# **251HS-339F-1 : Advanced Topics in Health Care Data Analytics and Data Mining**

# **GROUP 1**

*Doreen Accorley, Yash Puri, Puja Yadav, Peter Mukiibi, Biniyam Abebe*

Instructor: Professor Moaven Razavi

# 

# 

# 

# 

# 

# 

# 

# 

# **Smart ICU Bed Allocation System (SIBAS)**

## **Predictive Analytics for Risk Profiling and Triaging COVID-19 Patients**

### **Executive Summary**

This report presents the development and validation of predictive models to identify COVID-19 patients at highest risk of adverse outcomes. Our Smart ICU Bed Allocation System (SIBAS) aims to help healthcare providers efficiently allocate critical resources during pandemic surges.

Using a dataset of 370,633 COVID-19 patients with comprehensive medical history encoded in ICD-10 format, we developed and evaluated two primary approaches: Logistic Regression with class balancing and Random Forest Classification. Both models demonstrated strong performance, with the Logistic Regression achieving 94.4% accuracy with an AUC score of 0.97, and Random Forest Classification achieving 97% accuracy with an AUC score of 0.96. These results demonstrate robust predictive power for identifying high-risk patients.

## **Problem Statement**

The COVID-19 pandemic created unprecedented strain on healthcare systems worldwide, with demand for ICU beds and ventilators frequently exceeding available supply. Healthcare providers faced the critical challenge of identifying which patients required immediate intensive care versus those who could be safely managed in less resource-intensive settings.

This project addresses this challenge by developing a predictive model that:

* Processes patient demographic and medical history data
* Calculates individual risk scores for adverse outcomes
* Supports evidence-based triage decisions
* Optimizes allocation of limited healthcare resources

## **TOOLS USED**

* Python
* Rstudio
* SAS

## **Data Overview and Preprocessing**

**Data Source:** Our analysis utilized a comprehensive dataset of 370,633 COVID-19 patients with their demographic information and medical histories.

**Data Exploration Findings:**

* The target variable "Mortality\_1\_death\_" shows significant class imbalance: 94.3% of instances belong to class 0 (no adverse outcome) and only 5.7% to class 1 (adverse outcome)
* No missing values were detected in the dataset
* All features are binary (0/1) indicators representing either demographic characteristics or presence of specific health conditions

**Preprocessing Steps:**

* Transformed detailed ICD-10 codes (69,000+ possible codes) into 227 broader clinical categories using the DGL\_3\_Extend taxonomy
* Created binary variables for demographics (age groups, gender) and health conditions
* Removed irrelevant columns (Patient\_ID was identified as not predictive)
* Standardized numerical features using StandardScaler
* Split data into training (80%) and testing (20%) sets using stratified sampling to maintain class distribution

## **Modeling Approach**

We explored multiple modeling approaches to address the class imbalance issue:

### **Logistic Regression with Class Balancing**

* Implemented using class\_weight='balanced' parameter
* Solver='liblinear', max\_iter=1000
* Standardized features before training
* Training and testing performed using stratified sampling

### **Standard Random Forest Classifier**

* 100 decision trees
* Default hyperparameters
* Achieved 97% accuracy, but with lower recall for the minority class

### **Balanced Random Forest Classifier**

* Used class weights to address imbalance
* Maintained high accuracy (97%) while improving minority class detection
* AUC score of 0.96

### **SMOTE with Random Forest**

* Applied Synthetic Minority Over-sampling Technique
* Sacrificed some overall accuracy (95%) for better balance
* Slightly lower AUC score of 0.93

Both the Logistic Regression and Balanced Random Forest Classifier showed strong performance, with each having specific advantages.

## **Models Evaluated**

* Weighted Logistic Regression
* XGBoost with Class Weighting
* SMOTE + Random Forest

## **Model Performance**

### **Random Forest Performance:**

**Confusion Matrix Analysis:**

* True Negatives: 69,529 (correctly identified low-risk patients)
* False Negatives: 1,495 (high-risk patients incorrectly classified as low-risk)
* False Positives: 343 (low-risk patients incorrectly classified as high-risk)
* True Positives: 2,759 (correctly identified high-risk patients)

**Performance Metrics:**

* Accuracy: 97.52%
* Precision: 97.90%
* Recall: 99.51%
* F1-score: 98.69%
* AUC score: 0.9648

### **Logistic Regression Performance:**

**Confusion Matrix Analysis:**

* True Negatives: 69,519 (correctly identified low-risk patients)
* False Negatives: 1,037 (high-risk patients incorrectly classified as low-risk)
* False Positives: 353 (low-risk patients incorrectly classified as high-risk)
* True Positives: 3,217 (correctly identified high-risk patients)

**Performance Metrics:**

* Accuracy: 98.12%
* Precision: 98.53%
* Recall: 99.49%
* F1-score: 98.99%
* AUC score: 0.9678

### **SMOTE + Random Forest Performance:**

**Confusion Matrix Analysis:**

* True Negatives: 69,149 (correctly identified low-risk patients)
* False Negatives: 1,053 (high-risk patients incorrectly classified as low-risk)
* False Positives: 723 (low-risk patients incorrectly classified as high-risk)
* True Positives: 3,201 (correctly identified high-risk patients)

**Performance Metrics:**

* Accuracy: 97.6%
* Precision: 98.50%
* Recall: 98.97%
* F1-score: 98.73%
* AUC score: 0.9647

## **Key Risk Factors**

Our analysis identified the following medical conditions as the strongest predictors of adverse COVID-19 outcomes:

### **Cardiovascular Conditions**

* CVASC\_Cardiac\_B (15.3% importance)
* CVASC\_Other\_Nos\_B (7.5% importance)
* CVASC\_Heart\_Rhythm\_A (4.2% importance)

### **Metabolic Disorders**

* ENDOC\_MET\_Diabetes (6.3% importance)
* ENDOC\_MET\_Metabolic\_A (4.8% importance)

### **Age**

* Age\_75\_99 (3.9% importance)

### **Respiratory Conditions**

* CHEST\_Airway\_Lungs\_A (2.5% importance)

### **General Health Services Utilization**

* GENRL\_UNSP\_Service (2.6% importance)

### **Hematological Conditions**

* HEMO\_LYMPH\_Other\_Nos (1.7% importance)

## **Implementation Recommendations**

### **Integration with Electronic Health Records**

* Create API connections to receive patient data
* Automate ICD-10 code categorization
* Generate risk scores in real-time

### **Clinical Decision Support**

* Display risk scores with confidence intervals
* Highlight key contributing factors for each patient
* Provide suggested triage pathways based on risk levels

### **Resource Allocation Guidance**

* Set configurable thresholds based on local resource availability
* Dynamically adjust recommendations as capacity changes
* Track outcomes to continuously improve predictions

### **Deployment Strategy**

* Start with pilot implementation in limited settings
* Validate performance with local data
* Scale gradually with ongoing evaluation

## **Ethical Considerations**

### **Bias and Fairness**

* Examined model performance across demographic groups
* Ensured predictions don't disadvantage vulnerable populations
* Maintained transparency in feature importance

### **Decision Support vs. Automation**

* Model designed to augment, not replace, clinical judgment
* Final decisions remain with qualified healthcare providers
* System includes clear explanation of prediction factors

### **Resource Allocation Ethics**

* Model designed for priority-setting, not care rationing
* Focus on matching appropriate care level to patient needs
* May require additional ethical frameworks if used during extreme resource constraints

### **Continuous Evaluation**

* Regular assessment of real-world performance
* Monitoring for unexpected consequences
* Updating with new clinical knowledge about COVID-19

## **Limitations and Future Work**

### **Data Limitations**

* Based on historical data that may not reflect viral mutations
* Limited to information available in electronic health records
* May not capture all relevant social determinants of health

### **Model Complexity**

* Balancing interpretability with performance
* Potential for overfitting to training data
* Challenge of maintaining performance as disease patterns evolve

### **Implementation Challenges**

* Integration with diverse hospital systems
* Training requirements for clinical users
* Resistance to algorithm-assisted decision making

### **Future Research Directions**

* Incorporate real-time vital signs and lab values
* Develop specialized models for specific patient subgroups
* Integrate with treatment effectiveness data
* Explore explainable AI approaches for better clinical acceptance

## **Model Selection and Tradeoffs**

Both models demonstrated strong predictive performance, but with different tradeoffs:

**Logistic Regression:**

* Higher AUC score (0.97)
* Better recall for the minority class (94.4%)
* Higher false positive rate (more low-risk patients incorrectly classified as high-risk)
* More interpretable model structure
* Lower computational requirements
* Faster inference time

**Balanced Random Forest:**

* Higher overall accuracy (97%)
* Lower false positive rate
* Better precision for the minority class (86%)
* More complex model structure
* Better handling of non-linear relationships
* Requires more computational resources

**Recommendation:**

* If minimizing false negatives is the highest priority (ensuring high-risk patients aren't missed), the Logistic Regression model might be preferred despite its higher false positive rate.
* If resource constraints are severe and false positives must be minimized, the Balanced Random Forest might be the better choice.
* In a production environment, an ensemble approach that combines both models could potentially leverage the strengths of each.

## **Conclusion**

The Smart ICU Bed Allocation System represents a powerful tool to support clinical decision-making during pandemic conditions. By accurately identifying patients at highest risk of adverse outcomes, SIBAS can help optimize the allocation of limited healthcare resources while potentially saving lives.

Our models adjusted for class imbalance demonstrate the importance of addressing skewed outcome distributions. With only 5.7% of patients in our dataset experiencing adverse outcomes, class imbalance posed a serious risk of biasing models toward the majority class. Without adjustment, high-risk patients could be overlooked—undermining the very goal of timely ICU triage.

Among the models evaluated:

* **Weighted Logistic Regression:** 94.36% accuracy, AUC 0.9672
* **SMOTE + Random Forest:** 97.6% accuracy, AUC 0.9647
* **XGBoost with Class Weighting:** 95.4% accuracy, AUC 0.9665

**Comparison and Recommendation:** While Weighted Logistic Regression offered strong interpretability and AUC, the **SMOTE + Random Forest** model consistently delivered the best overall performance achieving high accuracy, precision, and recall. This balance is particularly valuable in clinical settings where missing high-risk patients (false negatives) could result in severe consequences. By synthetically balancing the minority class through SMOTE and leveraging the power of ensemble decision trees, this approach excels in identifying critical cases without overwhelming resources.

Therefore, **SMOTE + Random Forest** is the most suitable model for deployment in SIBAS when both predictive performance and sensitivity to minority outcomes are top priorities. **XGBoost** remains a strong secondary option, particularly in scenarios demanding model scalability across high-dimensional data environments.

Implementation of SIBAS should proceed with careful attention to ethical considerations, continuous performance monitoring, and recognition that the model serves to support, not replace, clinical judgment. With these safeguards in place, SIBAS offers significant potential to improve patient outcomes during pandemic surges.