Take a way cat.

1. Compile COVID-19 related Data(preferably data from Kenya) from relevant online sources such as https://coronavirus.jhu.edu/

https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distributioncovid-19-cases-worldwide
https://www.worldometers.info/coronavirus/#countries
https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6
https://vincentarelbundock.github.io/Rdatasets/datasets.html
https://africaopendata.org/group/kenya http://www.opendata.go.ke/

1. Ingest the data into Hadoop DFS Data lake
2. Use pyspark package to extract the data from the data lake
3. Choose appropriate techniques to Pre- process the extracted data
4. Apply one predictive analytics technique to generate a model for predicting any of the following cases:
   a) Number of Death cases or Mortality rate
   b) Number of confirmed cases
   c) Number of recovery cases or Recovery rate

5. Visualize the model
6. Test the model
7. ***Validate the Model***
8. Compile pdf processed document that has the following content:

(i) Describe how the data was compiled in task 1 and include Screen captures of both code &and output) (3 Marks)

(ii) Describe how the data was ingested into Hadoop data lake and include screen shots. (3 Marks)

(iii) Describe how data was extracted using pyspark and include associated screen shots (3 Marks)

(iv) Describe pre-processing tasks/techniques used to prepare the data (include screen shots) and give reason (s) to justify your choices (3 Marks).

(vi) Test results and interpretations (3 Marks)
(vii) Validation Results and interpretations (3 Marks)
(viii) Potential applications of the interpreted results (3 Marks)

10. Present your work in class on 23rd NOV. 2023 (5 Marks)

11. Host your **PDF**-processed **document** a **text file** of list of commands used to GitHub and submit your details and link by filling in the form here: by 1st DEC 2023.