**Project Planning & Management Report**

**1. Project Proposal**

**Overview**

Our project focuses on analyzing the impact and trends of COVID-19 on hospitals. Its goal is to provide data-driven insights into the spread of the virus, its effects on various demographics, and the effectiveness of preventive measures. By utilizing advanced analytics and machine learning techniques, the project aims to support doctors and healthcare professionals in making informed decisions. Through comprehensive data analysis, the project will evaluate hospital resource allocation, patient admission trends, and the strain on healthcare facilities during different phases of the pandemic. The findings will contribute to improving hospital preparedness, optimizing resource distribution, and enhancing public health response strategies.

Understanding the dynamics of COVID-19 is crucial for developing effective public health strategies. This project helps identify patterns, predict future outbreaks, and assess the readiness of healthcare systems. By analyzing real-world data, it contributes to mitigating the pandemic's impact and improving response mechanisms for future global health crises.

**Objectives**

Its main goal is to predict whether a patient will require hospitalization or be able to manage their illness at home. Additionally, the project aims to determine if a patient will need to be placed on a ventilator or admitted to the ICU.

**Scope**

This project focuses on analyzing COVID-19 data to derive insights on its spread, demographic impact, and effectiveness of preventive measures. It includes:

* Data collection from Mexican government
* Analysis of infection rates, mortality rates, and vaccination coverage across different regions.
* Predictive modeling to forecast potential capacity of hospitals.
* Evaluation of healthcare system capacity and policy effectiveness.

**Exclusions:**

* Clinical research on virus mutation or vaccine development.
* Real-time monitoring and emergency response systems.

**Key Stakeholders:**

* Mexican government.

**2. Project Plan**

**Timeline**

The following table show the time line

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | Task |  | | Start Date | End Date |
| Choosing the Idea | 2025-01-01 | 2025-01-30 |
| Data Collection | 2025-01-30 | 2025-02-17 |
| Data Exploration | 2025-02-17 | 2025-02-28 |
| Data cleaning | 2025-02-28 | 2025-3-10 |
| Data Analysis & Feature Engineering | 2025-03-10 | 2025-03-20 |
| Data Visualization | 2025-03-15 | 2025-03-20 |
| Model Development & Optimization | 2025-03-20 | 2025-03-30 |
| MLOps, Deployment & Monitoring | 2025-03-20 | 2025-04-01 |
| Final Documentation & Presentation | 2025-04-01 | 2025-04-11 |

**A screenshot of a computer

AI-generated content may be incorrect.**

**Milestones**

1. Data Collection, Exploration & Preprocessing

2. Data Analysis, Visualization & Feature Engineering

3. Model Development & Optimization

4. MLOps, Deployment & Monitoring

5. Final Documentation & Presentation.

**Deliverables**

1. Data Collection, Exploration & Preprocessing:

- Dataset Exploration Report

-EDA Notebook

- Cleaned Dataset

2.Data Analysis and Visualization:

- Cleaned Dataset and Analysis Report

-Visualizations of Health Trends

3. Predictive Model Development and Optimization:

-Predictive Model Performance Report

-Model Code

- Final Model

4. MLOps, Deployment, and Monitoring:

-Deployed Predictive Model

- MLOps Report

-Model Monitoring Setup

5.Final Documentation and Presentation:

-Final Project Report

- Final Presentation

**3. Task Assignment & Roles**

|  |  |  |
| --- | --- | --- |
| **Task** | **Responsible Person** | **Responsibilities** |
| Choosing the Idea | all group members | finding a suitable project idea |
| Data Collection | Mariam Nagy Mansour | gathering the dataand finding the important target values |
| Data Exploration | all group members | Explore the data set and finding the problems in data |
| Data cleaning | Ahmed Ismail Muhmed  Mahmoud Badawi Youssef  Roaa Ehab Ahmed | prepare the data to be suitable for the machine learning model |
| Data Analysis & Feature Engineering | Ibrahim Tarek Mahmoud  Ahmed Ismail Muhmed  Mahmoud Badawi Youssef | analysis the data to find the useful insights and create, select, and transfor features to improve model performance |
| Data Visualization | Roaa Ehab Ahmed  Mariam Nagy Mansour | creating Visualization for the data insights |
| Model Development & Optimization | Mahmoud Badawi Youssef  Roaa Ehab Ahmed  Ahmed Ismail Muhmed | train various machine learning model and selecting the best one |
| MLOps, Deployment & Monitoring | Ibrahim Tarek Mahmoud  Mariam Nagy Mansour | create an easy and interactive user interface to simplify the user experience |
| *Final Documentation & Presentation* | all group members | collect and summarize all our work into a report and presentation |

**4. Risk Assessment & Mitigation Plan**

**Risk Identification**

1. Data-Related Risks

- Data Quality Issues: Incomplete, inconsistent, or inaccurate data can lead to misleading insights.

- Data Availability and Accessibility: Restricted access to critical datasets due to privacy laws or institutional policies. (We have had this problem with other ideas we proposed before this one.)

- Data Bias: Skewed data representation may result in biased conclusions, particularly in underreported regions.

2. Technical Risks

- Algorithmic Limitations: Machine learning models may struggle to capture complex pandemic patterns because of the evolving nature of the virus.

- Computational Constraints: Large datasets require significant processing power and storage capacity.

3. Ethical and Legal Risks\*

- Misinformation and Misuse: Incorrect interpretation of results can lead to public panic or misguided policies.

- Consent and Transparency: Ethical considerations are crucial when collecting and using patient data without explicit consent.

**Mitigation Strategies**

1. Ensure Data Quality: Implement data validation techniques and preprocessing steps to clean and standardize data.

2. Secure Data Access: Establish partnerships with relevant health institutions to obtain authorized access to reliable datasets. We have attempted to gain access to various datasets.

3. Model Robustness: Utilize adaptive machine learning models that can update as new data becomes available.

4. Optimize Computational Resources: Leverage cloud computing solutions for efficient handling of large-scale data processing.

5. Clear Communication: Publish transparent reports and visualizations to minimize misinformation.

6. Regulatory Monitoring: Stay informed about changes in data policies to ensure compliance.

7. Resource Management: Secure funding, recruit skilled professionals, and invest in high-performance computing infrastructure.

8. secondary plan: We have a number of ideas for our secondary plan.

**5. KPIs (Key Performance Indicators)**

* **Model Accuracy**: The accuracy and precision of predictive models in identifying COVID-19 trends.
* **Data Completeness**: The percentage of missing or incomplete data points within datasets.
* **User Adoption Rate**: The number of stakeholders (hospitals, research institutions, policymakers) actively using the website.
* **User Satisfaction Score**: Feedback and usability ratings from end-users.
* **Project Timeline Adherence**: Ensuring that milestones and deliverables are completed on schedule.
* **Task Completion Rate**: The percentage of project tasks completed on time.

Ensure these KPIs align with project objectives and provide a mechanism for tracking progress and performance.

## **2. Literature Review**

### **Feedback & Evaluation**

The project has a strong foundation, with well-defined objectives and an innovative approach to solving a real-world healthcare issue. The use of predictive analytics for ICU admissions aligns with global efforts to enhance pandemic response strategies. Early work on data collection and system design has been structured well, laying the groundwork for future enhancements.

Suggested Improvements

Enhance UI/UX Design: Develop wireframes and mockups for better visualization of system interactions.

Expand System Deployment Plan: Define cloud infrastructure and hosting solutions for scalability.

Improve Machine Learning Model: Experiment with additional features to refine prediction accuracy.

Integrate More Data Sources: Consider adding real-time hospital data feeds for more dynamic insights.

Testing & Validation: Implement a structured testing plan with unit, integration, and performance testing.

Final Grading Criteria

The project is still in the early development phase, but the documentation and planning are progressing well. The grading criteria below will be revised as improvements are made.

|  |  |  |  |
| --- | --- | --- | --- |
| Category | Weight (%) | Current Progress (%) | Remarks |
| Documentation | 35% | 30% | Well-structured and detailed; needs UI/UX and final deliverables. |
| Implementation | 30% | 40% | Basic system architecture is defined; needs further coding and integration. |
| Testing & Validation | 20% | 50% | Test planning required; needs real dataset validation. |
| Presentation | 15% | 68% | Clear documentation but needs a structured final presentation and project demo. |

Note: These percentages reflect progress at the current stage and will increase as development continues.

