report tanit

iyed.mekki2023

December 2025

1 Data Preprocessing

This section outlines the initial steps taken to clean and prepare the raw dataset for analysis, ensuring data integrity and patient privacy.

1.1 Duplicate Removal

To maintain the quality of the dataset, I identified and removed duplicate entries. In this clinical context, a record was defined as a duplicate based on the following criteria:

- A patient having the same **Protocol** and **Cycle Number** within a single year.

These redundant rows were eliminated to prevent data leakage and bias in the model.

1.2 Patient De-identification

To comply with data privacy standards, all patient names were de-identified using a systematic mapping strategy:

- Each patient was assigned a unique numerical ID based on their row index.
- The mapping follows the format $25x$, where x is the row number (e.g., row 32 becomes Patient ID `2532`).

1.3 Protocol Standardization

The `Protocol` column contained various inconsistencies (typos and abbreviations). These were standardized by mapping all variations to three distinct clinical categories:

1. **Agonist**
2. **Fixed Antagonist**
3. **Flexible Antagonist**

This standardization ensures that the machine learning model receives consistent categorical inputs.

2 Missing Value Imputation

2.1 Correlation Analysis

To determine the most effective imputation strategy, we conducted a study of the correlations between clinical features. As illustrated in Figure 1, we observed significant dependencies among the biomarkers. Notably, there is a strong positive correlation between the **number of follicles**, **AMH**, and **AFC** (correlation coefficients of 0.92 and 0.82, respectively).

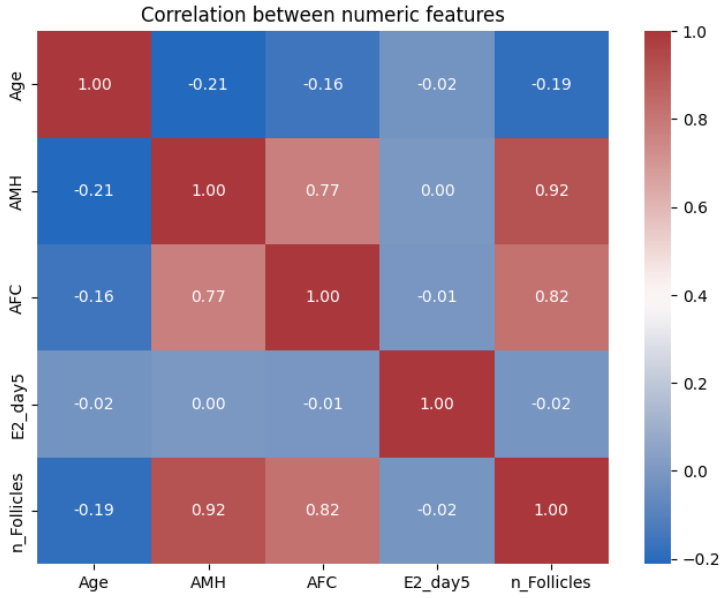

Figure 1: Correlation Matrix of Clinical Features

2.2 Imputation Methodology

Based on these findings, we implemented a hybrid imputation strategy:

2.2.1 Multivariate Imputation by Chained Equations (MICE)

We selected the MICE algorithm to impute the `AMH` and `AFC` variables. This decision was driven by the high missing rate in the `AFC` column (51.4%).

- **Why MICE?** Unlike K-Nearest Neighbors (KNN), which can become unstable when over half the data is missing, MICE utilizes the iterative,

predictive power of a regression model (Bayesian Ridge). This allows it to explicitly leverage the strong correlations established above to accurately estimate missing values.

- **Feature Retention:** Consequently, we retained `AFC` in the dataset rather than dropping it, identifying it as a critical predictor for the final patient response.

2.2.2 K-Nearest Neighbors (KNN) for Age

For the `Age` feature, we assumed a biological link with the `number of Follicles`. We applied KNN imputation to fill missing values based on this relationship. Given the low dimensionality of this specific imputation task, feature scaling was deemed unnecessary.

