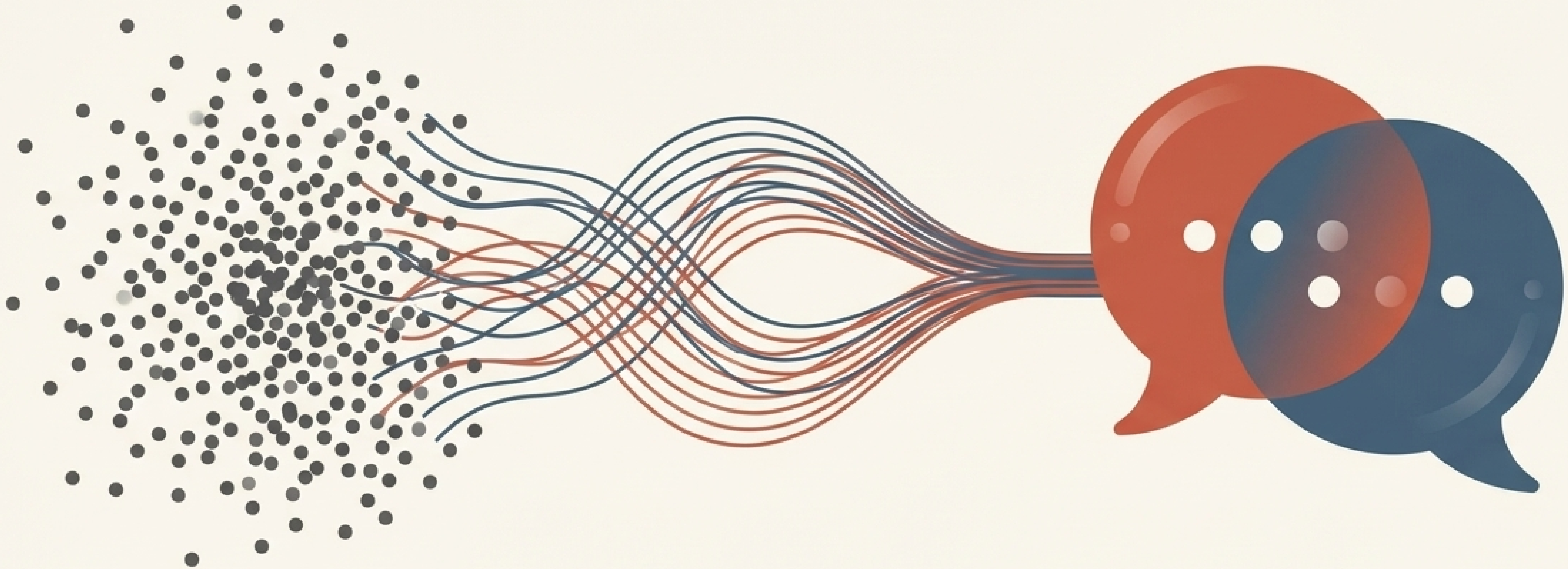


From Data to Dialogue: Predicting the Need for Mental Health Treatment with Machine Learning

A technical deep-dive into our end-to-end data science solution.

Team Members: Yahya Mustafa, Adham Fouad, Abdelkader, Ahmed Fahmy, Muhanad Mustafa, Marwan Gaber

An Intelligent System for Early Identification

The Purpose

To develop a system that predicts whether an individual may benefit from mental health treatment based on a comprehensive questionnaire.