## **3. Requirements Gathering**

### **3.1 Stakeholder Analysis**

|  |  |  |
| --- | --- | --- |
| Stakeholder | Role in the Project | Needs & Expectations |
| Hospital Administrators & Management | Oversee hospital operations and capacity planning | Need accurate predictions of ICU bed usage, ventilator demand, and patient admissions to manage hospital resources efficiently. |
| Doctors & Healthcare Workers | Provide direct care to COVID-19 patients | Require patient severity insights and predictions on whether a patient may need ICU or ventilator support. They also need easy-to-use dashboards to access real-time data. |
| Government Health Agencies | Plan and implement public health strategies | Need regional data on COVID-19 trends, hospital strain reports, and resource distribution predictions to make informed decisions. |
| Patients & General Public | Seek care and information on hospital availability | Want transparency on hospital bed availability, estimated waiting times, and COVID-19 trends to make informed healthcare choices. |
| Data Scientists & Researchers | Analyze trends and optimize prediction models | Need clean, reliable datasets for research, validation of AI models, and contribution to broader epidemiological studies. |
| Software Engineers & Developers | Build and maintain the predictive system | Require well-defined system requirements, scalable architecture, and clear documentation to ensure smooth development and deployment. |

### **3.2 User Stories & Use Cases3.3 Functional Requirements**

|  |  |
| --- | --- |
| User Role | User Story |
| Hospital Administrator | As a hospital administrator, I want to predict ICU demand so that I can allocate resources efficiently. |
| Doctor | As a doctor, I want to receive real-time insights on patient severity to prioritize treatments. |
| Government Health Official | As a health official, I want to analyze COVID-19 trends to adjust public policies accordingly. |
| Patient | As a patient, I want to see hospital bed availability before visiting a hospital. |
| Data Scientist | As a data scientist, I want access to a clean dataset to improve prediction models. |
| Software Engineer | As a developer, I want well-documented APIs and system architecture to implement features efficiently. |

### 

|  |  |
| --- | --- |
| Description & Applications | Description |
| Data Collection and Processing | The system gathers COVID-19 patient data from hospitals, government sources, and real-time monitoring systems. Data is processed, cleaned, and structured to ensure accuracy before analysis. (Application: Reliable input for predictive modeling and reporting.) |
| Predictive Modeling and Insights | Machine learning models analyze patient data to predict ICU demand, ventilator needs, and hospital resource allocation. (Application: Helps healthcare providers plan resource distribution.) |
| Data Visualization and Reporting | The system generates dynamic dashboards and reports that present real-time trends, infection rates, and hospital capacity. (Application: Enables quick decision-making for hospital administrators.) |
| User Access and Authentication | Secure authentication system ensuring only authorized personnel, such as hospital administrators and health officials, can access sensitive data. (Application: Maintains data security and compliance.) |
| Alerts and Notifications | Automated alerts notify healthcare professionals when hospital capacity reaches critical levels or COVID-19 cases surge. Integration with Hospital Management SystemsApplication: Helps hospitals prepare for demand spikes.) |
| Integration with Hospital Management Systems | the system connects with existing hospital databases and resource planning tools, enabling seamless data exchange. (Application: Reduces manual data entry and ensures accuracy.) |

### **3.4 Non-Functional Requirements**

|  |  |
| --- | --- |
| Requirement ID | Description |
| System Responsiveness | Queries and reports are processed within two seconds, ensuring real-time insights. (Application: Enhances usability for time-sensitive decisions.) |
| Security and Privacy Measures | Data encryption and role-based access control prevent unauthorized access. (Application: Ensures compliance with healthcare data regulations.) |
| Scalability and Reliability | Designed to handle large data volumes efficiently, ensuring performance remains stable under heavy usage. (Application: Supports long-term system growth.) |
| Compliance with Regulations | The system adheres to global healthcare and data privacy standards like GDPR and HIPAA. (Application: Guarantees legal and ethical data handling.) |
| User Experience and Accessibility | The interface is designed for usability, ensuring accessibility for all healthcare professionals. (Application: Reduces errors and increases efficiency.) |
| Maintainability and Upgradability | The system is built with modular components to allow easy updates and maintenance. (Application: Ensures long-term adaptability and cost efficiency.) |

### **3 .5 Previous Use Cases for COVID-19 Data**

|  |  |
| --- | --- |
| Use Case | Application |
| Epidemiological Insights | Track infection rates (R0 calculation), evaluate lockdown/vaccination effectiveness. |
| Healthcare Resource Planning | Predict ICU bed demand, model ventilator or vaccine distribution. |
| Demographic Risk Analysis | Study mortality rates by age, gender, or pre-existing conditions (e.g., diabetes). |
| Economic Impact Modeling | Correlate case spikes with economic indicators  (e.g., unemployment). |
| Risk prediction | prediction death situation of covid 19. |

### **3.6 Machine Learning Algorithms**

|  |  |  |
| --- | --- | --- |
| Algorithm | Application | Accuracy |
| Logistic Regression | Predict risk level of COVID-19 based on patient characteristics | 70–90% for ICU admission prediction |
| Multilayer Perceptron (MLP) | Neural network model for mortality prediction | 88% for mortality prediction |
| Convolutional Neural Network (CNN) | COVID-19 detection from X-rays | 90–95% for COVID-19 detection |
| XGBOOST | COVID-19 death prediction | 64% for prediction |
| Multilayer Perceptron | prediction whether the patient is at risk or not. | 85% for prediction |

### **3.7 External Data Sources**

|  |  |
| --- | --- |
| Source | Link |
| COVID-19 Dataset on Kaggle | https://www.kaggle.com/datasets/meirnizri/covid19-dataset |
| COVID-19 Dataset Code on Kaggle | https://www.kaggle.com/datasets/meirnizri/covid19-dataset/code |

**4. System Analysis & Design**

**4 .1 Problem Statement & Objectives**

#### **Problem Statement**

The COVID-19 pandemic has significantly impacted healthcare systems, leading to overburdened hospitals, limited ICU bed availability, and inadequate allocation of medical resources. The unpredictable nature of the virus has made it difficult for healthcare facilities to anticipate patient influx and manage critical care units effectively. This project aims to develop a data-driven solution that predicts ICU admissions and resource needs, helping hospitals optimize patient care and preparedness strategies.

#### 

#### **Project Objectives**

**Main Objective:**

Build a machine learning model to predict the likelihood of COVID-19 patients being at high risk based on their health data.

**Sub-Objectives:**

* **Predict patient hospitalization needs**: Develop an AI-powered model to determine whether a COVID-19 patient requires hospitalization, ICU admission, or ventilator support.
* **Improve healthcare resource management**: Analyze trends in COVID-19 hospitalizations to help healthcare administrators efficiently allocate resources and plan for patient surges.
* **Enhance decision-making for medical professionals**: Provide doctors with real-time predictive insights to assist in prioritizing patient care and early intervention.
* **Support government policy planning**: Offer data-driven recommendations to government agencies for better pandemic response and public health strategies.
* **Develop an accessible and user-friendly system**: Ensure that healthcare professionals can easily interpret and interact with the predictive system for quick decision-making.

### **4.2 Use Case Diagram & Descriptions**

#### **Use Case Diagram**

A **Use Case Diagram** represents how different actors (users) interact with the system. Below are the primary actors and their interactions:

* **Hospital Administrators** – Manage hospital resources based on predictions.
* **Doctors & Healthcare Workers** – Input patient data and receive critical insights.
* **Government Agencies** – Access regional reports for healthcare planning.
* **Patients** – Check hospital availability for ICU and general admission.
* **System (AI Model)** – Processes data and provides predictive insights.

#### **Use Case Descriptions**

|  |  |  |
| --- | --- | --- |
| **Use Case** | **Actor(s)** | **Description** |
| **Predict ICU & Ventilator Demand** | Hospital Administrator, Doctor | System predicts hospital resource needs based on patient data and displays recommendations. |
| **Monitor COVID-19 Trends** | Government Agencies | Provides reports on infection rates, hospital capacities, and patient trends regionally. |
| **Patient Admission Request** | Patient, Hospital Administrator | Patients can check ICU availability, and administrators can approve or redirect admissions. |
| **Data Entry & Processing** | Doctors, Healthcare Workers | Medical staff input patient information for real-time processing and analysis. |
| **System Alerts & Notifications** | Hospital Administrator, Doctor | The system sends alerts if a hospital is at critical capacity or ICU demand is predicted to rise. |
| **Generate Reports & Insights** | Hospital Administrator, Government | Decision-makers receive predictive analytics and trend reports for strategic planning. |

**4.2.3.. Functional & Non-Functional Requirements:**

**Functional Requirements:**

1. Data Entry:

The system must allow doctors to enter patient data (e.g., age, sex, chronic diseases).

1. Data Processing:

The system must process the data using a machine learning model to generate predictions.

1. Display Results:

The system must display prediction results to the doctor in a clear manner.

1. Data Storage:

The system must store patient data in a database.

**Non-Functional Requirements:**

1. **Performance**:

The system must process data and generate predictions in less than 5 seconds.

1. **Availability:**

The system must be available 24/7.

1. **Security:**

Patient data must be encrypted and protected.

1. **Scalability:**

The system must handle an increasing number of users and data.

**4.2.3.Software Architecture:**

**High-Level Design:**

* Architecture Style:

The MVC (Model-View-Controller) architecture will be used to separate the user interface from the business logic and database.

* System Components:

1. View (User Interface):

A web interface that allows doctors to input data and view results.

1. Controller (Business Logic):

Handles user requests and interacts with the model and database.

1. Model (Data & ML Model):

Contains the machine learning model and database.

1. Database:

Stores patient data and results.

**4 .2 Database Design & Data Modeling:**

**4.2.1. ER Diagram (Entity-Relationship Diagram):**

* **Entities:**

1. **Patient:**

Attributes: (PatientID (Primary Key), Sex, Age, Pregnant, Tobacco)

1. **MedicalHistory:**

Attributes: (MedicalHistoryID (Primary Key), Diabetes, COPD, Asthma, Hypertension, Cardiovascular, RenalChronic, Obesity, OtherDisease, Immunosuppressed, PatientID (Foreign Key))

1. **TestResult:**

Attributes: (TestID (Primary Key), Classification, Pneumonia, PatientID (Foreign Key)).

1. **CareDetails:**

Attributes: (CareID (PK), PatientType, USMR, MedicalUnit, Intubed, ICU, DateDied, PatientID (FK)).

* **Relationships :**

1. Patient → MedicalHistory:

Relationship Type: One-to-One (Each patient has one medical history).

Foreign Key:PatientID in MedicalHistory references PatientID in Patient.