3 Feature Engineering

3.1 Age Discretization

To enhance the model's ability to capture non-linear biological patterns, we opted to discretize the continuous `Age` variable. Rather than treating age as a single numerical value, we categorized patients into clinically significant groups:

- **¡ 35 years**
- **35–37 years**
- **38–40 years**
- **¿ 40 years**

This transformation is justified by our correlation analysis (see Figure 2), which demonstrated that these specific age buckets capture relationships with other clinical features more effectively than the raw continuous variable alone.

3.2 Feature Evaluation and Multicollinearity

The analysis of feature relationships provided further insights for dimensionality reduction:

- **Exclusion of Non-Predictive Features:** The correlation heatmap revealed that `E2_day5` exhibits near-zero correlation with other key variables. Consequently, this feature was dropped to reduce noise.
- **Handling Multicollinearity:** We observed a high degree of collinearity between `n_Follicles`, `AFC`, and `AMH`. While strong collinearity can potentially hinder model performance (specifically in linear models), we decided to retain `n_Follicles` at this stage. It will be subjected to rigorous automated feature selection later in the pipeline to determine if its inclusion provides a net benefit over `AFC` alone.

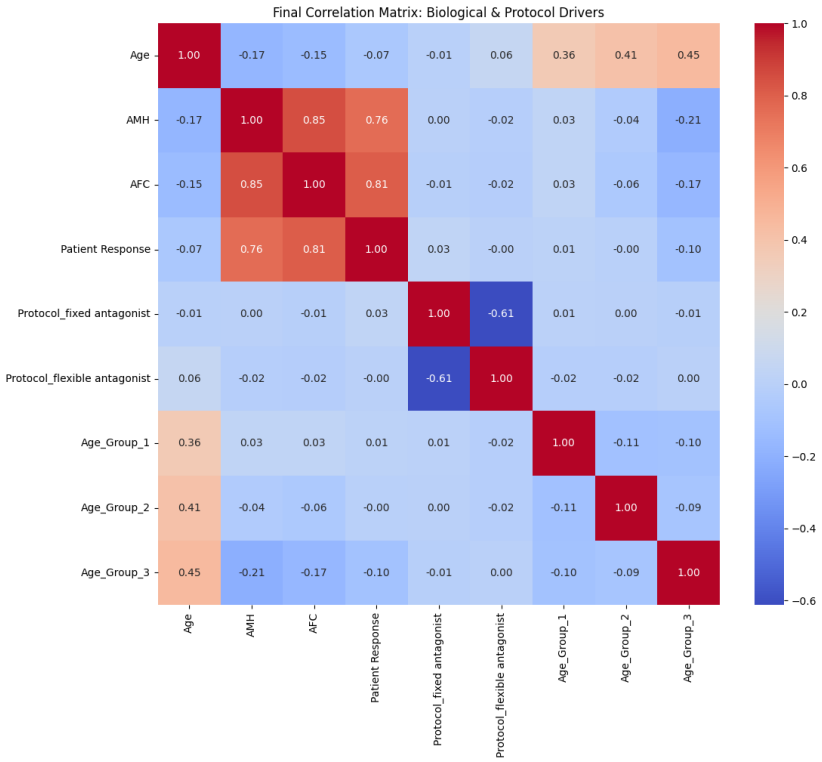

Figure 2: Correlation Heatmap of Binned Features

4 Encoding Categorical Variables

To prepare the dataset for machine learning algorithms, we applied specific encoding techniques to categorical variables:

4.1 One-Hot Encoding

We utilized One-Hot Encoding for the categorical features to prevent the model from assuming an inherent ordinal relationship where none exists (in the case of Protocol) or to allow the model to learn distinct weights for each category (in the case of Age Groups). This transformation was applied to:

- **Age Groups:** The binned age categories (*¡35*, *35-37*, *38-40*, *¿40*).
- **Protocol:** The stimulation protocol types (*Agonist*, *Fixed Antagonist*, *Flexible Antagonist*).