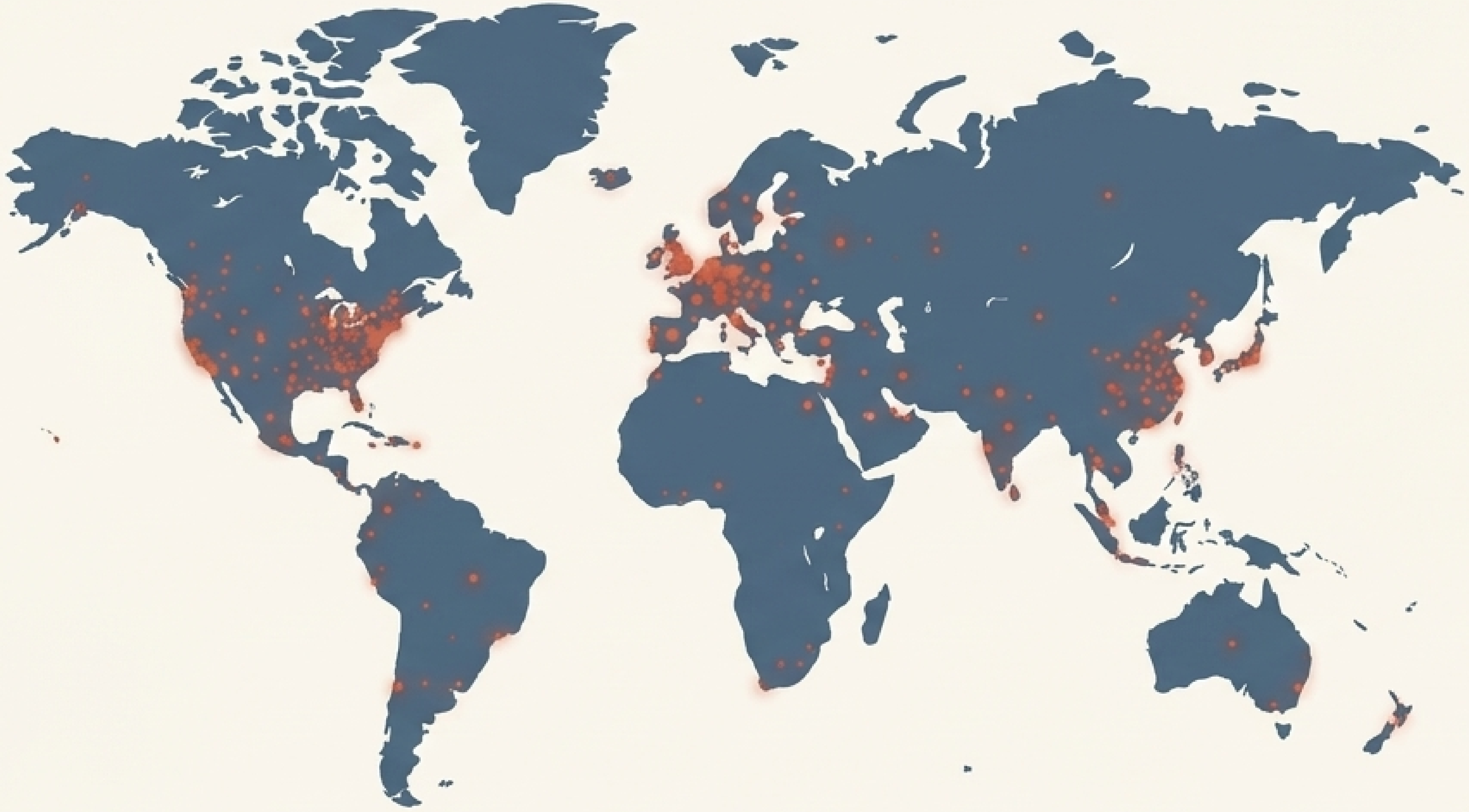
The Problem It Solves

Helps identify potential mental health concerns early, acting as a valuable screening tool where issues might otherwise go undetected.

The Beneficiaries

Healthcare Providers, Employers, Individuals, and Researchers.

Our Raw Material: The Global Mental Health Dataset



Key Statistics

Responses: ~290,000

Countries: 34

Target Variable: `treatment`

Key Feature Categories

Demographics: Country, Occupation

History: family_history, Mental_Health_History

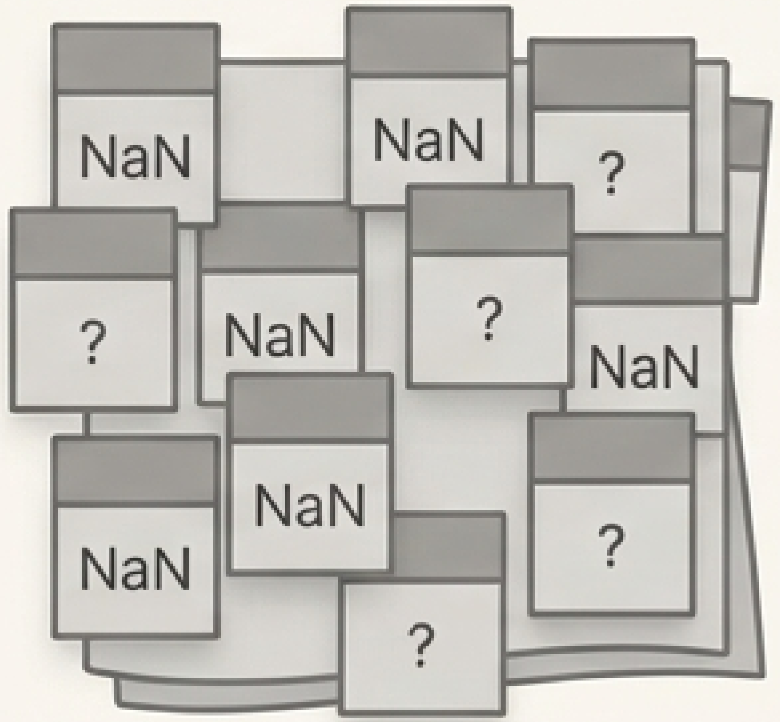
Behavioral Indicators: Days_Indoors, Growing_Stress, Mood_Swings