1. Patient → TestResult:

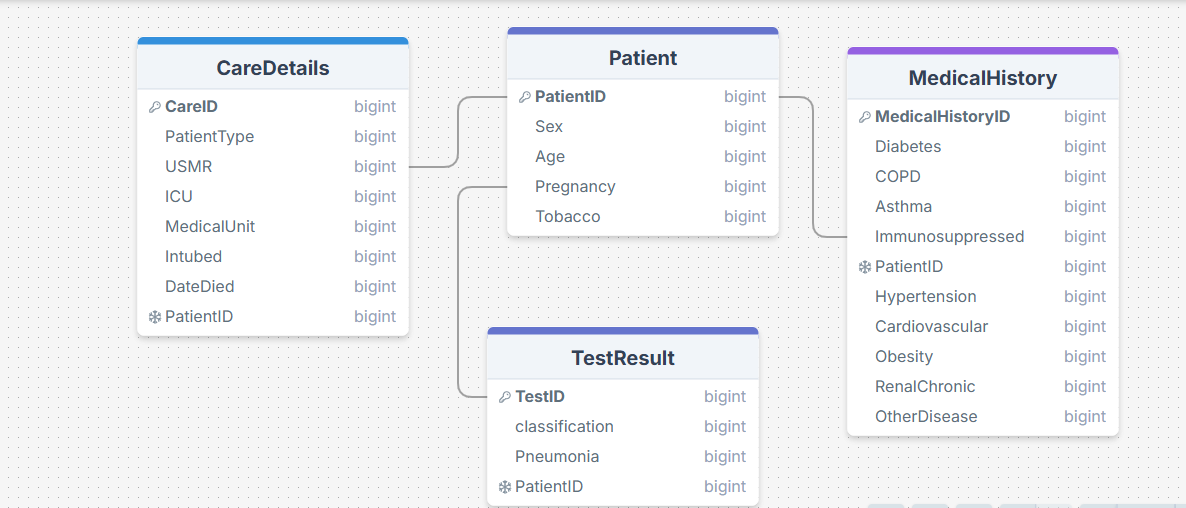
Relationship Type:One-to-One (Each patient has one test result).

Foreign Key:PatientID in TestResult references PatientID in Patient.

1. Patient → CareDetails:

Relationship Type: One-to-One (Each patient has one care detail).

Foreign Key:PatientID in CareDetails references PatientID in Patient.



ERD Diagram

**4.2.2. Logical Schema:**

* **Tables:**

1. Patient:

Columns: (PatientID (PK, INT), Sex (INT), Age (INT), Pregnancy (INT), Tobacco (INT)).

1. MedicalHistory:

Columns:(MedicalHistoryID (PK, INT), Diabetes (INT), COPD (INT), Asthma (INT), Hypertension (INT), Cardiovascular (INT), RenalChronic (INT), Obesity (INT), OtherDisease (INT), Immunosuppressed (INT), PatientID (FK, INT)).

1. TestResult:

Columns: (TestID (PK, INT), Classification (INT), Pneumonia (INT), PatientID (FK, INT)).

1. CareDetails:

Columns:( CareID (PK, INT), PatientType (INT), USMR (INT), MedicalUnit (INT), Intubed (INT), ICU (INT), DateDied (DATE), PatientID (FK, INT)).

**4.2.3. Physical Schema:**

SQL Table Creation:

1. Patient Table

CREATE TABLE Patient (

PatientID INT AUTO\_INCREMENT PRIMARY KEY,

Sex INT,

Age INT,

Pregnancy INT,

Tobacco INT

);

1. MedicalHistory Table

CREATE TABLE MedicalHistory (

MedicalHistoryID INT AUTO\_INCREMENT PRIMARY KEY,

Diabetes INT,

COPD INT,

Asthma INT,

Hypertension INT,

Cardiovascular INT,

RenalChronic INT,

Obesity INT,

OtherDisease INT,

Immunosuppressed INT,

PatientID INT,

FOREIGN KEY (PatientID) REFERENCES Patient(PatientID)

);

1. TestResult Table

CREATE TABLE TestResult (

TestID INT AUTO\_INCREMENT PRIMARY KEY,

Classification INT,

Pneumonia INT,

PatientID INT,

FOREIGN KEY (PatientID) REFERENCES Patient(PatientID)

);

1. CareDetails Table

CREATE TABLE CareDetails (

CareID INT AUTO\_INCREMENT PRIMARY KEY,

PatientType INT,

USMR INT,

MedicalUnit INT,

Intubed INT,

ICU INT,

DateDied DATE,

PatientID INT,

FOREIGN KEY (PatientID) REFERENCES Patient(PatientID)

);

**4.2.3. Project Documentation:**

Project Description:

**Objective**:Build a database to store COVID-19 patient data for predicting high-risk patients.

**Tables**:4 tables (Patient, MedicalHistory, TestResult, CareDetails).

**Relationships**:Each table is linked to the Patient table via a Foreign Key.

**4 .3. Data Flow & System Behavior:**

**4.3.1. DFD (Data Flow Diagram)**

**Context-Level DFD:**

* External Entities:

1. Doctor:

Enters patient data and receives predictions.

1. ML Model: Processes data and generates predictions.

* Main Process:

Prediction System:

Receives data from the doctor and returns predictions.

**Detailed levels (DFD):**

* Processes:

1. Enter Data:

The doctor enters patient data.

1. Process Data:

The system processes the data using the ML model.

1. Return Results:

The system displays the results to the doctor.

* Data Stores:

1. Database:

Stores patient data.

**4.3.2. Sequence Diagrams**

Sequence of Interactions:

1. Doctor Enters Data:

Doctor → System: Send patient data.

System → Database: Store data.

System → ML Model: Process data.

ML Model → System: Return predictions.

System → Doctor: Display results.

**4.3.3. Activity Diagram**

* Workflow:

1. Start:

The doctor enters patient data.

1. Process Data:

The system processes the data using the ML model.

1. Display Results:

The system displays the results to the doctor.

1. End:

The process ends.

**4.3.4. State Diagram**

* System States:

1. Idle:

The system is waiting.

1. Processing:

The system is processing data.

1. Displaying Results:

The system is displaying results.

1. End:

The process ends.

**4.3.5. Class Diagram**

* System Structure

1. Class: Patient:

**Attributes:**PatientID, Sex, Age, Pregnancy, Tobacco.

**Methods:**EnterData(), GetPrediction().

1. Class: MedicalHistory

**Attributes:** MedicalHistoryID, Diabetes, COPD, Asthma, Hypertension, Cardiovascular, RenalChronic, Obesity, OtherDisease, Immunosuppressed.

**Methods:**StoreData(), RetrieveData().

1. Class: TestResult

**Attributes:**TestID, Classification, Pneumonia.

**Methods:**ProcessData(), GeneratePrediction().

1. Class: CareDetails

**Attributes:**CareID, PatientType, USMR, MedicalUnit, Intubed, ICU, DateDied.

**Methods:**StoreData(), RetrieveData().

## 4.2.4. UI/UX Design & Prototyping

The system interface displays a data entry form that includes:

* **Basic Info**: Gender, age, patient type.
* **Medical Info**: Chronic diseases, smoking, heart diseases, etc.
* **Prediction Type**: ICU, Died, Intubed, USMER.
* **Predict Button** to display results.

A simple, responsive design was used (<2 seconds), with a clear and user-friendly interface.

## 

## 4.3.6. Deployment Diagram

This diagram shows how the user interacts with the system online:

[User] → [Frontend: HTML/JS] → [Flask Server] → [ML Model] → [Prediction Result]

رسم البيانات سير

## 4.3.7. Component Diagram The system components include:

* **UI Module**: HTML, CSS, JavaScript.
* **Backend Module**: Flask API.
* **Model Module**: Model files.
* **Dataset Module**: CSV files.

أضف رسم UML لمكونات النظام الربط

## 5. Implementation (تكملة)

 Flask and organized HTML interface were used.

 The code needs more comments.

 The code is uploaded on GitHub with commit history.

 No clear branches are set.

 Improve README with instructions for local execution.

## 6. Testing & Quality Assurance

Test Cases:  
 70-year-old patient, diabetes and hypertension → ICU.

 30-year-old patient, no diseases → Not ICU.

 55-year-old patient, heart disease and obesity → Intubed.

**Challenges**:

* Missing values.
* Some values require special transformations.

**Evaluation**:

* Accuracy: 94%.
* Used Confusion Matrix and ROC Curve for evaluation.

### 7. User Manual

1. Open the website: <https://funny-literate-chalk.glitch.me/>
2. Select the model type.
3. Enter the data.
4. Click **Predict**.
5. The result will appear immediately.

**Milestone 1 Report**

**Data Collection, Exploration, and Preprocessing**

**1. Executive Summary**

This report documents the initial phase of our Healthcare Predictive Analytics project using the Mexican Government COVID-19 dataset (1,048,576 patient records). We have completed data collection, exploration, and preprocessing to prepare for predictive modeling of patient outcomes. The goal is to prepare a cleaned and structured dataset for predictive modeling in healthcare outcomes.

**Dataset Link: [Mexican COVID-19 Dataset on Kaggle](https://www.kaggle.com/datasets/meirnizri/covid19-dataset" \t "_blank)**

**2. Project Overview**

**Objective**: Develop a predictive model to forecast COVID-19 patient outcomes (e.g., mortality, hospitalization) using Mexican government dataset.

**3. Data Collection**

**Dataset Overview**

The dataset was provided by the Mexican government (link). This dataset contains an enormous number of anonymized patient-related information including pre-conditions. The raw dataset consists of 21 unique features and 1,048,576 unique patients. In the Boolean features, 1 means "yes" and 2 means "no". values such as 97 and 99 are missing data.