4.2 Target Label Encoding

For the target variable, `Patient_Response`, we applied Label Encoding to map the clinical outcomes to numerical values suitable for classification:

- **Low Response** $\rightarrow$ `0`
- **Optimal Response** $\rightarrow$ `1`
- **High Response** $\rightarrow$ `2`

5 Feature Normalization

To ensure that all features contribute equally to the model parameters—particularly for distance-based algorithms like SVM—we applied normalization techniques tailored to the specific distribution of each variable.

5.1 Standard Scaling

For the features `Age` and `AMH`, our distribution analysis revealed no significant skewness. Consequently, we applied Standard Scaling (Z-score normalization) directly to center the data around a mean of 0 with a standard deviation of 1.

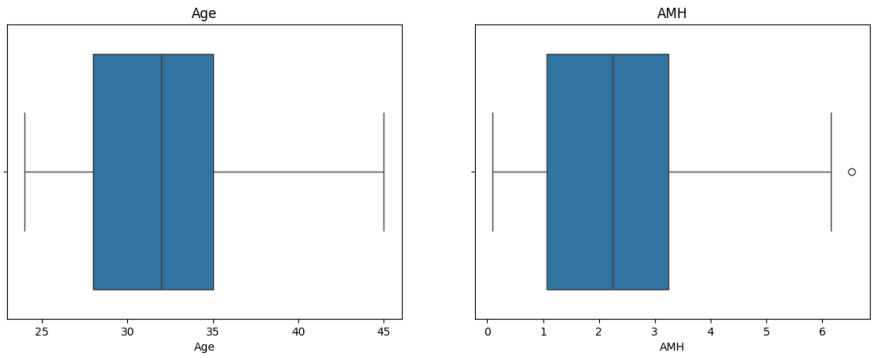

Figure 3: Boxplots of Age and AMH (Normal Distribution)

5.2 Skewness Correction

Conversely, the features `n_Follicles` and `AFC` exhibited noticeable skewness. To mitigate the impact of outliers and bring these distributions closer to normality before scaling, we employed a two-step transformation:

1. **Square Root Transformation:** Applied first to stabilize variance and reduce right-skewness.
2. **Standard Scaling:** Applied subsequently to the transformed values.

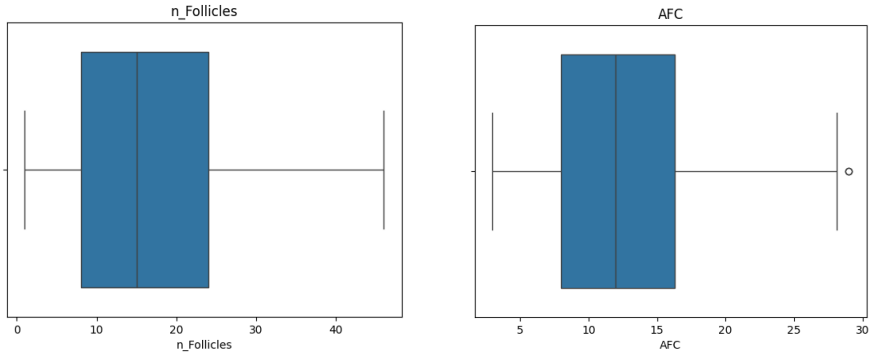

Figure 4: Boxplots of n_Follicles and AFC (Skewed Distribution)

6 Model Selection and Training

6.1 Data Distribution Analysis

Before initiating the modeling process, we inspected the distribution of the target classes to understand the dataset's balance. As illustrated in Figure 5, we observed a clear imbalance:

- **Optimal Response:** 225 samples
- **Low Response:** 153 samples
- **High Response:** 122 samples

To mitigate the bias toward the majority class, we addressed this imbalance by applying class weights (penalization) during the training of our models.

