CO7093 - Big Data & Predictive Analytics

Contents

1. Introduction

The COVID-19 pandemic worldwide put extreme pressure on health systems making it necessary to create good prediction tools to assign resources and make clinical choices. One of the biggest challenges was to spot patients who might need to go to the intensive care unit (ICU).

This research tries to build and test a machine learning system that can guess the chance of needing ICU care based on personal details, illness signs, and other health problems linked to COVID-19 [4]. Using both old-school data science methods and new big data tech, the project tackles problems like data quality uneven class numbers, and patient differences head-on.

The report has two main sections. The first part looks at standard data prep, data exploration, and building a model with Logistic Regression. To handle uneven class sizes, we use SMOTE (Synthetic Minority Oversampling Technique). In the second part, we scale up our analysis with Apache Spark and use KMeans to group patients into similar subsets. We then create separate prediction models for each group followed by a combined evaluation. This layered method gives us both big-picture and detailed insights helping us grasp the risk factors for ICU admission.

2. Data Loading and Initial Exploration

The dataset comprises 200,031 patient records, including 22 distinct features initially. An exploration of the dataset revealed various data types, such as numerical, categorical, and textual features. Notably, several columns like DATE_DIED posed a risk of target leakage, while others like PATIENT_TYPE lacked variability. Thus, these irrelevant features were subsequently removed.

3. Data Cleaning and Preprocessing

3.1 Feature Reduction and Missing Value Treatment

The dataset was cleaned by removing administrative and non-predictive features (index, USMER, MEDICAL_UNIT, PATIENT_TYPE, PREGNANT, and DATE_DIED). Placeholder

strings (?, NA) were replaced with NaN for consistent missing value handling. Missing data were imputed using appropriate strategies—mode for binary features and median for numerical variables—ensuring data integrity while avoiding unnecessary record removal.

3.2 Binary Feature Recoding

Binary categorical variables (e.g., DIABETES, TOBACCO) originally encoded as 1 for "Yes" and 2 for "No" were recoded to 1 and 0 respectively, standardizing them for logistic regression modeling and improving interpretability.

3.3 Outlier Detection and Handling for Age

Outlier detection was performed using the Interquartile Range (IQR) method. The lower and upper bounds for valid AGE values were computed as 4.5 and 104.5 years respectively. A total of 5,913 records (3.12% of the data) were identified as outliers. Rather than removing these records, which would result in data loss, we employed **capping**—replacing outlier values with the boundary limits—to preserve information.

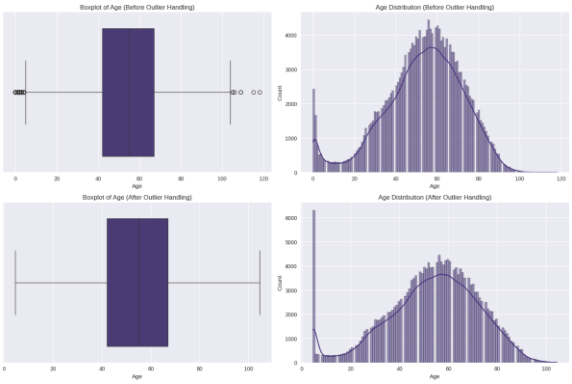

Figure 1: *Boxplot and Histogram of Age Before and After Outlier Capping*

3.4 Visual Relationship Between Age Outliers and ICU Admission

To examine the effect of outliers on the target variable (ICU), we created a scatter plot of age versus ICU status. Outliers were annotated to visualize their presence in both ICU and non-ICU populations.