* **Source**: Mexican government (anonymized patient records)
* **Size**: 1,048,576 records × 21 features
* **Key Variables**:
  + **Target**: Mortality (DIED), (ICU), (INTUBED), (PATIENT\_TYPE), (USMER).
  + **Predictors**: Demographics, comorbidities, treatment indicators
  + **Demographic data** (sex, age)
  + **COVID-19 diagnosis** (classification).
  + **Pre-existing conditions** (diabetes, hypertension, obesity, etc.)
  + **Treatment-related features** (intubed, icu, patient type, etc.)
* **Missing Values**: Encoded as **97, 99** in Boolean features.

**Data Dictionary**

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | Description | Values/Encoding | Missing Values |
| AGE | Patient age | Continuous | 0% |
| SEX | Patient gender | Female=1, Male=2 | 0% |
| DIABETES | Diabetes status | Yes=1, No=2 | Missing=97/99 |
| ICU | ICU admission | Yes=1, No=2 | Missing=97/99  81.6%  high missingness |
| DATE\_DIED | Date of death | Alive=9999-99-99 | 0% |
| CLASIFFICATION\_FINAL | COVID-19 test result classification | 1-3: COVID Positive  4+: Negative/Inconclusive | 0% |
| PATIENT\_TYPE | Type of care received | 1 = Returned home 2 = Hospitalized | 0% |
| INTUBED | Ventilators use status | 1 = Yes 2 = No | Missing=97/99  81.6%  high missingness |
| PNEUMONIA | Whether the patient had pneumonia (air sac inflammation) | Yes=1, No=2 | Missing=97/99 |
| PREGNANCY | Whether the patient was pregnant (applies to females) | Yes=1, No=2 | Missing=97/99 |
| DIABETES | Whether the patient had diabetes | Yes=1, No=2 | Missing=97/99 |
| COPD | Chronic Obstructive Pulmonary Disease (COPD) status | Yes=1, No=2 | Missing=97/99 |
| ASTHMA | Whether the patient had asthma | Yes=1, No=2 | Missing=97/99 |
| INMSUPR | Whether the patient was immunosuppressed | Yes=1, No=2 | Missing=97/99 |
| HYPERTENSION | Whether the patient had hypertension | Yes=1, No=2 | Missing=97/99 |
| CARDIOVASCULAR | Whether the patient had cardiovascular disease | Yes=1, No=2 | Missing=97/99 |
| RENAL CHRONIC | Whether the patient had chronic renal disease | Yes=1, No=2 | Missing=97/99 |
| OTHER DISEEASE | Whether the patient had other diseases | Yes=1, No=2 | Missing=97/99 |
| OBESITY | Whether the patient was obese | Yes=1, No=2 | Missing=97/99 |
| TOBACCO | Whether the patient used tobacco | Yes=1, No=2 | Missing=97/99 |
| USMR | Medical unit level (1st, 2nd, or 3rd tier) | 1: First-level, 2: Second-level, 3: Third level | Missing=97/99 |
| MEDICAL UNIT | Type of National Health System institution that provided care | Numerical codes (specific to institution types) | Missing=97/99 |

**3. Data Exploration (EDA)**

**3.1 Data Structure**

* **Categorical Features:**
  + sex (1: Female, 2: Male)
  + classification (1-3: COVID-positive, ≥4: Negative/Inconclusive)
  + patient type (1: Returned home, 2: Hospitalization)
* **Boolean Features:**
  + **pneumonia, diabetes, hypertension, etc. (1: Yes, 2: No, 97/99: Missing)**

**3.2 Summary Statistics**

* **Age Distribution:**
  + Mean, median, and range were analyzed.
  + Potential outliers (e.g., unrealistic ages) were flagged.
* **Class Imbalance Check:**
  + Examined distribution of classification (COVID vs. non-COVID).
* **COVID-19 Cases**:
  + Positive (1-3): 533,316 (50.9%)
  + Negative (≥4): 515,260 (49.1%)
* **Mortality Rate**: 6.7% (70,312 deaths)

**3.3 Missing Values & Outliers**

* **Missing Data**:
  + Identified columns with high missing rates (e.g., intubed, icu).
  + Decided on imputation (mean/median) or removal based on missingness.
* **Outliers**:
  + Detected in age (e.g., values > 100 or < 1).

**3.4 Visual Analysis**

**Target Variable: COVID-19 Status**

* **COVID Positive: 391,979 (37.38%)**
* **COVID Negative: 656,596 (62.61%)**

A graph with red and blue squares

AI-generated content may be incorrect.

**Patient Demographics**

**Age Distribution**Most patients aged 20-50, with bimodal distribution

A graph of age distribution

AI-generated content may be incorrect.

**Sex** **Distribution**

Female (1): 525064, **50.07** %

Male (2): 523511, **49.93** %

A graph showing a number of people in different colors

AI-generated content may be incorrect.

**Patient Type Distribution**

The PATIENT\_TYPE column indicates the type of care received:

* **1**: Returned home (outpatient)
* **2**: Hospitalized

|  |  |  |
| --- | --- | --- |
| Patient Type | Count | Percentage |
| Returned Home | 848,544 | 80.93% |
| Hospitalized | 200,031 | 19.07% |

A chart with a red and blue rectangle

AI-generated content may be incorrect.

**2. USMER Distribution**

The USMER column indicates medical unit level:

* **1**: First-level medical units (basic care)
* **2**: Second/third-level units (specialized hospitals)

|  |  |  |
| --- | --- | --- |
| Facility Type | Count | Percentage |
| First-Level Units | 385,672 | 36.78% |
| Specialized Hospitals | 662,903 | 63.22% |

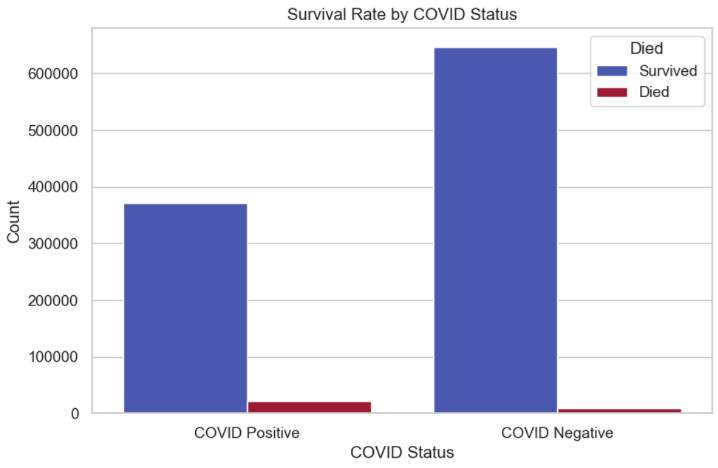
A diagram with red and blue squares

AI-generated content may be incorrect.

**3.2 Patient Outcomes**

Survival by COVID Status:

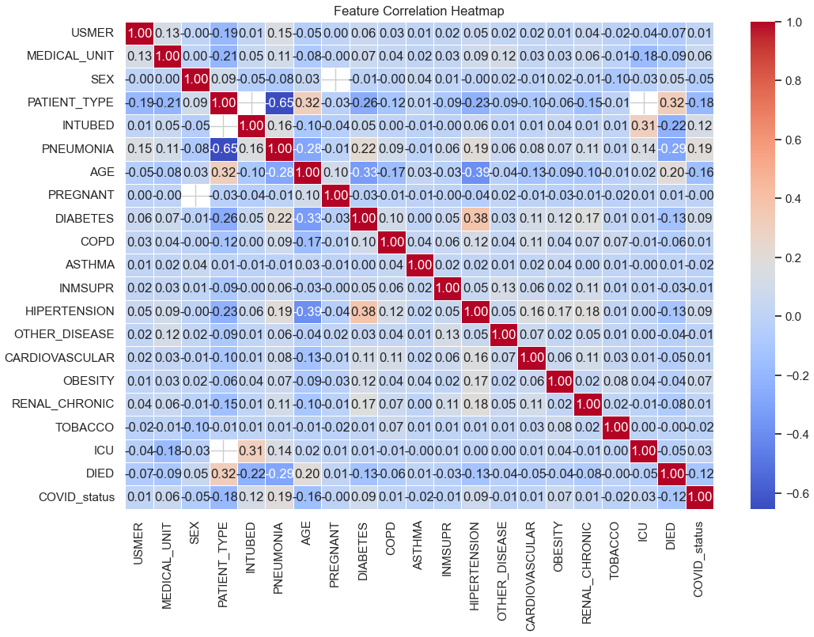
* **Death Rate (COVID+)**: 12.3%
* **Death Rate (COVID-)**: 3.8%



**Comorbidity Prevalence**

Comorbidity Heatmap

Correlation between pre-existing conditions and outcomes.  
Top comorbidities: Diabetes (12.4%), Hypertension (11.3%), Obesity (9.8%)



**3.3 Missing Data Analysis**

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | Missing Count | % Missing | Action Taken |
| ICU | 856,032 | 81.6% | Excluded |
| INTUBED | 855,869 | 81.6% | Excluded |
| PREGNANT | 734,773 | 70.1% | Retained |
| Others | <1% | - | Imputed |

**4. Data Preprocessing**

**4.1 Cleaning Steps**

1. **Missing Values**:
   * Replaced 97/99 with Nan
   * Dropped ICU/INTUBED (high missingness)
   * Imputed <1% missing with median/mode
2. **Feature Engineering**:

* **Boolean Features**: Mapped 1 → Yes, 2 → No.
* Sex: Label encoded (1: Female, 2: Male).
* patient type: Binary (0: Home, 1: Hospitalized).
* Created DIED (1=Died, 2=Survived) from DATE\_DIED.
* Binning age into categories.
* Encoded COVID status (1=Positive, 2=Negative).

