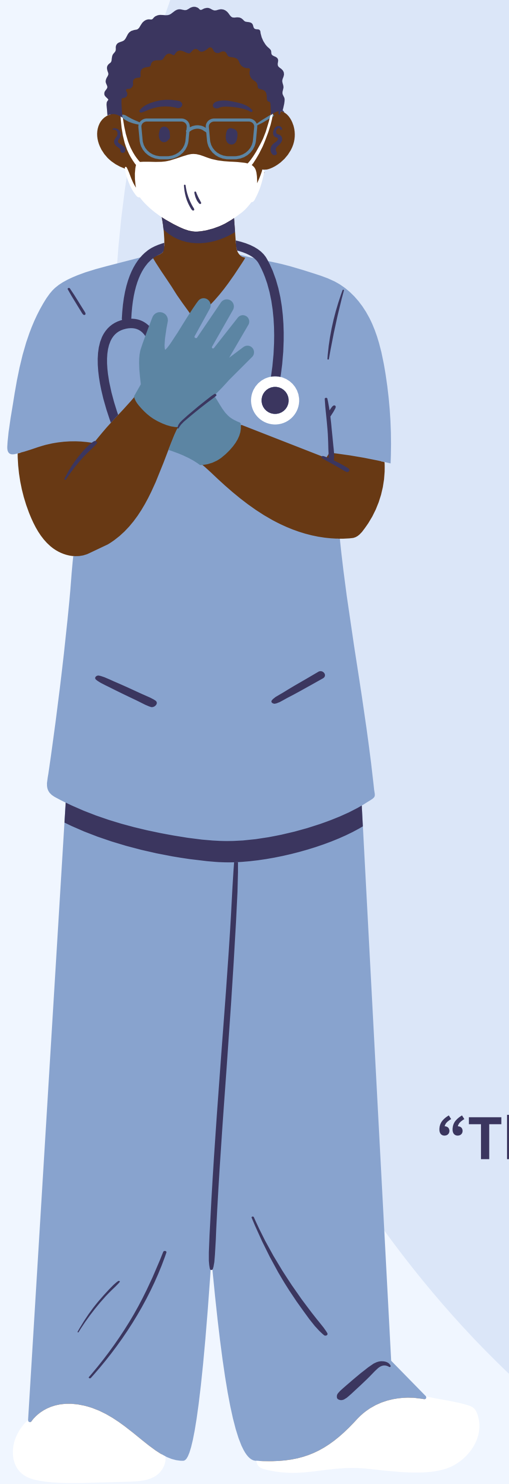
Predicting Burn Patient Mortality Using Machine Learning

Project Background



- Burn mortality is a critical metric in emergency and trauma care.
- Traditional scores (e.g., Baux) are simple but lack flexibility and personalization.
- ML offers the ability to learn complex patterns from large datasets.
- Goal: Develop and evaluate ML models for mortality prediction and compare them against the Baux score.

“This project isn’t just about predicting death — it’s about enabling faster, smarter, and fairer clinical decisions when time matters most.”

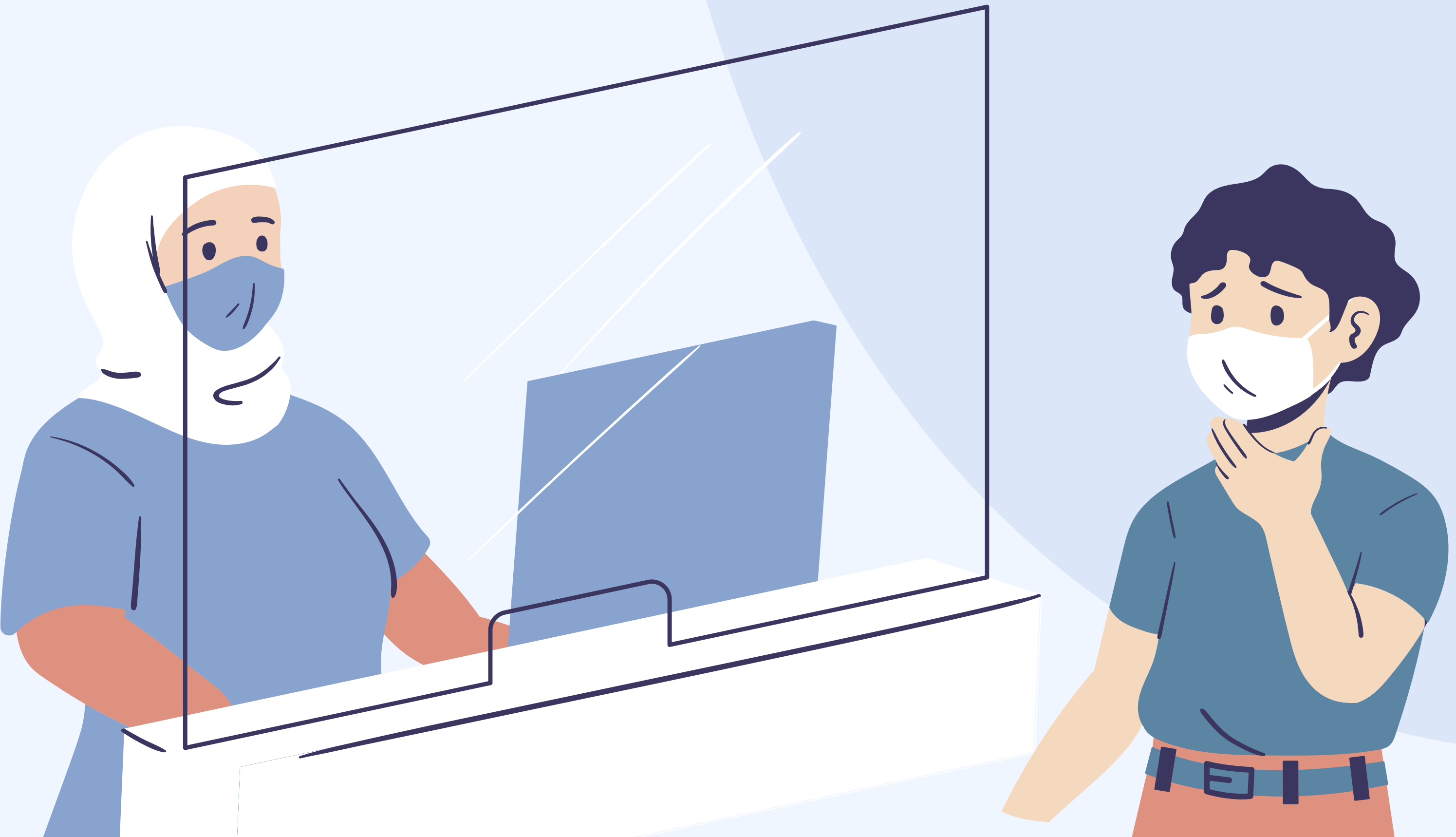
Problem Statement

- Traditional burn mortality assessment tools, such as the Baux score, have limitations in their effectiveness.
- These tools do not adequately consider the nonlinear interactions between clinical and injury variables.
- This oversight leads to less accurate predictions.

There is a clear need for a more adaptive, data-driven approach to improve the accuracy of these assessments.

Objectives

- Build ML models for mortality prediction using comprehensive patient data.
- Compare ML models with Baux score performance.
- Optimize recall and AUC to identify high-risk patients early.
- Apply resampling, encoding, and feature engineering to enhance performance.
- Ensure interpretability to support clinical usage.



