The Hunt for Hidden Risk

Decoding the Complex Link Between Lifestyle and Disease with Machine Learning

Team Members:
Prithvi Seshadri, Vamsi Krishna N K, Anaya Choudhari

Our Investigation: 100,000 Individuals, One Critical Question

The Problem

Lifestyle-driven diseases are prevalent, but the relationship between daily habits and health outcomes is complex and non-linear.

Our goal is to move beyond simple correlations to build a robust predictive tool.

The Evidence (Dataset)

A comprehensive dataset of 100,000 records featuring:

- **Lifestyle Factors:** `daily_steps`, `sleep_hours`, `water_intake_l`, `smoker`, `alcohol`
- **Health Metrics:** `bmi`, `systolic_bp`, `diastolic_bp`, `cholesterol`, `resting_hr`
- **Target Variable:** `disease_risk` (Binary: 1 = High Risk, 0 = Low Risk)

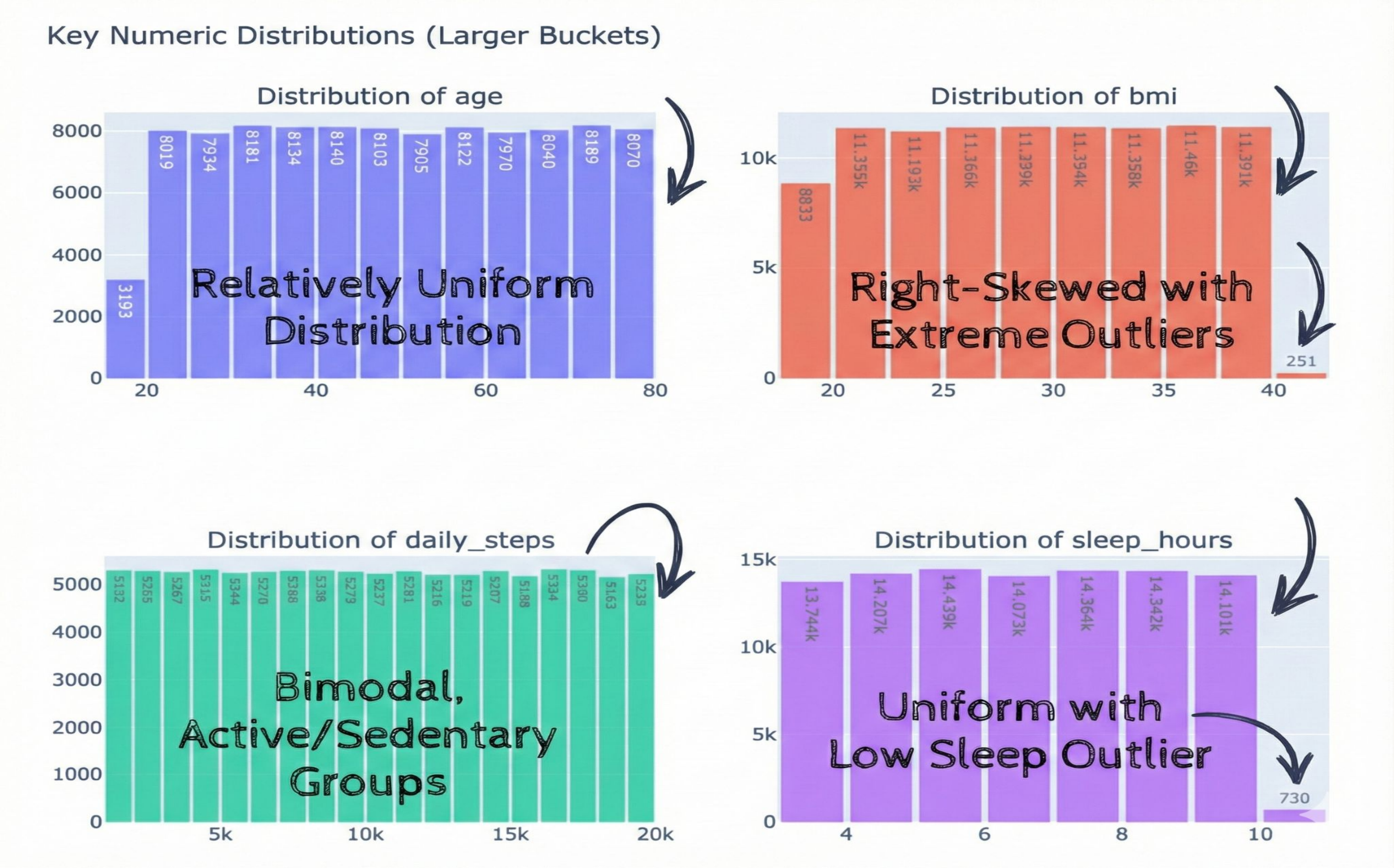
The Objectives

1. **Explore:** Uncover hidden patterns using advanced visualizations.
2. **Predict:** Build a classifier to predict `disease_risk` using ensemble methods.
3. **Strategize:** Optimize the model for a practical, real-world clinical use case.