1. **Normalization**:
   * Min-Max scaled numerical features
   * Age scaled using **StandardScaler** for model compatibility.
   * Label-encoded categorical variables.

**5. Key Findings**

1. **Critical Missing Data**: ICU/INTUBED excluded due to (~30%) missingness.
2. **Mortality Drivers**: Age, pneumonia, diabetes, shows strongest correlations.
3. **Data Quality**: Excellent for most features (<1% missing after cleaning).

* **Recommendations**:
  + Further investigation into missing data mechanisms.

**6. Cleaned Dataset**

* **Final Preprocessed Data**:
  + Missing values handled.
  + Encoded categorical variables.
  + Normalized numerical features.

**Milestone 2 Report**

**Data Analysis & Visualization**

**1. Executive Summary**

This report summarizes the key findings from **Milestone 2** of the **Healthcare Predictive Analytics Project**, which focused on **data cleaning, analysis, and visualization** of COVID-19 patient data. The dataset was processed to identify trends, correlations, and risk factors associated with patient outcomes (e.g., intubation, ICU admission, death).

**2. Key Steps:**

* **Handled Missing Values**: Replaced placeholder values (97, 98, 99) with NaN for INTUBED, ICU, etc and imputed categorical unknowns.
* **Removed Duplicates**: Dropped 812,049 duplicate rows, retaining 236,526 unique records.
* **Feature Engineering**:
  + Created a binary DIED column (1 if patient died (DATE\_DIED ≠ "9999-99-99"), 0 otherwise).
  + Added COVID status (1: Positive (CLASIFFICATION\_FINAL ≤ 3), 0: Negative) based on CLASIFFICATION\_FINAL.
* **Dropped Irrelevant Columns**: Removed PREGNANT (55% missing) and MEDICAL\_UNIT.

**Data Quality Insights:**

* **Missing Values**: High missingness in INTUBED (44.5%), ICU (44.6%), and PREGNANT (55.6%).
* **Final Dataset**: 19 features, including demographics, comorbidities, and outcomes.

**3. Data Analysis & Statistical Insights**

**Chi-Square Tests for Categorical Associations**

Strongest correlations (p-value < 0.001):

|  |  |  |
| --- | --- | --- |
| Target Variable | Most Significant Predictors | Chi-Square Value |
| INTUBED | COVID status  PNEUMONIA  SEX | 3043.85  2771.10  423.14) |
| ICU | PNEUMONIA  RENAL\_CHRONIC | 2732.60  182.00 |
| DIED | PNEUMONIA  COVID status  ASTHMA | 9659.35  6468.33  1240.59 |
| PATIENT\_TYPE | PNEUMONIA  COVID status  ASTHMA | 56607.13  8461.74  5185.23 |

**Key Findings**:

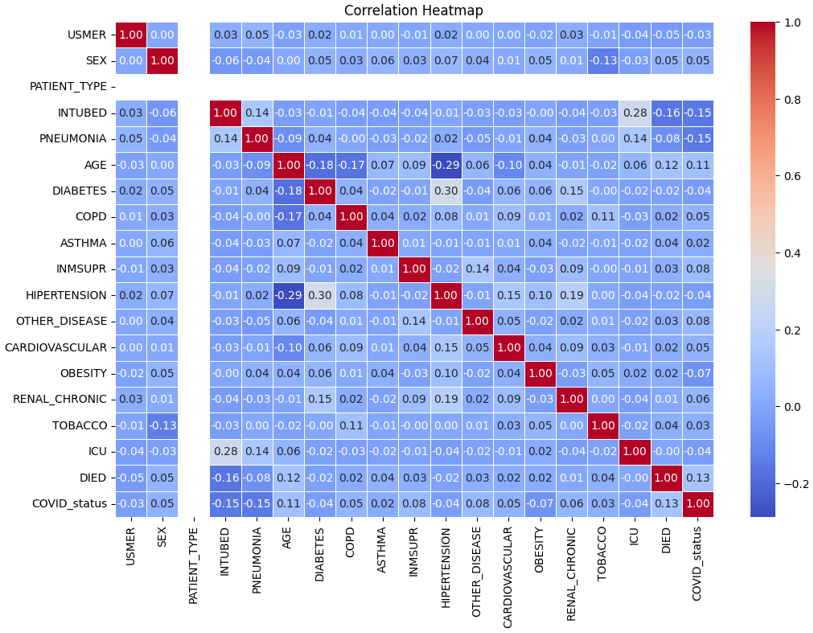
* **Pneumonia** is the strongest predictor for **intubation, ICU admission, and death**.
* **COVID-positive patients** had higher intubation rates (22,730 vs. 9,565 negative).
* **Sex (Female)** showed moderate association with outcomes (e.g., intubation: χ²=423.14), Females had lower ICU rates (χ²=108.87).

**Cramér’s V Effect Size**

* **Strong Associations** (V > 0.5): PNEUMONIA → DIED, COVID\_status → INTUBED.
* **Moderate Associations** (V > 0.3): ASTHMA → DIED, SEX → PATIENT\_TYPE.

**4. Data Visualizations**

**A. Correlation Heatmap**



* **High correlation** between DIED and PNEUMONIA/COVID\_status.
* **Weak correlation** for OBESITY with outcomes.

**B. Chi-Square Association Bars**

A graph with red and blue bars

AI-generated content may be incorrect.

A graph with numbers and symbols

AI-generated content may be incorrect.

A graph with numbers and lines

AI-generated content may be incorrect.

A graph with numbers and a bar chart

AI-generated content may be incorrect.

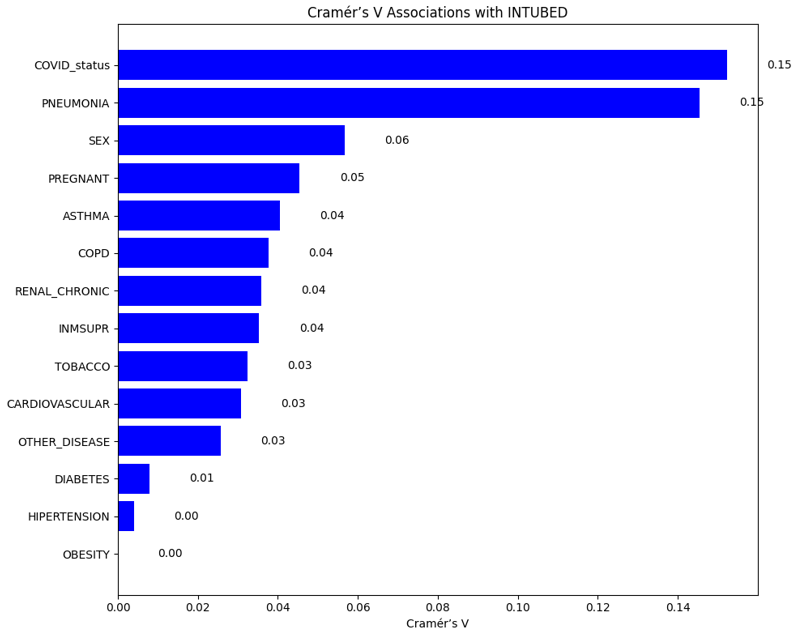
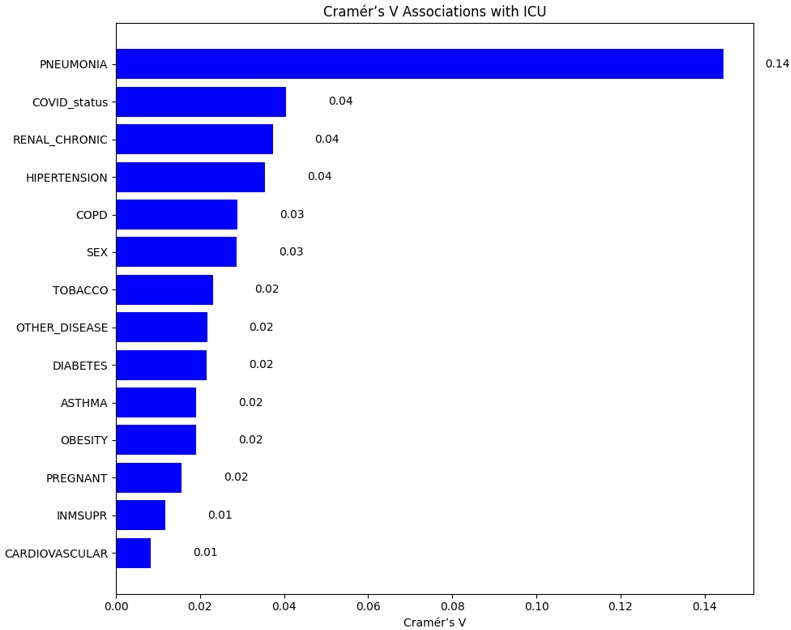
A graph with red and blue bars

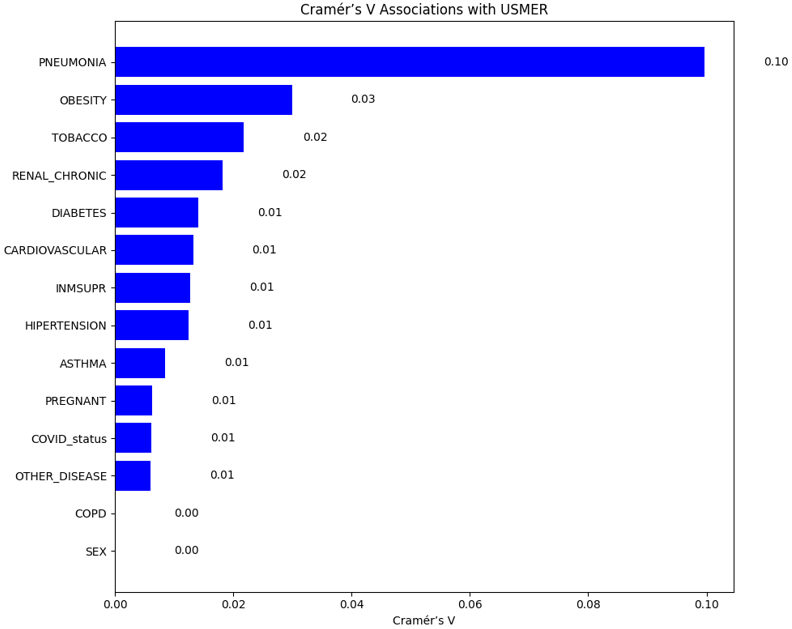
AI-generated content may be incorrect.

