COSC 3337 - Data Science I

Contributors:
Naomi Ayub (1868125)
Elyjaiah Durden (1977092)
Nirmal John (2052312)

Group Project Proposal: COVID-19

Title Of The Project Report:

Predicting COVID-19 Hospitalization Risks Using Data Science Models

Thoughts On What Tasks To Perform:

We plan to undertake three primary data-science tasks:

Exploratory Data Analysis (EDA):

We will start by cleaning and visualizing the data to understand temporal trends in cases, deaths, and hospital/ICU admissions. This includes identifying any missing values, outliers, and potential anomalies. We also intend to explore correlations among features such as total_cases, new_deaths, and icu_patients.

Classification:

We plan to classify time periods into "Low," "Medium," or "High" hospitalization risk, leveraging the dataset's derived category hospitalization_need. By building a predictive classification model (e.g., decision tree or random forest), we can examine which features (like total_cases or total_deaths) are most influential in predicting hospitalization_need.

Regression / Forecasting:

To anticipate healthcare requirements, we aim to develop a regression model to predict future ICU admissions (icu_patients or weekly_hosp_admissions) based on current infection rates, death rates, and other derived features.

If time allows, we may explore a simple time-series forecast (e.g., ARIMA or LSTM) on key metrics such as new_cases or new_deaths to see if short-term predictions are feasible.

Through these tasks, we hope to gain insight into COVID-19 trends and produce actionable predictions for ICU/hospital utilization.

Motivation Of The Tasks:

Public Health Relevance: The COVID-19 pandemic heavily impacted healthcare systems. Analyzing real data can reveal which factors drove hospital demand and where peaks in infection rates occurred.

Predictive Insights: By focusing on classification and regression, we can see if short-term risk forecasting is possible. Hospitals or policymakers might benefit from models that suggest upcoming surges in ICU admissions.

Skill Development: Our tasks align with core data science competencies: data cleaning, EDA, model building, and evaluation. This project provides hands-on practice with real-world, time-series data.

Our Understanding About The Data:

The dataset originates from Our World in Data and covers various COVID-19 metrics in the United States, including:

- Original Variables: total_cases, new_cases, total_deaths, new_deaths, icu_patients, hosp_patients, and others.

- Derived Features: daily_case_change_rate, hospitalization_rate, 7day_avg_new_cases, 7day_avg_new_deaths, etc.

These time-series metrics will allow us to see both short-term fluctuations (daily or weekly) and longer-term trends throughout the pandemic. Because certain dates or states might have missing or incomplete data, we will need to perform data cleaning or imputation where appropriate.

Description Of The Data & How It Supports The Analysis:

Time-Series Analysis: The "date" column enables longitudinal tracking of the pandemic's progression.
Hospitalization and ICU Statistics: Knowing hosp_patients and icu_patients helps us gauge healthcare system strain.
Categorical Labels: The dataset already includes derived fields like hospitalization_need (Low, Medium, High), making it straightforward to create supervised classification models without extensive feature engineering.
Rolling Averages: Variables like 7day_avg_new_cases smooth daily fluctuations, which can help in trend detection and reduce noise in predictive models.
Because these features are highly relevant to understanding health outcomes, they can support classification and forecasting tasks effectively.

Justification Of Usefulness For Our Analysis:

Healthcare Planning: Insights about when ICU demands spike help in resource allocation.
Risk Stratification: Classifying hospitalization risk categories can offer quick decision support for policy.
Predictive Power: Understanding how well we can predict near-future ICU admissions can inform readiness for potential surges.

High-Level Ideas About The Methods That Will Be Used In Our Analysis:

EDA:
We will use Python (pandas, NumPy) for data wrangling, and matplotlib for visualization to illustrate trends in cases, deaths, and hospital usage. We'll look for correlations and potential outliers using correlation heatmaps and box plots.
Classification:
We will likely use scikit-learn. After splitting the data into training and test sets (e.g., 80:20), we may train a decision tree, random forest, or logistic regression model to predict hospitalization_need.
Performance Measures: We'll use accuracy, precision, recall, and F1-score.
Regression / Forecasting:
For basic regression: we might use linear regression or random forest regression to predict icu_patients.
For time-series forecasting: if time allows, we'll explore ARIMA or an LSTM model (in Keras) on daily new_cases.
Performance Measures: Mean Squared Error (MSE) or Mean Absolute Error (MAE).

Expected Outcomes:

- Cleaned Dataset & Analysis: A well-documented notebook illustrating the data cleaning steps and EDA.
- Model Results: Confusion matrices, classification metrics, and feature importance plots for classification tasks; error metrics (e.g., MSE) for regression tasks.

- Visualizations: Graphs showing case/death trends over time, side-by-side with hospitalization or ICU admissions, plus any time-series forecasts.

Final Deliverables:

- A PDF final report (5–10 pages) detailing our methods, findings, and visualizations.
- All code (Python notebooks) in a zip file with a README.
- Reproducible figures (plots) referenced in the report.

Main Contributions Of The Project:

Integrative Approach: Combining EDA, classification, and regression to offer a comprehensive view of COVID-19 trends.

Predictive Modeling: Evaluating how accurately we can predict ICU admissions, which is a critical outcome for healthcare systems.

Practical Insights: Providing visual summaries and classification outcomes that can be intuitively understood by non-technical audiences.

How We Will Collaborate:

We will hold weekly Zoom or Discord calls to coordinate tasks and share progress. Communication will occur primarily via a private Discord channel for day-to-day updates.

- Naomi Ayub:
  - Responsible for the initial data exploration, data cleaning, and part of the EDA visuals.
  - Helps with documentation and summarizing findings in the final report.
- Elyjaiah Durden:
  - Takes the lead on classification model selection and development.
  - Evaluates model performance (precision, recall, F1-score) and contributes to final discussion sections.
- Nirmal John:
  - Focuses on regression/time-series approaches and scripts for predictive modeling.
  - Assists in writing the methodology section and final conclusions.

Collective: Each member will review one another's code and results to ensure clarity and correctness.

Expected Workload Of Each Group Member:

- Naomi: Approximately 30% of total effort (data cleaning, EDA, partial write-up).
- Elyjaiah: Approximately 30% of total effort (classification modeling, performance evaluation).
- Nirmal: Approximately 30% of total effort (regression/time-series modeling, code integration).

Note: Remaining 10% of tasks (like finalizing the report, editing, ensuring reproducibility) will be split evenly or handled collaboratively during our weekly calls.