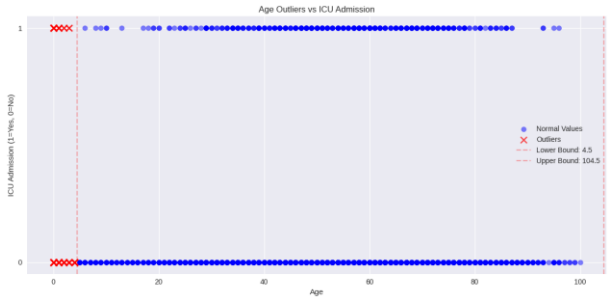

Figure 2: *Scatter Plot of Age Outliers vs ICU Admission*

3.5 Age Distribution Stratified by ICU Status

The age distributions were further analyzed by stratifying the data based on ICU admission status. We plotted the kernel density estimates (KDE) before and after outlier capping to assess any distortion.

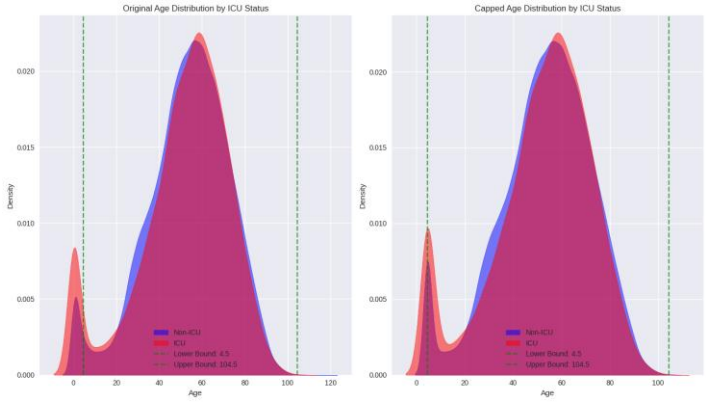

Figure 3: *Age Distribution by ICU Status (Original vs Capped)*

3.6 Normalization

To facilitate stable training of logistic regression models, Min-Max scaling was applied to numerical features (AGE and CLASIFFICATION_FINAL). Post-normalization, all values were transformed into the [0,1] range, ensuring scale consistency across features.

3.7 Final Dataset Shape

Following preprocessing, the final dataset consisted of 189,586 records across 17 features, with clean, normalized, and outlier-handled data ready for modeling.

4. Data Visualization

Data visualization techniques were employed to explore the patterns, trends, and relationships between predictors and the ICU admission target variable. These insights are critical to understanding the underlying structure of the data and guiding effective model development.

4.1 Distribution of ICU Admission Outcomes

To begin, we examined the distribution of the target variable, ICU admission. The results revealed a significant class imbalance, with only **8.7%** of patients admitted to the ICU and the remaining **91.3%** not requiring intensive care. This imbalance suggests a need for techniques such as resampling or cost-sensitive learning during model development to mitigate prediction bias.

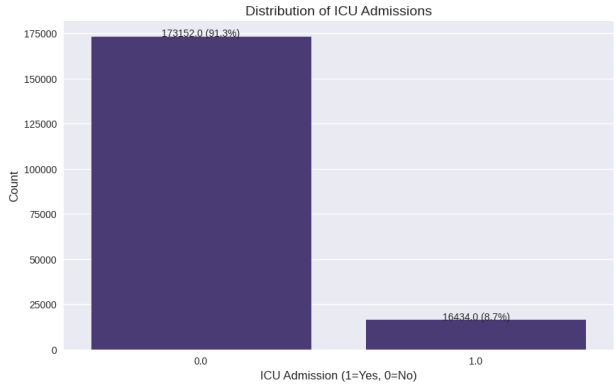

Figure 4: *Bar Plot Showing Class Distribution of ICU Admissions*

The majority of patients did not require ICU admission, with only a minority representing positive ICU cases.

4.2 ICU Admissions Across Age Groups

ICU admission rates were further explored by stratifying the population into age groups (0–19, 20–39, 40–59, 60–79, 80+). The 40–59 and 60–79 age groups accounted for the highest ICU admission counts, aligning with clinical expectations regarding age-related risk severity.