Key Challenges Faced During the Project

Severe Class Imbalance

- Most patients survived, causing models to overfit the majority class
- Required advanced resampling (SMOTE, ENN, Tomek) to improve recall for deaths

Multi-label and High-Cardinality Features

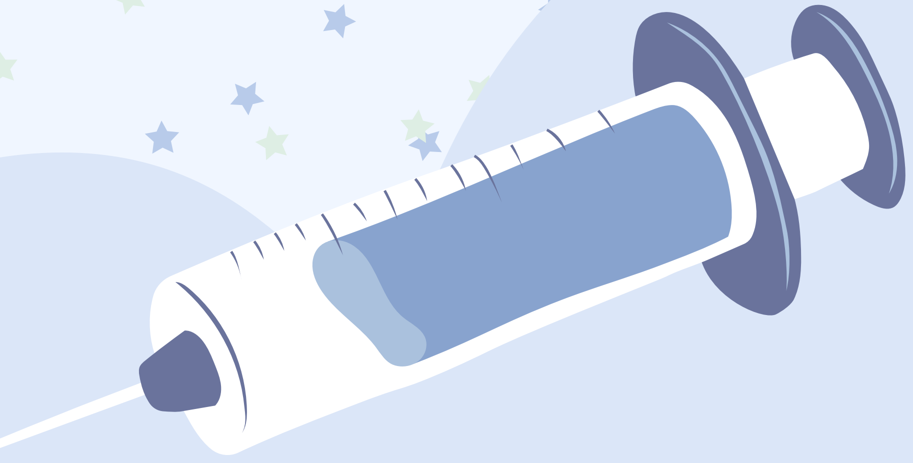
- Injury-related columns contained composite values (e.g., "Face;Neck")
- Demanded binarization, grouping, and dimensionality reduction

Noisy and Skewed Data Distributions

- Features like hospital stay had extreme outliers and long-tailed distributions
- Required winsorization and robust preprocessing

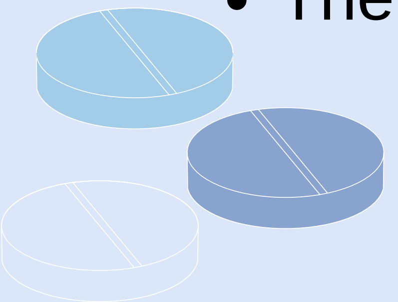
Baux Score Comparison

- Traditional score needed to be converted into probability for fair ML comparison
- Required logistic transformation and AUC-based head-to-head analysis

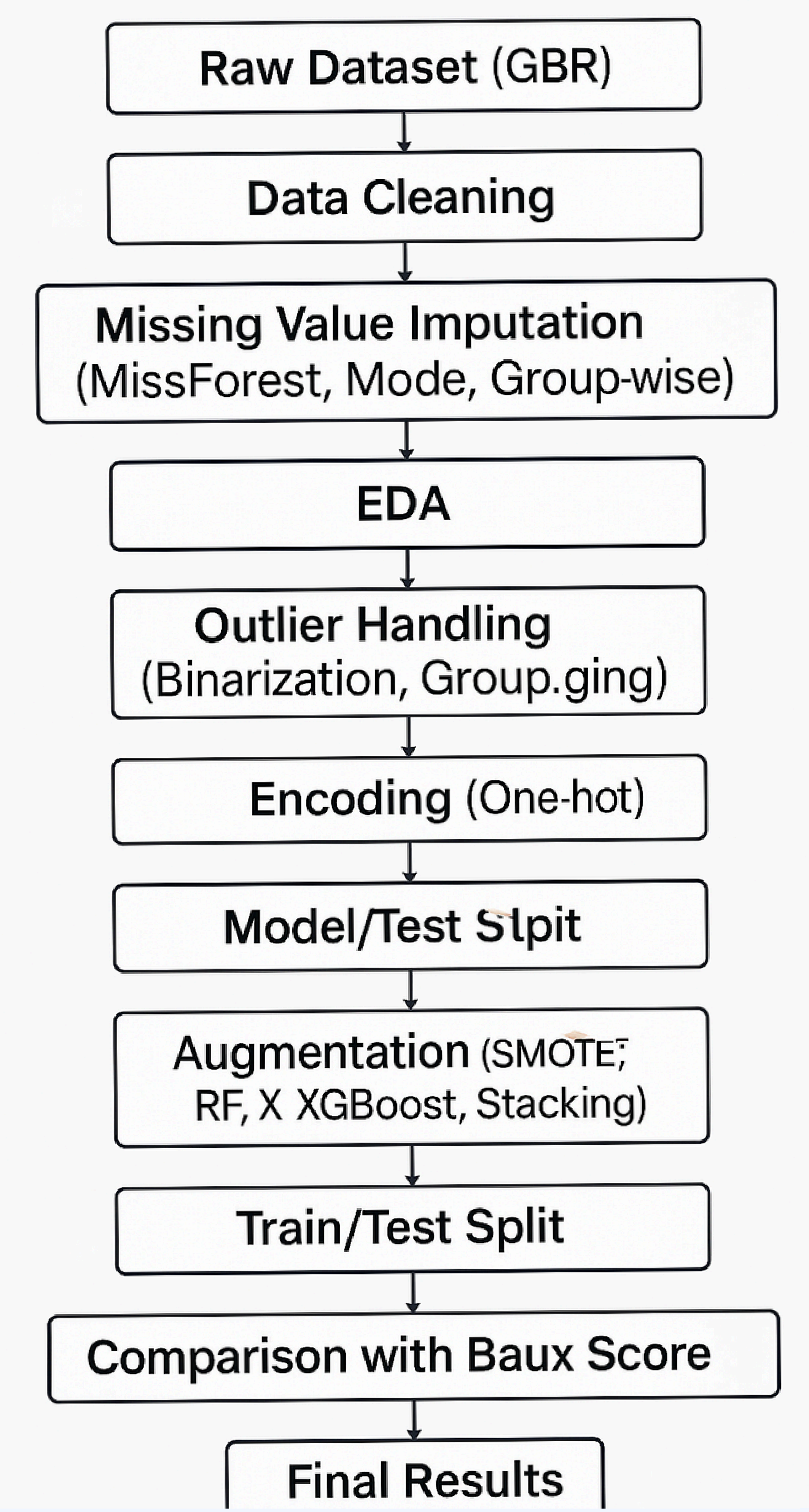


Dataset Overview

- The **Global Burn Registry (GBR)** dataset contains about **5,400 record**s of adult burn patients aged 18 and over.
- The target variable is the patient's condition upon discharge, classified as either "Dead" or "Alive."
- Key features include:
  - Demographics: age, sex
  - Injury details: Total Body Surface Area affected, smoke inhalation, associated injuries, affected body parts
  - Contextual: cause of the burn, contributing factors
  - Time-related: pre-hospital delay, length of hospital stay
  - Surgery status
- Initial data cleaning involved:
  - Removing records with missing ages
  - Excluding rows with an "Unknown" outcome
  - Addressing approximately 2,500 missing values in the "RelatedTo" field
- The dataset is valuable for analyzing factors influencing burn patient outcomes and improving care strategies.



Machine Learning Pipeline



Data Cleaning & Processing

Missing Data Handling:

- RelatedTo: 2,575 missing values filled with 'Unknown'.
- Visualized missingness using missingno bar and matrix

Multi-label and High-Cardinality Features

- Injury-related columns contained composite values (e.g., "Face;Neck")
- Demanded binarization, grouping, and dimensionality reduction

Noisy and Skewed Data Distributions

- Features like hospital stay had extreme outliers and long-tailed distributions
- Required winsorization and robust preprocessing