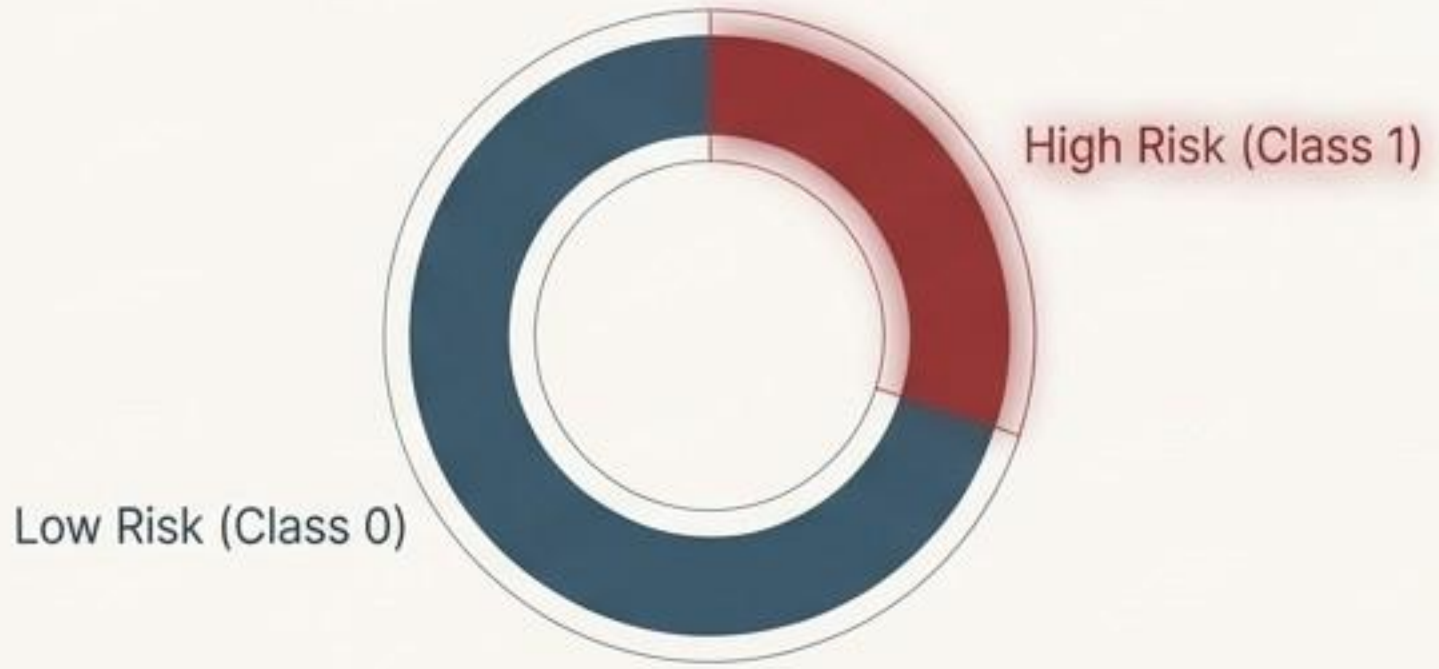
First Look at the Evidence: Distributions and Imbalance

Population Characteristics

The population shows a right-skewed BMI, indicating a prevalence of overweight individuals—a known risk factor. Daily steps show a wide variance, suggesting distinct "active" and "sedentary" groups.