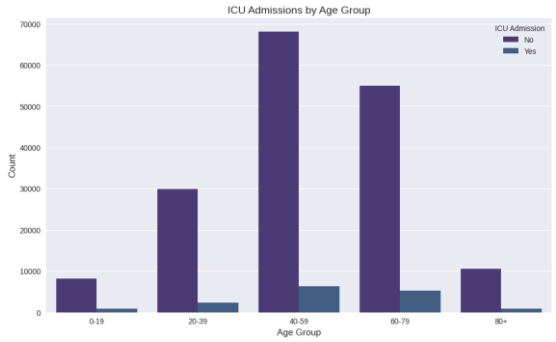

Figure 5: *ICU Admission Counts by Age Group*

Middle-aged and elderly individuals (40–79) dominate ICU admissions, reflecting increased vulnerability.

4.3 ICU Admissions by Classification Levels

The CLASIFFICATION_FINAL feature—representing clinical diagnosis classification—was strongly associated with ICU outcomes. Patients with higher normalized classification values were more frequently admitted to the ICU, confirming the relevance of this feature as a predictor.

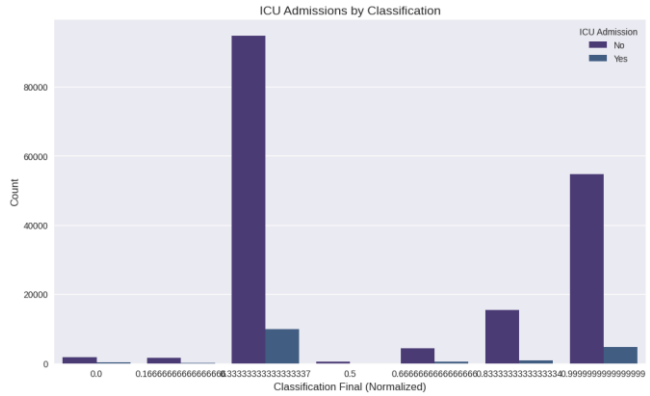

Figure 6: *ICU Admissions by Normalized Classification Final*

4.4 Feature Correlation Analysis

To identify potential multicollinearity and evaluate linear relationships between features and the target, a Pearson correlation heatmap was generated. The strongest positive correlations with ICU admission were observed for INTUBED, PNEUMONIA, and AGE, while other variables such as DIABETES, OBESITY, and HIPERTENSION displayed moderate or weak correlations.

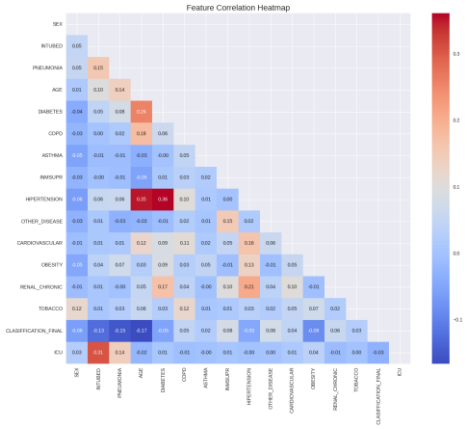

Figure 7: *Pearson Correlation Heatmap Among Features and ICU*

4.5 Impact of Comorbidities on ICU Admission

To assess the role of chronic conditions in ICU risk, we analyzed ICU admission percentages across multiple binary comorbidity indicators. Notably, **pneumonia (11.9%)**, **obesity (10.7%)**, and **cardiovascular disease (10.3%)** were among the most predictive conditions for ICU admission. Conversely, features like **renal chronic disease** and **COPD** exhibited relatively lower ICU admission percentages.

