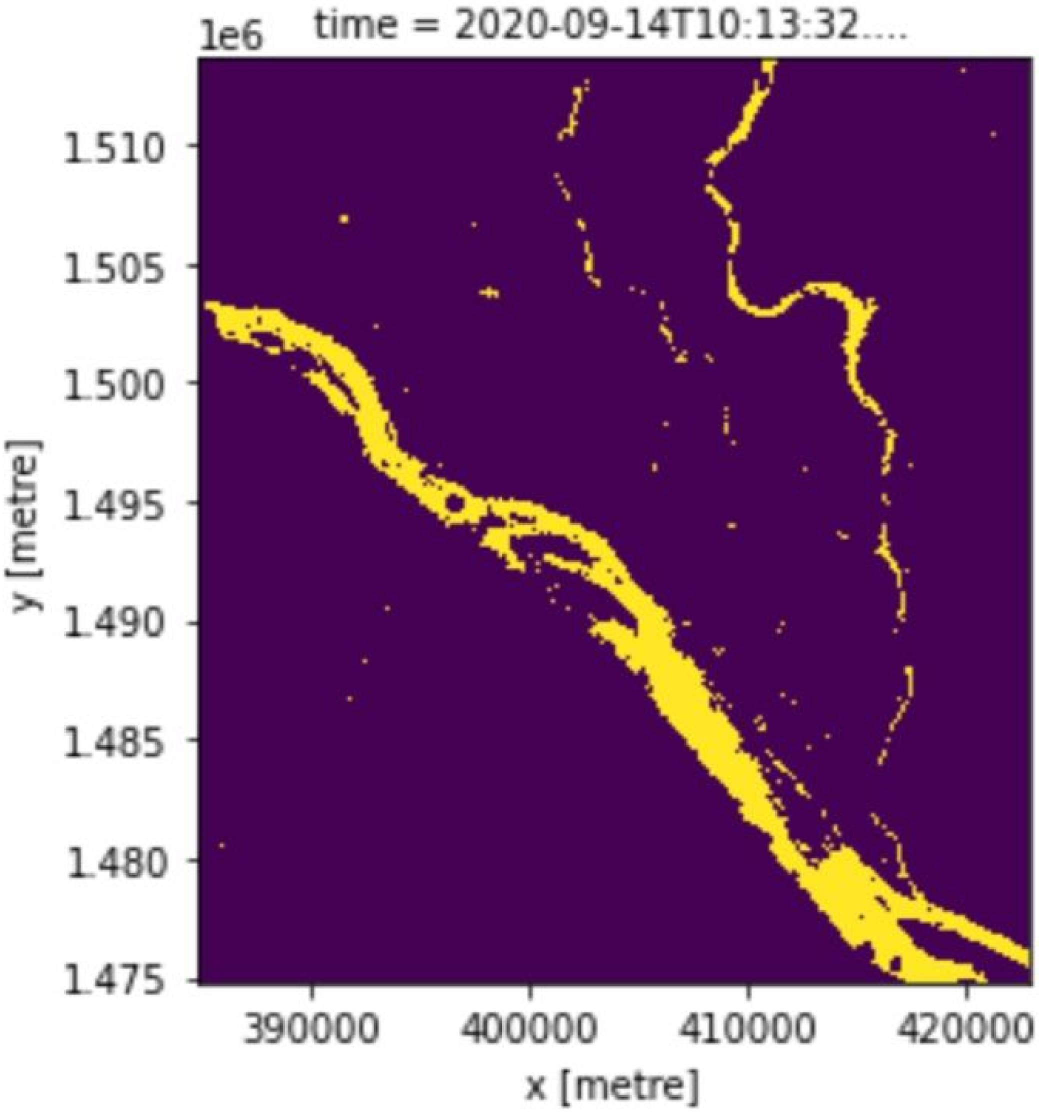


Figure 2.September 14th, 2020
The flooded water body of river Niger

Data preprocessing:

There is always a chance that the data will have missing values or might be having false values. Data preprocessing is the step where the data has been adjusted to make it more realistic, to get along with the actual data. There are plenty of ways and approaches to solve this data inconsistency, like by using the mean values to fill the missing data. Another common problem with training data is class imbalance. This can occur when one of your classes is relatively rare and therefore the rare class will comprise a smaller proportion of the training set. When imbalanced data is used, it is common that the final classification will under-predict less abundant classes relative to their true proportion. For solving these types of issues we generate or multiply such low available data to match with the other features.

Deafrica_tools is a package that contains custom util methods like collect_training_data, to collect the test data or train data. This method is written with logic to pre-process our training data by stacking the arrays into a useful format and removing any NaN or inf values.

Machine Learning Part:

We will be preparing a model that can predict the crop pixels and non crop pixels from a given image. Below are the steps incurred to prepare a highly accurate machine learning model.