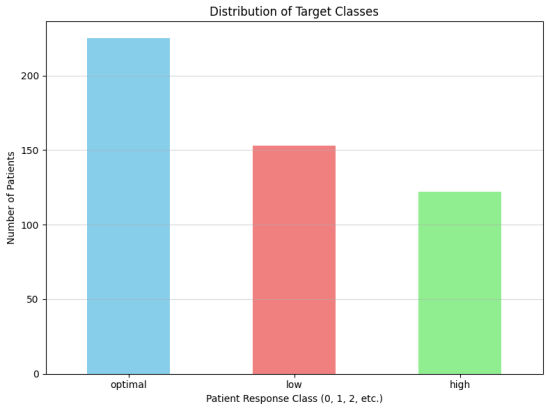

Figure 5: Distribution of Target Classes (Low, Optimal, High)

6.2 Training Methodology

The data was split using a standard ratio of **80% for training** and **20% for testing**. To prevent the model from memorizing patient identities or relying on non-predictive noise, the `patient_id` and `E2_day5` columns were discarded prior to training.

We selected four distinct algorithms for evaluation:

1. Random Forest
2. Logistic Regression
3. Support Vector Machine (SVM)
4. XGBoost Classifier

Each model underwent a rigorous Grid Search to identify the optimal hyperparameters.

6.2.1 Performance Metrics

Given the imbalance in our dataset, accuracy is an unreliable metric, as high scores can be achieved simply by predicting the majority class. Instead, we prioritized the following metrics:

- **Weighted ROC AUC:** This was our primary metric. It provides a robust, single measure for multi-class problems. By weighting the score based on class support, it offers a fair assessment of overall performance while ensuring strong discrimination for minority classes.
- **F1-Score:** We utilized the F1-Score (particularly Macro F1) to ensure a balance between precision and recall, validating that the model performs well across all classes, not just the frequent ones.

6.3 Initial Model Performance

The models were trained exclusively on the training set (X_{train}) to strictly prevent data leakage. Based on the initial evaluation, the Support Vector Machine (SVM) demonstrated superior performance.

Table 1: Model Performance Comparison on Training Data

Model	**Weighted ROC AUC**
Support Vector Machine (SVM)	**0.9318**
XGBoost Classifier	0.9252
Random Forest	0.9274
Logistic Regression	0.9176

As shown in Table 1, the SVM is currently in the lead with a Weighted ROC AUC of 0.9318. While we will retain this as our primary candidate, given the

small dataset size ($n = 500$) and low computational cost, we will continue to track the other models during the optimization phase.

7 Feature Selection

To identify the most predictive clinical markers and reduce noise, we adopted a multi-stage feature selection strategy consisting of Pearson Correlation, Hypothesis Testing, and Model-Based Selection.

7.1 Pearson Correlation Analysis

7.1.1 Methodology

In the first stage, we analyzed the Pearson correlation coefficients between the target variable (`Patient_Response`) and all candidate features. The underlying hypothesis is that features exhibiting stronger linear correlations with the target signal contribute more effectively to the classification task, while those with weak correlations likely introduce noise.

To determine the optimal cutoff, we performed a Grid Search on the correlation threshold. This process allowed us to retain only the strongest predictors while maximizing the Weighted ROC AUC metric.

7.1.2 Results

This selection step yielded immediate improvements. As illustrated in Table 2, the Support Vector Machine (SVM) maintained its lead and improved its performance:

- **Metric Improvement:** The Weighted ROC AUC increased from **0.9318** to **0.9322**.
- **Dimensionality Reduction:** This performance gain was achieved using only **4 features** (half the original feature set size).

The specific features selected by this method were:

- `Age`
- `AMH`
- `AFC`
- `Protocol_flexible antagonist`

While other models also showed performance gains after this filtration, the SVM remained the superior classifier.

Table 2: Model Performance Comparison

Model	Weighted ROC AUC
Support Vector Machine (SVM)	**0.9322**
XGBoost Classifier	0.9256
Random Forest	0.9304
Logistic Regression	0.9223

7.2 Hypothesis Testing (ANOVA)

7.2.1 Methodology

To statistically validate our feature selection and identify the most predictive clinical markers, we implemented a rigorous pipeline using **Analysis of Variance (ANOVA)**. Specifically, we utilized the `SelectKBest` algorithm with the F-test scoring function (`f_classif`).

This technique tests the hypothesis that the mean value of a specific feature (e.g., `AMH` or `Age`) differs significantly across the three outcome groups (Low, Optimal, High). Rather than arbitrarily selecting a fixed number of features, we treated the number of features (k) as a hyperparameter and optimized it via a Grid Search pipeline.