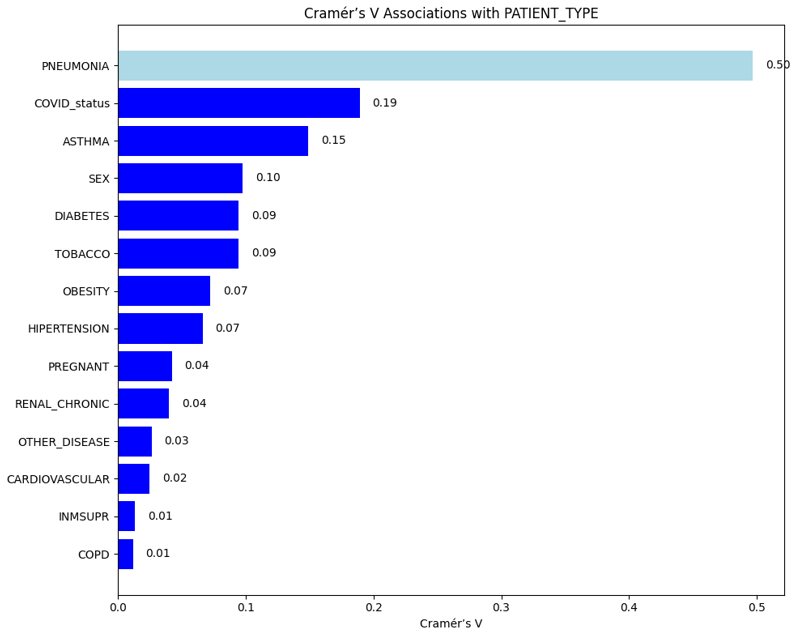


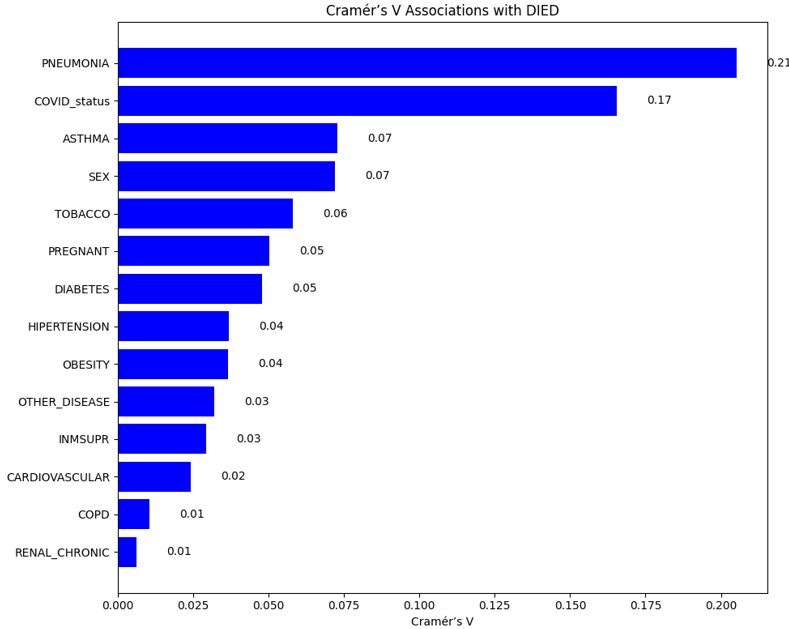
* **Top predictors** for each outcome visualized by χ² magnitude

