**Project Proposal**

**Title**: **Epidemic Outbreak Prediction using Artificial Intelligence.**

**Abstract:** Reliable predictions of the dynamics of infectious diseases are extremely valuable to public health organizations planning interventions to reduce or prevent the spread of disease. With the growth of big data in the fields of health and biomedicine, accurate analysis of such data helps early detection of diseases and better patient care. With enormous computing power, it is now highly feasible to use "Click here to enter text.big data" to predict and manage outbreaks. Our idea is to analyze and determine the spread of epidemics in villages and suburbs, where medical care may not be available. We want to build a machine learning model that can predict the dynamics of an epidemic and tell us where the next epidemic is most likely to break out. Our method takes into account the geography, climate, and population distribution of the affected area because these are relevant characteristics and subtly contribute to the dynamics of the disease epidemic. Our model will help health authorities take appropriate measures to ensure that there are sufficient resources to meet demand and, where possible, to curb the emergence of such epidemics.

**Group Size**: 3

**Timeline**: 5 Weeks

**Planned Start Date: 23rd August 2021**

# Planned End Date: 27th September 2021

**Expected Member Profile**: Skilled in Python, Knowledge of ML concepts, deep learning like ***ADABoost, XGBoost, SVM, Multi-Layer Perceptron, Logistic Regression***, CNNs, RNNs, and Hands-on with libraries like scikit-learn, pandas, NumPy. Knowledge of clouds like AWS/ GCP, data visualization frameworks

# Recommended Preparation and Study Material (from Packt Library):

* [The Data Science Workshop](https://courses.packtpub.com/courses/data-science)
* [The Deep Learning with Keras Workshop](https://courses.packtpub.com/courses/deep-learning-with-keras)
* [The Deep Learning with PyTorch Workshop](https://courses.packtpub.com/courses/deep-learning-with-pytorch) (Chapter 6)
* [The Deep Learning Workshop](https://courses.packtpub.com/courses/deep-learning) (Chapter 5 and 6)
* [The Supervised Learning Workshop](https://courses.packtpub.com/courses/supervised-learning) (Chapter 2,3,4, and 6)

For AWS/GCP

* [Engineering MLOps](https://www.packtpub.com/product/engineering-mlops/9781800562882)
* [Machine Learning with AWS](https://www.packtpub.com/product/machine-learning-with-aws/9781789806199)
* [Mastering Machine Learning on AWS](https://www.packtpub.com/product/mastering-machine-learning-on-aws/9781789349795)

**Technology Stack:** Python, Pandas, Numpy, scikit-learn, Tensorflow (for advanced sophisticated models), and visualization frameworks, deep learning, AWS/GCP.

**Datasets:**

* Zika Data Repository maintained by the Centre for Disease Control and Prevention contains publicly available data for the Zika epidemic. ( <https://github.com/cdcepi/zika> )
* Google Geolocation API for procuring latitude and longitude of places associated with an outbreak
* Worldwide airport location data retrieved from [Falling](http://www.fallingrain.com/world/index.html) [rain](http://www.fallingrain.com/world/index.html)
* The weather data scraped from Wunderground.com by the nearest airport code
* Population density of different regions was extracted from the gridded map via NASA (SEDAC)

[(https://earthdata.nasa.gov/about/daacs/daac-sedac](https://earthdata.nasa.gov/about/daacs/daac-sedac))

* Vector agents (Aedes albopictus, Aedes aegypti) occurrences from the global compendium of

*Aedes aegypti*and *Ae. albopictus* occurrence

(https://datadryad.org/resource/doi:10.5061/dryad.47v3c)

* GDP/ GDP PPP data from IMF World Economic Outlook

**GitHub Repo Link**: <https://github.com/TeamEpicProjects/Epidemic-Outbreak-Prediction-using-Artificial-Intelligence>

**Process flow:**

A screenshot of a computer

Description automatically generated with medium confidence

**Architecture diagram**:

Diagram

Description automatically generated

**Recommended System Setup:**

**Hardware Requirements**

Laptop/Desktop with i5/8GB RAM and 128Gb storage. GPU if available, is not mandatory.

**Software Requirements**

Python, TensorFlow, pandas, NumPy, scikit-learn and editors like PyCharm, visualization frameworks like tableau, or matplotlib, seaborn as per team members skills, etc.

