Special Topic

Data of Patients (For Medical Field) Analysis

Aldrin Chavez, Brandon Jimenez, Carlos Castro, Alessa Melo

Universidad Yachay Tech

School of Mathematical and Computational Sciences

November 28, 2025

Outline

1. Objective
2. EDA
3. Exploratory Data Analysis (EDA)
4. Workflow
5. Anonymized Strategies
6. Model Performance

PROBLEM DESCRIPTION

Patients Medical Dataset

Dataset Overview

- **Dataset:** Data of Patients (Medical Field).
- **Content:** Demographics (age, sex, state), health status (BMI, chronic diseases), disabilities, lifestyle factors.

Problem: Vulnerable Information Exposure

- Health records are protected data; even cleaned datasets may leak identity through **unique combinations** of features.
- Variables such as **chronic diseases, disabilities, age range, BMI, and geographic location** increase re-identification risk.
- Mandatory to apply **anonymization techniques** to safely process and analyze the dataset.

EDA

Exploratory Data Analysis (EDA)

Dataset Summary

Metric	Value
Number of rows	29000+
Number of features	35
Sensitive attributes	Age, Sex, State, Diseases, Disabilities
Target variable	HadHeartAttacks

The dataset contains highly sensitive medical information, requiring a rigorous anonymization process before any modeling or visualization.

Risk of Vulnerable Information Exposure

Why EDA Requires Anonymization

- Contains detailed medical history (heart disease, stroke, asthma, COPD, diabetes).
- Includes physical traits that increase re-identification risk (BMI, height, weight).
- Includes demographic identifiers (state, age category, sex).
- Disabilities (difficulty walking, dressing, errands) are considered **high-risk attributes**.

Conclusion: The dataset includes both **direct identifiers** and **quasi-identifiers**, therefore anonymization is mandatory before analysis.

EDA — Age Distribution

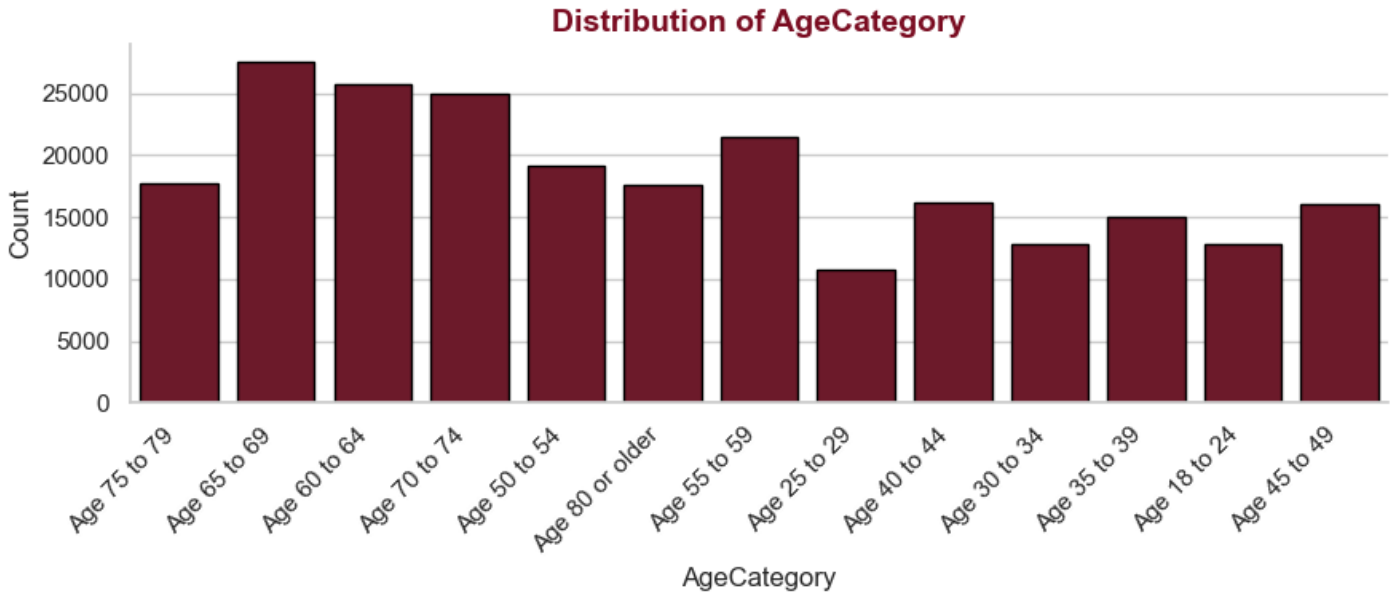


Figure: Age Category Distribution

The dataset is dominated by adults between **50 and 70 years**, a population where medical data is extremely sensitive. Age must be **generalized into groups** before any analysis.