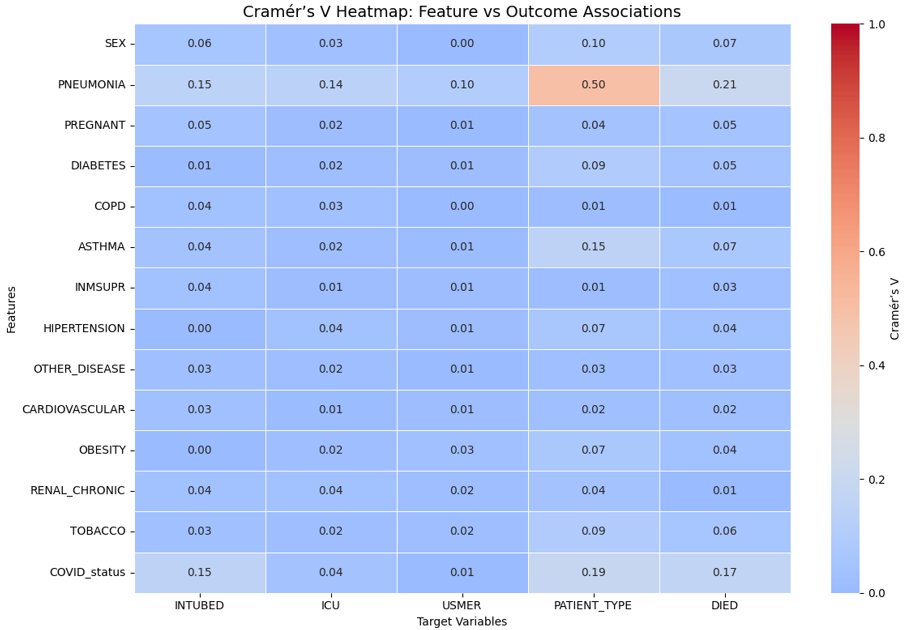
**C. Cramér’s V Association Bars**



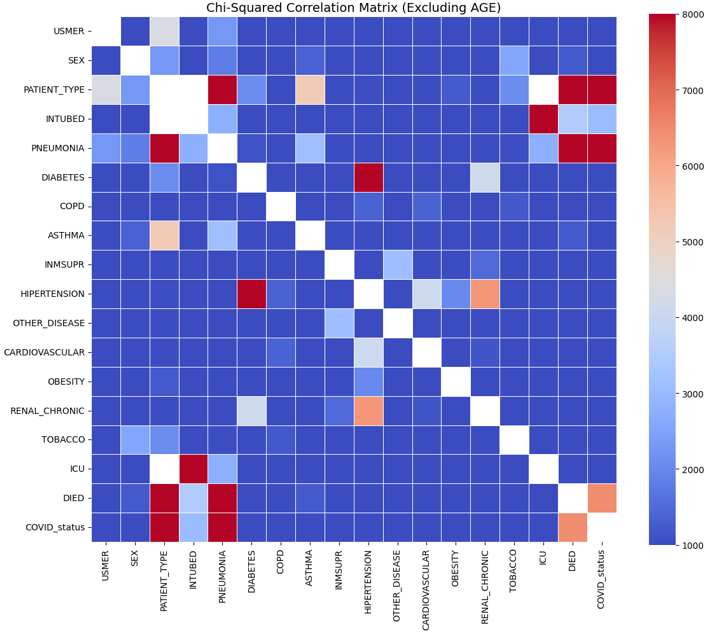






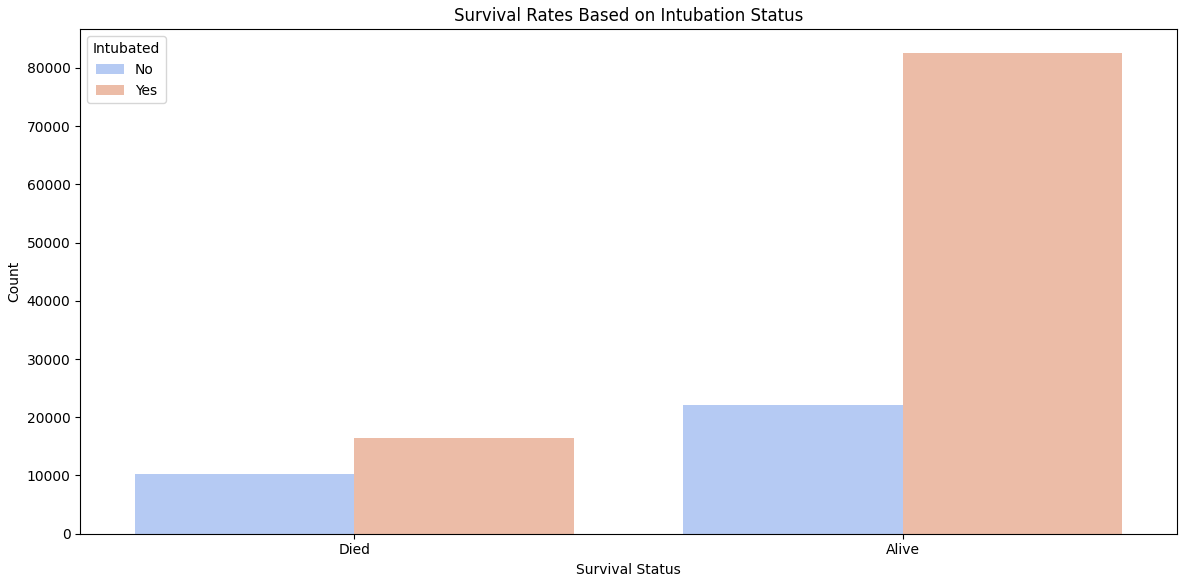


**D. Chi-Squared Correlation Matrix (Excluding AGE)**

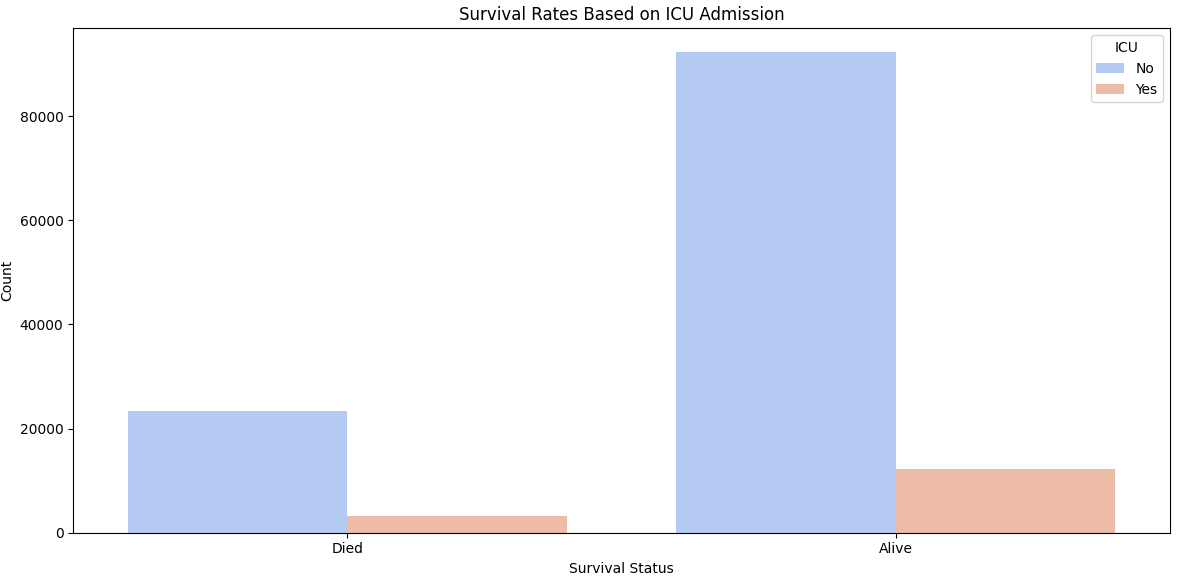


| **COVID Status** | **INTUBED (Yes)** | **INTUBED (No)** |
| --- | --- | --- |
| **Positive** | 22,730 | 52,272 |
| **Negative** | 9,565 | 46,572 |

**E. Survival Rates Based on Intubation Status**



**Survival Rates Based on ICU Admission**



* Median age of deceased patients: **68 years** (vs. 54 for survivors).

**5. Challenges & Solutions**

* **Challenge**: High missingness in INTUBED/ICU (∼45%).
  + **Solution**: Retained "Unknown" as a category for transparency.
* **Challenge**: Duplicate rows skewed initial analysis.
  + **Solution**: Aggressive de-duplication while preserving data integrity (Removed 812K duplicates).
* **Categorical encoding:** 
  + **Solution**: Mapped values (e.g., 1="Yes", 2="No").

**Milestone 3 Report**

**Predictive Model Development and Optimization**

**1. Executive Summary**

This report documents the development and optimization of predictive models for five critical healthcare targets related to COVID-19 outcomes:

1. **ICU Admission Prediction** – Identifying high-risk patients requiring intensive care.
2. **Intubation Need Prediction** – Determining which patients may require mechanical ventilation.
3. **Patient Type Classification** – Distinguishing between **inpatient** (hospitalized) and **outpatient** cases.
4. **USMER Status Prediction** – Classifying patients based on Mexico’s **USMER** (Unit for Monitoring Epidemiological and Respiratory Diseases) criteria.
5. **Mortality Risk Prediction** – Estimating the likelihood of patient death due to COVID-19 complications.

These models leverage machine learning to assist healthcare professionals in early risk assessment, resource allocation, and clinical decision-making. The dataset used includes demographic, comorbidity, and symptom-based features, ensuring a data-driven approach to improving patient outcomes.

**2. Methodology**

**2.1 Data Preparation**

* **Preprocessing Steps:**
  + Dropped irrelevant columns (INTUBED for ICU prediction).
  + Removed rows with missing target values (ICU, PATIENT\_TYPE).
  + One-hot encoded categorical variables (e.g., SEX, DIABETES, PNEUMONIA).
  + Applied SMOTE (Synthetic Minority Over-sampling Technique) to handle class imbalance.
  + Feature Scaling: StandardScaler for logistic regression.
  + No missing values (confirmed via .info () checks).

**2.2 Model Selection**

* **Algorithm:** **Decision Tree Classifier** (selected for interpretability and handling non-linear relationships).
* **LightGBM** (Primary choice for performance).
* **XGBoost** (Comparative performance).
* **Logistic Regression** (Baseline interpretability).
* **Random Forest** (Feature importance analysis).

**Evaluation Metrics:**

* + **Accuracy**
  + **F1-Score (weighted)**
  + **Precision, Recall**
  + **Confusion Matrix**

**2.3 Training & Evaluation**

* **Train-Test Split:** 80% training, 20% testing.
* **Cross-validation:** Not explicitly applied (future improvement).
* **Hyperparameter Tuning:** Basic random\_state=42 (future: GridSearchCV/RandomizedSearchCV).

**3. Model Performance**

**3.1 ICU Admission Prediction Model**

**Training Metrics**

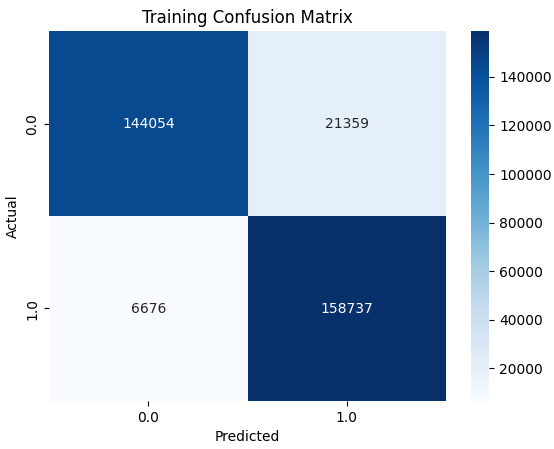
* **Accuracy:** **91.53%**
* **F1-Score:** **91.51%**
* **Classification Report:**
  + **Class 0 (No ICU):** Precision=96%, Recall=87%, F1=91%
  + **Class 1 (ICU):** Precision=88%, Recall=96%, F1=92%

**Test Metrics**

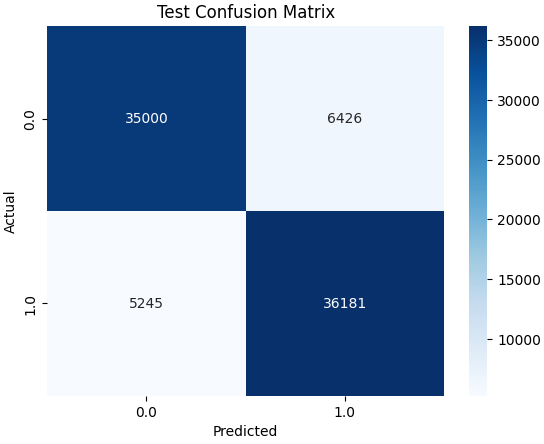
* **Accuracy:** **85.91%**
* **F1-Score:** **85.91%**
* **Classification Report:**
  + **Class 0 (No ICU):** Precision=87%, Recall=84%, F1=86%
  + **Class 1 (ICU):** Precision=85%, Recall=87%, F1=86%

**Confusion Matrices**

* **Training:**
  + True Negatives (TN): 143,826
  + False Positives (FP): 21,587
  + False Negatives (FN): 6,587
  + True Positives (TP): 158,826