**Libraries**

Python, TensorFlow, pandas, NumPy, scikit-learn

**Problem Statement**

COVID19 has shown the real face of our healthcare infrastructure, planning, and readiness to address a pandemic with limited resources, unprepared staff, and broken supply chains. Government agencies if provided with this information well in advance, can plan and execute them in a well-orchestrated manner optimizing the use of staff, resources at their disposal. In this project, we aim is to analyze and build a multimodal model to predict the likelihood of an area having an outbreak.

# Expected Solution

AI solution to help health authorities take appropriate measures to ensure that there are sufficient resources to meet demand and, where possible, to curb the emergence of such epidemics.

**Work Package**

**Week 1: Environment setup and ETL Milestone / Feature**

Environment Setup, building a pipeline to load the data from sources and process

**User Story**

As a data engineer, I should be able to load and transform the data from multiple sources into a data store making it easy for the data scientists to use this data for processing, gain insights and make predictive models to predict the likelihood of epidemics in localities.

**Acceptance Criteria**

ETL pipeline to process the data from different sources and store it for modeling.

# Individual Task

ALL: Setup environment

Member 1: Build ETL Pipeline to load the data and process

Member 2: Schedule jobs to periodically load the data from the sources, using airflow

Member 3: Build Dashboard to serve these analyses.

# Expected Outcome

ETL pipeline to load and process the data, scheduling for running the pipeline in intervals automatically, and tests to ensure data quality

**Week 2: Exploratory Data Analysis Milestone / Feature**

Explore the processed data and find insights that can be used for modeling

# User Story

As a data scientist, I will be utilizing the data processed by ETL jobs and stored in the data store and build pipelines to gain insights, build visualizations useful to stakeholders.

**Acceptance Criteria**

Dashboard / Visualizations on insights from the data.

# Individual Task

Member 1: EDA on the relation of Population, mobility on disease spread and write test cases to ensure data quality

Member 2: EDA on the relation of Weather on disease spread and write test cases to ensure data quality

Member 3: Find additional data sources that can be used and run EDA and write test cases to ensure data quality

**Expected Outcome**

Find insights that can be useful to model the solution, build visualizations/dashboards to present the insights

**Week 3: Modelling Milestone / Feature**

Build models to predict the disease outbreak

# User Story

As a data scientist, I want to build predictive models using the insights gained from the previous week's exercise. I want to build a baseline model using simple methods and then iterate using more sophisticated approaches and build models that can predict the likelihood of epidemics.

**Acceptance Criteria** models to predict the disease outbreak

# Individual Task

Member 1: Experiment and come up with models to predict outbreaks using Population, Mobility, and Weather at a location using ML approaches.

Member 2: Experiment and come up with models to predict outbreak using Population, Mobility, and Weather a location at a location using DL approaches

Member 3: Experiment and build hybrid / multimodal models

**Expected Outcome**

Build Models to predict the outbreak.

**Week 4: Modelling Milestone / Feature**

Deploy models to make online predictions

# User Story

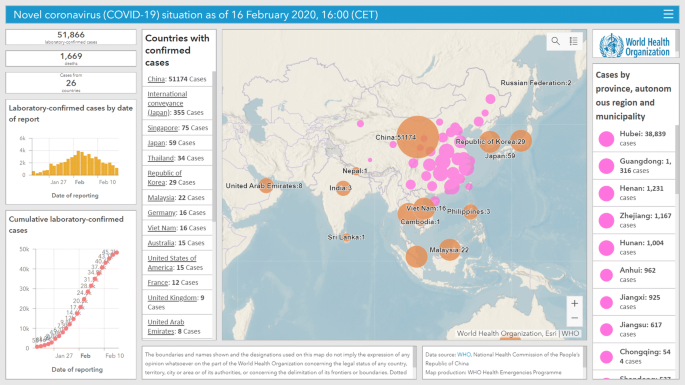
Now with the models built for predicting the likelihood of epidemics, I want an end-to-end pipeline for deploying and maintaining the models scale allowing me to retrain and scale the predictions.

**Acceptance Criteria**

models deployed on the cloud to make interfaces

# Individual Task

All members: Build end to end pipeline to deploy the model and integrate with dashboard developed in week 2



**Expected Outcome** Models deployed to make predictions and integrated with dashboard along with documentation and report.

Week 5 is buffer time to accommodate for any delays, time required for team members to learn and deliver items.