7.2.2 Results

As shown in Table 3, this method yielded mixed results relative to the correlation baseline. The Support Vector Machine (SVM) slightly improved its performance but did not surpass expectations based on prior analysis. Importantly, the feature selection identified **Age_Group_3** (the > 40 years cohort) as a statistically significant predictor. This feature was included alongside the core biomarkers, highlighting that the "biological cliff" after age 40 is a critical discriminator for the model.

Table 3: Comparison of Weighted ROC AUC for Different Models

Model	Weighted ROC AUC
Support Vector Machine (SVM)	**0.9322**
XGBoost Classifier	0.9270
Random Forest	0.9304
Logistic Regression	0.9218

7.3 Model-Based Selection: Tree-Based Metrics

7.3.1 Methodology

To complement the statistical filter methods, we employed an embedded feature selection strategy using Tree-Based metrics. When fitting decision trees,

the algorithm greedily selects the optimal split at each node to maximize purity (typically utilizing the Gini impurity metric). By aggregating these split improvements across all trees in an ensemble, a Random Forest estimator inherently calculates a "Feature Importance" score for every variable.

We integrated this logic into our pipeline by using a Random Forest classifier as the feature selector. We performed a Grid Search to determine the optimal importance threshold (e.g., "mean", "median"), allowing the model to dynamically discard features that contributed less than the average predictive value.

7.3.2 Results

This approach yielded the best performance of all selection methods. The Support Vector Machine (SVM) improved its Weighted ROC AUC score from **0.9322** to **0.9332**.

The final optimized feature set selected by this method was:

- `Age`
- `AMH`
- `AFC`
- `Age_Group_3`

Table 4: Best Weighted ROC AUC Scores for Different Models

Model	**Weighted ROC AUC**
Random Forest	0.9294
Logistic Regression	0.9218
Support Vector Machine (SVM)	**0.9332**
XGBoost Classifier	0.9272

7.3.3 Clinical Interpretation

The selection of this specific feature set provides a strong medical justification for the model's validity. It effectively captures the "Gold Standard" trifecta of ovarian reserve assessment alongside a critical non-linear age factor:

- **AMH (Biochemistry):** Provides a quantitative measure of the pre-antral follicle pool, offering a long-term view of ovarian reserve.
- **AFC (Ultrasound):** Represents the visible, functional count of follicles ready for immediate recruitment.
- **Age (Biology):** A fundamental determinant of oocyte quality and the expected rate of response.

- **Age_Group_3 (> 40 Years):** The inclusion of this binary feature is critical. Biologically, fertility does not decline linearly; it remains relatively stable before dropping off a "biological cliff," typically after age 35 and steeply after 40.

  Standard algorithms might treat Age as a purely linear variable (assuming the risk difference between age 25 and 30 is the same as 35 and 40). By explicitly selecting `Age_Group_3`, the model accounts for this non-linear threshold, recognizing that patients in this advanced maternal age cohort respond fundamentally differently to stimulation protocols compared to younger cohorts.

7.4 Model-Based Selection: Linear Models

7.4.1 Methodology

To address potential multicollinearity and enforce sparsity, we implemented a Linear Embedded Feature Selection strategy using Regularized Logistic Regression. This technique leverages the mathematical properties of L1 (Lasso) and L2 (Ridge) regularization to penalize model complexity.

By tuning the regularization strength (C) and penalty type, the Logistic Regression estimator acts as a "judge," forcing the coefficients of weak or redundant features toward zero:

- **L1 Regularization (Lasso):** Effectively "switches off" irrelevant variables by shrinking their coefficients to exactly zero, performing rigorous feature elimination.
- **L2 Regularization (Ridge):** Handles highly correlated predictors (such as `AMH` and `AFC`) by shrinking their weights together, reducing noise without necessarily eliminating the signal.