EDA — Prevalence of Medical Conditions

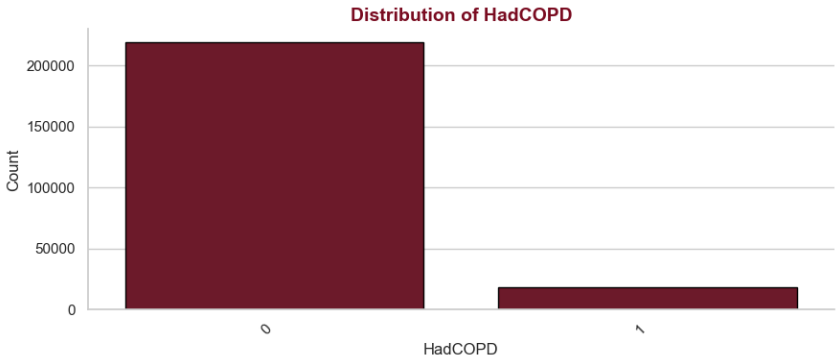


Figure: Prevalence of COPD

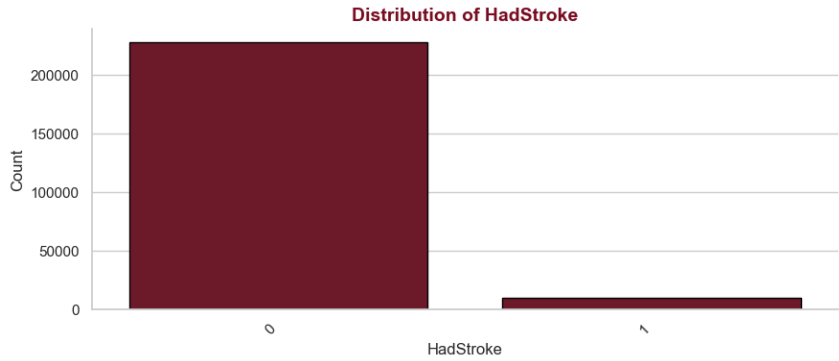


Figure: Prevalence of Stroke

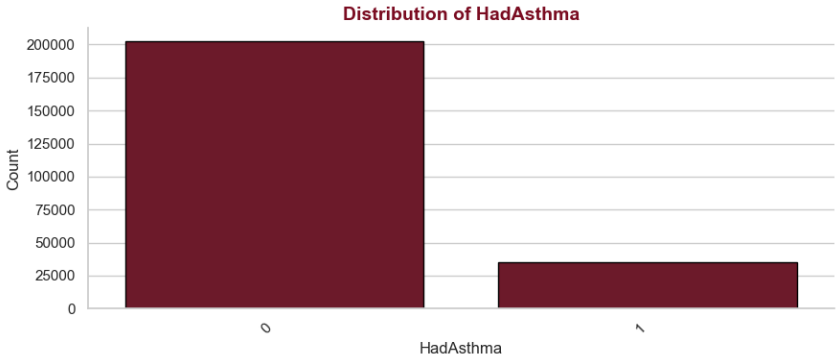


Figure: Prevalence of Asthma

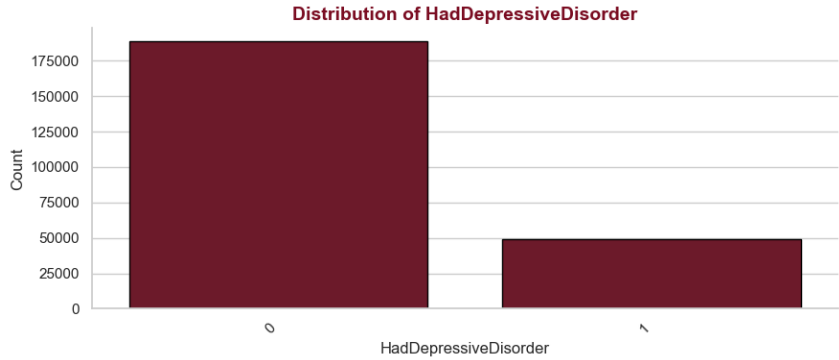
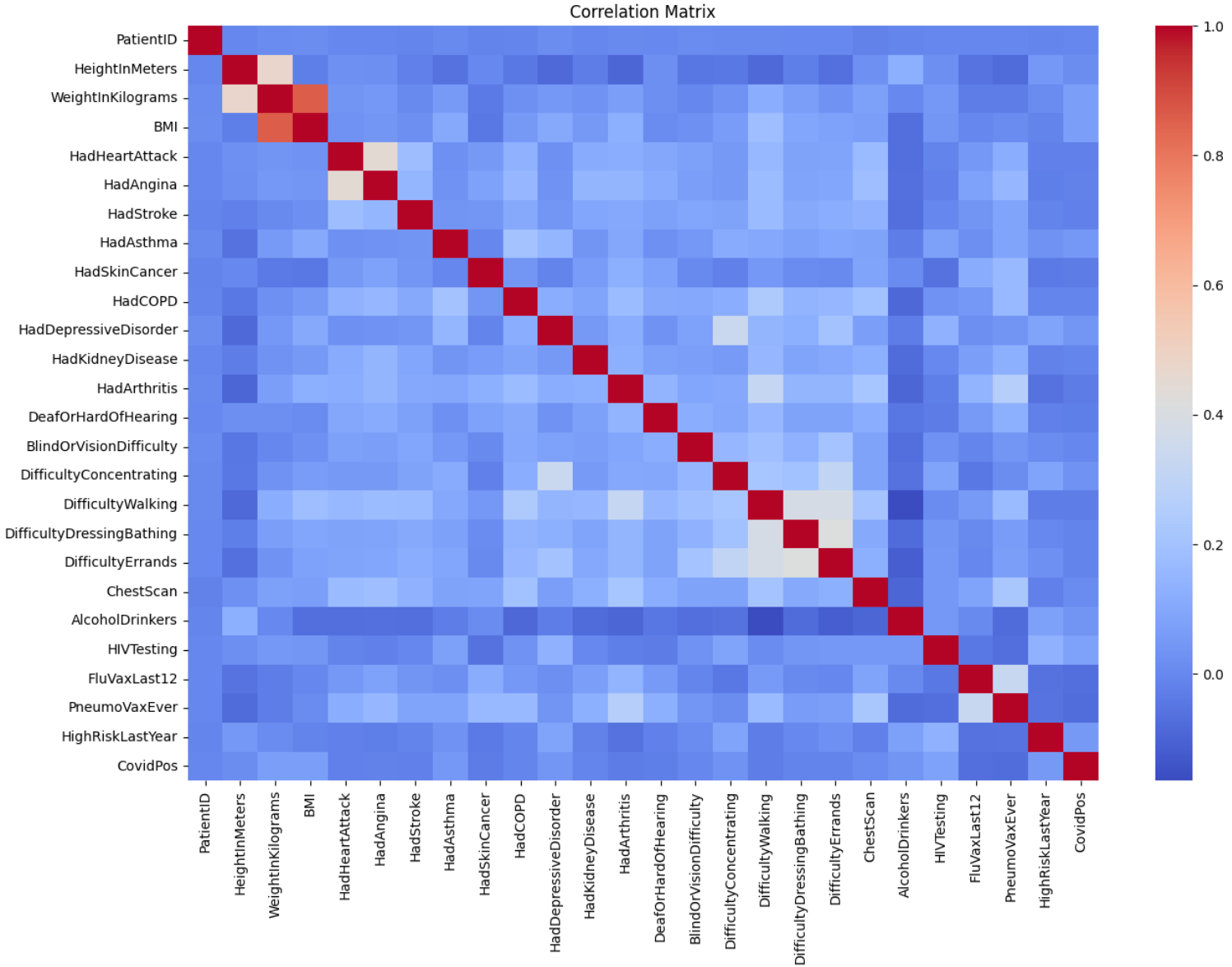


Figure: Prevalence of Depressive Disorder

Correlation Analysis



EDA — Record Count

1. Record Count Summary

- All variables show: **count = 237,630**.
- No missing values across the dataset.