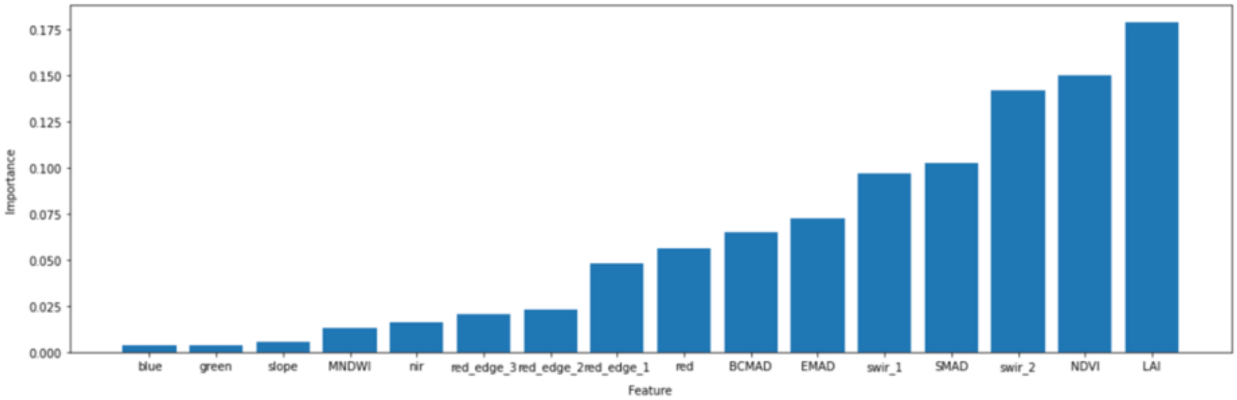
1) Extract the training data:

   We need to get the data for training the model as to get fit with, so that it can better predict the new test data that we feed later with high accuracy.

   We will be using crop_training_egypt.geojson data for training the model. We use collect_training_data, a custom method from the Deafrica_tools package. This extracts the data that are geometry data type and is polygon specific. This function requires query parameters, zonal_stats(mean,max,min), feature layer function as parameters. Based on these params the data will be fetched and will be written into the physical file in the disk.

2) Inspect Training data:

In this phase the extracted data will be read and inspected. The data is wrangled in such a way that each class in the training data array assigns it to a dictionary. After extraction feature wise it is plotted and analyzed which features are best required to train the model. We use the RandomForrestClassifier method to calculate the importance of the features.



The above graph shows swir_2, NDVI and LAI are 3 important features to be considered. Then PCA(Principal Component Analysis) is conducted, this will calculate and plot the first two principal components of our training dataset. This will transform our dataset with lots of variables (16 features) into a dataset with fewer variables, this allows us to visualize our dataset in a relatively intuitive and straightforward manner.

3) Evaluate, optimize and fit

In this stage we have seen that the training data size is just 156, this is not sufficient enough to split the data into training and testing serts. Hence we use all the extracted data to train and to measure the predicting power of the model we use nested k-fold cross-validation. We use nested CV here where the outer CV is for the accuracy of the classifier and inner CV is to optimize the hyper parameters. The range of the split will be 5-10. Now that with help of this method we get hyper parameters that can best produce the accuracy, we now fit the model with those params and will save the model in the physical file.

4) Classifying satellite data

Now that the model is ready, we will import the model and get the test data with some query parameters. The test data is selected in such a way that we only select the small regions. Once the model is giving the best results then we predict for the large region by re-entering a new latitude, longitude and larger buffer size.The results are stored in tif files in the disk.

Covering up the analytics:

We have updated our repository with our work in which we have tried to establish descriptive analytics of the areas around River Niger that were flooded. Also using the digital earth Africa we could create one more visual aid to substantiate the situation before and after September 14th, 2020, which can be found on the link:

Water bodies comparisons Between August 05 and September 14th

The jupyter notebook consister of descriptive analytics has been updated and committed to the git repository. The daily observations of surface water bodies have been updated to "Individual_Day_Satellite_Pictures.ipynb" and the flood observations have been described in "Flood_Observations.ipynb".
We can clearly observe the cloud formation between August 06th, 2020 to September 09th, 2020 which indicates weather abnormality confirmed by occurrence of Torrential Rain and followed by flooding.

Future Scope:

As a next level of analysis, we aim to map the effects of the flood on the vegetation on the flooded areas and also, conduct a predictive analysis on the areas which are prone to become barren as a consequence of subsequent floods.

Along with it, the **cumulonimbus** cloud which has been the primal agent in the floods of Niamey, can also be detected by WOfS Data. We aim to target such data to predict if any such calamity can be foreseen without the need of Weather forecasting and high end technology.

References used:

https://docs.digitalearthafrica.org/en/latest/data_specs/GeoMAD_specs.html

https://registry.opendata.aws/deafrica-geomad.