7.4.2 Results

Using this method, the Support Vector Machine (SVM) achieved a Weighted ROC AUC of **0.9323**. This score is slightly lower than the performance achieved with the Random Forest estimator (0.9332). However, this method was more aggressive in its selection, retaining only **3 features**:

- `Age`
- `AMH`
- `AFC`

Table 5: Best Weighted ROC AUC Scores for Different Models

Model	Weighted ROC AUC
Random Forest	0.9316
Logistic Regression	0.9222
Support Vector Machine (SVM)	0.**9323**
XGBoost Classifier	0.9263

7.4.3 Conclusion on Selection Strategy

While the Linear selection resulted in a more parsimonious model (3 features vs. 4), it came at a slight cost to predictive performance. In high dimensional or computationally constrained environments, this trade-off might be favorable. However, in our clinical context where the dataset is small and computational overhead is negligible-we prioritize the slight performance gain provided by the Tree-based selection (which included `Age_Group_3`) over extreme feature reduction.

7.5 Model-Based Selection: Geometric Models (LinearSVC)

7.5.1 Methodology

To provide a distinct perspective from the probabilistic (Logistic Regression) and statistical (ANOVA) methods used earlier, we implemented a Geometric Embedded Feature Selection strategy using **Linear Support Vector Classification (LinearSVC)**.

Unlike previous methods that evaluate features based on variance or probability, this technique evaluates features based on their contribution to the **decision boundary margin**. By applying L1 regularization (Lasso) within the Linear SVM, the algorithm penalizes non-informative features by forcing their coefficients to zero, theoretically isolating the specific clinical markers that define the "widest gap" between patient groups.

7.5.2 Results

Contrary to expectations, this geometric selection method did not yield performance improvements over the baseline or the Tree-based approach. The SVM trained on features selected by LinearSVC did not beat previous records. Furthermore, the selection process was ineffective at dimensionality reduction in this specific context, as the model retained **all 8 original features**, suggesting that the linear geometric margin could not easily distinguish redundant features from critical ones without the non-linear flexibility of an RBF kernel.

Table 6: Best Weighted ROC AUC Scores for Different Models

Model	**Weighted ROC AUC**
Random Forest	0.9274
Logistic Regression	0.9219
Support Vector Machine (SVM)	**0.9318**
XGBoost Classifier	0.9252

8 Final Model Training and Evaluation

8.1 Retraining with Selected Features

Having identified the optimal algorithm (SVM) and the most predictive feature subset (`['Age', 'AMH', 'AFC', 'Age_Group_3']`), we proceeded to the final training phase. We subsetted the original dataset to include only these selected variables and performed a stratified split (80% training, 20% testing) to ensure robust evaluation.

The final model performance was evaluated using the Weighted ROC AUC score and a detailed classification report.

Table 7: Final Classification Report

Class	**Precision**	**Recall**	**F1-score**	**Support**
low	0.83	0.94	0.88	31
optimal	0.84	0.82	0.83	45
high	0.86	0.75	0.80	24
Accuracy	0.84			
Macro Avg	0.84	0.84	0.84	100
Weighted Avg	0.84	0.84	0.84	100

Weighted ROC-AUC Score: 0.9220

8.2 Threshold Optimization for Clinical Safety

8.2.1 Medical Justification: The Risk of OHSS

By default, the classifier assigns a class label based on the highest probability (or a default threshold of 0.5). However, in this clinical context, the cost of misclassification is not symmetric. Specifically, missing a **High Response** is far more critical than missing a Low or Optimal response due to the severe risks associated with **Ovarian Hyperstimulation Syndrome (OHSS)**.

[Image of ovarian hyperstimulation syndrome symptoms diagram]

When ovaries are over-stimulated, they can swell and leak fluid into the body.

- **Mild Cases:** Cause discomfort, bloating, and nausea.

- **Severe Cases:** Affect 1-5% of cycles and are potentially life-threatening. Complications include thrombosis (blood clots), kidney failure, severe electrolyte imbalance, and respiratory distress.

8.2.2 Threshold Tuning

Given these stakes, we prioritized **Recall** (sensitivity) for the "High Response" class over Precision. It is clinically safer to flag a patient as a potential High Responder (even if it turns out to be a False Positive) than to miss a true High Responder who might subsequently develop OHSS.