The Critical Imbalance



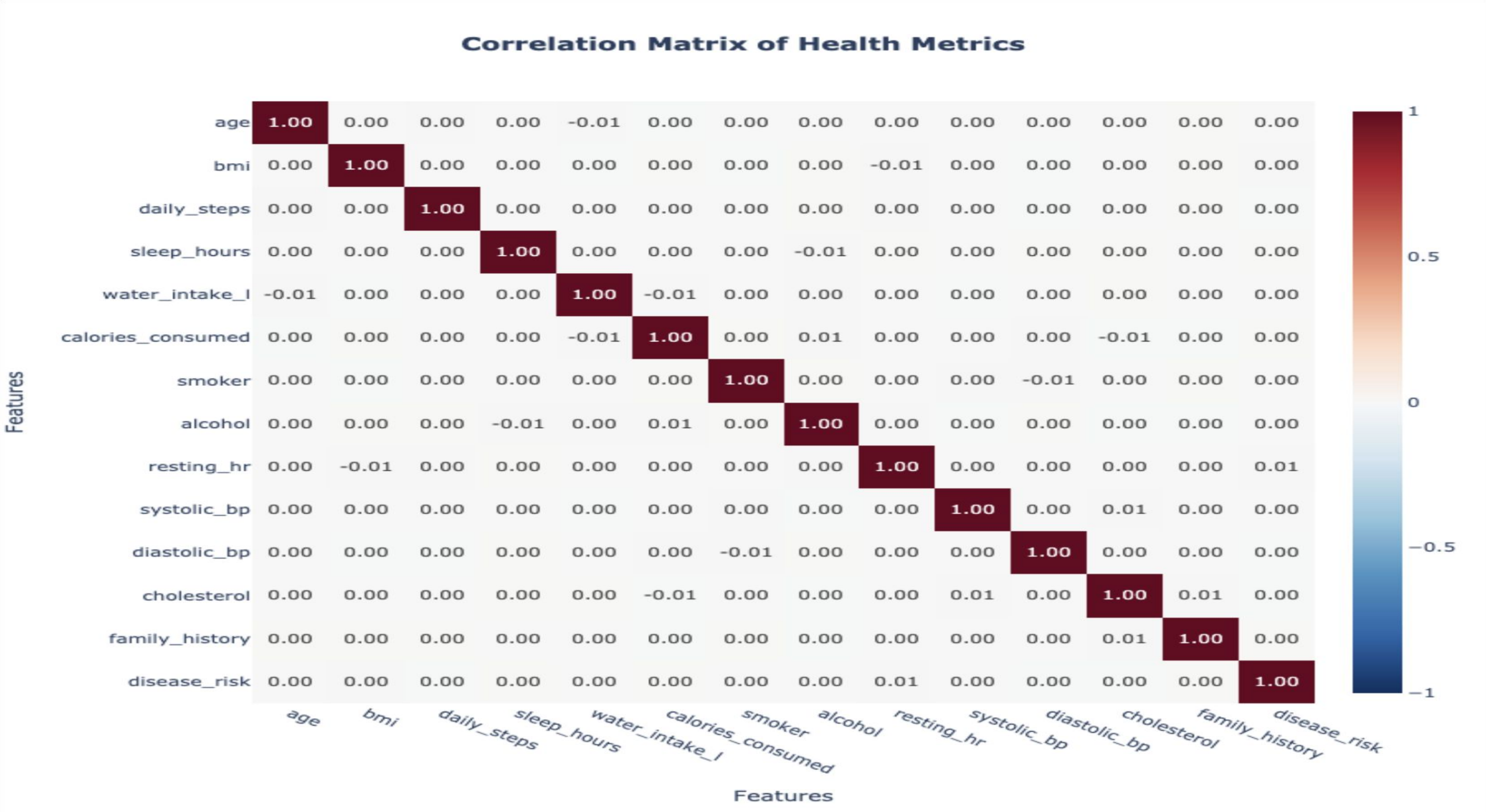
A crucial finding is the class imbalance. The "High Risk" class is a significant minority. This means a naive model could achieve high accuracy by simply guessing "Low Risk" every time, making it dangerously ineffective.

This discovery mandates the use of advanced techniques like SMOTE to ensure our model learns from the minority group.

The Case of the Missing Signal: No Single ‘Smoking Gun’ Exists

A standard correlation heatmap reveals the absence of strong, linear relationships between any single lifestyle factor and `disease_risk`. The risk is determined by complex, non-linear interactions, not a simple, direct cause.

Incredibly low correlations (near 0.0) with `disease_risk`. This proves a simple linear model like Logistic Regression will struggle.



Correlation Matrix of Health Metrics

	age	bmi	daily_steps	sleep_hours	water_intake_l	calories_consumed	smoker	alcohol	resting_hr	systolic_bp	diastolic_bp	cholesterol	family_history	disease_risk
age	1.00	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
bmi	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	0.00	0.00
daily_steps	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
sleep_hours	0.00	0.00	0.00	1.00	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	0.00	0.00	0.00
water_intake_l	-0.01	0.00	0.00	0.00	1.00	-0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
calories_consumed	0.00	0.00	0.00	0.00	-0.01	1.00	0.00	0.01	0.00	0.00	0.00	-0.01	0.00	0.00
smoker	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	-0.01	0.00	0.00	0.00
alcohol	0.00	0.00	0.00	-0.01	0.00	0.01	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
resting_hr	0.00	-0.01	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.01
systolic_bp	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.01	0.00	0.00
diastolic_bp	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	1.00	0.00	0.00	0.00
cholesterol	0.00	0.00	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	0.01	0.00	1.00	0.01	0.00
family_history	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	1.00	0.00
disease_risk	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	1.00

