Documentation

Dataset Overview

- Source: COVID19 Pediatric Dataset
- Format: Excel (.xlsx)
- Size: 5644 rows × 111 columns
- Content: Patient demographics, medical test results, COVID-19 outcomes

Data Processing Pipeline

1. Data Cleaning

- Removed columns with all missing values
- Imputed partial missing values with column means
- Applied zero imputation for columns with NaN means

2. Feature Engineering

- Selected numeric features only
- Applied standardization (zero mean, unit variance)
- Handled negatively skewed distributions through normalization

3. Model Implementations

A. Logistic Regression

- **Objective**: Predict SARS-CoV-2 test results
- **Features**: 18 key medical parameters including:
  - Patient age
  - Hospital admission status
  - Blood test results (Hematocrit, Hemoglobin, etc.)
  - Blood cell counts
- **Preprocessing**:
  - Removed features with high missing values
  - Standardized numeric features
  - Encoded categorical variables

B. Random Forest Classification

- **Objective**: Predict COVID-19 test outcomes
- **Implementation**:
  - Used RandomForestClassifier from sklearn
  - Train-test split for model validation
  - Feature importance analysis
- **Key Features**:
  - Blood markers and cell counts

    - Patient demographics
    - Hospital admission data
- **Performance Metrics**:
    - Confusion matrix analysis
    - Classification report with precision, recall
    - Feature importance ranking

C. K-Means Clustering

- **Method**: k=3 clusters (based on elbow method)
- **Features**: Standardized numeric variables
- **Cluster Characteristics**:
    1. Young patients with elevated blood markers
        - High: Blood gas levels, platelets
        - Low: Age, hemoglobin levels
    2. Patients with distinct blood gas profiles
        - High: MCH, lactic acid, calcium
        - Low: pH levels, oxygen saturation
    3. Older patients with elevated blood counts
        - High: Blood cell counts, age
        - Low: Platelets, blood gas levels

D. Principal Component Analysis (PCA)

- **Purpose**: Dimensionality reduction and feature analysis
- **Implementation**:
    - Applied to standardized features
    - Analyzed variance explained by components
    - Visualized feature relationships
- **Results**:
    - Identified key contributing features
    - Revealed patterns in blood marker correlations
    - Reduced feature space while preserving information

4. Visualization

- Generated plots for:
    - Cluster relationships and distributions
    - Feature correlations (heatmaps)
    - Model performance metrics
    - PCA component analysis
    - Feature importance in Random Forest

5. Model Performance

- Successfully identified distinct patient groups
- Achieved balanced prediction accuracy
- Revealed important feature relationships

- Provided interpretable medical parameter groupings

6. Future Improvements

- Implement ensemble methods combining multiple models
- Explore feature selection techniques
- Add cross-validation for robust evaluation
- Consider deep learning approaches
- Investigate temporal patterns in patient data