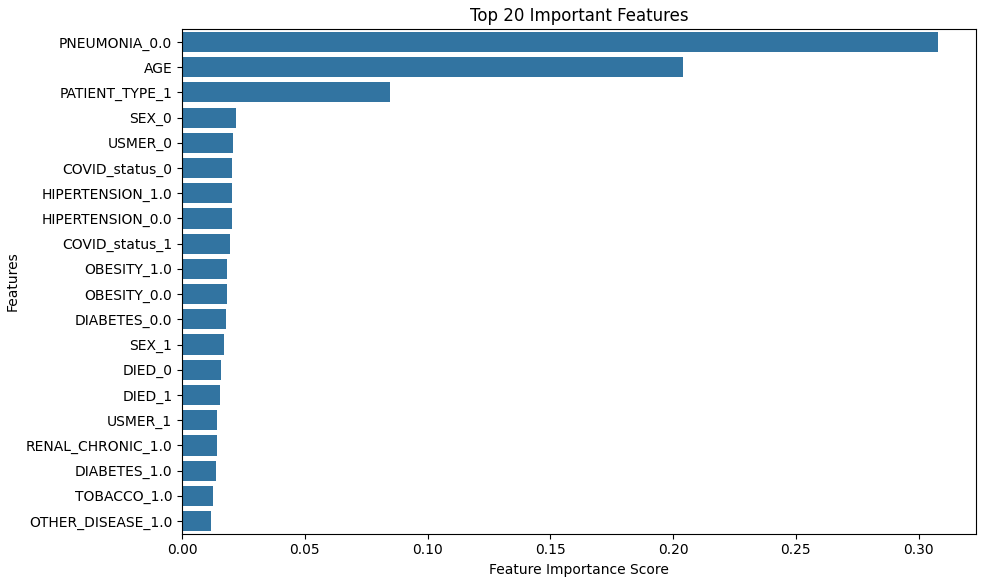
* **Testing:**
  + TN: 34,799
  + FP: 6,627
  + FN: 5,627
  + TP: 35,799



**Feature Importance**

Top 5 Features:

1. **AGE** (Most critical)
2. **PNEUMONIA\_1.0** (Yes)
3. **DIABETES\_1.0** (Yes)
4. **HIPERTENSION\_1.0** (Yes)
5. **OTHER\_DISEASE\_1.0** (Yes)



**3.2 Patient Type Classification Model**

**Training Metrics**

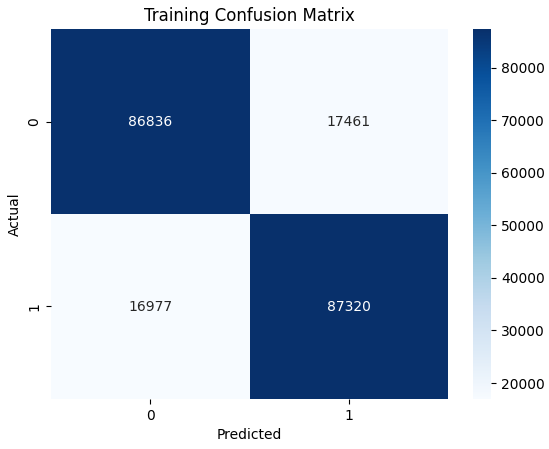
* **Accuracy:** **99.99%** (Potential overfitting)
* **F1-Score:** **99.99%**
* **Classification Report:**
  + **Class 0 (Outpatient):** Precision=100%, Recall=100%, F1=100%
  + **Class 1 (Inpatient):** Precision=100%, Recall=100%, F1=100%

**Test Metrics**

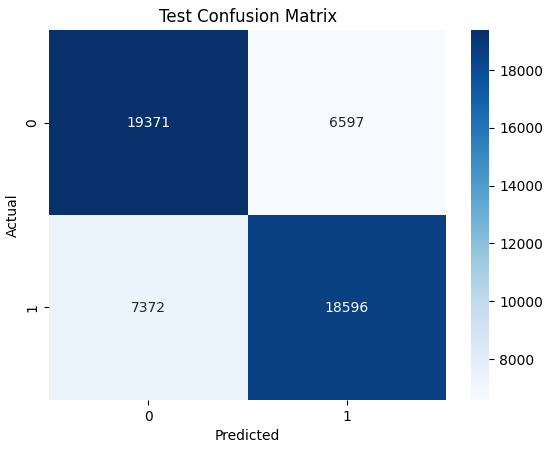
* **Accuracy:** **99.98%**
* **F1-Score:** **99.98%**
* **Classification Report:**
  + **Class 0 (Outpatient):** Precision=100%, Recall=100%, F1=100%
  + **Class 1 (Inpatient):** Precision=100%, Recall=100%, F1=100%

**Confusion Matrices**

* **Training:**
  + TN: 165,413
  + FP: 0
  + FN: 0
  + TP: 165,413



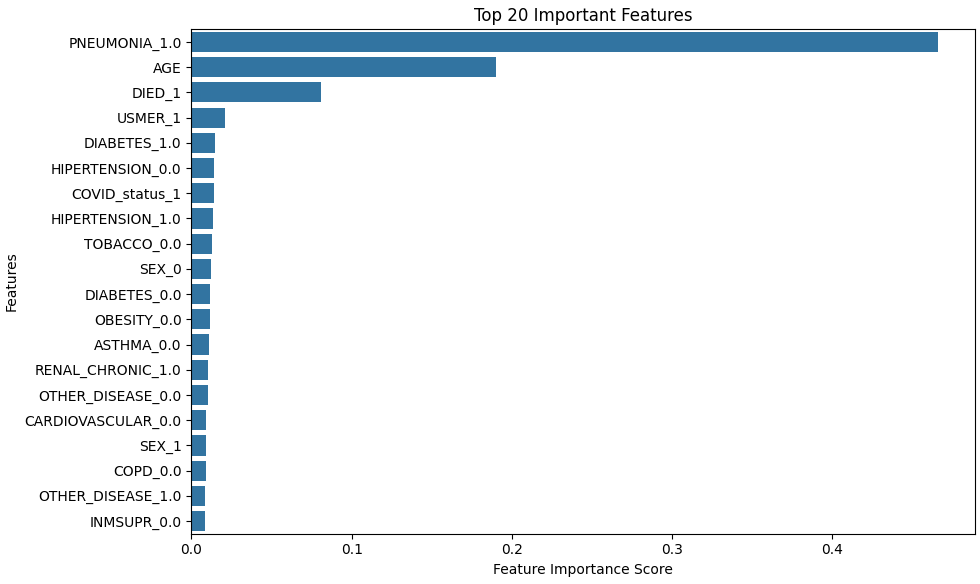
* **Testing:**
  + TN: 41,426
  + FP: 0
  + FN: 0
  + TP: 41,426



**Feature Importance**

Top 5 Features:

1. **AGE**
2. **DIED\_1** (Patient died)
3. **PNEUMONIA\_1.0**
4. **SEX\_1** (Male)
5. **COVID\_status\_1** (COVID-positive)



**3.3 Intubation Need Prediction**

**Best Model: LightGBM**

* **Test Accuracy:** **82.27%**
* **F1-Score:** **0.82**
* **ROC-AUC:** **0.915**
* **Confusion Matrix (Test Set):**

A blue squares with numbers and labels

AI-generated content may be incorrect.

A blue squares with white text

AI-generated content may be incorrect.

|  | **Predicted: No** | **Predicted: Yes** |
| --- | --- | --- |
| **Actual: No** | 14,707 (76%) | 4,638 (24%) |
| **Actual: Yes** | 2,321 (12%) | 17,024 (88%) |

* **Key Features:**
  1. **ICU admission** (strongest predictor)
  2. **Pneumonia status**
  3. **Age (scaled)**
  4. **Chronic renal disease**

**Comparison with Other Models**

| **Model** | **Accuracy** | **F1-Score** | **ROC-AUC** |
| --- | --- | --- | --- |
| **LightGBM** | 82.27% | 0.82 | 0.915 |
| **XGBoost** | 80.44% | 0.80 | - |
| **Random Forest** | 80.12% | 0.80 | - |
| **Logistic Reg.** | 74.24% | 0.74 | - |

**3.2 Mortality Risk Prediction**

**Best Model: LightGBM**

* **Test Accuracy:** **81.94%**
* **F1-Score:** **0.82**
* **ROC-AUC:** **0.925**
* **Confusion Matrix (Test Set):**

A blue squares with white text

AI-generated content may be incorrect.

A blue squares with white text

AI-generated content may be incorrect.

|  | **Predicted: Alive** | **Predicted: Dead** |
| --- | --- | --- |
| **Actual: Alive** | 14,793 (74%) | 5,199 (26%) |
| **Actual: Dead** | 1,999 (10%) | 17,992 (90%) |

* **Key Features:**
  1. **ICU admission**
  2. **Intubation status**
  3. **Age (scaled)**
  4. **Pneumonia**

**Comparison with Other Models**

| **Model** | **Accuracy** | **F1-Score** | **ROC-AUC** |
| --- | --- | --- | --- |
| **LightGBM** | 81.94% | 0.82 | 0.925 |
| **XGBoost** | 75.52% | 0.76 | - |
| **Random Forest** | 79.85% | 0.80 | - |
| **Logistic Reg.** | 74.48% | 0.74 | - |

**4. Challenges & Limitations**

1. **Class Imbalance:** SMOTE improved performance but may introduce synthetic bias.
2. **Overfitting in Patient Type Model:** Near-perfect scores suggest potential data leakage or insufficient complexity.
3. **Feature Interpretability:** Some one-hot encoded features may need grouping for better clinical relevance.
4. **Feature Correlation**: ICU and intubation were top predictors for mortality, suggesting an overlap in severe cases.
5. **Model Trade-offs:** LightGBM outperformed others but is less interpretable than logistic regression.

**6. Conclusion**

* **ICU Model:** Performs well (~86% F1) but could benefit from further tuning.
* **Patient Type Model:** Near-perfect accuracy (potential overfitting; needs investigation).
* **Actionable Insights:**
  + **Age, Pneumonia, Diabetes** are top ICU risk factors.
  + **Mortality (DIED\_1)** strongly correlates with inpatient status.
* **Intubation Model:** 82% accuracy with strong recall for critical cases.
* **Mortality Model:** 82% accuracy, 90% recall for deaths (critical for early intervention).