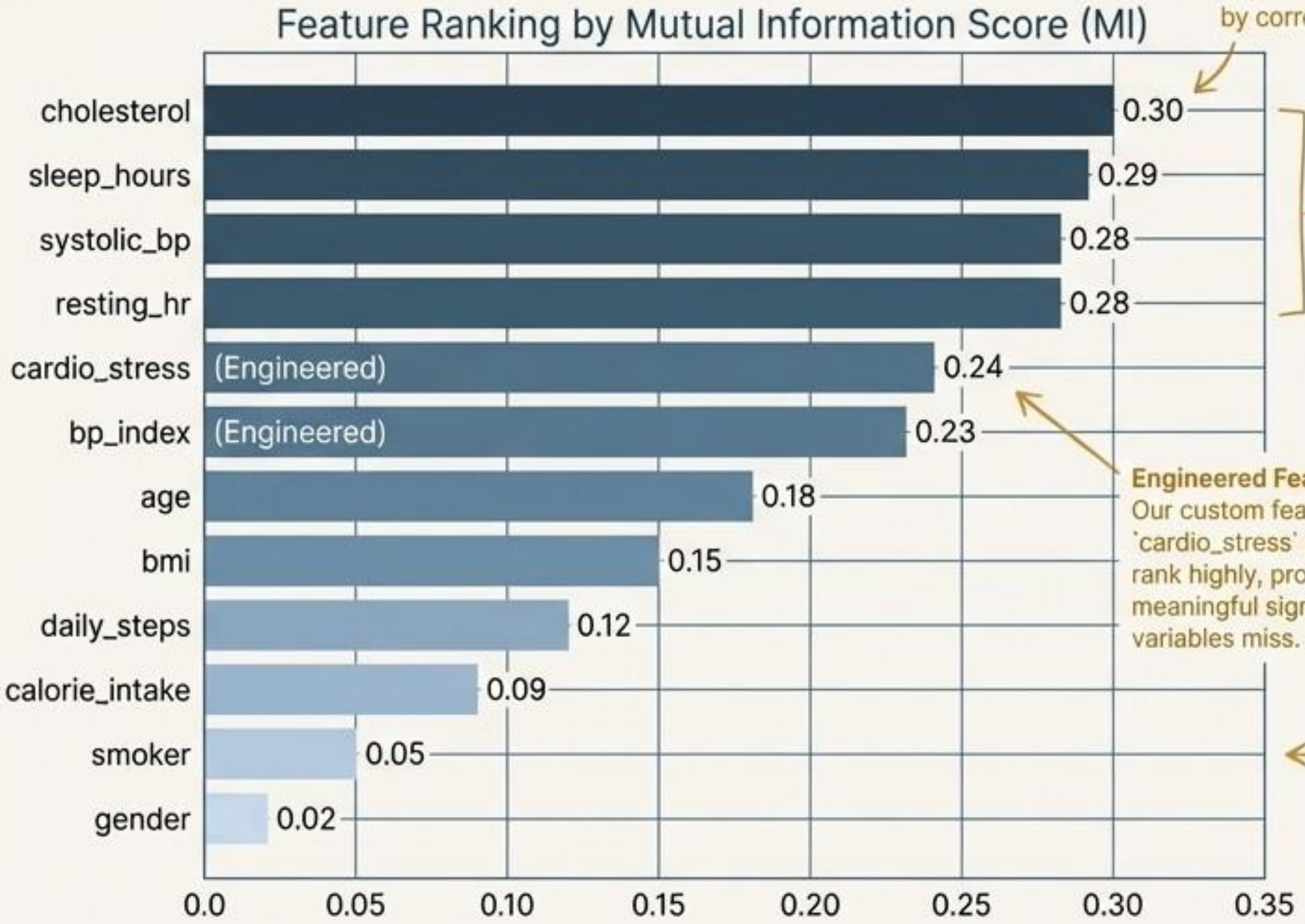
Virtually zero multicollinearity. Every feature provides unique, independent information, even if its linear signal is weak.

Strategic Conclusion

The flat heatmap is not a failure of analysis, but a clear directive. We must abandon the search for a simple cause and instead deploy models capable of capturing complex, non-linear thresholds, such as Random Forest and XGBoost.

Advanced Forensics: Using Mutual Information to Detect Hidden Patterns

Correlation only finds straight lines. **Mutual Information (MI) detects *any* kind of relationship.** This allows us to rank features by their true predictive power before modeling.



Identifies non-linear dependencies missed by correlation.

Engineered Features Succeed: Our custom features like `cardio_stress` and `bp_index` rank highly, proving they capture meaningful signals that raw variables miss.

Key Findings

1. **Physiological Metrics Dominate**: `cholesterol`, `sleep_hours`, `systolic_bp`, and `resting_hr` show the highest MI scores. The model will rely on objective clinical measurements.

2. **Engineered Features Succeed**: Our custom features like `cardio_stress` and `bp_index` rank highly, proving they capture meaningful signals that raw variables miss.

3. **Lifestyle is a Weaker Signal**: Habits like `smoker` and `gender` have low MI scores. This suggests the *outcomes* of a habit (e.g., high BP) are more predictive than the habit itself.