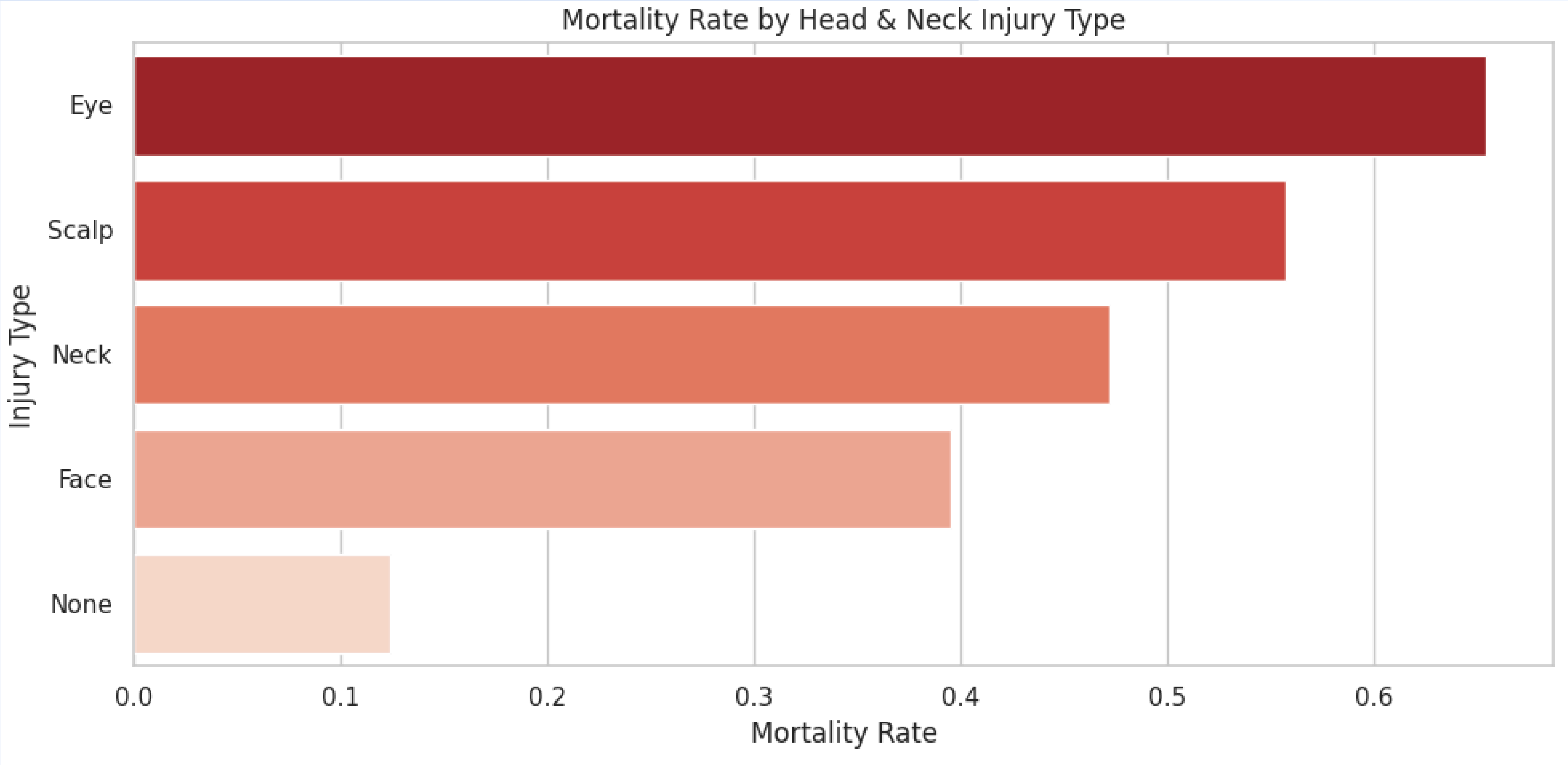
EDA- Studying Mortality rate by injury type

Method:

- Converted composite injury columns into binary flags.
- Calculated mortality rate for each body part and sub-location.
- Visualized patterns using horizontal bar plots.

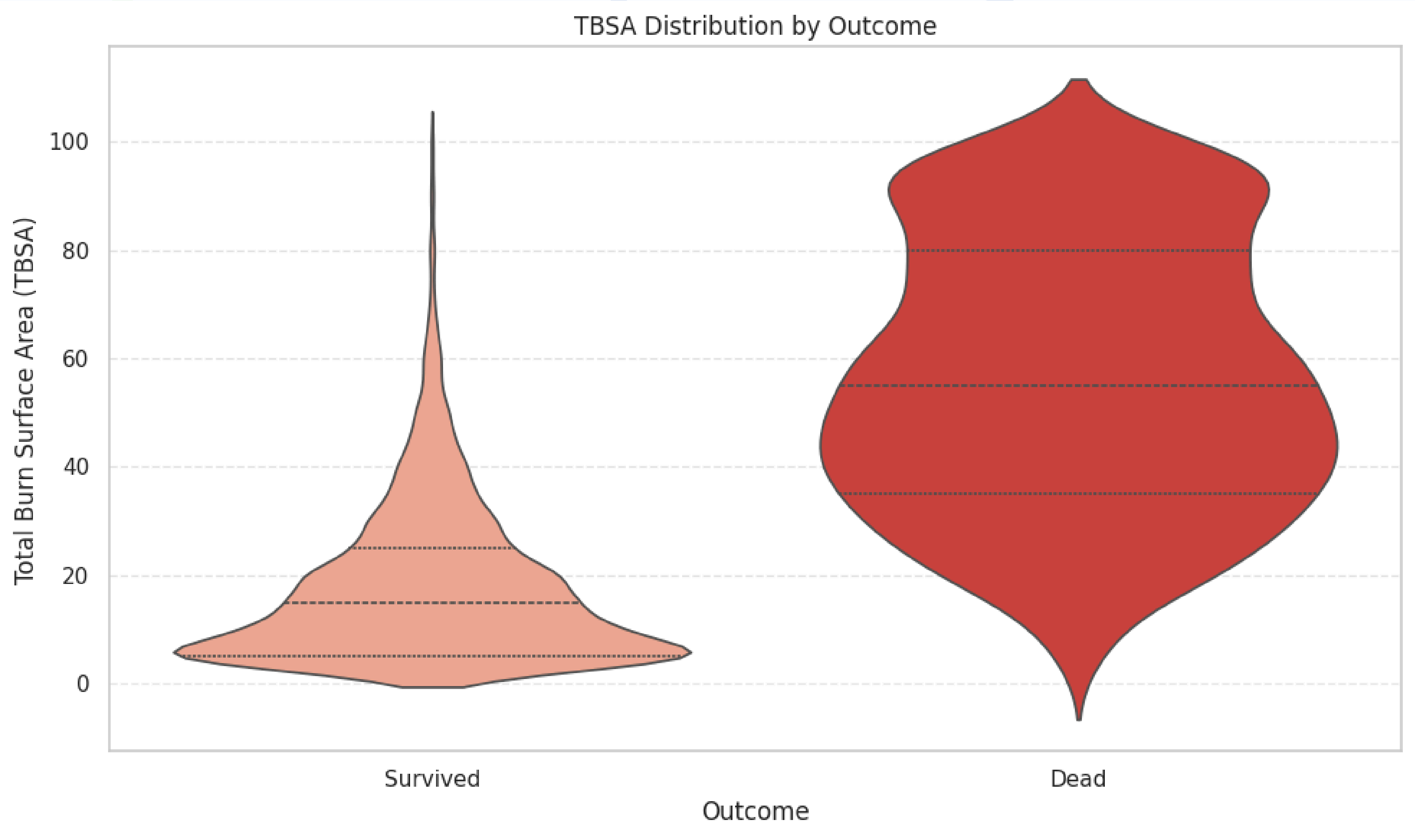
Key Findings:

- Trunk injuries (Perineum/Genitals) showed the highest mortality rate: 68%.
- Head (Eye: 66%), Legs (Knee: 51%), and Arms (Shoulder: 49%) are critical areas.
- "None" categories consistently showed the lowest mortality, serving as a baseline.
- This insight helps emphasize localized injury severity in ML models.