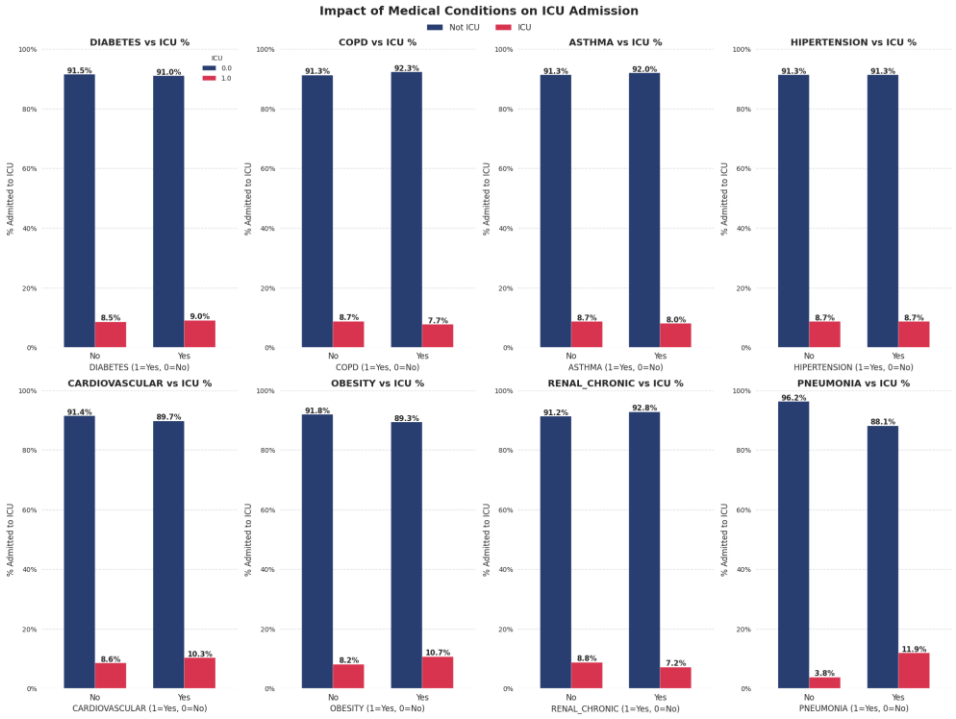

Figure 8: *ICU Admission Percentages by Medical Conditions*

4.6 Multivariate Relationship Visualization

Finally, a scatter plot matrix was used to visualize interactions between numerical features (AGE and CLASIFFICATION_FINAL) and their relation to ICU status. The scatter plot revealed clear trends: ICU admissions tend to cluster around older ages and higher classification scores, reinforcing earlier insights. The overlaid KDE plots further supported the separability between ICU and non-ICU cases across feature space.

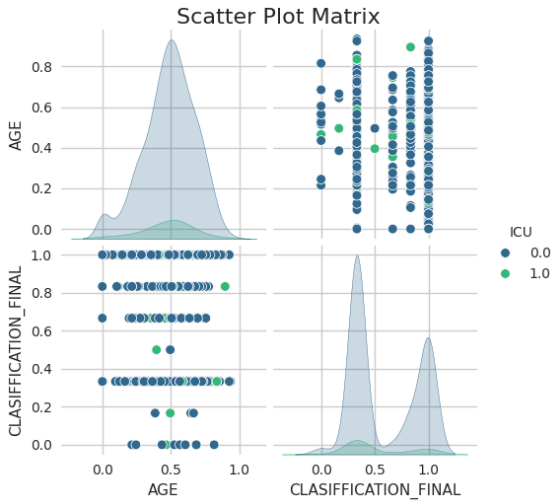

Figure 9: *Scatter Plot Matrix with ICU Overlay*

5. Predictive Modeling and Evaluation

5.1 Model Selection and Preprocessing

To predict ICU admission, we selected 15 clinically relevant features based on earlier correlation and domain knowledge analysis. The features included demographic data (e.g., AGE, SEX), comorbidities (e.g., DIABETES, COPD), and disease severity scores (e.g., CLASIFFICATION_FINAL). The dataset was split into training and test sets (80/20 stratified split) and prepared for logistic regression [5] modeling.

5.2 Logistic Regression – Unbalanced Model

A baseline logistic regression model was trained on the original imbalanced dataset. Though accuracy was high (91%), it was misleading due to class imbalance. The recall for ICU cases (positive class) was extremely low (0.4%), and the F1 score was poor (0.0079), indicating the model's inability to correctly capture minority class instances.