We varied the decision threshold and selected an optimal cutoff of **0.24** for the High Response class as shown by the figure below

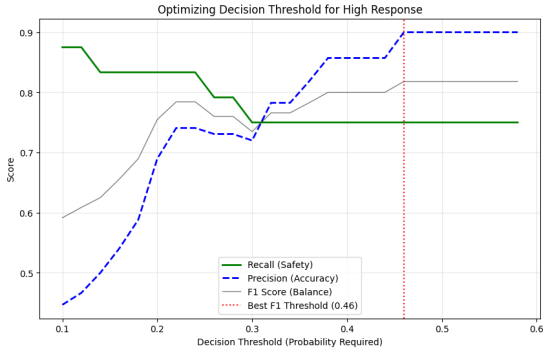

Figure 6: Your image caption here.

- **Result:** As illustrated in Figure 7, lowering the threshold significantly improved the model's ability to catch High Responders, effectively acting as an early warning system for the clinical team.

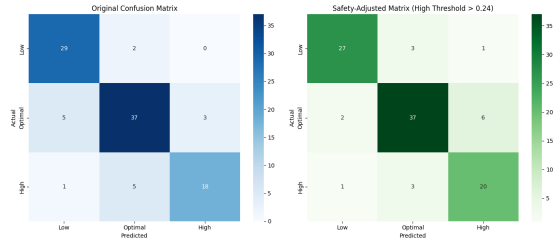

Figure 7: Confusion Matrices: Default Threshold vs. Optimized Threshold (0.24)

9 Explainable AI (XAI)

To validate the safety of our Champion SVM model, we employed SHAP to interpret its decision-making logic, specifically for the critical "High Response" outcome. As shown in Figure 8, the analysis confirms the model prioritizes **AFC** and **AMH** as the dominant risk factors, aligning with established biological principles where elevated biomarkers directly increase hyperstimulation risk. The visualization reveals that high values (red dots) of these markers consistently push predictions toward a "High Response," while Age plays a secondary, protective role. This transparency ensures the model is not merely pattern-matching but relying on valid physiological signals for clinical risk assessment.

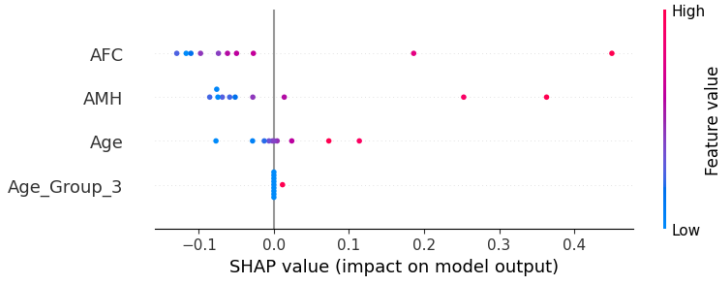

Figure 8: SHAP Summary Plot: Feature Impact on High Response Prediction

9.1 Local Prediction Logic

The SHAP Force Plot (Figure 9) provides a micro-level view of the model's reasoning for an individual patient, illustrating how specific features push the probability from the base value to the final output (0.34). In this instance, elevated **AFC** and **AMH** (red bars) act as primary drivers increasing the risk of a High Response, while **Age** (blue bar) serves as a protective counter-force. This demonstrates the model's capacity to balance conflicting biological signals-weighing reserve markers against demographic factors-to generate a nuanced, patient-specific risk assessment.