Step 1: Forging a Clean Foundation for Analysis

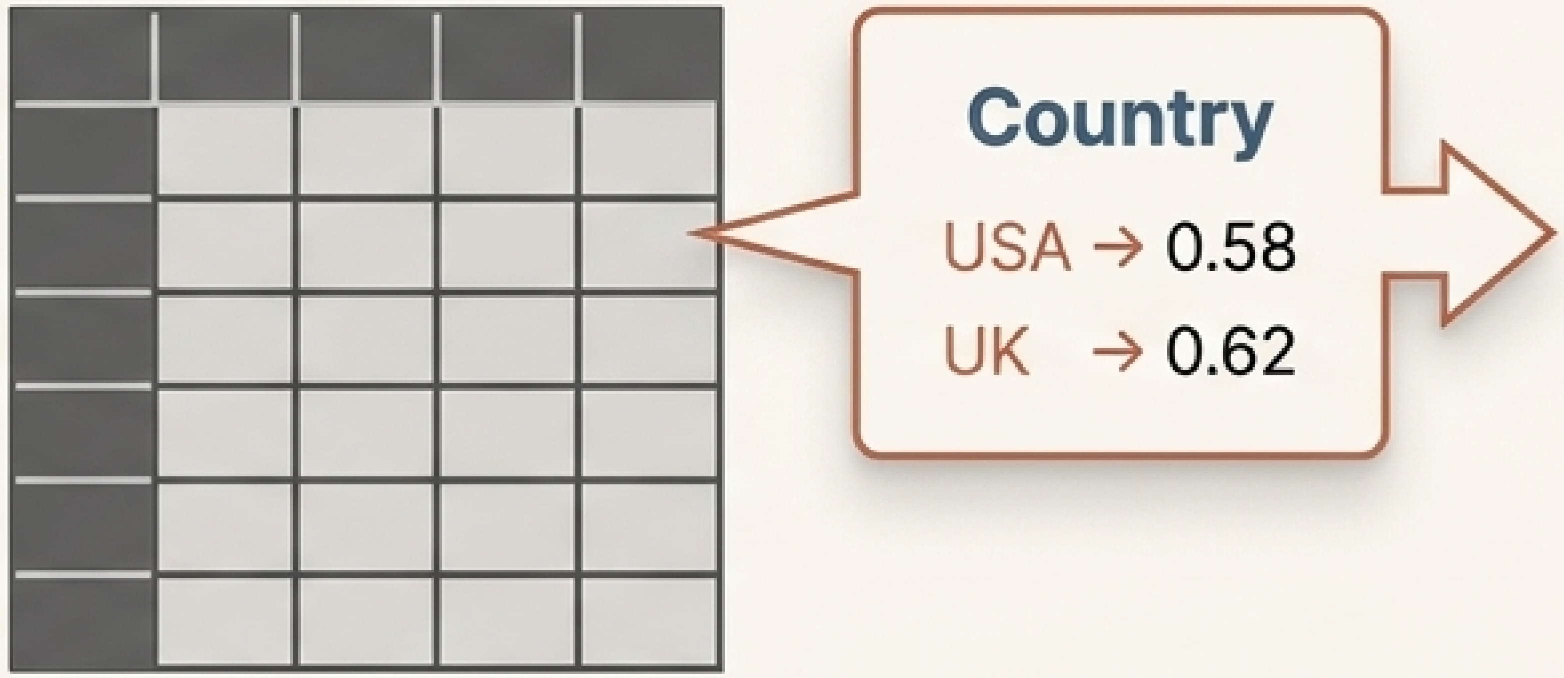
Key Actions

- **Cleaning:** Removed all duplicate entries.
- **Missing Values:** Systematically handled missing data in the `self_employed` column.

