University of Hertfordshire UH

School of Physics, Engineering and Computer Science

Project and Data Management (PDM) Plan.

My Project Title Name: Covid-19: Comparative Data Analysis and Prediction for the World and Bangladesh.

My Name is - Farid Hossain.
My Student ID: 23006446
Email Address:faridhossain7600@gmail.com
My project github link:https://github.com/Fariduk/PROJECT-DATA-SCIENCE-COVID-19-Comparative-Data-Analysis-and-Prediction-for-the-World-and-Bangladesh.git

My Project Supervisor Name is - William Bate

Email Address: w.bate@herts.ac.uk

1. Introduction:

1.2.Project Overview: The project titled "Covid-19: Comparative Data Analysis and Prediction for the World and Bangladesh" aims to analyze historical Covid-19 data and build a predictive model to forecast future cases. The project will use various machine learning techniques to identify trends and provide insights into the pandemic's impact.

1.2 Objectives:

. To collect and preprocess Covid-19 data from multiple sources (Kaggle, WHO, Bangladesh Health Ministry, etc.).

. To apply machine learning techniques for predictive modeling.

. To evaluate and compare predictive models.

. To visualize trends and provide insights.

. To ensure ethical considerations in data handling and predictions.

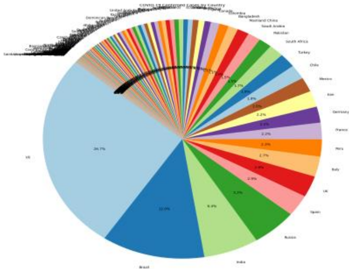
2. Research Questions:

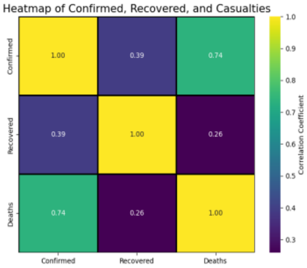
1. **How do Covid-19 trends in Bangladesh compare to global trends?**
   - Bangladesh follows similar global trends but with unique variations in case surges, recovery rates, and mortality influenced by population density, healthcare infrastructure, and government interventions.
2. **What are the key factors influencing Covid-19 case surges in Bangladesh?**
   - Factors include population density, urbanization, healthcare capacity, testing rates, lockdown effectiveness, public compliance, and vaccination rollout.
3. **Can we build an accurate predictive model for future Covid-19 cases?**
   - Yes, using machine learning techniques like time-series forecasting, regression models, and deep learning (e.g., LSTM in TensorFlow), we can develop an accurate predictive model for future case trends.

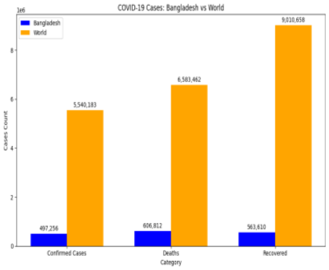
3. Objectives:

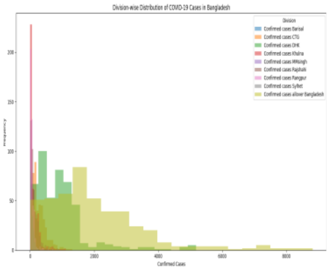
- Conduct exploratory data analysis on global and Bangladesh-specific datasets.
- Develop predictive models using machine learning techniques.

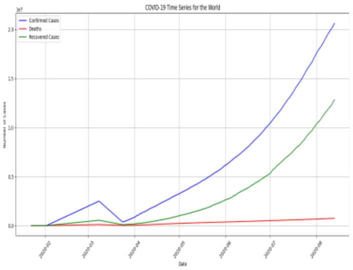
- Implement data visualization for better trend interpretation.
- Assess the accuracy and performance of different models.
- Provide insights for policymakers and healthcare professionals.