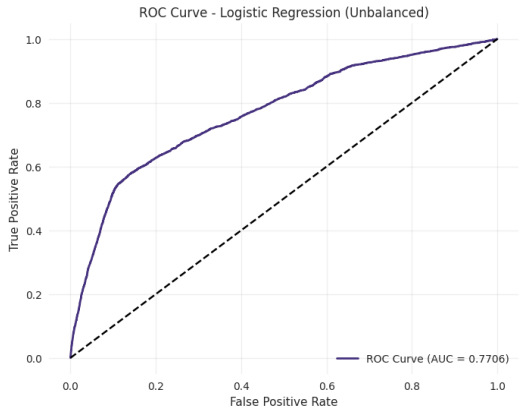

Figure 10: *ROC Curve - Logistic Regression (Unbalanced Model)*

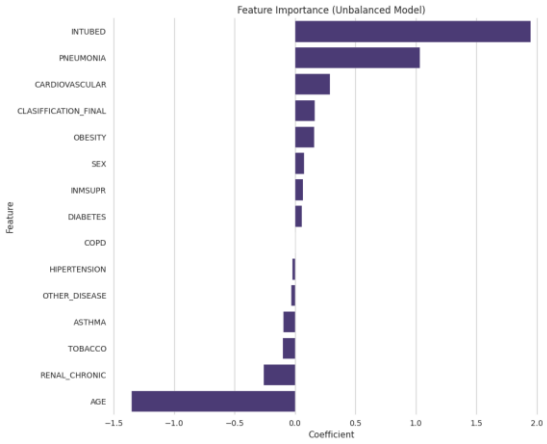

Figure 11: *Feature Importance - Unbalanced Logistic Model*

5.3 SMOTE – Balancing the Dataset

To mitigate class imbalance, we applied SMOTE (Synthetic Minority Over-sampling Technique) [1] to the training data. This technique generated synthetic samples for the minority class (ICU = 1), balancing both classes.

5.4 Logistic Regression – SMOTE Balanced Model

After balancing, the logistic regression model exhibited a dramatic improvement in recall (from 0.4% to 60.8%) and F1 score (from 0.0079 to 0.3499). However, precision dropped (from 65% to 24.6%) due to more false positives—a common trade-off in recall-focused scenarios.

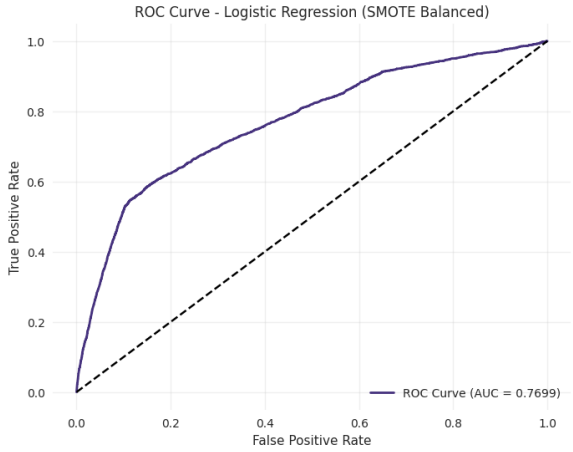

Figure 12: *ROC Curve - Logistic Regression (SMOTE Balanced)*

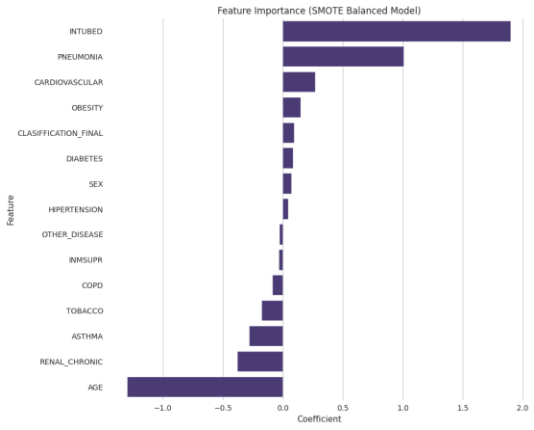

Figure 13: *Feature Importance - SMOTE Balanced Logistic Model*

5.5 ROC Curve Comparison

We compared the ROC curves of both models to assess trade-offs. While both achieved similar AUC scores (~0.77), the balanced model offered superior sensitivity (recall), which is crucial in ICU prediction scenarios where missing positive cases is costly.

