**COVID-19 Tracker Data Analysis & Predictive Modeling Report**

**Author:** B  
**Role:** Data Analyst  
**Date:** [Insert Date]

**Executive Summary**

This project involves a comprehensive exploratory data analysis (EDA) and predictive modeling effort on a COVID-19 tracker dataset. Our goal was to extract actionable insights from demographic and clinical data, quantify key outcomes, and develop preliminary predictive models for hospitalization and mortality. Despite challenges with data quality and class imbalance, the analysis delivers critical findings about patient demographics, disease severity, reporting delays, and model performance, setting the stage for future improvements and impactful applications.

**1. Data Overview & Preprocessing**

* **Dataset Size:** 9,019 patient records with 13 variables including demographics, dates, and clinical outcomes.
* **Data Cleaning:** Addressed missing values, inconsistent date formats, and aligned relative file paths for reproducible workflows.
* **Key Variables:** Sex, Age Group, Hospitalization (hosp\_yn), ICU admission (icu\_yn), Death (death\_yn), and case report timing.

**2. Exploratory Data Analysis**

**2.1 Demographic Distributions**

* The dataset is roughly balanced by sex: 52.8% female, 47.2% male.
* The age distribution clusters around two groups: adolescents (10-19 years) and middle-aged adults (50-59 years).
* These distributions reflect key subpopulations affected and help contextualize outcome disparities.

**2.2 Outcome Rates**

* Hospitalization rate: **4.3%** of cases.
* ICU admission rate: **2.6%**.
* Mortality rate: **0.5%**.
* These low percentages indicate a largely mild cohort but emphasize the need to focus on predicting severe cases early.

**2.3 Reporting Lag Analysis**

* The average delay between symptom onset and case reporting is 78.74 days, with a high standard deviation (188 days), indicating significant reporting delays and potential data entry issues.
* Negative minimum lag (-467 days) flags possible data errors, requiring cautious interpretation.
* Median delay is 6 days, suggesting most reports occur within a reasonable timeframe.

**2.4 Temporal Trends**

* Peak case day identified as January 5, 2022, with 65 reported cases.
* This peak timing aligns with broader epidemiological trends and offers an anchor point for temporal modeling.

**3. Predictive Modeling**

**3.1 Approach**

* Developed classification models to predict hospitalization and death using demographic and clinical features.
* Used stratified train-test splits to preserve outcome class proportions.
* Applied Logistic Regression with class balancing to mitigate imbalanced data issues.

**3.2 Hospitalization Prediction Model**

* **Performance:** Accuracy at 41%, with high recall (0.89) for hospitalization but very low precision (0.06), indicating many false positives.
* **Confusion Matrix:** Model correctly identifies only 24 of 27 hospitalized patients but misclassifies many non-hospitalized cases.
* **Limitations:** Severe class imbalance (134 hospitalized vs. 2949 non-hospitalized), limited feature set, and data quality constraints reduce model efficacy.

**3.3 Death Prediction Model**

* **Performance:** Accuracy at 41%, with perfect recall (1.00) but extremely low precision (0.01) for death cases.
* **Confusion Matrix:** Captures all 4 death cases but with overwhelming false positives.
* The rarity of death events (only 20 cases) severely limits model learning capacity.

**4. Challenges & Limitations**

* **Data Quality:** Significant missing data and reporting inconsistencies necessitate enhanced cleaning and validation.
* **Class Imbalance:** Critical outcomes like hospitalization and death are rare, requiring advanced techniques like SMOTE, anomaly detection, or cost-sensitive learning.
* **Feature Engineering:** Current features are limited; inclusion of clinical history, comorbidities, and symptom severity would enrich models.
* **Temporal Dynamics:** Delays in reporting and varying epidemic phases complicate modeling and interpretation.

**5. Recommendations & Next Steps**

* **Data Enrichment:** Collaborate with healthcare providers to integrate richer clinical and temporal data.
* **Advanced Modeling:** Experiment with ensemble methods, gradient boosting, and neural networks tailored for imbalanced classification.
* **Validation & Deployment:** Validate models on external datasets and explore deployment for real-time risk stratification.
* **Automated Reporting:** Develop dashboards and automated reports to monitor key metrics and model outputs, enabling proactive interventions.

**6. Conclusion**

This project demonstrates a robust analytical workflow, from data ingestion and cleaning through insightful EDA to foundational predictive modeling. While current models require refinement, the insights generated provide a clear roadmap for actionable improvements in COVID-19 patient risk prediction. This work showcases expertise in handling complex, messy real-world health data and applying sound data science principles to deliver valuable clinical insights.

**Appendix: Key Metrics Summary**

| **Metric** | **Value** |
| --- | --- |
| Total Records | 9,019 |
| Female (%) | 52.8% |
| Male (%) | 47.2% |
| Hospitalization Rate | 4.3% |
| ICU Admission Rate | 2.6% |
| Death Rate | 0.5% |
| Average Reporting Lag | 78.74 days |
| Peak Case Date | 2022-01-05 |
| Peak Case Count | 65 |