Before



After



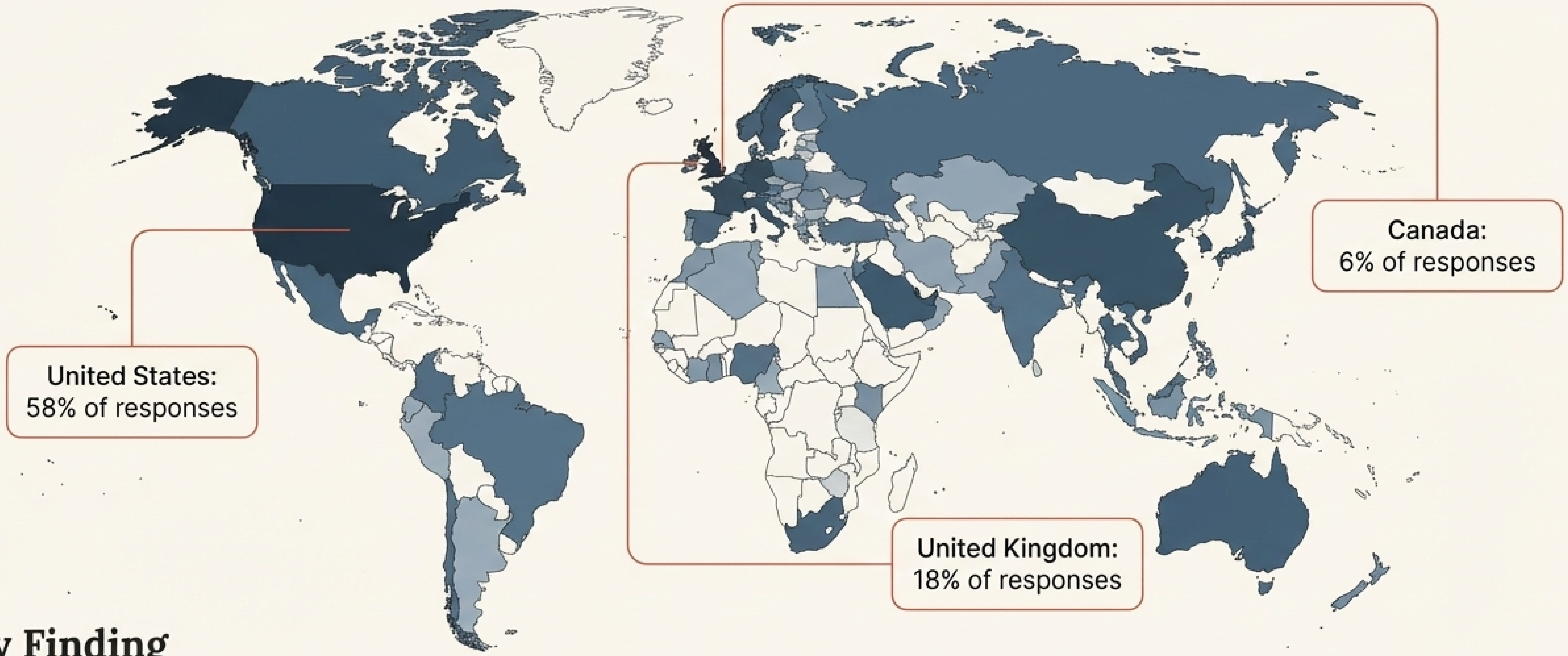
Core Technique: Target Encoding

- **What:** Converted categorical features into numerical values based on their relationship with the target variable (`treatment`).
- **Strategic Application:** Applied specifically to the `Country` column. This captured the influence of geographic location on treatment-seeking behavior without creating hundreds of new, sparse features.

The Goals of Our Exploration

- Analyze the distributions of **key features.**
- Identify the most significant **correlations between variables.**
- Examine the relationship between **specific behaviors** and the need for **treatment.**

Treatment Need Varies Significantly Across the Globe



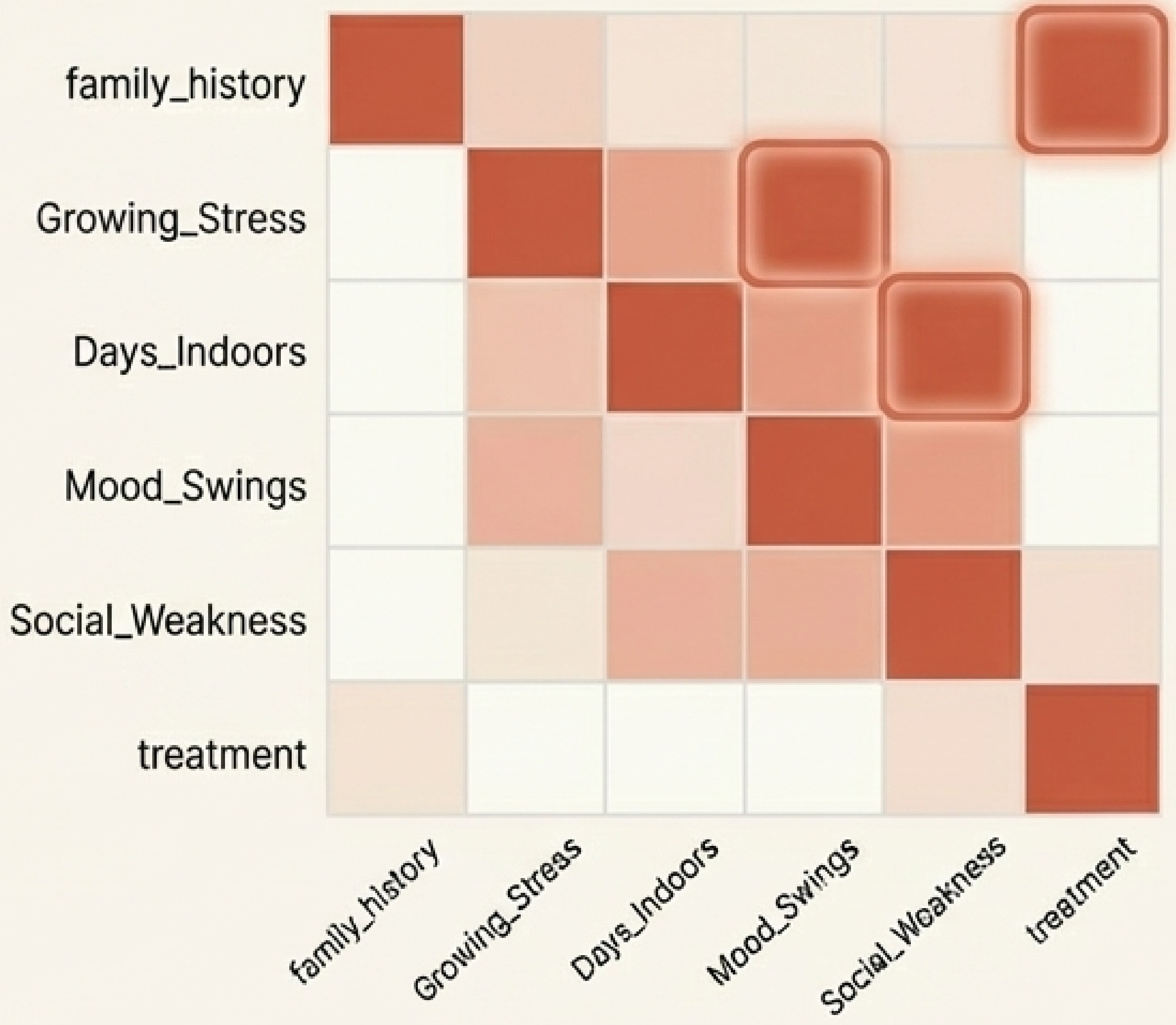
Key Finding

Our analysis reveals distinct geographic concentrations in both survey responses and treatment rates, with the United States (58%), United Kingdom (18%), and Canada (6%) being the most represented.