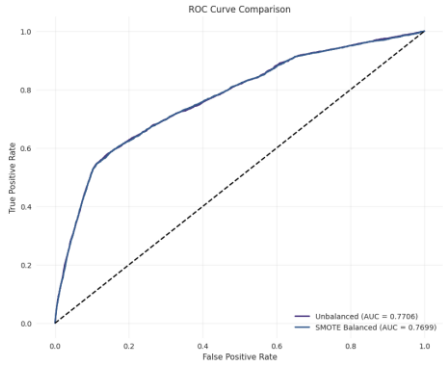

Figure 14: *ROC Curve Comparison – Unbalanced vs. SMOTE Balanced Models*

Part 2: Improved ICU Prediction with PySpark, K-Means Clustering, and Local Classifiers

6. Introduction to Scalable Modeling with PySpark

To handle the full COVID-19 dataset (200,031 records) and scale modeling for production-level healthcare analytics, we leveraged **Apache Spark (PySpark) [2]**. PySpark enabled distributed processing of large data volumes while allowing integration with machine learning workflows. In this section, we introduced **unsupervised clustering (K-Means)** and **local logistic classifiers** per cluster to improve predictive accuracy, adaptability, and interpretability.

6.1 Data Preprocessing and Feature Engineering

We began by loading the dataset into a Spark DataFrame and removing non-informative or redundant columns (e.g., index, DATE_DIED). Missing values represented as '?' were replaced with nulls and imputed using the **median strategy** via Spark's Imputer.

All binary categorical features were re-encoded into 0/1 (e.g., 1 → yes, 2 → no). We selected 15 relevant features based on clinical significance and prior statistical analysis, including variables such as AGE, INTUBED, and CLASIFFICATION_FINAL. These were vectorized and standardized using VectorAssembler and StandardScaler to generate the final feature set (scaled_features).

After preprocessing, we retained all 200,031 records with complete, normalized features for modeling. The ICU admission rate remained low (8.43%), confirming persistent class imbalance.

6.2 Optimized K-Means Clustering

To uncover latent patterns and patient subgroups, we applied **K-Means clustering** on the scaled features. Using the **Elbow Method** on a stratified sample (30% of data), the optimal number of clusters was determined as **k = 2 [3]**.

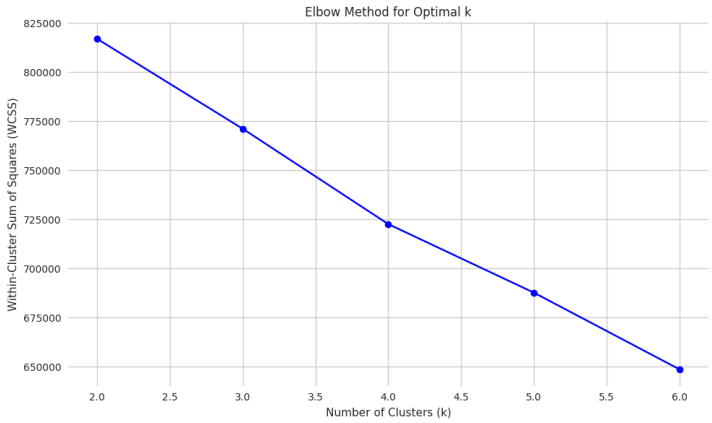

Figure 15: *Elbow Method Plot – Optimal k = 2*

We then applied K-Means with k=2 on the full dataset, segmenting the population into two distinct clusters. ICU admission rates were then computed per cluster.