Injury Area	Subtype	Mortality Rate	Key Observation
Associated Injury	Abdominal trauma	**0.62**	Highest mortality in this category
	Traumatic brain injury	0.50	Also very lethal
	Spinal cord injury	0.14	Surprisingly lower mortality
	None	0.26	Baseline without injury
Head & Neck	Eye	**0.66**	Eye injury has highest mortality
	Scalp	0.56	Followed by scalp injuries
	None	0.13	Lowest risk
Trunk	Perineum or genitals	**0.68**	Highest mortality overall across all body parts
	Chest/Abdomen/Back/Buttocks	0.40	Still significantly high
	None	0.04	Very low mortality when trunk not affected
Arm	Shoulder and/or axilla	**0.49**	Highest arm-related mortality
	Elbow	0.47	Close to shoulder
	None	0.08	Safest in this area
Hand & Wrists	Fingers and/or thumb	**0.44**	Most lethal area in hand injuries
	Wrist	0.43	Slightly lower
	None	0.12	Least fatal
Legs	Knee	**0.51**	Most lethal part of the leg
	Ankle	0.45	Next highest risk
	None	0.09	Safest zone in leg injuries

EDA-Burn Surface Area (TBSA) Distribution by Outcome



Why This Matters:

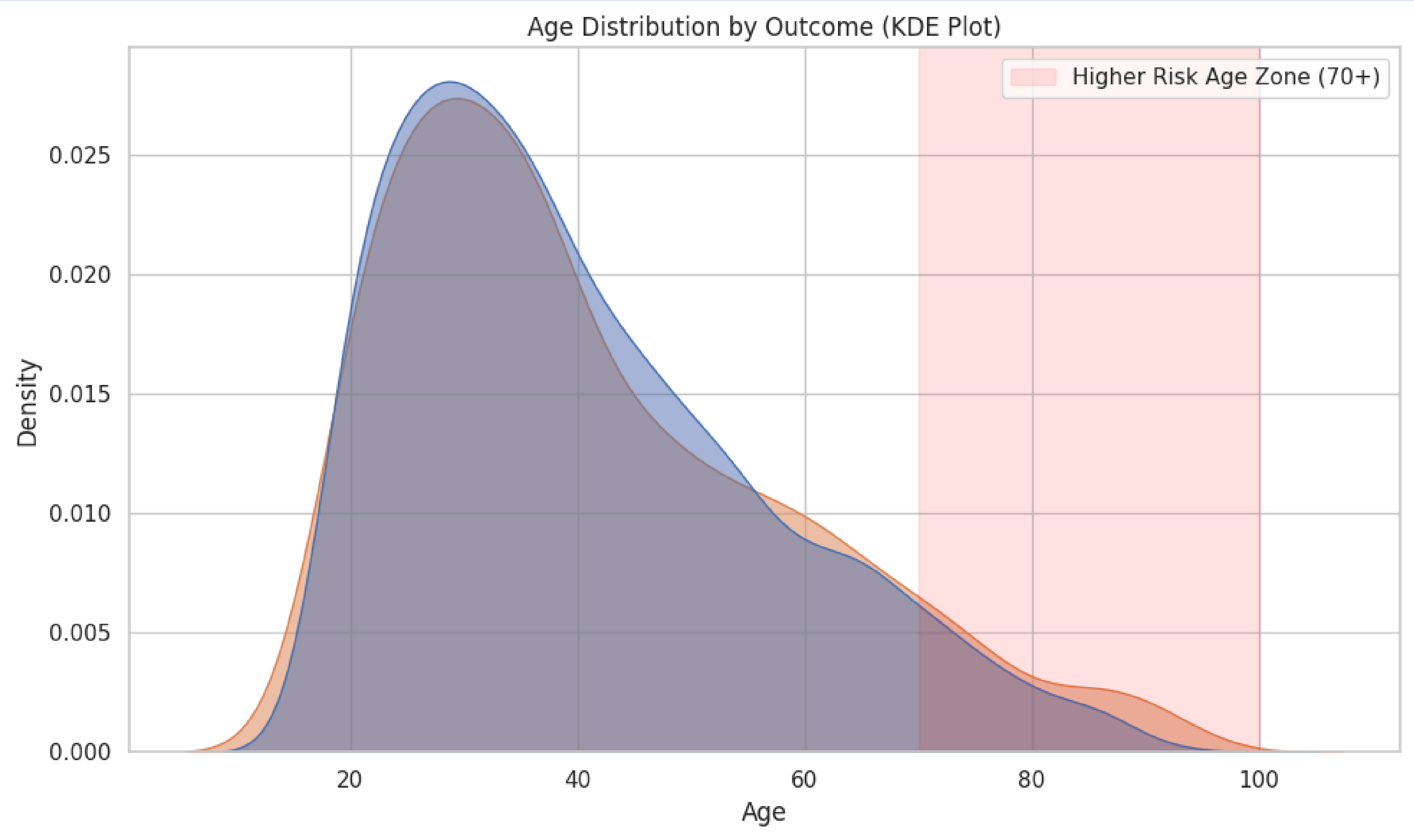
Total Body Surface Area (TBSA) burned is one of the most clinically important indicators of burn severity and patient outcome.

Key Observations:

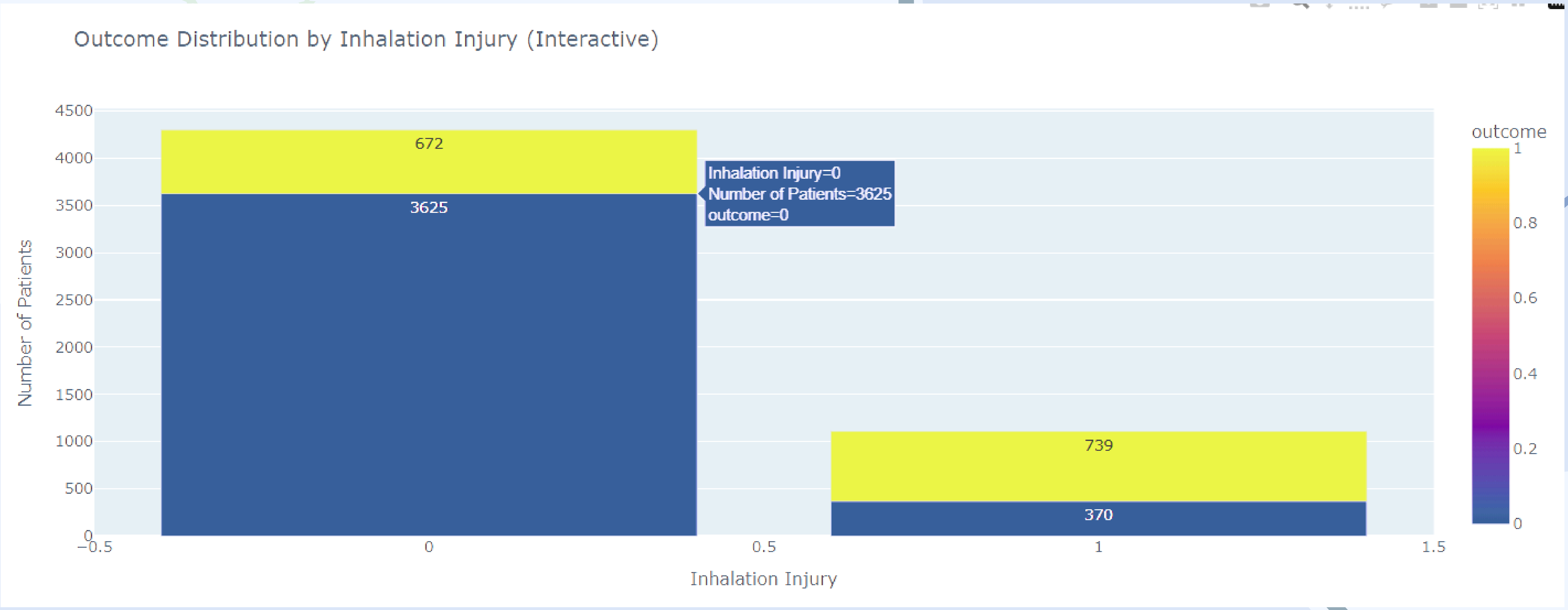
- Higher TBSA → Higher Mortality:
- Deaths are concentrated around 40–60% TBSA, with many even above 80%.
- Survivors Have Lower TBSA:
- Most survivors fall under 30% TBSA, especially around 5–10%.
- Clear Visual Separation:
- The plot shows distinct, non-overlapping peaks between outcome groups — reinforcing TBSA's role as a strong predictive feature.

