Heart Disease Prediction

Logine Rashed

Habiba Ayman

Arwa Elewa

Seifeldein gamal

Mohamed Ayoub

Ahmed

Heart Disease Prediction Project

This project outlines the development of an end-to-end machine learning pipeline designed to predict heart disease risk from clinical data. We cover everything from initial data analysis to model deployment and continuous monitoring.

Data Analysis

In-depth exploration of clinical features.

Modeling

Developing robust predictive algorithms.

Deployment

Bringing the model to a production environment.

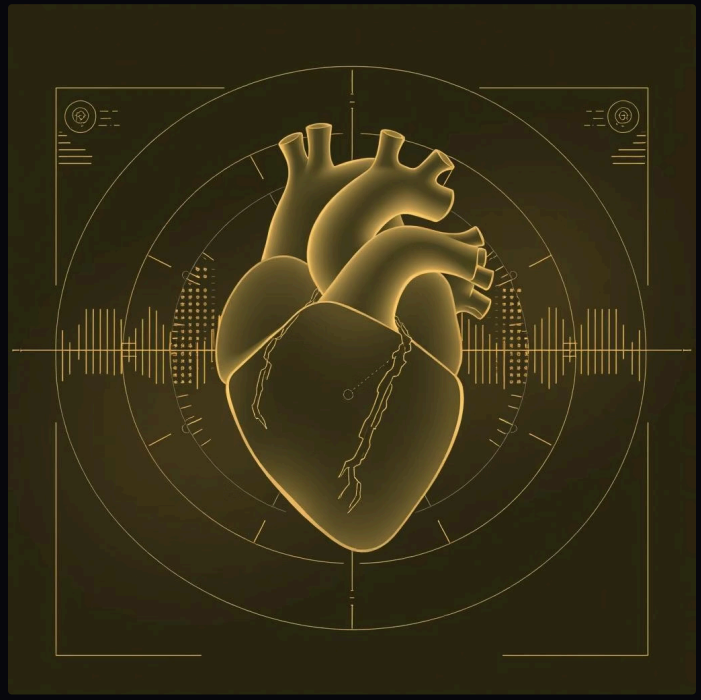
Monitoring

Ensuring continuous performance and reliability.

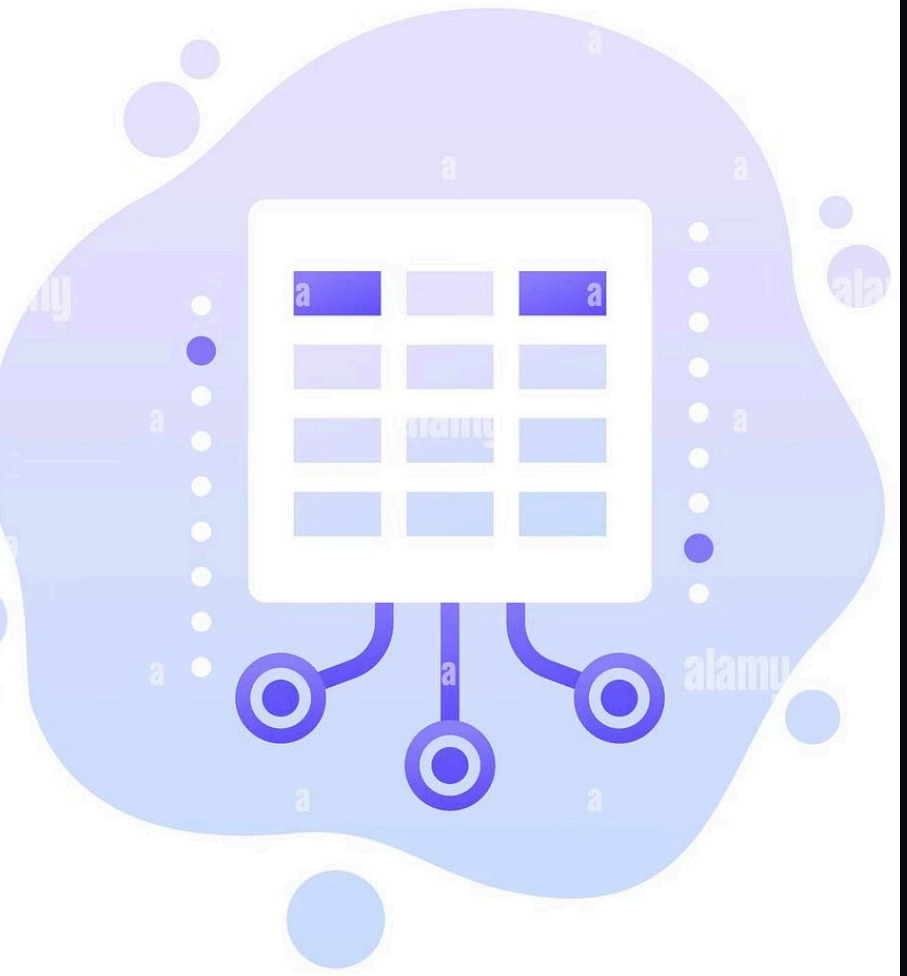
Problem & Dataset Overview

Our objective is a binary classification: predicting the presence (1) or absence (0) of heart disease. The dataset comprises critical clinical features such as:

- **Age:** Patient's age in years.
- **Sex:** Male or female.
- **Chest Pain Type:** Categorization of chest pain.
- **Blood Pressure:** Resting blood pressure.
- **Cholesterol:** Serum cholesterol levels.
- **ECG:** Resting electrocardiographic results.
- **Heart Rate:** Maximum heart rate achieved.
- **Exercise Angina:** Exercise-induced angina.
- **ST Depression:** ST depression induced by exercise relative to rest.
- **Vessels:** Number of major vessels colored by fluoroscopy.
- **Thalassemia:** Thalassemia type.



Before analysis, the dataset underwent thorough checks for size, data types, missing values, and duplicate records to ensure data quality and integrity.



Milestone 1 – Data Loading & Cleaning

01

Load Raw Data

The initial step involves loading the raw CSV dataset into a structured table format, ready for manipulation.

02

Inspect & Summarize

We inspect the first few rows, confirm data types, and generate basic statistical summaries to understand the data's characteristics.

03

Rename Columns

Technical column names are standardized to more intuitive and human-readable labels for clarity and ease of use.

04

Verify Cleanliness

A final verification ensures the absence of any missing values or duplicated rows, guaranteeing a clean dataset for further analysis.

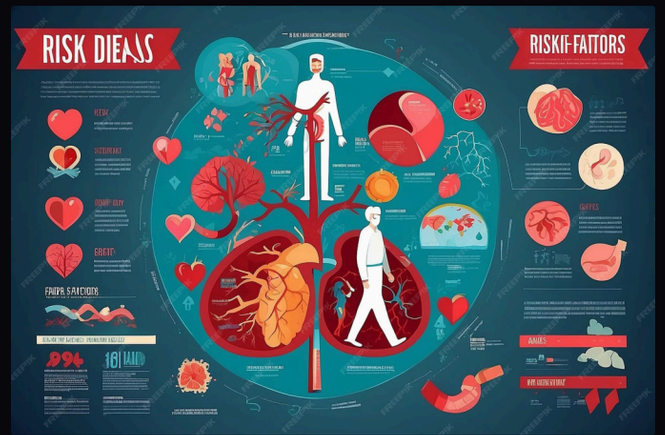
Milestone 2 – Exploratory Data Analysis (EDA)

Deep Dive into Distributions

We meticulously examine the distributions of critical variables such as age, cholesterol, and blood pressure. This helps in understanding their inherent patterns and ranges within the patient population.

Risk Categorization

To gain segmented insights, we categorize patients into age groups (young, middle, old) and define cholesterol risk categories (desirable, borderline, high). This allows for a granular analysis of risk factors.



Visualizing Relationships

Key relationships are visualized to identify potential correlations. This includes:

- Sex vs. Heart Disease
- Chest Pain Type vs. Heart Disease
- Age Group vs. Heart Disease

These visualizations provide immediate insights into which demographic and clinical factors are most strongly associated with heart disease.

Milestone 3 – Correlation & Feature Insights

Quantifying Relationships

A correlation matrix is computed to quantitatively assess the relationships between numerical features and the target variable (heart disease). This matrix highlights positive and negative correlations.

Identifying Key Predictors

Features showing strong correlation with heart disease are identified as key predictors. This helps in prioritizing the most impactful variables for model development.

Guiding Questions & Answers

- **Who is at higher risk?** By analyzing correlations, we determine which demographic groups or clinical profiles are statistically more prone to heart disease.
- **Which factors matter most?** The strength of correlations informs us about the most significant contributing factors to heart disease risk, guiding subsequent feature engineering and model focus.

Milestone 4 – Preprocessing & Feature Engineering

Data Split

The dataset is split into 80% training and 20% testing sets using stratified sampling to maintain class proportions.

Standardization

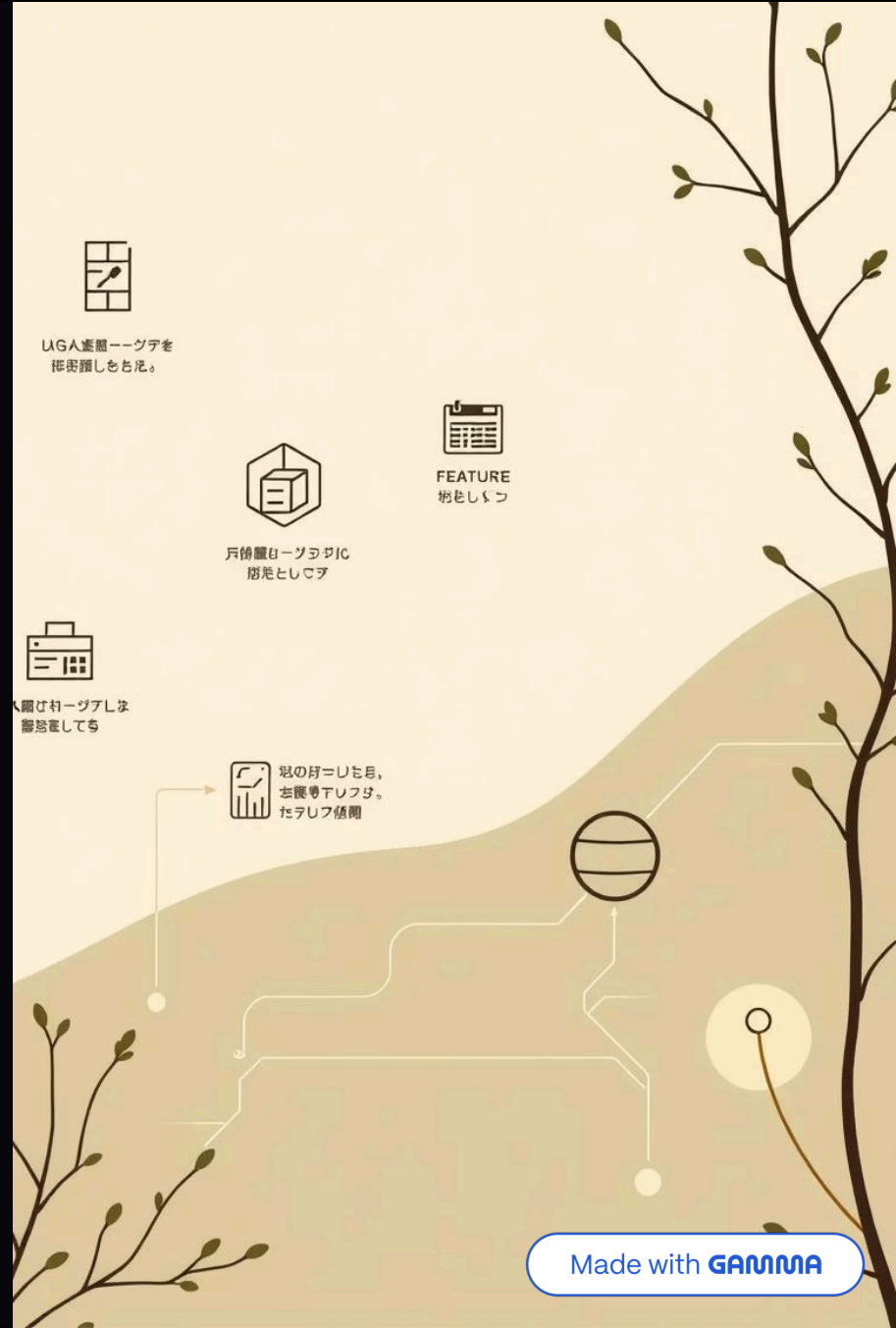
Features are standardized to ensure all variables contribute equally to the model, preventing dominance by larger-scale features.

Class Imbalance

SMOTE (Synthetic Minority Over-sampling Technique) is applied to generate synthetic samples for the minority class, addressing class imbalance.

Dimensionality Reduction

Principal Component Analysis (PCA) is used to reduce the number of features while preserving the majority of the dataset's variance.



Milestone 5 – Model Training & Comparison

Model Selection

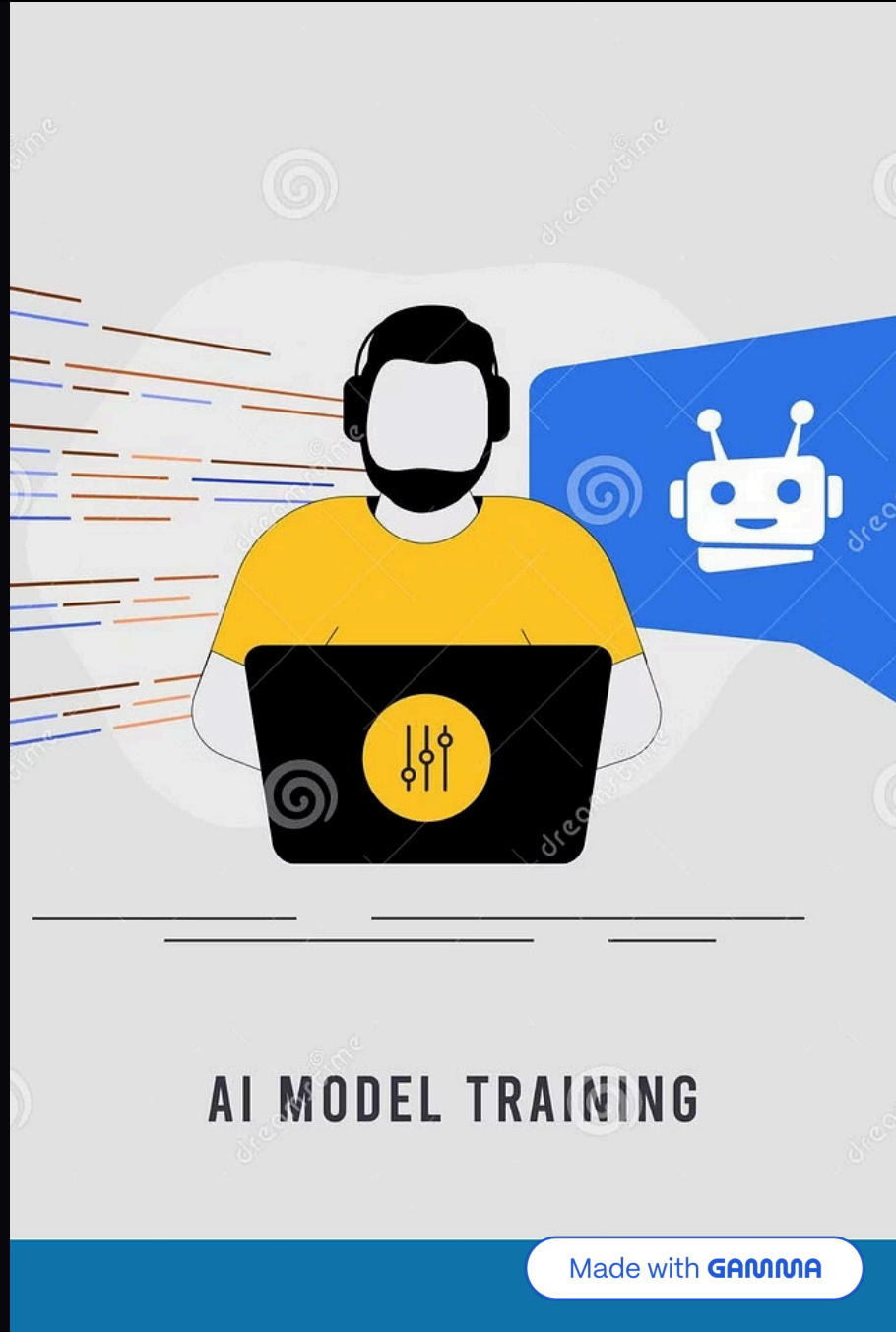
We trained a diverse set of machine learning models including Support Vector Machine (SVM), Random Forest, XGBoost, and a Multi-Layer Perceptron (MLP) Neural Network.

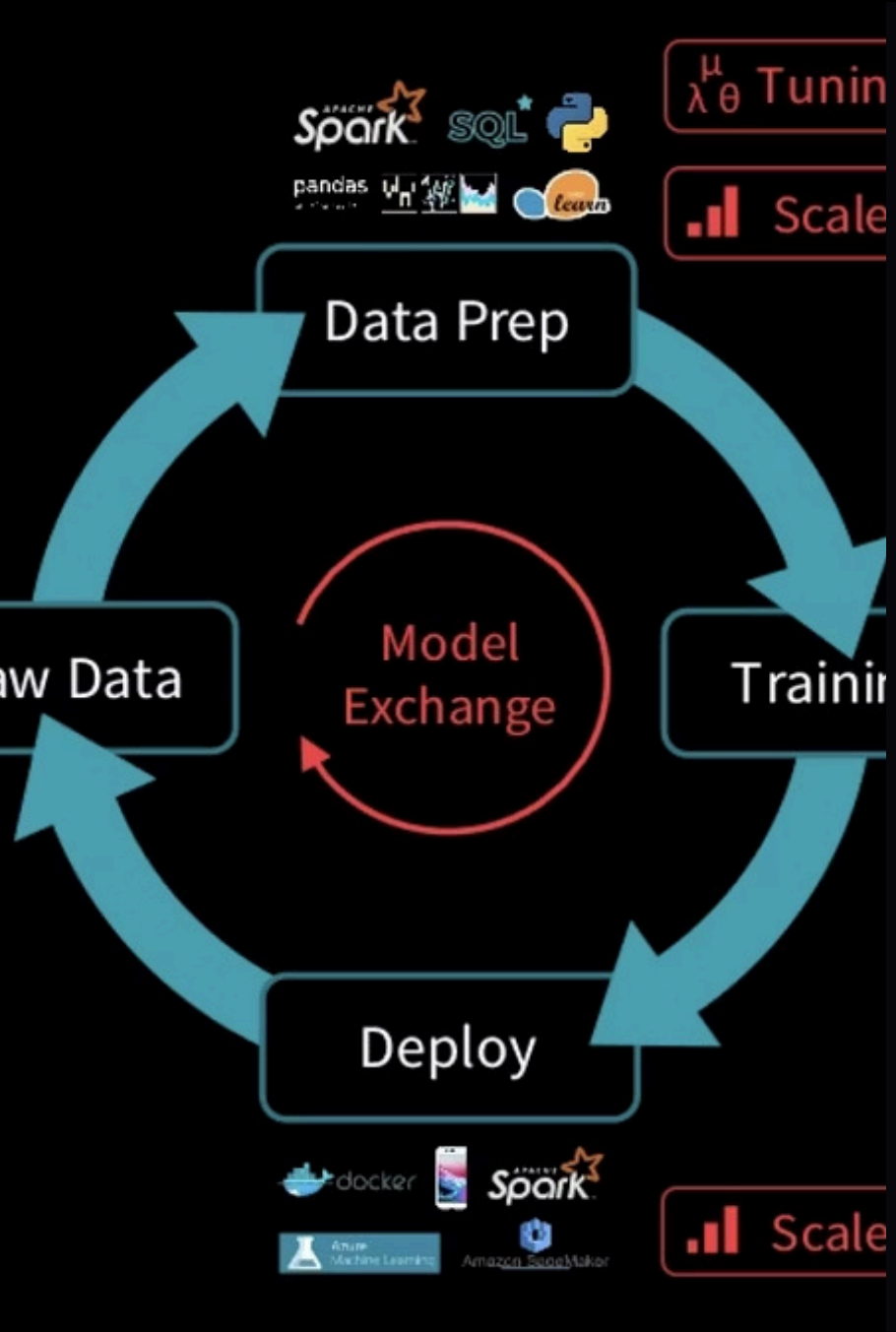
Performance Evaluation

Each model's performance was rigorously evaluated using key metrics: accuracy, precision, recall, F1-score, and ROC-AUC to provide a comprehensive understanding of their effectiveness.

Model Selection & Analysis

A detailed comparison table helps in selecting the best-performing model. The confusion matrix of this model is analyzed to understand its specific error patterns.





Milestone 6 – Hyperparameter Tuning & Experiment Tracking

1 — **GridSearch & Cross-Validation**

Optimal hyperparameters for candidate models are identified using GridSearch combined with cross-validation to ensure robust performance.

2 — **F1-Score Optimization**

The primary optimization metric is the F1-score, chosen to achieve a balanced performance between precision and recall, crucial for clinical predictions.

3 — **MLflow Logging**

MLflow is utilized to log all experiment details: model configurations, performance metrics, and the final selected model, ensuring full traceability.

4 — **Reproducible Records**

This systematic tracking creates a reproducible record of all experiments, facilitating collaboration and future model improvements.

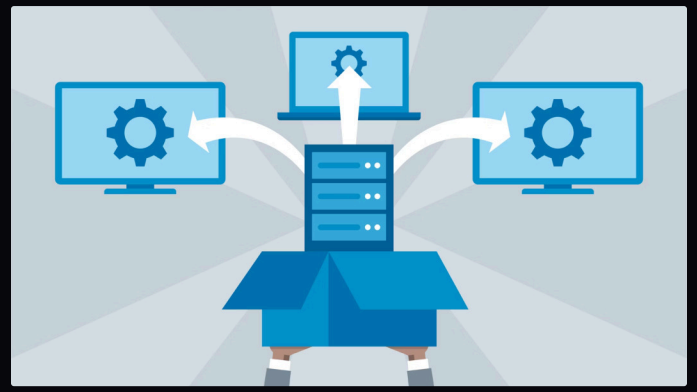
Milestone 7 – Final Pipeline & Deployment

Unified Predictive Pipeline

A unified machine learning pipeline is constructed, integrating the best preprocessing steps (scaling, PCA) with the chosen classifier. This streamlined pipeline ensures consistent and efficient predictions.

Model Persistence

The final pipeline is trained on the entire available dataset to maximize its learning potential and then saved as a persistent model file for future use and deployment.



API Exposure

The trained model is exposed as a REST API using Flask. A dedicated `/predict` endpoint allows external applications to submit patient data and receive predictions.

Interactive Web Application

A simple, user-friendly Streamlit web application is developed. This app provides an intuitive interface for healthcare professionals to input patient data and instantly view heart disease risk predictions.

Milestone 8 – Monitoring & Model Drift

Prediction Logging

Every model prediction, including its timestamp, input features, predicted outcome, and probability, is meticulously logged to a CSV file for auditing and analysis.

Monitoring Reports

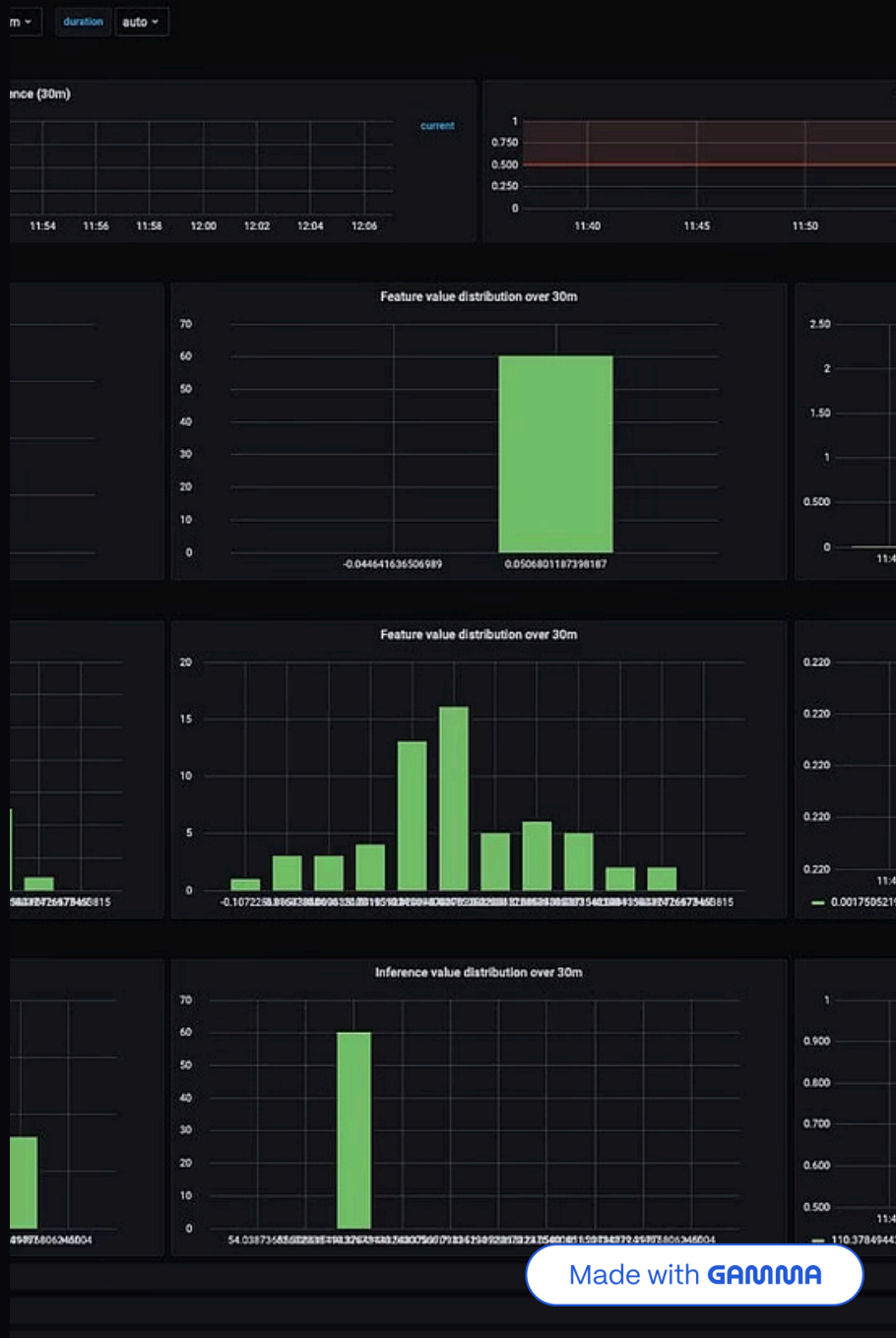
Automated reports are generated to visualize the distribution of predicted risk probabilities over time, providing a high-level overview of model behavior.

Data Drift Detection

Statistical methods, such as tracking the standard deviation of predicted probabilities, are employed to detect any signs of data drift, indicating changes in the input data distribution.

Retraining Triggers

Insights from monitoring reports and drift detection serve as crucial indicators to determine when the model requires retraining to maintain its accuracy and relevance.



Thank You