The Lineup: A Linear Baseline vs. A Non-Linear Specialist

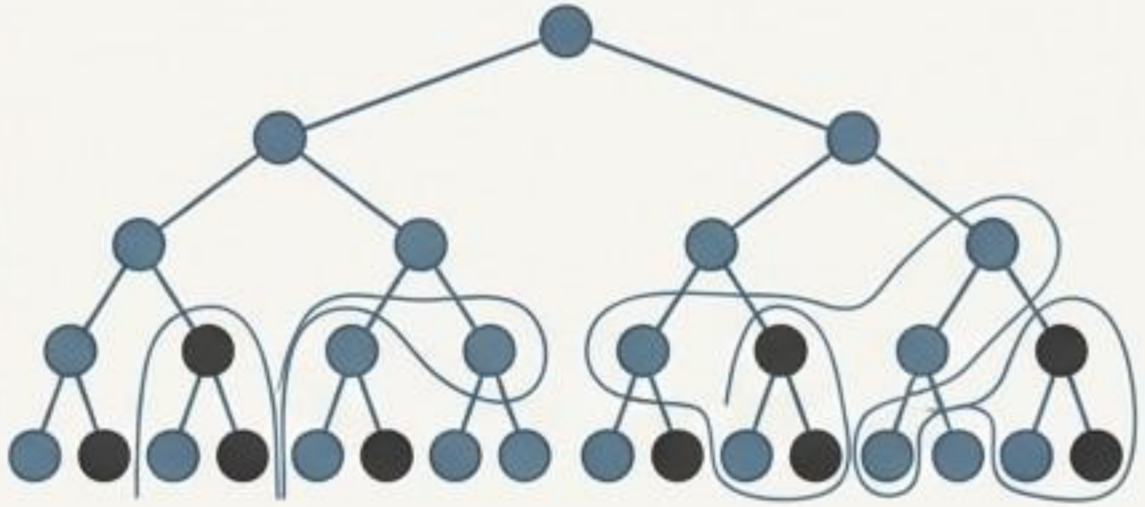
The Baseline - Logistic Regression

Role: Our 'sanity check.' A simple, fast, and highly interpretable model that establishes a performance benchmark.

How it Thinks: Assumes linear relationships (e.g., risk increases steadily with BMI).

Expectation: Given the flat correlation matrix, we anticipate this model will struggle to capture the true complexity of the data. Its accuracy is expected to be below 75%.

The Challenger - Tuned Random Forest



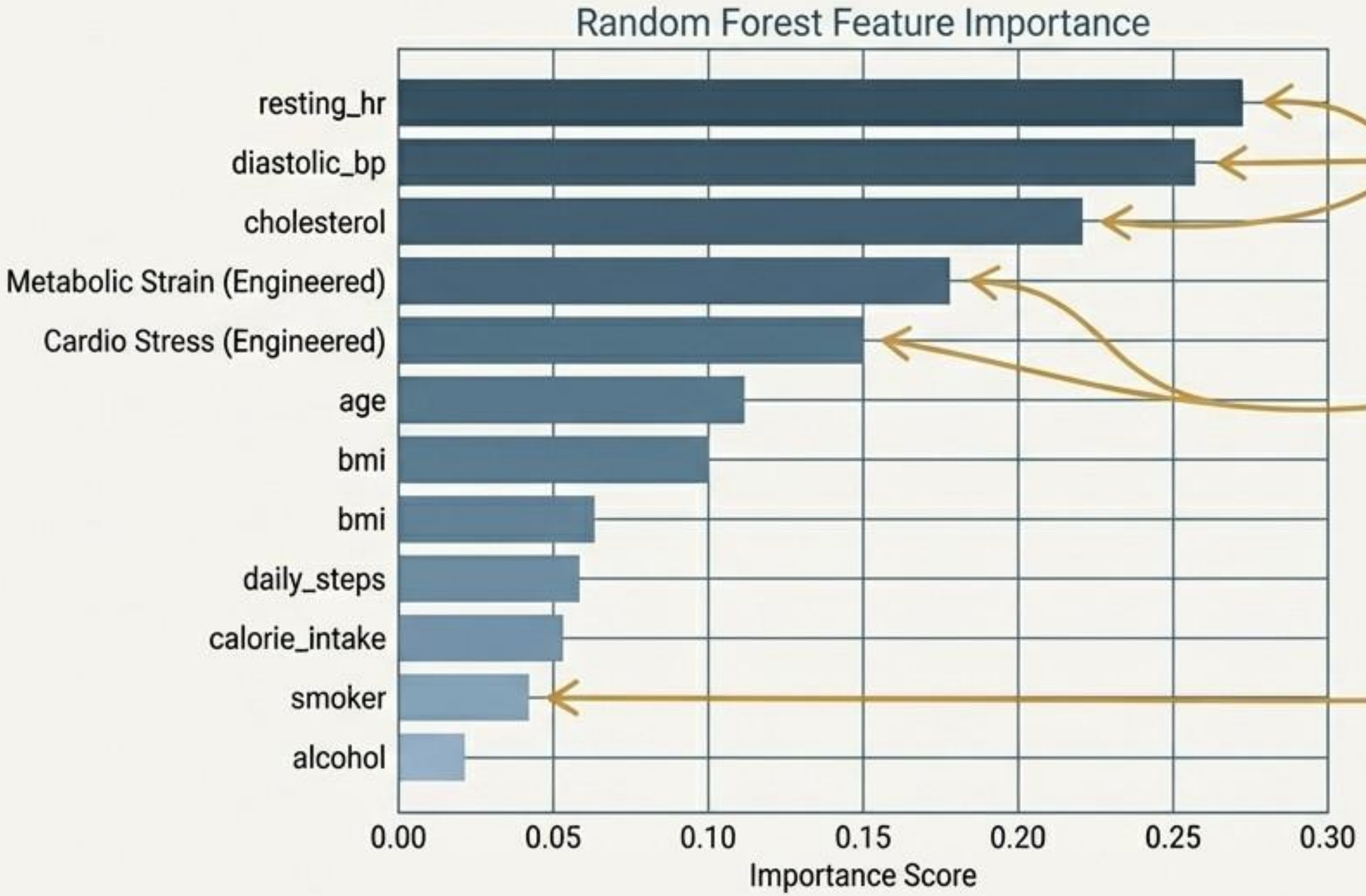
Role: Our non-linear specialist. Designed to capture the complex interactions and thresholds our EDA revealed.

How it Thinks: Builds hundreds of decision trees to learn 'if-then' rules (e.g., 'IF BMI > 30 AND Steps < 3000, THEN risk is high').