EDA-Age as a Predictor of Mortality

- **Main Overlap (20–60):** Both survivors and deaths peak here, indicating age alone may not discriminate well in this range.
- **Elderly Risk Zone (70+):** The curve for deaths shifts right with a notable tail beyond age 70, aligning with clinical intuition that older adults are highly vulnerable.
- **Young Adult Advantage:** Majority of survivors are aged 20–40, showing better prognosis in younger patients.



EDA–Mortality Impact of Inhalation Injury on Mortality



Without Inhalation Injury:

Survived: 3,625 patients

Died: 672 patients

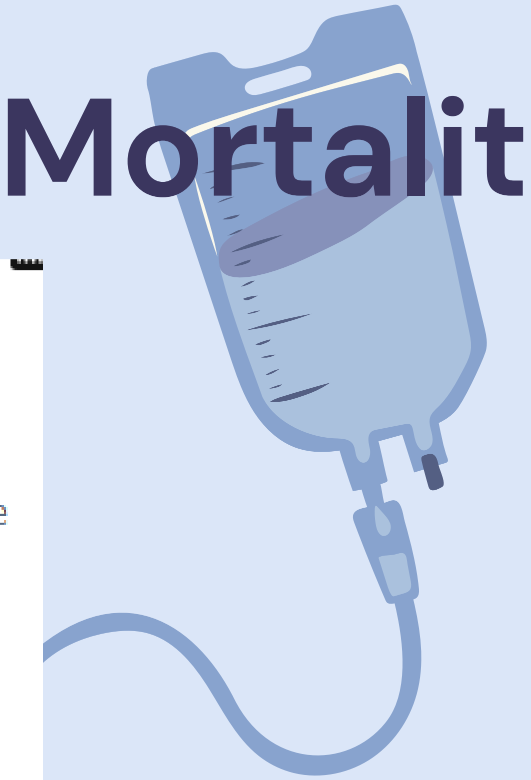
Mortality Rate: ~15.6%

With Inhalation Injury:

Survived: 370 patients

Died: 739 patients

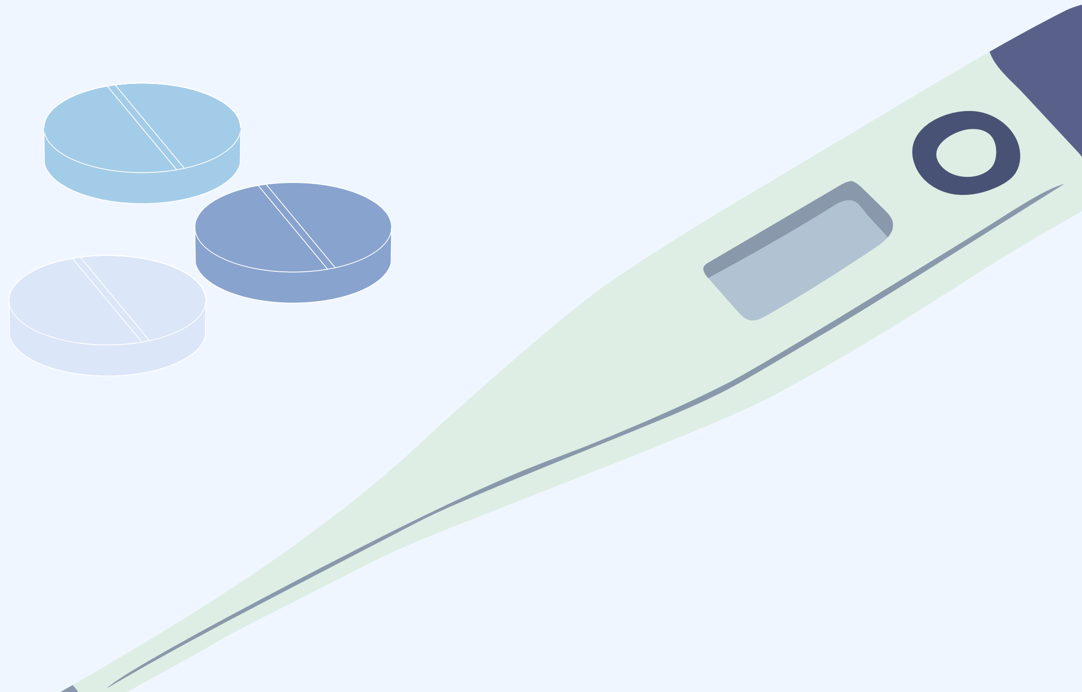
Mortality Rate: ~66.6%



EDA-Mortality Insights by Burn Cause & Sex

- **Female patients:** 33.2% mortality (732 of 2204)
- **Male patients:** 21.2% mortality (679 of 3202)

- Flame injuries dominate in both frequency and fatality (~36% mortality).
- Electrical burns show unexpectedly high mortality (~12.7%) despite fewer cases.
- Hot liquids & steam are common but have lower mortality (~4.1%).
- Inhalation as a burn cause is rare but shows serious risk.
- Cooling & radiation burns seem non-lethal in this dataset.



Handling Outliers

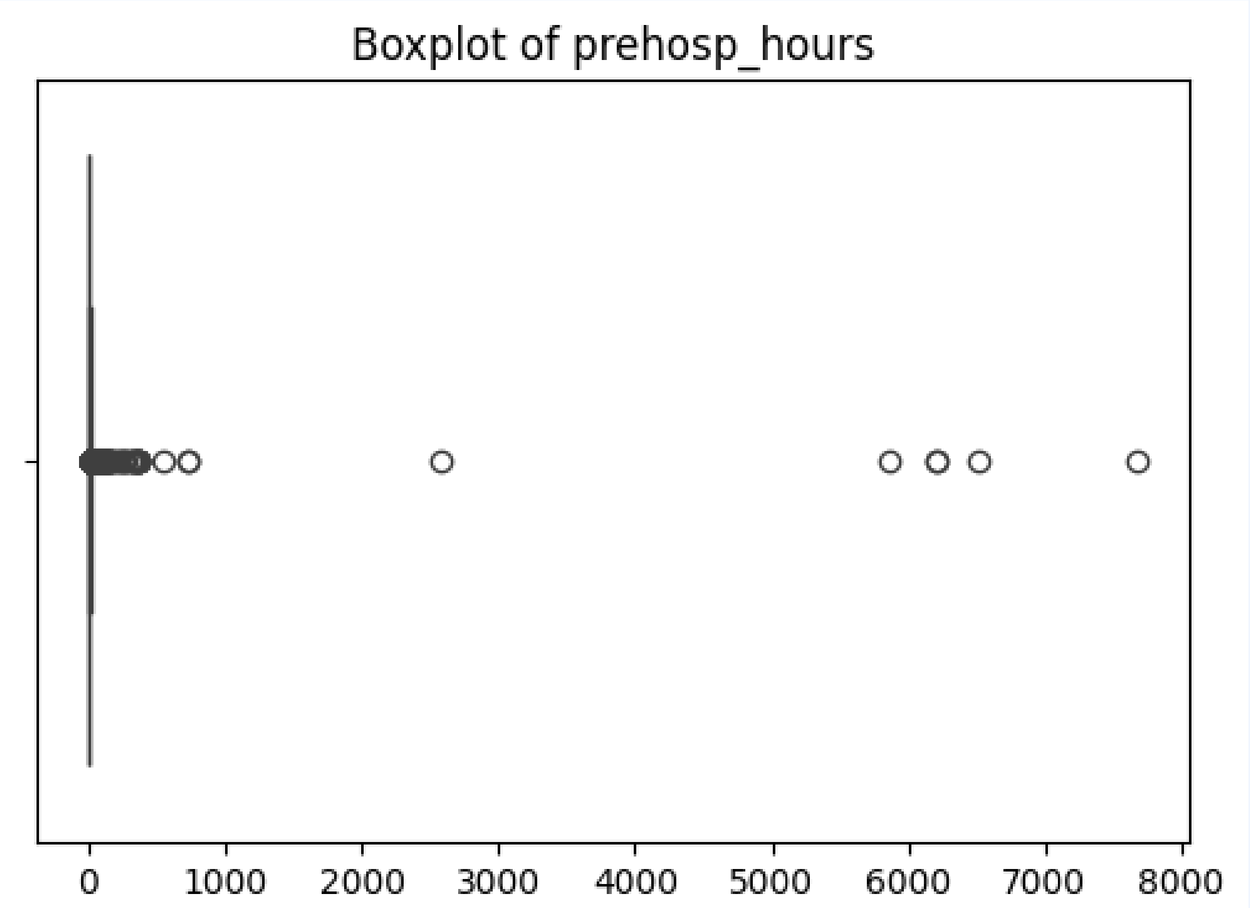
Features Reviewed: age, tbsa, pre_hosp_hours