Strong Links Between History, Stress, and Social Behavior

Critical Insights Discovered

- **Family History:** The single strongest predictor of an individual's need for treatment.
- **Stress & Mood:** A powerful positive correlation exists between Growing_Stress and Mood_Swings.
- **Isolation & Social Function:** Days_Indoors is strongly linked to a higher score in Social_Weakness.



Step 3: Crafting Smarter Features to Boost Predictive Power

Our Rationale

We moved beyond raw data to engineer composite and latent features. These new features are designed to capture more complex human behaviors and patterns that are not immediately obvious in the original variables.

Engineering Composite Scores from Raw Signals

Stress Score

```
mean(Days_Indoors, Growing_Stress,
Coping_Struggles, Mood_Swings, ...)
```

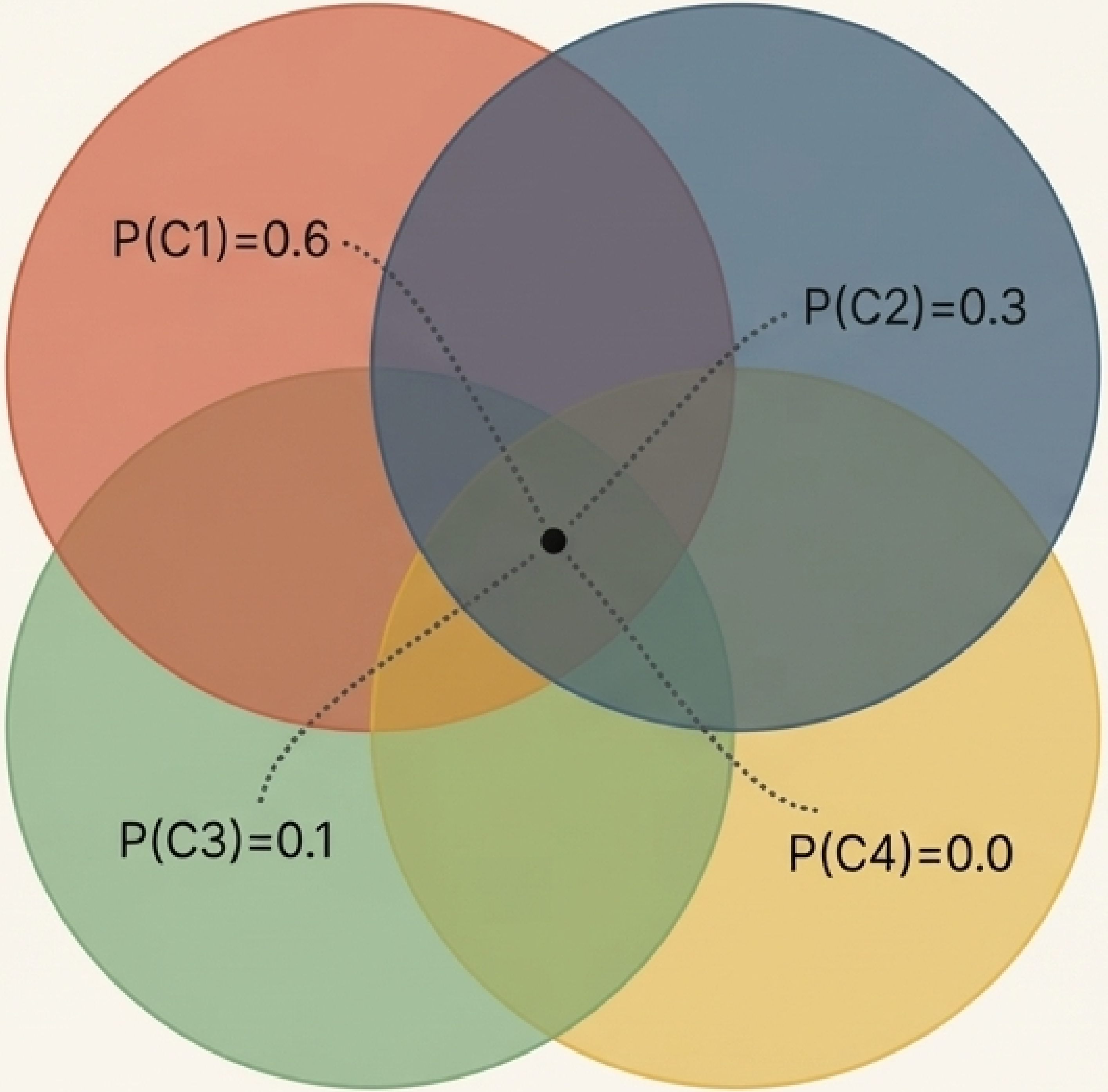
Purpose: Aggregates multiple stress indicators into one potent, comprehensive measure.

Social Function Score

```
Work_Interest - Social_Weakness
```

Purpose: Captures the critical balance between professional engagement and social difficulties.

Uncovering Latent Groups with Fuzzy Clustering



The Technique

We applied a Gaussian Mixture Model (GMM) to identify 4 latent clusters within the dataset.

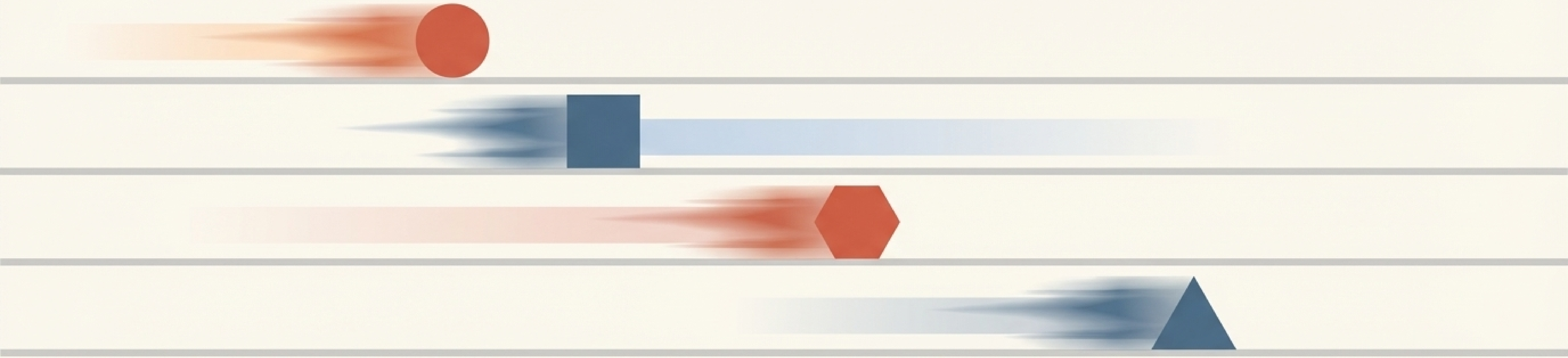
The "Fuzzy" Output

Instead of assigning each person to a single group, we created features like `Cluster_0_Score`, `Cluster_1_Score`, etc. Each feature represents the ***probability*** of an individual belonging to that cluster.

The Advantage

This method captures nuanced, overlapping characteristics that a simple classification would miss, providing a richer signal for our model.

Step 4: A Gauntlet of Models to Find the Best Performer



Our Approach

To ensure our final model was truly superior, we first benchmarked a diverse range of powerful algorithms.

The Contenders

CatBoost, Multi-Layer Perceptron (MLP), Stacked Ensemble, XGBoost.

Methodology

We employed a 5-fold cross-validation strategy to ensure our performance evaluation was robust and generalizable.

How the Baseline Models Performed

Model	Accuracy	ROC-AUC	Key Characteristic
CatBoost	92.23%	0.98+	Handles raw categorical data natively
MLP	91.14%	0.98+	Deep learning approach for complex patterns
Stacked Ensemble	91.46%	~98.23%	Combines the strengths of multiple models
XGBoost	92.01%	~98.59%	Fast, robust, and highly accurate

Step 5: Why XGBoost Was the Clear Champion

A Multi-Faceted Decision

Peak Performance

Achieved the highest ROC-AUC score (98.59%), demonstrating superior class discrimination.

Excellent Balance

Offered an exceptional balance between Precision (92.56%) and Recall (91.49%), which is crucial for a reliable screening tool.

Deployment-Ready Speed

Fast prediction times are essential for the real-time, interactive user application.

Built-in Interpretability

Provides clear feature importance scores, allowing us to understand *why* the model makes its predictions.