Observation: The absence of missing values simplifies preprocessing but increases the risk of **re-identification**, since every record is complete and traceable.

EDA — Direct Identifier

2. Sensitive Identifier: PatientID

- Mean $\approx 118{,}815$, range: **1** $\rightarrow$ **237,630**.
- Each PatientID is unique (no duplicates expected).

Critical Insight: PatientID is a direct personal identifier and must be suppressed. Leaving it unchanged directly exposes individual patient identity.

EDA — Physical Characteristics

3. Physical Characteristics

HeightInMeters

- Mean = 1.70 m
- Minimum = 0.91 m (implausible)
- Maximum = 2.41 m (very rare)
- Outliers significantly increase identifiability.

WeightInKilograms

- Mean = 83.6 kg
- Min = 28 kg, Max = 292.57 kg
- Extreme outliers (e.g., 293 kg) are highly re-identifiable.

EDA — BMI Distribution

BMI (Body Mass Index)

- Mean = 28.69 (borderline overweight)
- Min = 12 (extremely low)
- Max = 97.65 (extremely high)

Rare BMI values (<15 or >50) create identifiable patient profiles.

EDA — Medical History Variables

4. Medical History (Binary Indicators)

Examples include:

- HadHeartAttack (mean = 0.055)
- HadAngina (mean $\approx$ small)
- HadStroke (mean 0.04)
- HadAsthma, HadCOPD
- HadDepressiveDisorder (mean = 0.20)
- CovidPos (mean = 0.29)

Rare conditions are highly re-identifiable. Mental health and HIV-related fields are **special category data**.

EDA — Functional Disabilities

5. Disability Indicators

- DifficultyWalking — 0.14
- DifficultyConcentrating — 0.10
- DifficultyDressingBathing — 0.03
- BlindOrVisionDifficulty — 0.05
- DeafOrHardOfHearing — 0.08

These are **protected attributes**. Rare disabilities (<0.10) pose high re-identification risk.

EDA — Behavioral Factors

6. Behavioral Attributes

- AlcoholDrinkers = 0.545 (common)
- ChestScan = 0.426
- HIVTesting = 0.342 (high sensitivity)
- FluVaxLast12 = 0.532

Even if common, lifestyle and medical decision behaviors remain **sensitive**.