Decision:

- age & tbsa: Kept as-is due to clinical importance.
- pre_hosp_hours: Extreme values (>7000 hrs) were winsorized to ~80 hrs.

Justification:

- Emergency cases are expected to be admitted within hours.
- Retains interpretability while eliminating outlier noise.



Rare Category grouping

Columns Processed: burn_cause, related_to, contrib_factors

Approach:

- Any category <1% frequency grouped into "Other"
- Cleaned long-tail categories like 'Playing with fire', 'Lightning', 'Dementia', etc.

Why?

- Reduces feature sparsity
- Improves model generalization & prevents noise from rare outliers

Feature Engineering
Age Grouping & Injury Score

- **Binned raw age values into meaningful clinical brackets:**
  - 18–30, 31–45, 46–60, 61–75, 76+
- **New column**: age_group
- **Helps visualize mortality trends across age cohorts and improves model interpretability.**

Created a **composite injury_burden_score**:
Sum of all binary flags from:

- **associated_injury_**
- **head_neck_**
- **trunk_**
- **arm_**
- **hand_wrists_**
- **legs_**

Feature Encoding

Binary Encoding (Label Encoding)

- Applied to strictly binary columns:
  - sex, inhalation_injury
- Mapped categorical labels to 0 or 1 using LabelEncoder
- Efficient for tree-based models that can naturally handle integers

One-Hot Encoding (OHE)

- Applied to multi-class nominal features:
  - burn_cause, related_to, contrib_factors, age_group
- Used pd.get_dummies(..., drop_first=True) to:
  - Avoid multicollinearity
  - Ensure clean, interpretable feature set

Target Label Encoding

- Final mapping:
  - 'Survived' → 0
  - 'Dead' → 1
- Ensures binary classification with 0 = Alive, 1 = Deceased

Feature Selection – Variance-Based Strategy

Technique: VarianceThreshold (Unsupervised)

- Used sklearn.feature_selection.VarianceThreshold()
- Filters out features with low variance, which carry less discriminative power
- Does not rely on the target — useful as a fast, unsupervised dimensionality filter

Rank	Feature Name	Insight
1	`tbsa`	Burn size — clinically relevant
2	`age`	Mortality risk ↑ with age
3	`prehospt_hours_winsor`	Delay before hospital entry
4	`injury_burden_score`	Cumulative severity score
5	`head_neck_Face`	High-risk burn area
6	`hand_wrists_Back of hand`	Indicator of severe trauma
7	`head_neck_None`	Absence of head/neck injury
8	`related_to_Unknown`	Context often unavailable
9	`legs_Thighs and/or lower leg`	Critical for mobility recovery
10	`hand_wrists_Palm`	Daily function impact

Feature Selection – Statistical Analysis

Methods Applied:

- ANOVA F-test: Chosen for continuous variables (e.g., TBSA, age).
- Chi-Square Test: Used for categorical binary features.
- Random Forest Feature Importance: Captures non-linear patterns.

🔝 **Top Predictive Features (Consistently Selected):**

Feature	Key Reason
TBSA	Strongest signal (F = 4853)
Injury Burden Score	Aggregated severity across injury sites
Inhalation Injury	Extremely high χ^2 score, mortality-linked
Prehospital Hours	Time delay strongly correlated with outcom
Burn Locations (e.g., arm, trunk, scalp)	Highly significant by Chi²

Statistical Tests Reveal Deep Feature–Outcome Associations

🔬 ANOVA + T-Test (for continuous features):

Feature	F-Score	p-value	Comment
TBSA	4853.9	< 0.001	Strongest predictor
Injury Burden Score	2713.0	< 0.001	Aggregated severity indicator
Prehospital Hours	69.3	~0.0	Emergency response importance

🔎 Chi-Square Test (for categorical binary features):

Feature	χ^2 Score	p-value
Inhalation Injury	944.76	1.8e-207
Trunk: Perineum/Genitals	646.26	1.4e-142
Arm: Shoulder/Axilla	591.63	1.1e-130

- Very low p-values across most features — indicating statistical significance.
- Supports clinical validity of burn site, TBSA, and inhalation injury.
- Validates our inclusion of these features in the final modeling phase.

Data Splitting & Augmentation Strategy

Train-Test Split:

- Stratified split used to ensure class distribution consistency
- Test set remained untouched throughout all experiments.

Why Augment the Data?

- The dataset was highly imbalanced
- To improve model generalization and robustness

```
       Raw Data
          ↓
    Stratified Split
          ↓
 ┌────────┬─────────┬─────────┬─────────┬──────
 │        │         │         │         │
Original  SMOTE     Tomek     SMOTE+ENN  ADASYN+ENN
 Data      Data      Data      Data       Data
```

Augmentation Techniques Applied:

- **SMOTE**: Generated synthetic samples for the minority class
- **Tomek Links:** Removed borderline noisy samples
- **ENN (Edited Nearest Neighbors)**: Removed ambiguous points
- **SMOTE + ENN**: Combined synthetic oversampling with noise filtering
- **ADASYN + ENN**: Adaptive sampling based on data density + noise cleaning

Strategy:

- Created multiple copies of the training data, each using a different augmentation technique
- Applied same modeling pipeline on each version for fair comparison
- Final results were compared across all versions to determine best-performing setup

Scaling and Preprocessing for Modeling

Objective:

- Ensure fair comparison across models by scaling feature ranges
- Prevent dominance of large-scale features (like TBSA or pre-hospital hours)

Techniques Used:

- StandardScaler applied to all datasets (Original and Augmented)
- Scaling was fit only on training data, then applied to test data (to avoid leakage)

Baux Score Modeling – Clinical Baseline

What is the Baux Score?

- Traditional clinical metric to estimate burn mortality.
- Formula: **Age + TBSA + 17 × Inhalation Injury (Yes=1)**
- Widely used in emergency care to quickly assess patient risk.

Approach Taken:

- Computed Baux Score for each patient.
- Trained a Logistic Regression model using only the Baux Score.
- Predicted probability of death from this score.

Why this Matters:

- Acts as a strong clinical baseline for comparison with ML models.
- Helps answer: "Does our model significantly outperform what doctors already use?"

Modeling Overview – Our Approach to Predicting Mortality

As we moved into modeling, our goal was simple: start with something easy to trust, then grow into more powerful techniques. Each model we used had a purpose in our journey:

Logistic Regression

We began here because it's transparent and clinically interpretable. It helped us benchmark against traditional scoring methods like Baux.

Random Forest

This gave us a step up. It captures complex patterns without overfitting easily — and we didn't have to worry about outliers or feature scaling.

XGBoost

Once we had stronger data and more features, XGBoost allowed us to push performance. It's fast, handles imbalance well, and is great for structured medical data.