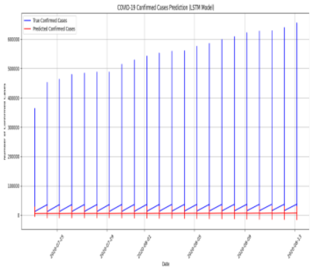


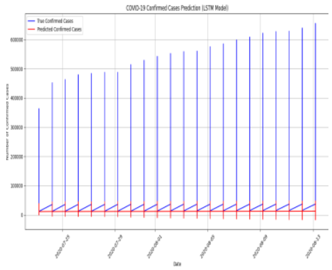


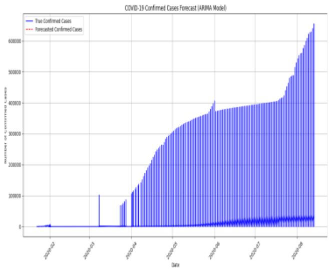












4. Project Plan & Timeline:

Task	Description	Duration	Tools
Data Collection	Gather datasets from Kaggle, WHO, and Bangladesh government sources	3 days	Kaggle API, CSV ,excel,files
Data Cleaning	Handle missing values, remove duplicates, and preprocess data	4 days	Pandas, NumPy
Exploratory Data Analysis	Visualize trends and correlations	5 days	Matplotlib, Seaborn
Model Selection	Implement classification models and predictive models	6 days	TensorFlow, Scikit-Learn
Model Training & Evaluation	Train models and assess performance metrics	5 days	TensorFlow, Pandas
Report & Documentation	Summarize findings and generate final report	3 days	LaTeX, MS Word
Presentation Preparation	Create slides for final presentation	2 days	PowerPoint

5. Data Management Plan -

5.1 Data Storage & Organization:

- **Storage Location:** All datasets will be stored on GitHub for version control.
- **Backup:** Daily backups on cloud storage (Google Drive/OneDrive).
- **Data Format:** CSV format for raw data and processed data And excel formate.
- **File Naming Convention:** bgd-covid19-subnational.xlsx, Covid Dataset of Bangladesh divisionwise.csv, covid_19_data.csv, covid_19_data1.csv, COVID_DataSet_Bangladesh_Gender_Age_Analysis.csv, COVID_DataSet_Bangladesh_QuarentineData.csv, COVID_DataSet_Bangladesh_Test_Confirm_Death_Recovery.csv,

5.2 Data Processing Workflow:

1. Load raw data from GitHub repository.
2. Clean and preprocess datasets.
3. Conduct exploratory data analysis.
4. Apply machine learning models for prediction.
5. Validate results and visualize findings.

5.3 Version Control:

- **Platform:** GitHub repository for tracking code and data changes.
- **Branching Strategy:** Separate branches for data preprocessing, model training, and visualization.
- **Commit Frequency:** Daily commits to maintain project progress.

5.4 Ethical Considerations:

- Ensure compliance with WHO and government data privacy policies.
- Maintain transparency in data sources and avoid misrepresentation.
- Ethical reporting and unbiased model interpretation.

6.Model Implementation Plan:

6.1 Model Selection

The following models will be tested:

.Linear Regression (Baseline model for trend analysis)

.LSTM (Long Short-Term Memory) (Deep learning model for time-series forecasting)

.Random Forest (For feature importance and robust predictions)

6.2 Model Training

.Splitting dataset into train (80%) and test (20%) sets.

.Using TensorFlow/Keras for deep learning implementation.

.Hyperparameter tuning for model optimization.

6.3 Model Evaluation

.Mean Absolute Error (MAE)

.Mean Squared Error (MSE)

.Root Mean Squared Error (RMSE)

.Accuracy comparison across models

7. Expected Outcomes:

- A comprehensive comparative analysis of Covid-19 trends.
- A predictive model with high accuracy for future Covid-19 cases.
- Data visualizations for better trend interpretation.
- A well-documented report and presentation for academic evaluation.

8. References :

- World Health Organization (2024). 'Covid-19 Global Data'. Available at: www.who.int
- Kaggle (2024). 'Covid-19 Dataset'. Available at: www.kaggle.com
- Bangladesh Government Health Ministry (2024). 'Covid-19 Division-wise Data'. Available at: www.dghealth.gov.bd, https://data.humdata.org/dataset/district-wise-quarantine-for-covid-19, https://data.mendeley.com/datasets/b98d8mj2xk/1 .