Codefest Datathon 2020

Evaluation - Initial Rounds

Introduction

The head of research has requested you to do an analysis on the current ***COVID 19 Pandemic*** situation in Sri Lanka and build a ***predictive model that can predict the number of infected people***. You may refer to any such studies done on global context and your goal should be to derive as many insights and predictions as possible with respect to local context. You may be able to provide predictions *island wide/ province wise/ district wise/ city wise/ electorate wise* but not limited to. (*city level/electorate will be more appropriate to authorities to take actions*). The value of the insights and predictions you generate (both breadth and depth) would be considered favorably in your evaluations. You are required to use only ***open source technologies*** for implementation of your system.

You may use the below evaluation criteria as a guideline when designing your system.

	Evaluation Criteria
Data Engineering	1. Identification of data sources 1. Local context data 2. Global context data 2. Identifying the required fields 3. Data extraction mechanisms(as flat files, via API, streams, through web scraping etc.)

Data Science Engineering	1. Data Pipeline Architecture (Possible Steps) a. Use of data ingestion processes b. Use of a data staging strategy c. Use of a clear ETL/ELT design d. Provision for data warehousing e. Insight generation component integration f. Visualization component integration 2. Accommodating a flow orchestration component 3. Technology selection and justification
Data Science	1. Data Analysis 1. Identified patterns in the dataset such as, 1.Province wise 2.Age wise 3.Profession wise 4.Cluster wise 5.If any other 2. Feature Engineering 1. New Features generated 2. Best feature set identification 3. Model Implementation 1. Model design and architecture(Mathematical Model, Statistical Model, ML Models, DL Models etc.) 2. Models that have tried out(Time series, Classification, Regression etc.) 3. Model evaluation criteria 4. Accuracy improvement mechanisms used 4.Model Operationalize 1. How predictions will be made available 2. Data pipeline for predictions 3. Model deployment mechanisms

Predictions	1. Visualization of predictions 2. Alerts or actions 3. Recommendations if any

Data Extraction

- You can explore all the available data sources with respect to local context in processed/semi processed or unprocessed formats. You may have to do an initial analysis in order to identify the scope, granularity and type of data you require in order to perform the analysis.

- Where local stats are not available on certain aspects (e.g. profession of the deceased) you may explore the possibility of using suitable substitutes from global context.

Data Science Engineering

- Once the data extraction process is defined, you are required to design the architecture of the data pipeline for the above requirement.

- The design should cover the basic aspects of a data pipeline to obtain data from different sources, preprocessing, warehousing, insight generation and visualization.

- The predictions could be generated on a daily schedule and could be integrated into the pipeline.

- The design should elaborate the Open Source technologies and tools you plan to use and any justifications as required.

Data Science

- Possible pattern identification would depend on the dataset you select. You may consider the categorizations given in the evaluation criteria but not be limited to it.

- Feature Engineering - Generating new features or alternating existing features

Deliverables

- Dataset you used for the analysis - you may provide a url(s) for the source(s) you used if / when applicable.

- High Level design of your system - can be in the form of a report (not more than two pages) with supporting diagrams where applicable. You may include the justification of the design decisions you made.

- Your code base - code covering entire pipeline including,
    - Data extraction
    - ETL
    - Staging or warehousing if used
    - Data Science analysis, model implementation and evaluation

- Visualizations based on the insights you derived from your dataset.

Resources

- You may get an idea about predictions and simulations that have been done on the following links, but not limited to. .
    - Epidemic Calculator
    - Modeling COVID 19 Spread vs Health Care Capacity
    - Agent based simulation of COVID 19 Health and Economical Effects

You have to submit your solutions, reports, visualizations **on or before 2nd of November 2020, 4pm** to the respective OneDrive folders shared with you. .