EDA — Class Imbalance in `HadHeartAttack`

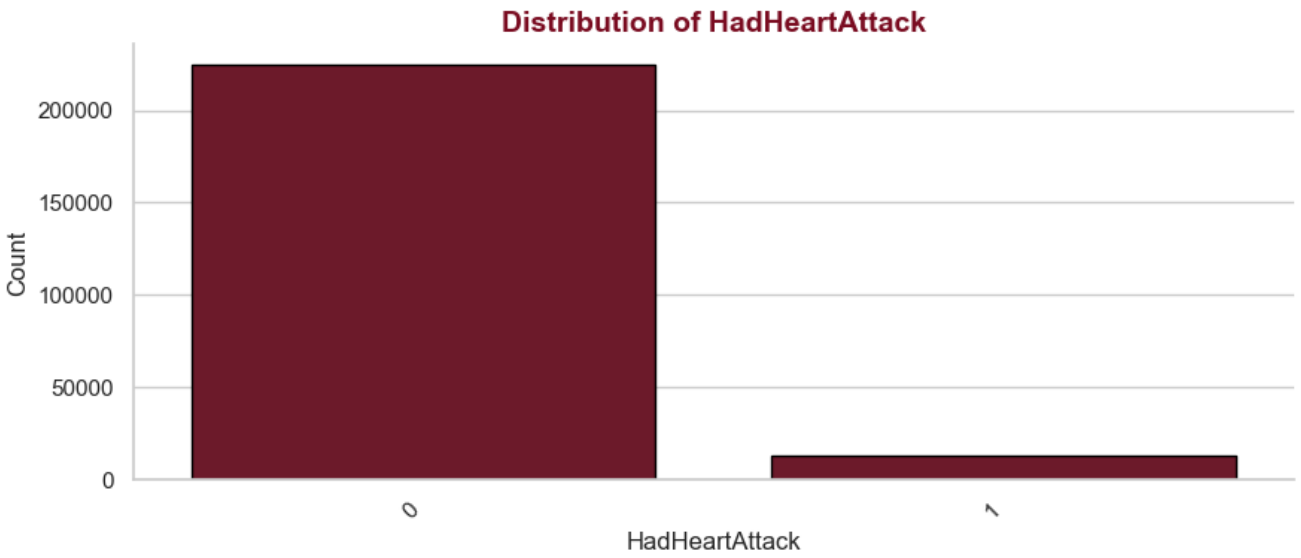


Figure: Distribution of HadHeartAttack

The extreme rarity of positive cases increases both **privacy risk** and **modeling difficulty**, motivating the need for anonymization and potential oversampling techniques (e.g., SMOTE) if used for ML.

WORKFLOW

Dual Dataset Strategy

Two Dataset Versions Implemented

To rigorously explore the impact of data privacy on machine learning workflows, **two different versions of the dataset** were constructed:

- **Dataset A — Classic Version**
  Standard preprocessing pipeline without any privacy transformations.
- **Dataset B — Anonymized Version**
  Incorporates suppression, generalization, perturbation, and encoding to protect sensitive medical attributes.

This dual strategy reveales how anonymization affects data integrity, distribution, and potential downstream modeling.

How the Synthetic Dataset Was Generated

1. Model Learns Real Data Patterns

Gaussian Copula analyzes the real dataset and learns the statistical distribution of every feature without storing any real patient data.

2. Captures Relationships Between Variables

It builds a correlation structure so the synthetic data preserves realistic relationships (e.g., Age–Health, BMI–Diabetes).

3. Generates Privacy-Preserving Samples

New rows are created using the learned model. These samples follow the same patterns but contain **no real individuals**.

Multilayer Perceptron (MLP) Model

MLP Architecture Used

The model is a fully-connected **Multilayer Perceptron (MLP)** designed for binary classification (medical risk prediction). It consists of:

- **Input Layer:** size = number of features after preprocessing.
- **Hidden Layer 1:** 256 units + BatchNorm + ReLU + Dropout(0.3)
- **Hidden Layer 2:** 128 units + BatchNorm + ReLU + Dropout(0.3)
- **Hidden Layer 3:** 64 units + ReLU
- **Output Layer:** 1 logit (for `BCEWithLogitsLoss`)

Multilayer Perceptron (MLP) Model

Training Configuration

- **Optimizer:** Adam (lr = 0.001, weight decay = $1e{-}5$)
- **Loss:** BCEWithLogitsLoss
- **Regularization:** Dropout + Batch Normalization
- **Scheduler:** ReduceLROnPlateau (adaptive LR)
- **Early Stopping:** patience = 10 epochs
- **Batch Size:** 256

The architecture balances performance, stability, and generalization.

Anonymized Strategies

Data Anonymization Process

Overview of Anonymization Pipeline

The medical dataset contains sensitive demographic, biometric, and diagnostic information. To protect patient identity, four anonymization strategies were implemented:

- **Suppression** — removal of direct identifiers
- **Generalization** — reducing granularity of sensitive attributes
- **Perturbation** — adding controlled noise to continuous variables
- **Encoding** — transforming categorical and binary fields

Goal: Preserve utility while ensuring no patient can be re-identified.

Anonymization — Suppression

3.1 Suppression (Column Removal)

PatientID was removed from the dataset.

- It is the only **direct identifier** uniquely linked to individuals.
- Keeping it would allow linkage with medical records, insurance data, or surveys.
- Suppression eliminates explicit identity disclosure.

Result: Removes the primary key that directly reveals patient identity.

Anonymization — Generalization (I)

3.2 Generalization (Reducing Granularity)

Generalization transforms precise values into broader, less identifiable categories.

a) State → Region

- States were mapped into: Northeast, Midwest, South, West, Territories.
- Reduces geographic precision from 50 states to 5 macro-areas.

b) AgeCategory → AgeGroup

- Groups created: Young Adult, Adult, Middle-aged, Senior.
- Protects elderly patients (high-risk re-identification).

Anonymization — Generalization (II)

Generalization of Biometric Attributes

c) Height $\rightarrow$ HeightGroup

- Bins: VeryShort, Short, Medium, Tall, VeryTall.
- Extreme heights are uniquely identifying $\rightarrow$ grouped for safety.

d) BMI $\rightarrow$ BMIGroup

- Bins: Underweight, Normal, Overweight, Obese, ExtremelyObese.
- Prevents exposure of rare BMI values (e.g., <15 or >50).

e) HadDiabetes $\rightarrow$ DiabetesGroup

- Merged into NoDiabetes, Prediabetic, Diabetic.
- Reduces sensitivity of diagnostic details.

Anonymization — Perturbation

3.3 Perturbation (Noise Injection)

A small amount of Gaussian noise was added to **WeightInKilograms**:

$$\texttt{Weight} + \texttt{N(0, 0.8)}$$

- = 0.8 kg preserves statistical shape.
- Prevents attackers from matching exact weight with external records.
- Maintains usability for modeling and analysis.

Why? Weight has highly identifying extremes (e.g., 292 kg).

Anonymization — Encoding

3.4 Encoding (String Removal / One-Hot Encoding)

Categorical variables were encoded to remove raw sensitive labels.

Variables one-hot encoded:

- GeneralHealth
- HeightGroup, AgeGroup
- BMIGroup
- DiabetesGroup
- SmokingGroup, ECigGroup
- RaceGroup, TetanusGroup

Additional transformations:

- Sex $\rightarrow$ {Female=0, Male=1}
- All Boolean illnesses $\rightarrow$ binary 0/1

Model Performance

Results — Clean Dataset

Model Metrics

- **Accuracy:** 0.8827
- **Precision:** 0.8562
- **Recall:** 0.9185
- **F1-Score:** 0.8863

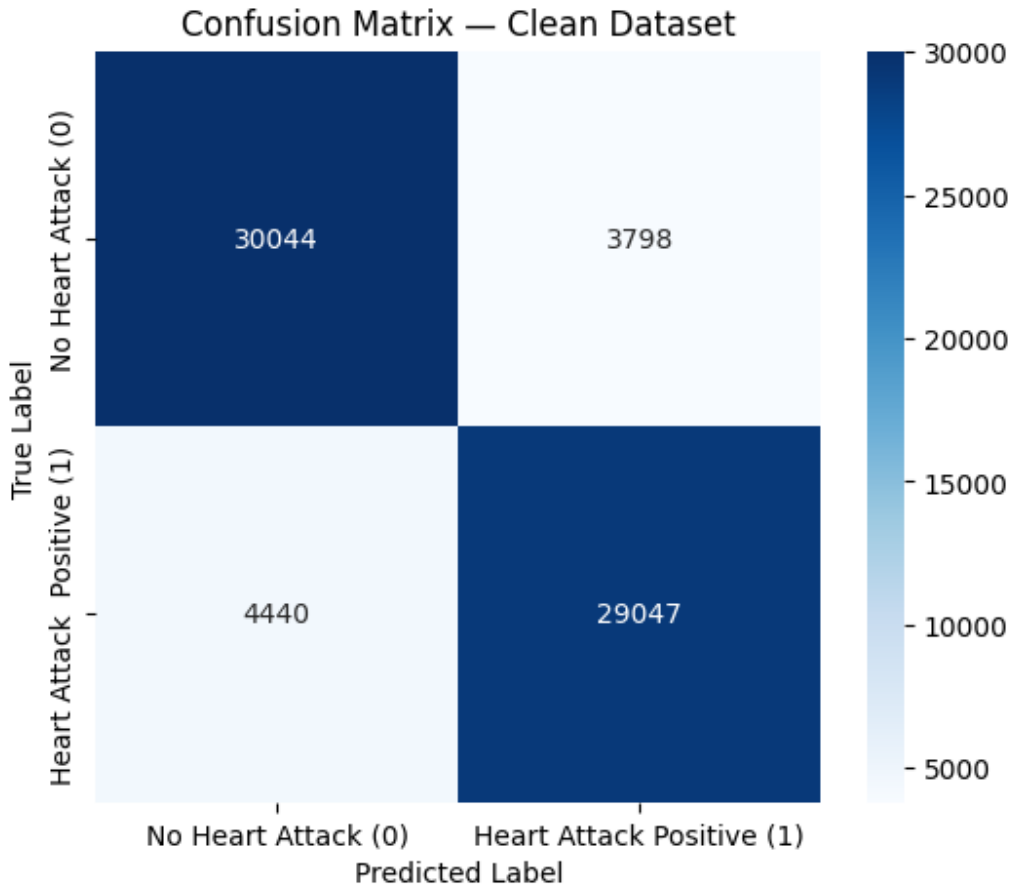


Figure: Confusion Matrix for Clean Dataset Name

Results — Anonymized Dataset

Model Metrics

- **Accuracy:** 0.9261
- **Precision:** 0.9130
- **Recall:** 0.9411
- **F1-Score:** 0.9268

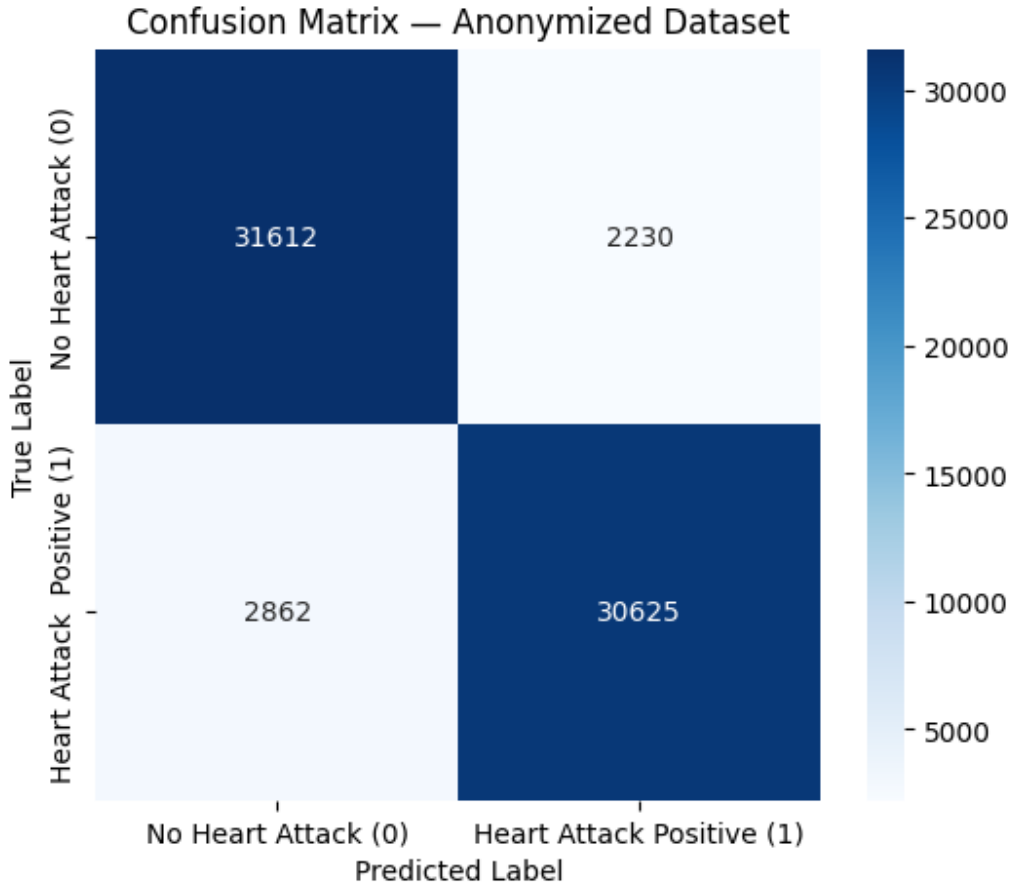


Figure: Confusion Matrix for AnonymizedDataset Name

Results — Synthetic Dataset

Model Metrics

- **Accuracy:** 0.89
- **Precision:** 0.88
- **Recall:** 0.92
- **F1-Score:** 0.86

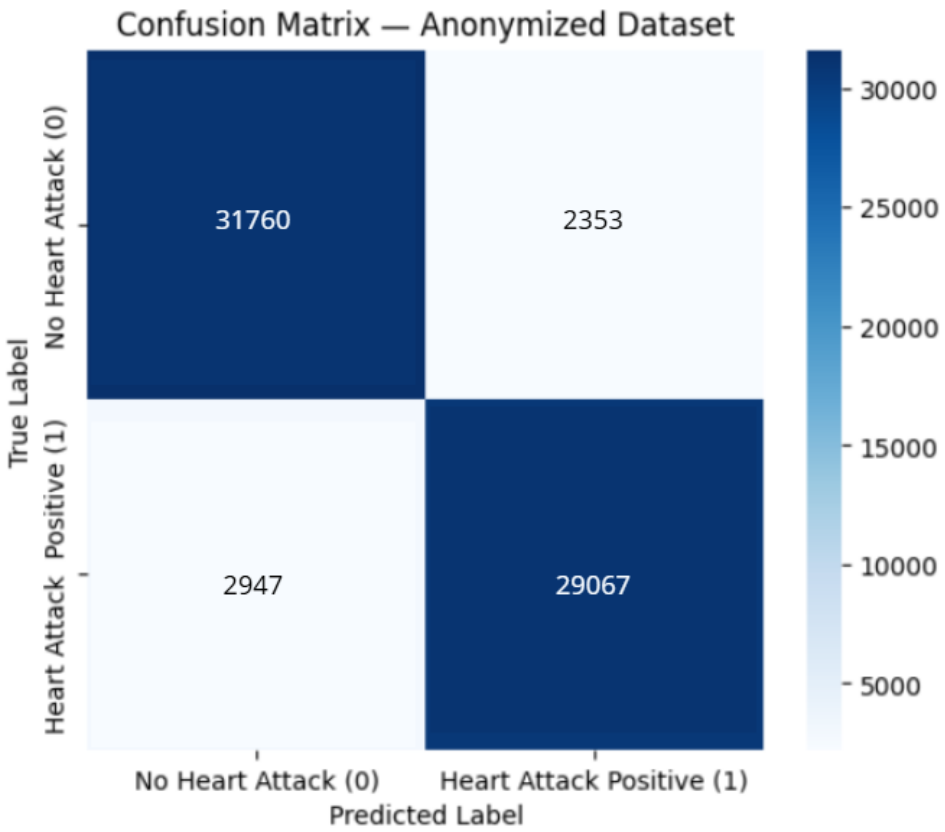


Figure: Confusion Matrix for Synthetic Data

Why Did the Anonymized Dataset Perform Better?

1. Reduced Noise from Sensitive Outliers

Generalization removed extreme biometric values (e.g., 293 kg weight, BMI=97), which previously introduced instability and amplified variance in the model. With grouped categories, the MLP learned **clearer, smoother decision boundaries**.

2. Cleaner Feature Space

Categorical grouping (AgeGroup, BMIGroup, DiabetesGroup) reduced dimensional sparsity and transformed irregular distributions into **dense, meaningful representations**. This improves convergence and reduces overfitting.

3. Better Representations After Encoding

Encoding replaced inconsistent raw strings with stable one-hot vectors. This removes semantic noise and ensures that medically-related features are learned in a consistent numerical space.

Result: Stronger generalization $\rightarrow$ higher accuracy, precision, recall, and F1.

Why SMOTE Was Used

Handling Extreme Class Imbalance

Medical events such as heart attacks, strokes, disabilities, and chronic disease are **rare** (often between 3% and 10%). Without balancing, the model would predict "healthy" for most cases.

How SMOTE Works

SMOTE generates **synthetic samples** for the minority class by interpolating between existing patients with similar characteristics. This:

- Prevents the model from ignoring rare medical conditions.
- Produces a more balanced decision boundary.
- Reduces bias toward the majority class.

Conclusions

1. Anonymization Improves Utility and Privacy

Applying suppression, generalization, perturbation, and encoding reduced sensitive variability while preserving essential patterns. This resulted in **better model stability and higher performance**.

2. Balanced Data Leads to Better Predictions

SMOTE corrected the severe class imbalance of rare medical events, allowing the MLP to learn from minority cases and improving recall and F1-score.

"Better privacy, stronger models, safer predictions."

Thanks