Executive Summary:

Introduction

The COVID-19 pandemic has underscored the critical importance of efficient resource allocation within healthcare systems worldwide. In response to this challenge, this project employs advanced machine learning (ML) techniques to develop predictive models aimed at optimizing the distribution of medical resources in the management of COVID-19 cases.

Leveraging anonymized patient data from the Mexican government, we aim to accurately predict the medical needs of COVID-19 patients. This endeavor not only seeks to improve patient outcomes but also to reduce the strain on healthcare infrastructures, ensuring that resources are allocated where they are needed most.

Purpose of the Study

This study provides healthcare authorities with actionable insights that can significantly enhance the allocation of medical resources in the face of the ongoing pandemic. By utilizing a range of ML models, including Logistic Regression, Random Forest, SVM, Decision Tree, KNN, Naive Bayes, Gradient Boosting, Neural Network, AdaBoost, and Easy Ensemble, our project strives to deliver a nuanced understanding of COVID-19 patients' medical requirements, facilitating a more effective response to the crisis.

Project Insights and Recommendations

Observations based on the exploratory data analysis:

1- Most of the patient population is in the mid-age range, with a skew towards younger ages, which could suggest a need for focused public health strategies for the most represented age groups.
2- The median age of patients who died is higher compared to those who survived, indicating that age is a significant factor in COVID-19 mortality, which can inform age-targeted treatment and prevention efforts.
3- both elderly men and women were similarly affected by COVID-19 in terms of mortality, which is important for understanding that age, not gender, was the more critical factor in deaths.
4- More males have died from COVID-19 than females, a crucial consideration for healthcare strategies that may need to address this gender disparity in outcomes.

5- Higher percentage of deaths among obese patients compared to non-obese patients, reinforcing the need for obesity to be a primary consideration in health planning and resource allocation for COVID-19 treatment.
6- Information and observations based on the correlation matrix is given below:

    - The correlation values can range from -1 to 1:

    - Positive Correlation (Above 0): Indicates that as one feature increases, the other feature tends to also increase. Values closer to +1 indicate a strong positive correlation.

    - Zero Correlation (Around 0): Indicates no linear relationship between the features. Values close to 0 suggest that changes in one feature do not reliably predict changes in the other.

    - Negative Correlation (Below 0): Indicates that as one feature increases, the other feature tends to decrease. Values closer to -1 indicate a strong negative correlation.

    These correlations suggest which factors may be related to each other and could inform medical resource planning.

    - Death & Pneumonia (0.47): Pneumonia significantly increases the risk of death from COVID-19, warranting intensive monitoring.

    - Hypertension & Diabetes (0.38): Co-occurrence of hypertension and diabetes calls for a combined treatment strategy.

    - Patient Type & Age (0.32): Higher hospitalization rates in older adults necessitate focused preventative healthcare.

    - Pneumonia & Patient Type (-0.65): Pneumonia patients are predominantly hospitalized, emphasizing the need for proactive care.

    - Death & Patient Type (-0.52): Greater mortality among hospitalized patients suggests a critical need for enhanced hospital care.

    - Hypertension & Age (-0.39): The unexpected decrease in hypertension with age signals potential data irregularities or unique patient characteristics.

Understanding the Metrics:

Looking at the metrics, models that have been oversampled generally show an improved balance between precision and recall for the minority class. This suggests that oversampling might have provided a more informative training process for the models, given the context of an imbalanced dataset where class 1 (death) is likely the minority (high-risk patients).

In predictive modeling, especially in healthcare applications like COVID-19 prediction, both precision and recall are crucial metrics. Precision is important to ensure that when we identify a patient as high risk, this prediction is accurate to avoid unnecessary interventions. Recall is vital to ensure we correctly identify as many high-risk patients as possible, given the potentially severe consequences of missing such cases.

A balanced model aims to achieve an optimal trade-off between precision and recall. This balance is particularly represented by the F1 score, a metric that harmonizes the two by taking their harmonic mean. A higher F1 score indicates a model effectively balances identifying true positives while minimizing false positives, crucial in healthcare settings for efficient and effective patient care.

Key Findings:

Oversampling Enhances Model Performance: Our analysis indicates that oversampling methods improve model outcomes by balancing precision and recall, particularly for predicting high-risk patients within an imbalanced dataset.

The Importance of Balanced Metrics: In healthcare applications like ours, achieving a balance between precision (correctly identifying high-risk patients) and recall (ensuring no high-risk patient is overlooked) is crucial. The F1 score, which harmonizes these metrics, serves as a key indicator of a model's efficacy in this context.

Recommendations:

For Resource-Constrained Times:

In situations where resources are limited, such as during a pandemic peak, maximizing recall for class 1 becomes critically important. A high recall for class 1 means the model is effective at identifying most of the patients who are at risk of death, which is paramount to prioritize and allocate medical resources effectively to save lives. Models with the highest recall for class 1 should be prioritized, even if it results in higher false positives (misclassifying some surviving patients as high-risk for death).

For Normal Times:

During normal times, when the healthcare system is not overwhelmed, a balanced approach might still be preferred to avoid unnecessary alarm and use of resources on patients not at high risk of death. In this scenario, models with a high F1 score for class 1 offer a good balance between precision (minimizing false alarms) and recall (not missing patients at risk of death), making them suitable choices.

Model Selection:

Normal Times (Balanced Model Selection for Death Prediction):

The models with the highest F1 scores for predicting death accurately would be most suitable. AdaBoost model, for instance, demonstrate a good balance, suggesting they might be effective for use in standard operational conditions.

Resource-Constrained Times (High Recall for Death Prediction):

The emphasis should be on models that demonstrate the highest recall for class 1 (death) to ensure that nearly all patients who could succumb to the condition are identified and given the necessary attention. **Gradient Boosting** and **Neural Network models** (with oversampling) show high recall rates for class 1, making them potentially more suitable for these critical conditions.

Conclusion

Our project exemplifies the potential of machine learning to revolutionize healthcare delivery, especially in crisis situations like the COVID-19 pandemic. By predicting patient needs with unprecedented accuracy, we can ensure that medical resources are allocated efficiently, ultimately saving lives, and mitigating the pandemic's impact on healthcare systems.