Advantage: Naturally captures the synergistic effects we observed (e.g., `Smoker + High BMI`) and is robust to the dataset's structure. We used `RandomizedSearchCV` to optimize its hyperparameters for a fair comparison.

Inside the Specialist’s Mind: The True Drivers of Risk

The Random Forest model’s **feature importance** plot provides a clear hierarchy of risk drivers, confirming that measurable vital signs are the most powerful predictors.



Key Insighgs

1. **The Big Three**
   The undisputed top predictors. The model prioritizes immediate, measurable red flags.

2. **Engineered Features Prove Their Worth**
   Our custom features rank high, outperforming many raw variables. This validates our domain-specific feature engineering strategy.

3. **The “Lifestyle Lag”**
   Reasoning: The model has learned that the *outcome* of a bad habit (like high blood pressure) is a more direct and powerful predictor than the habit itself.

A Shocking Revelation: Our Specialist Model Missed 92% of High-Risk Cases

Despite sophisticated feature engineering and a powerful non-linear model, the initial performance on the **most critical task** was a **near-total failure.**

Healthy Patient ID Rate (Recall Class 0)

~92%

Excellent

High-Risk Patient ID Rate (Recall Class 1)

~8%

Catastrophic Failure

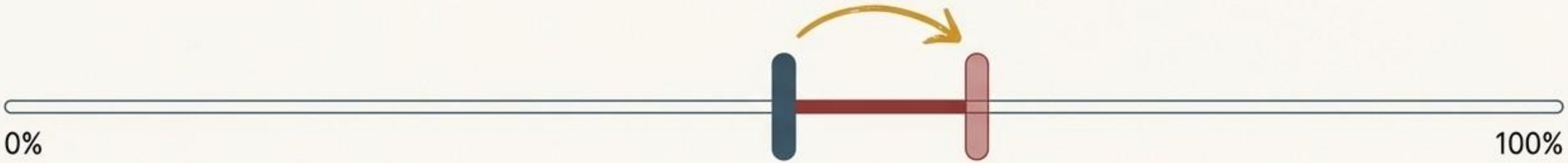
The model is exhibiting a severe "**Conservative Bias.**" It is so good at identifying the majority "Healthy" class that it's afraid to flag anyone as "**High Risk**" **unless it is** overwhelmingly certain. This is a common problem, but the severity here demands a strategic intervention, not just a technical one.

The Diagnosis: A Conservative Bias and the Strategic Fix

The model's internal “risk score” for many sick patients is likely high (e.g., 30-40%), but it falls short **of the default 50% cutoff required to label them “High Risk”.**

The Technical Solution: Threshold Tuning

We can change the rule the model uses to make a decision.



Default Rule: Predict “High Risk” only if probability is > 50%. (This results in high precision but catastrophically low recall).

50%

15%

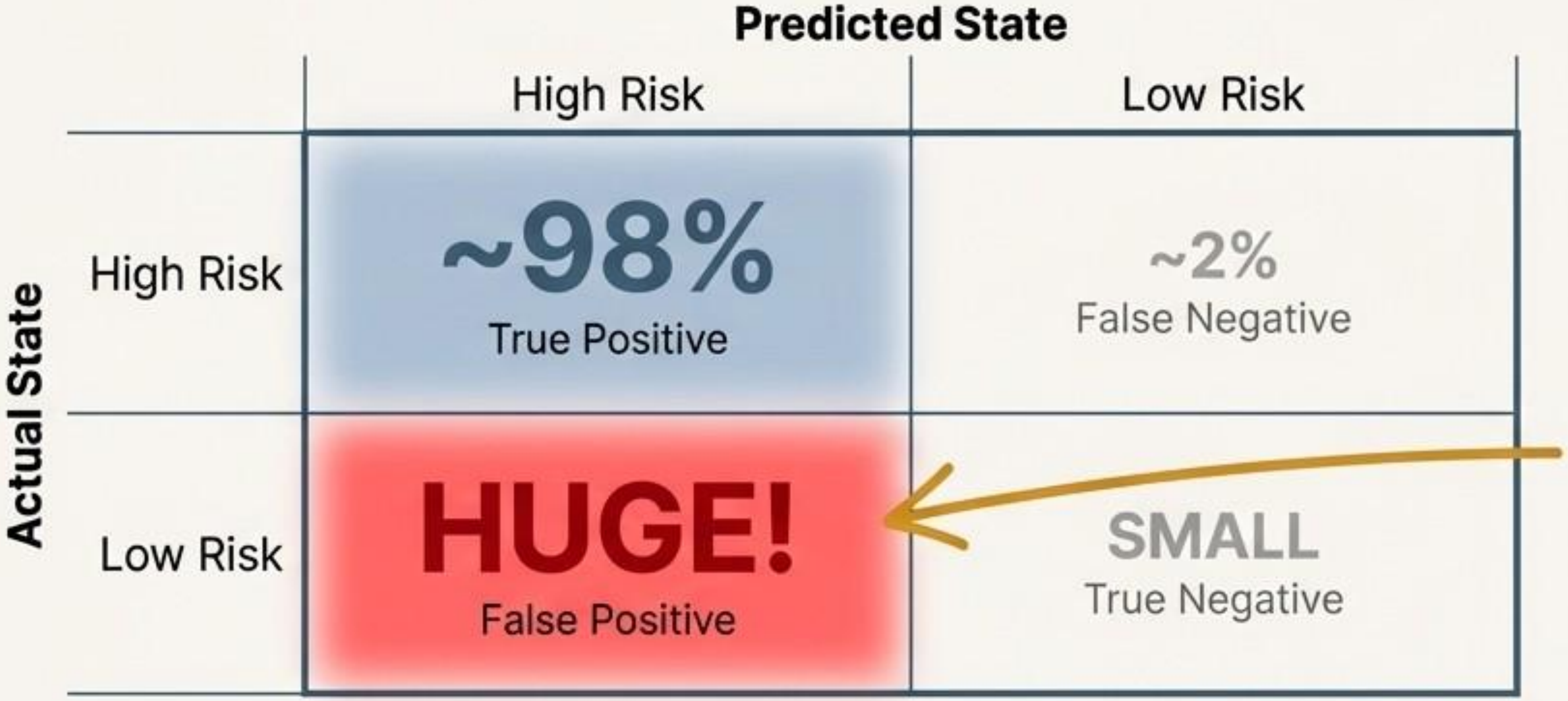
Proposed New Rule: Predict "High Risk" even if probability is only > 15%. (This is an aggressive screening strategy designed to catch everyone who might be at risk).

