Data Science

Capstone Project: Predictive Modelling for COVID-19 in Public Health

---

1. Case Scenario

In response to the COVID-19 pandemic, public health organizations have faced immense challenges in predicting the spread of the virus and understanding key factors that influence transmission and patient outcomes. Imagine you have been hired as a data scientist by a public health organization, "HealthGuard Analytics," to build a predictive modeling system. The organization requires actionable insights to inform policies, anticipate future outbreaks, and improve health resource allocation.

Using historical COVID-19 data, you will conduct data cleaning, perform exploratory data analysis (EDA), and develop predictive models to forecast COVID-19 trends. You will present your findings through visualizations and provide a final report summarizing insights and recommendations for public health responses.

---

2. Methodology

Data Collection

- **Dataset**: Access the COVID-19 Open Research Dataset (CORD-19) on Kaggle, which includes COVID-19 case counts, demographic data, and various health metrics.
- **Dataset Source**: CORD-19 on Kaggle

Data Preprocessing

- **Cleaning**: Address missing values, remove duplicates, and standardize date and location formats.
- **Transformation**: Normalize data for machine learning models, ensuring consistency across all numerical features.
- **Feature Engineering**: Create derived variables, such as daily growth rates, mortality ratios, and cases per population to enrich the dataset and strengthen model insights.

Exploratory Data Analysis (EDA)

- **Objective**: Conduct EDA to uncover trends, correlations, and outliers in the data.
- **Visualizations**: Use charts like line plots, bar charts, and scatter plots to analyze COVID-19 trends, such as case and mortality rates over time.
- **Key Insights**: Focus on identifying demographic and environmental factors that could influence the spread and severity of COVID-19 cases.

Model Development

- **Machine Learning Models**: Apply predictive models such as:
    - Time-Series Models.
    - Classification Models
- **Evaluation**: Assess model performance with accuracy, precision, recall, F1-score, or RMSE, as applicable.

Data Visualization and Reporting

- **Visualization**: Create clear, informative visualizations of model predictions and EDA findings, using libraries like Matplotlib and Seaborn.
- **Reporting**: Present results in a structured report, with narratives and visuals that effectively communicate insights to non-technical stakeholders.

---

3. Assessment Criteria

- **Data Preparation**: Proficiency in data cleaning, transformation, and feature engineering.
- **EDA and Insights**: Depth of EDA, relevance of insights, and quality of visualizations used to communicate findings.
- **Model Performance**: Evaluation of model suitability and accuracy using appropriate metrics.
- **Clarity and Presentation**: Quality of the final report, visualization clarity, and coherence of findings and insights.

---

4. Submission Details

Deliverables

1. **Technical Report**: A detailed report covering data preparation, EDA, model development, model evaluation, and key insights with visualizations.
2. **Code Repository**: A GitHub repository with organized, well-documented code for all phases of the project.
3. **Presentation**: A slide deck or recorded presentation (10–15 minutes) summarizing findings, model insights, and public health implications.

Submission Format

- **Report**: Submit as a PDF or Word document.
- **Code**: Provide a GitHub link or a .zip file containing the project code.
- **Presentation**: Submit as PowerPoint, Google Slides, or a video file (MP4).

Data Resources

All participants will use the COVID-19 Dataset (CORD-19) available on Kaggle.

Here is the URL to the dataset:
https://www.kaggle.com/datasets/imdevskp/corona-virus-report