Cost-Sensitive Modeling with Imbalanced Data

Metric	Logistic Regression	Random Forest	XGBoost	XGBoost Tuned	Baux Score
AUC	0.933	0.930	0.915	0.935	0.894
Recall (Dead)	0.85	0.82	0.79	0.89	0.59
F1 (Dead)	0.76	0.73	0.74	0.75	0.67

- Models were trained without SMOTE or resampling, but used class_weight='balanced' to penalize misclassification of minority (dead) class.
- Logistic Regression and Tuned XGBoost delivered the best AUC and recall, showing that internal weighting alone can be quite powerful.

Augmented Data Modelling

Model Performance on Augmented Data (SMOTE)

Model	ROC AUC	Recall (Dead)
Logistic Regression	0.917	0.81
Random Forest	0.919	0.75
XGBoost	0.911	0.73
Tuned XGBoost	0.912	0.72

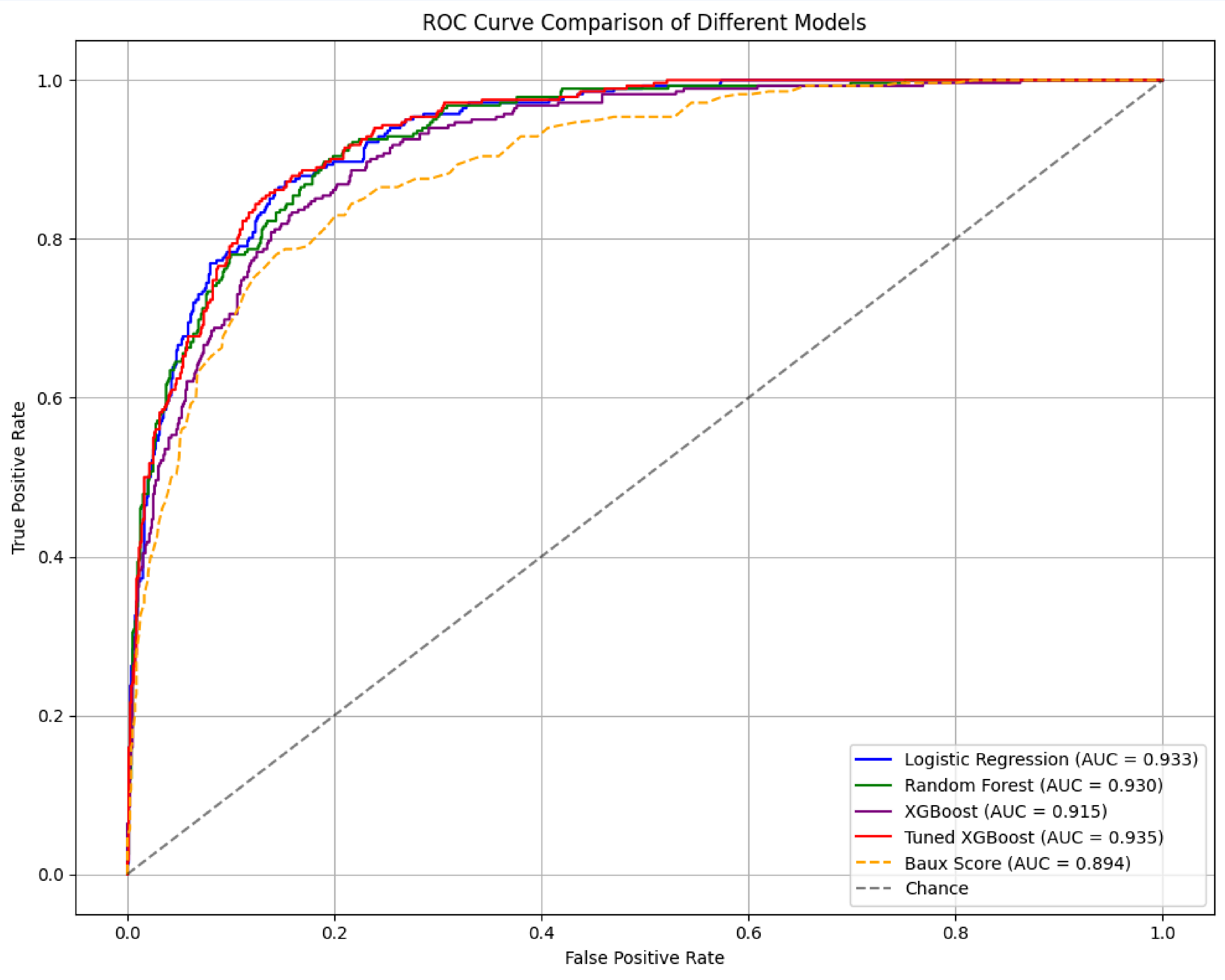
Model Performance on Augmented Data (ENN)

Model	ROC AUC	Recall (Dead)
Logistic Regression	0.933	0.86
Random Forest	0.927	0.90
XGBoost	0.920	0.89
Tuned XGBoost	0.932	0.93

- **ENN (Under-sampling)** resulted in the best overall recall, especially for Tuned XGBoost, indicating improved sensitivity to high-risk (dead) cases.
- **SMOTE (Over-sampling**) improved class balance but slightly lowered recall compared to ENN; however, Logistic Regression and Stacking maintained high AUC performance.

ROC Curve Comparison of Different Models:

- The ROC curve visualizes how well each model distinguishes between survivors and deceased patients.
- Tuned XGBoost leads with the highest AUC (0.935), followed by Logistic Regression (0.933) and Random Forest (0.930).
- Baux Score, a clinical baseline, shows an AUC of 0.894, highlighting the advantage of ML-based approaches.
- All models significantly outperform the diagonal “chance” line, showing valid learning signals.



Tuned XGBoost – Our Best Model

- Achieved highest AUC (0.935) – best at separating survival outcomes.
- Recall (0.89) ensures critical cases (deaths) are not missed.
- Tuned using GridSearchCV and scale_pos_weight to handle class imbalance directly.
- Performs consistently well across both original and augmented datasets.
- Suitable for real-world triage, ICU demand forecasting, and early alerts in clinical settings.

“When lives are on the line, missing a high-risk patient isn't an option. Tuned XGBoost gave us the best shot at catching them early.”

SHAP Analysis for Explainability

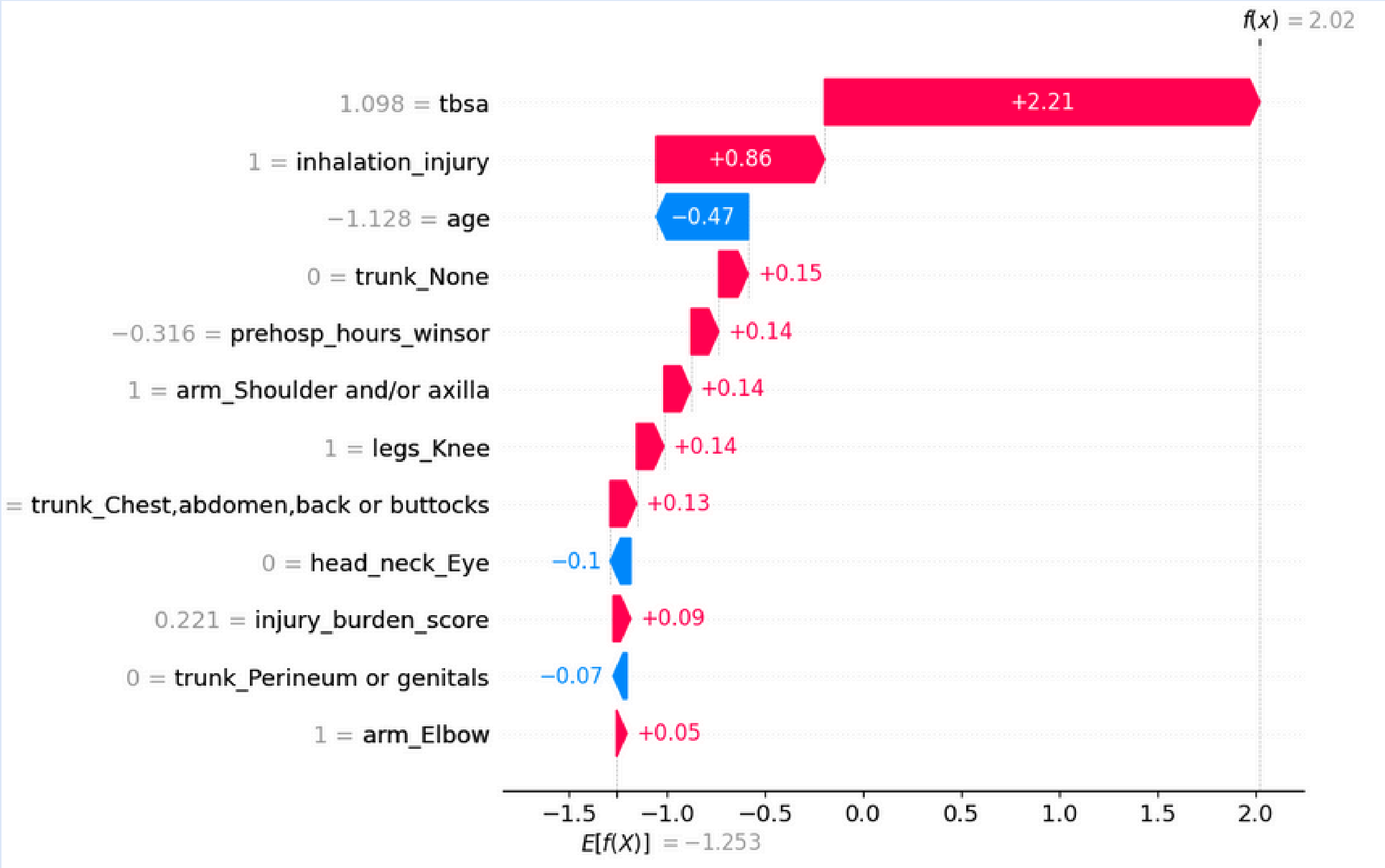
- **TBSA, Age, and Inhalation Injury remain dominant — aligning with the Baux Score.**
- **But the model goes further:**

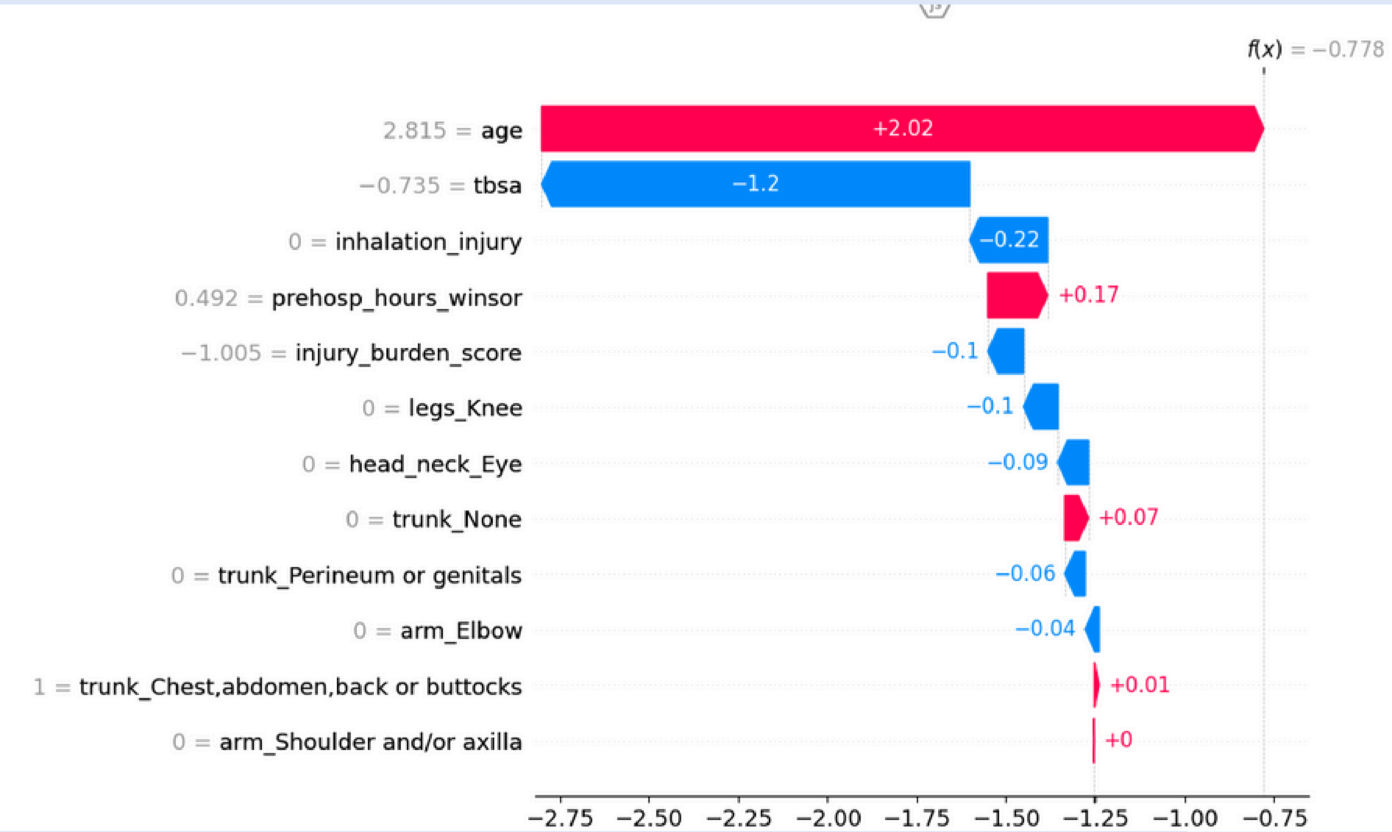
1. **Burn location** (e.g., chest/abdomen, eye, perineum) also strongly impacts prediction.
2. **Injury Burden Score** adds cumulative trauma insight — not captured in traditional scores.
3. Delayed hospital admission **(prehosp_hours_winsor)** clearly increases risk, supporting urgency-based interventions.

"SHAP shows that our model captures both well-established clinical indicators and subtle, under-recognized factors — making it a more holistic and interpretable risk predictor than the Baux Score alone."

SHAP Waterfall Plot – Patient-Level Explanation

- HAP Waterfall Plot – Patient-Level Explanation
- Red arrows represent features that increased the model's prediction toward death.
- Blue arrows represent features that pulled the prediction toward survival.
- For this patient:
- High TBSA, inhalation injury, and burns on critical body parts (e.g., chest/abdomen) pushed the prediction toward higher mortality.
- Meanwhile, younger age and shorter pre-hospital delay helped lower the risk.





These insights go beyond prediction — they help clinicians identify modifiable risk factors and support better triage in emergency care.

Post-Deployment Monitoring & Maintenance (For Future Implementation)

Although this project was research-focused, real-world clinical deployment would require ongoing monitoring to ensure the model stays reliable, fair, and useful.

What We Would Monitor

- **Model Usage**
 Track how frequently the model is used and when predictions are accepted or overridden.

- **Prediction Accuracy**
 Continuously evaluate AUC, recall, F1-score, and calibration using new patient data.

- **Data Drift**
 Detect shifts in feature distributions (e.g., age, TBSA, admission delays).

- **Concept Drift**
 Identify if the relationship between features and outcomes changes (e.g., new treatments).

- **Fairness**
 Ensure model performs consistently across patient subgroups (e.g., gender, age, injury type).

Tools & Techniques We Could Use

- EvidentlyAI, WhyLabs, or custom dashboards for live drift and performance tracking

- SHAP explanations for continuous interpretability and auditing

- Threshold alerts if recall or calibration drop below safety thresholds

Thank You...