The Expected Trade-off

Lowering the threshold will drastically increase our ability to find sick patients (Recall), but it will also inevitably flag more healthy people by mistake (a drop in Precision). The next slide shows the real-world impact of this trade-off.

The ‘Fix’ Creates a New Crisis: A False Alarm Epidemic

Lowering the decision threshold to 0.15 worked as intended—we are now catching almost all high-risk patients. However, this came at an unacceptable cost: the model's predictions have become almost meaningless due to a flood of false positives.



	Predicted State	
Actual State	High Risk	Low Risk
High Risk	~98% True Positive	~2% False Negative
Low Risk	HUGE! False Positive	SMALL True Negative

The model now incorrectly flags nearly 97% of all healthy individuals as ‘High Risk’.

This **renders the tool unusable** in a clinical setting. If a doctor receives a ‘High Risk’ alert, it is overwhelmingly likely to be a false alarm. This would destroy clinician trust and create immense patient anxiety.

The Strategic Verdict: We Must Prioritize Trust Over Raw Sensitivity

We will **reject** the aggressively tuned threshold (0.15) and deploy the original **Random Forest with the default 0.50 threshold**.

The Rationale: Credibility is the Most Important Metric

The primary goal is not to find every possible case, but to create a tool that clinicians can trust.

Unacceptable operational burden

- A 97% false alarm rate makes the tool untrustworthy and operationally useless.
- We are making a strategic choice to accept lower sensitivity (missing some cases) to ensure that when the model **does** issue a 'High Risk' warning, it is highly credible and warrants immediate attention.

Actionable signal

Our Deployed Model's Use Case

Avoids alert fatigue

✗ **It is NOT**
A general population screening filter.