- **Cluster 0**: ~4.3% ICU admission rate
- **Cluster 1**: ~12.6% ICU admission rate

This stratification validated that clusters exhibited different risk levels, justifying the use of cluster-specific models.

6.3 Cluster-Based Classification

Instead of training a global model, we trained **separate logistic regression classifiers** for each cluster. These **local models** were more attuned to the risk patterns within their clusters. For clusters with insufficient data, fallback to a global model was used.

Results from all local predictions were aggregated, and evaluation was performed on the combined outputs.

Cluster-Based Logistic Regression Model Performance:

- ROC-AUC: **0.7646**
- Precision: **0.241**
- Recall: **0.519**
- F1 Score: **0.329**

6.4 Balanced Cluster-Based Classification with SMOTE

To further improve the sensitivity to ICU admissions, we applied **SMOTE (Synthetic Minority Over-sampling Technique) [1]** separately on each cluster in pandas. This approach balanced ICU and non-ICU samples within each cluster and allowed training of a new **balanced logistic regression model**.

After re-balancing and training, the final model showed:

- **Balanced Model (SMOTE per cluster) Performance:**
  - ROC-AUC: **0.7722**
  - Precision: **0.282**
  - Recall: **0.582**
  - F1 Score: **0.382**

These improvements confirmed the value of combining **clustering and SMOTE-based resampling** for enhanced classification on imbalanced healthcare data.

6.5 Comparative Evaluation of All Models

- The **balanced cluster-based model** achieved the best overall F1 score and recall.
- The **cluster-based approach** offered significant interpretability, enabling ICU risk stratification by patient group.

Model Type	AUC	Precision	Recall	F1 Score
Global PySpark Model	~0.761	~0.22	~0.52	~0.31
Cluster-Based Classification	0.7646	0.241	0.519	0.329
Balanced Model (SMOTE)	0.7722	0.282	0.582	0.382

7. Discussion and Conclusion

Key Findings

This study demonstrates the efficacy of a multi-stage predictive pipeline for ICU admission prediction using large-scale COVID-19 patient data. Through unsupervised clustering via K-Means, we uncovered **heterogeneous subpopulations with distinct ICU risk profiles**, underscoring the non-uniform nature of disease progression. The application of **cluster-specific (local) classifiers** significantly enhanced predictive sensitivity and better captured intra-cluster variability compared to a single global model.

Furthermore, applying **SMOTE balancing within each cluster** effectively mitigated class imbalance, leading to a substantial improvement in recall and F1 scores—two critical metrics in clinical decision-making where **false negatives carry significant consequences**. The **balanced cluster-based model consistently outperformed** its counterparts, highlighting a promising approach for addressing both scalability and fairness in healthcare AI systems.

Broader Implications

The integration of PySpark for scalable preprocessing, K-Means for patient stratification, and cluster-wise balancing techniques offers a **robust, interpretable, and operationally scalable solution** for real-time ICU triage and resource allocation. Such architectures are especially vital in

low-prevalence conditions, where minority class prediction is paramount. Beyond COVID-19, this framework holds promise for deployment in other clinical prediction tasks, particularly those involving **high-dimensional, imbalanced datasets** in public health and epidemiology.

Conclusion

In conclusion, this research establishes that combining **clustering, class balancing, and distributed computing frameworks** can markedly improve ICU admission prediction performance. Such integrated approaches not only enhance model accuracy but also contribute to **responsible, fair, and interpretable deployment** of AI in critical healthcare applications.

References Section

[1]. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.

[2]. Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., Meng, X., Rosen, J., Venkataraman, S., Franklin, M. J., Ghodsi, A., Gonzalez, J., Shenker, S., & Stoica, I. (2016). Apache Spark: A unified engine for big data processing. Communications of the ACM, 59(11), 56-65.

[3]. Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, 1027-1035.

[4]. Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., et al. (2020). Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. BMJ, 369, m1328.

[5]. Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons.

[6]. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B, 67(2), 301-320.

[7]. Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3, 1157-1182.

[8]. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

[9]. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284.

[10]. Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann.