ICU Risk Prediction Project Report

1. Executive Summary

This project focuses on predicting whether a COVID-19 patient is at high risk of requiring Intensive Care Unit (ICU) admission. Using a dataset containing demographic information, symptoms, comorbidities, and clinical indicators, the goal was to develop a reliable risk-prediction model to support early clinical decision-making.

The project was completed in two major phases:

1. **Part 1 – Data Cleaning, Exploratory Data Analysis (EDA), and Baseline Modeling**
   An aggressive data-cleaning approach was used (removing rows with missing values). A baseline Logistic Regression and a SMOTE-balanced model were built and evaluated.

2. **Part 2 – Improved Modeling, Imputation, Clustering, and PySpark Implementation**
   A softer cleaning approach with imputation was applied to retain more data. Improved models such as Random Forest were built. K-Means clustering was used to explore natural grouping patterns, and PySpark pipelines were constructed for large-scale data handling.

The best models obtained strong predictive performance on recall and ROC-AUC, and cluster-based models provided additional interpretability about patient profiles.

2. Data & Methodology

2.1 Dataset Overview

The dataset consists of anonymized COVID-19 patient records with approximately 200,000+ samples and over 20 features, including:

- Demographics (AGE, SEX)
- Symptoms (PNEUMONIA, INTUBED, etc.)
- Comorbidities (OBESITY, ASTHMA, HYPERTENSION, etc.)
- Clinical classifications
- ICU admission (target variable)

The dataset was originally noisy and contained issues such as missing values marked as "?", inconsistent numeric encodings, and invalid category codes (97, 98, 99).

2.2 Part 1 – Aggressive Data Cleaning

The following steps were applied:

- Converted all missing markers to NaN
- Forced numeric columns into numeric types
- Removed invalid category codes (97/98/99)
- Converted DATE_DIED to datetime and created a DIED indicator
- Dropped all rows with any missing value
- Removed outliers (e.g., unrealistic AGE values)
- Normalized features

This reduced the dataset but produced a highly consistent subset for initial modeling.

2.3 Part 1 – Exploratory Data Analysis

EDA highlighted:

- ICU cases were severely imbalanced (<10% positive cases)
- Older patients had significantly higher ICU admission likelihood
- Certain comorbidities (OBESITY, HYPERTENSION, RENAL CHRONIC) were more prevalent in ICU patients
- Some medical variables were moderately correlated

Visualizations included distributions, count plots, scatter matrices, and correlation heatmaps.

2.4 Part 1 – Baseline Modeling

A Logistic Regression model was trained with stratified train-test split.
Key metrics were computed:

- Accuracy
- Precision
- Recall
- F1 Score

- ROC-AUC

Due to the imbalance, recall and F1 were initially low.

SMOTE Balancing

To address class imbalance, SMOTE was used on the training set.
This significantly improved recall and F1 scores, confirming that oversampling helps the model learn minority-class patterns more effectively.

2.5 Part 2 – Improved Data Preparation

Instead of dropping missing rows, imputation was used:

- Mode imputation for categorical numeric fields
- Retained nearly the entire dataset, improving model generalizability

This allowed for more robust modeling and better performance.

---

3. Models & Evaluation

3.1 Logistic Regression (Improved Data)

With more data retained, logistic regression showed higher stability and better calibration.
Improvements were observed across most performance metrics, particularly ROC-AUC.

3.2 Random Forest Classifier

The Random Forest model performed better than logistic regression on:

- F1 Score
- Recall
- ROC-AUC

This suggests that non-linear relationships and feature interactions are important in predicting ICU outcomes.

3.3 Cluster-Based Classifiers

K-Means clustering (k=3) revealed meaningful groups of patients with similar characteristics.

For each cluster, a separate classifier was trained.
Findings included:

- Some clusters contained predominantly low-risk patients
- Others contained mixed or high-risk profiles
- Local models sometimes outperformed the global model within their cluster

Cluster-based modeling improved interpretability and highlighted which patient subgroups are most vulnerable.

3.4 PySpark Pipeline

A PySpark implementation was developed to demonstrate distributed data handling:

- Loaded raw dataset in Spark
- Replaced missing markers
- Used VectorAssembler, StandardScaler, and LogisticRegression
- Built a scalable ML pipeline suitable for bigger datasets

This satisfies the requirement of handling data using big-data tools.

---

4. Insights

4.1 Key Predictive Features

Based on model coefficients, feature importance, and cluster patterns, the following features are strongly associated with ICU risk:

- AGE
- PNEUMONIA
- INTUBED
- OBESITY
- HYPERTENSION
- RENAL CHRONIC
- IMMUNOSUPPRESSION
- CARDIOVASCULAR DISEASE

4.2 Risk Patterns

- ICU admission probability increases significantly with age
- Patients with multiple comorbidities face higher risk
- A subset of patients (cluster 1) shows consistently high ICU rates
- Some features interact non-linearly (captured by Random Forest)

4.3 Class Imbalance Effects

- Logistic regression struggles without balancing
- SMOTE improves detection of high-risk patients
- Random Forest naturally handles imbalance better

---

5. Recommendations

5.1 For Clinical Use

- Models should prioritize **recall** to identify as many at-risk patients as possible
- Use Random Forest over Logistic Regression for non-linear patterns
- Use cluster insights to tailor monitoring strategies for specific patient groups

5.2 For Further Model Improvement

- Perform hyperparameter tuning (GridSearchCV)
- Try advanced algorithms (XGBoost, LightGBM)
- Use probabilistic calibration (Platt Scaling)
- Incorporate time-based features if available

5.3 For Operational Deployment

- Switch to PySpark or cloud-based ML for full-scale hospital data
- Build dashboards for real-time risk scoring
- Integrate models into patient management systems

---

6. Conclusion

This project successfully developed an ICU risk prediction system using patient data containing symptoms, medical history, and comorbidities.

- Part 1 built a baseline model using aggressively cleaned data

- Part 2 improved the process using imputation, full dataset size, and advanced models
- Random Forest emerged as the best-performing model
- Clustering provided deeper insights into patient subgroups
- PySpark ensured scalability for real-world healthcare datasets

Overall, the models demonstrate strong potential for early identification of high-risk COVID-19 patients and could assist clinicians in prioritizing care and allocating hospital resources.