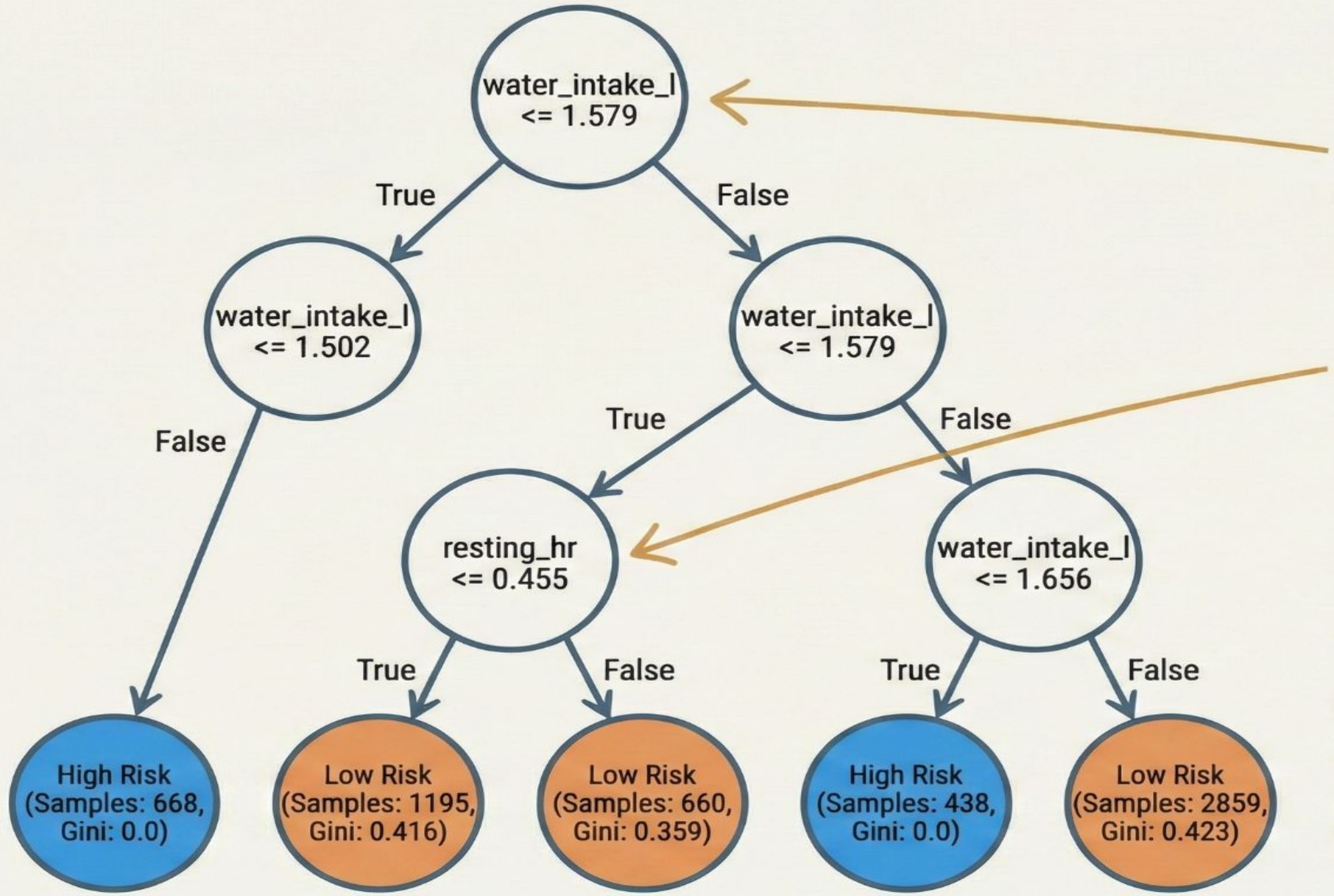
Directs focus to genuine risk

✓ **It IS**
A **'High-Certainty Confirmation Tool.'**

If this model flags a patient, the probability of risk is genuinely high, justifying further clinical investigation.

The Rules of Risk: A Glimpse Inside the Model's Logic

To understand *how* our model makes decisions, we visualized a shallow Decision Tree. It reveals the sharp, non-linear 'if-then' rules it uses to classify individuals, confirming why linear models were destined to fail.


water_intake_l <= 1.579
True
False
water_intake_l <= 1.502
water_intake_l <= 1.579
False
True
False
resting_hr <= 0.455
water_intake_l <= 1.656
True
False
True
False
High Risk (Samples: 668, Gini: 0.0)
Low Risk (Samples: 1195, Gini: 0.416)
Low Risk (Samples: 660, Gini: 0.359)
High Risk (Samples: 438, Gini: 0.0)
Low Risk (Samples: 2859, Gini: 0.423)

Interpreting the Rules

- **Primary Splitter**: The tree's first decision is based on `water_intake_l`, splitting the population at the specific normalized value of 1.579.
- **Secondary Splitter**: For those in a certain `water_intake_l` band, `resting_hr` becomes a deciding factor.

The Key Insight

The decision boundaries are extremely sharp and specific (e.g., `1.579` vs `1.502`). Risk in this dataset behaves like a series of cliffs, not a gentle slope. Our Random Forest excels because it is designed to find these exact cliffs.

What Our Investigation Revealed: Key Conclusions

1. Non-Linear Relationships Dominate.

Risk is driven by complex interactions, not simple correlations. This was confirmed by both the the flat correlation matrix and the high Mutual Information scores for physiological metrics.

Our **choice of a tree-based model** was essential.

Optimal for complexity

2. Physiological Metrics Outperform Self-Reported Habits.

The model consistently prioritized measurable vital signs ('blood_pressure`, `cholesterol`, `resting_hr`) over lifestyle choices (`smoker`). The **outcomes of habits** are more predictive than the habits themselves.

Hard data signals

Predictive reality

3. Business Logic Trumps Pure Performance Metrics.

The most important decision was not technical but strategic. We chose to optimize for model credibility and trust by accepting a lower sensitivity for high-risk cases to **avoid a 97% false alarm rate** on healthy individuals.

Prioritizing trust

The use case defines the 'best' model.

Strategic alignment

The Next Case: A Roadmap for a Smarter, More Proactive Model

This project provides a strong foundation. To further enhance predictive power and clinical utility, we recommend the following strategic initiatives:

Enhanced Data Collection

Longitudinal Data
Track patients over time to capture disease progression and identify ‘pre-clinical’ cases.

Wearable Device Integration
Incorporate continuous, real-time data (heart rate, sleep patterns) to move beyond static snapshots.

Genetic Markers
Add genetic predisposition data for a more holistic risk profile.

Advanced Modeling & Interpretability

Explore More Complex Architectures
Test deep learning models to capture even more subtle feature interactions.

Prioritizing trust

Implement Patient-Specific Explanations (SHAP)
Move beyond general feature importance to show clinicians *why* a specific patient was flagged as high-risk, building further trust.

Develop a Multi-Threshold Strategy
Deploy different model thresholds for different clinical needs (e.g., a low threshold for initial screening vs. a high threshold for diagnostic support).

Strategic alignment