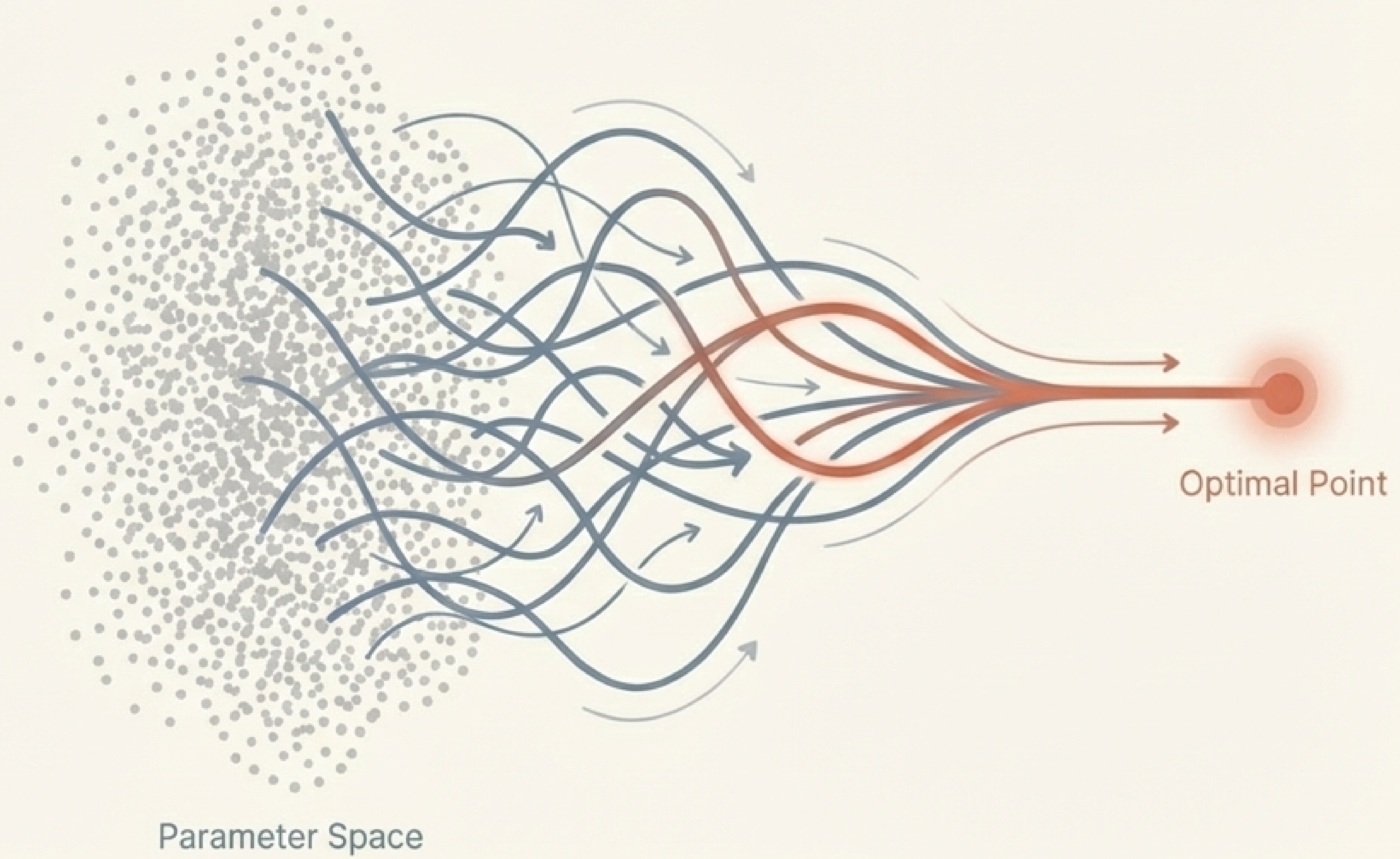
Step 6: Tuning XGBoost for Maximum Performance with Optuna

The Tool

We used Optuna, an advanced hyperparameter optimization framework, to fine-tune our champion model.

The Process

Optuna intelligently searched for the optimal combination of model settings (e.g., `learning_rate`, `max_depth`, `subsample`) over 30 trials, using 5-fold cross-validation within each trial to ensure robustness.



The Result: A Highly Optimized and Powerful Model

Final Optimized Parameters

n_estimators: 372

learning_rate: 0.070

max_depth: 10

subsample: 0.520

colsample_bytree: 0.871

Peak Performance Achieved

Accuracy
92.01%

F1-Score
92.02%

ROC-AUC
98.59%

Step 7: Ensuring a Robust Workflow with MLflow

Our Experiment Management System

We used MLflow, an open-source platform, to manage the end-to-end machine learning lifecycle.

Why We Used It

- **Reproducibility** : Automatically log every parameter, metric, and model file.
- **Comparison** : Easily compare dozens of model runs side-by-side.
- **Collaboration** : Provide a central, shared repository for all team experiments.

Our Experiment Dashboard: A Single Source of Truth

Comprehensive Tracking for Each Run

- **Parameters Logged:** `learning_rate`, `max_depth`, `n_estimators`, `gamma`, etc.
- **Metrics Logged:** Accuracy, Precision, Recall, F1-Score, ROC-AUC.
- **Artifacts Stored:** Trained model files (.pkl), visualizations (ROC curves), and environment specifications.

Final Model Run ID:

`8545643c6264451b886e4dd5c7ace671`

MLflow UI

Run Name	ROC-AUC	Accuracy	F1-Score
CatBoost_Raw	0.982	92.23%	92.15%
MLP_Classifier	0.981	91.14%	91.05%
XGBoost_Optuna_Tuned	0.9859	92.01%	92.02%