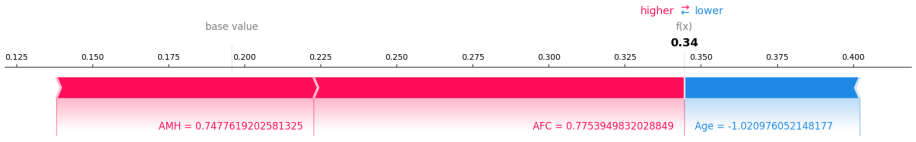

Figure 9: SHAP Force Plot: Local Explanation for a Single Patient

9.2 LIME Analysis

To cross-validate our findings, we applied LIME to a young patient (Standardized Age: -1.24), a demographic typically considered at risk for hyperstimulation. Despite this risk factor, the model correctly predicted an "Optimal Response" (86%) rather than "High" (4%). LIME reveals that this safety decision was primarily driven by the patient's moderate ovarian markers ($-0.72 < AFC \leq 0.03$ and $-0.83 < AMH \leq 0.12$), demonstrating the model's clinically nuanced understanding that youth alone is insufficient to trigger a high-risk warning without corroborating physiological evidence.

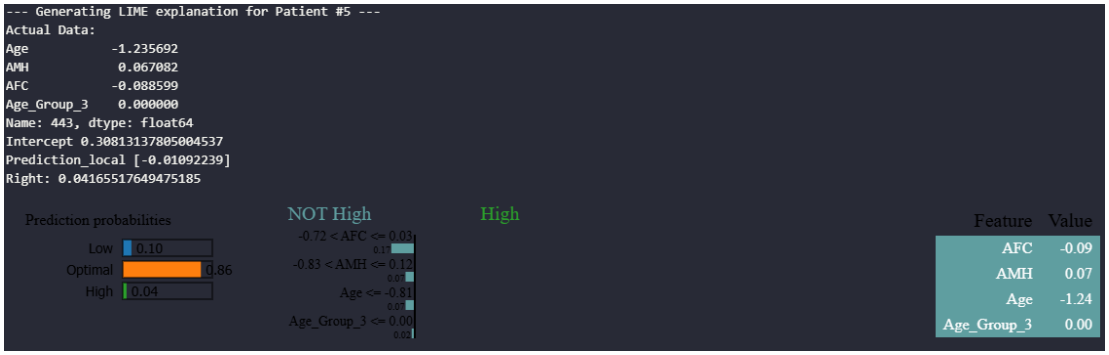

Figure 10: LIME Explanation for Patient #5: Reasons for "NOT High" Prediction

10 Trade-offs, Challenges, and Future Directions

Developing a clinical decision support tool for IVF involves navigating complex trade-offs between statistical performance and patient safety. This section outlines the critical challenges we faced and the strategic decisions made to address them.

10.1 The Precision-Recall Trade-off

The most significant challenge in this study was managing the inherent tension between Precision and Recall, particularly for the critical "High Response" class.

- **The Clinical Dilemma:** A high precision model minimizes false alarms (False Positives), ensuring that patients flagged as high risk are almost certainly high responders. However, optimizing for precision typically results in a lower Recall, meaning some actual high-risk patients might be missed (False Negatives).
- **Safety-First Strategy:** In the context of Ovarian Hyperstimulation Syndrome (OHSS), a False Negative is potentially life-threatening, whereas

a False Positive merely results in a more conservative (safer) stimulation protocol. Therefore, we consciously chose to prioritize **Recall** over Precision.

- **Quantifiable Impact:** By lowering the decision threshold to **0.24**, we accepted a decrease in Precision (resulting in more false alarms) to maximize the capture rate of true High Responders. This strategic sacrifice ensures the model acts as a highly sensitive "safety net," effectively minimizing the risk of a patient developing severe OHSS unnoticed.

10.2 Limitations and Potential Solutions

While the current model performs robustly, we identified specific limitations that offer avenues for future improvement:

- **Advanced Feature Engineering:** Our current feature set focused on individual biomarkers. Future work should investigate explicit **feature interactions** (e.g., AMH/Age). While our non-linear SVM captures these relationships implicitly, explicitly engineering these interaction terms could help linear models or simpler tree-based estimators capture the synergistic effect of biomarkers more effectively.
- **Ensemble Methods:** Although the SVM was the single best performer, a **Voting Classifier** (combining the predictions of the SVM, Random Forest, and XGBoost) could potentially offer superior stability. By aggregating the decision boundaries of multiple algorithms, a soft-voting ensemble might smooth out individual model biases and yield marginal but valuable performance gains in this sensitive clinical domain.