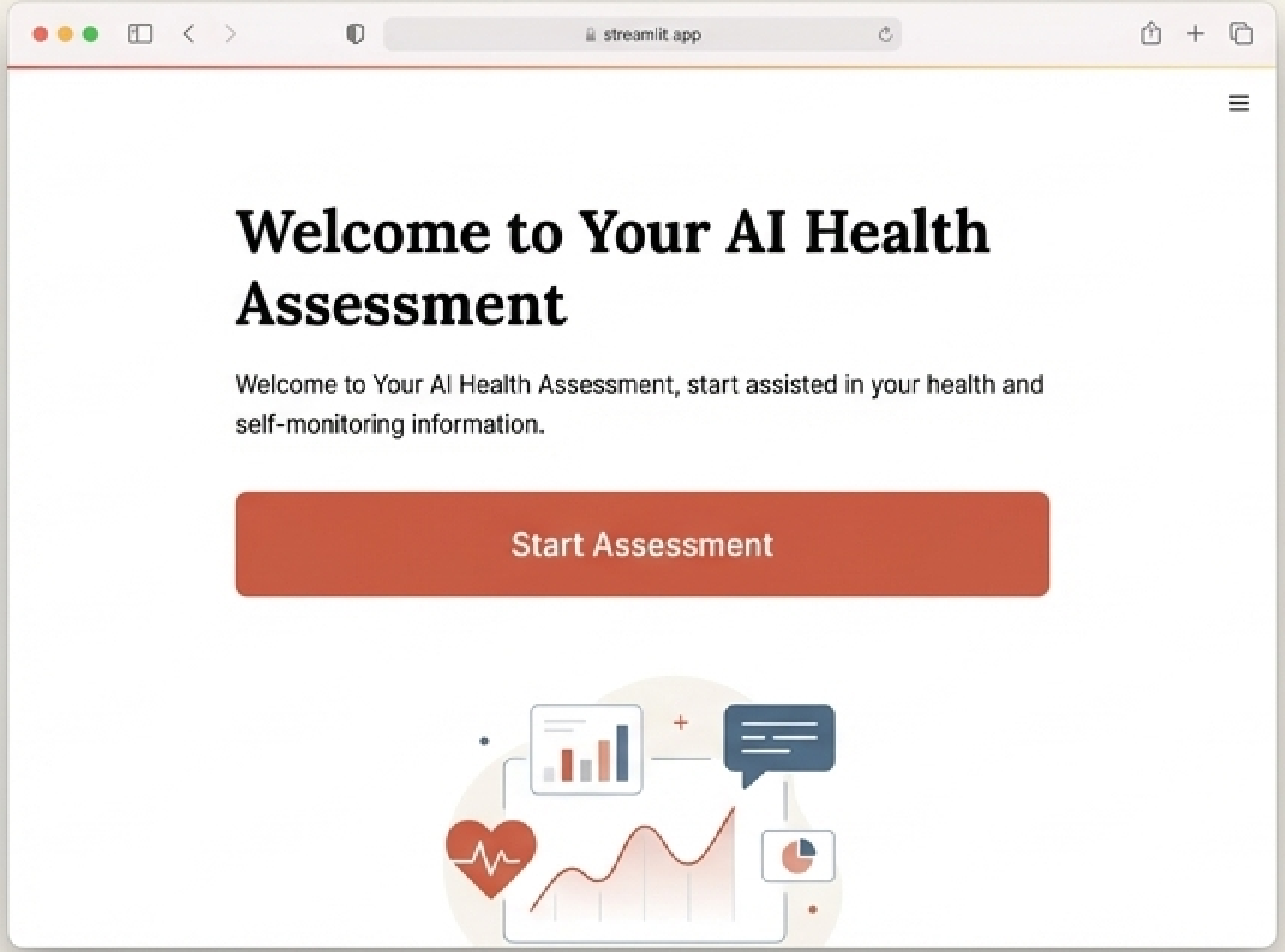
Step 8: The Interactive AI Clinic

Transforming the Model into a User-Facing Tool

We built a web application using Streamlit to bring our system to life.

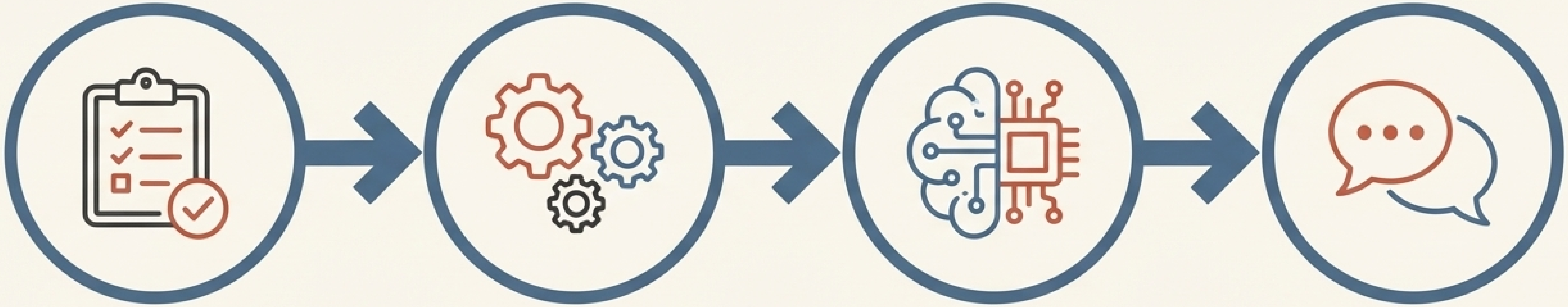
Key Components

- **Interactive Dashboard:** For data exploration and visualization.
- **Prediction System:** A multi-step questionnaire that feeds user input directly to our trained XGBoost model.
- **AI Chatbot:** A conversational interface, powered by Google Gemini, to help users understand and discuss their results.



The Real-Time Prediction Flow

A User's Journey in Four Steps



User Input

The user answers questions across four categories (Basics, Medical History, Lifestyle, Social & Coping).

Live Preprocessing

Responses are encoded, and engineered features like the Stress Score are created on the fly.

XGBoost Prediction

The optimized model generates a prediction and a probability score in real-time.

Result & Dialogue

The prediction is displayed, and the Gemini-powered chatbot engages the user to discuss the result in a supportive context.

A Complete System for Insight and Support

We have successfully built and deployed an end-to-end system that translates complex data into a supportive, interactive dialogue, demonstrating the power of machine learning in mental health screening.

Team

Yahya Mustafa
Adham Fouad
Abdelkader
Ahmed Fahmy
Muhanad Mustafa
Marwan Gaber

NotebookLM