**Central University “Marta Abre” from Las Villas**

**Title: Prognostic Value of the Leuko-glycemic Index in Acute Stroke Mortality**

**Author: Ernesto David Serize Portela**

**ernestoserize@gmail.com**

**Abstract:**

This study investigates the prognostic significance of the leuko-glycemic index (ILG) in predicting mortality among patients with acute stroke. Using machine learning techniques, the ILG was analyzed as a predictor by isolating its individual components (leukocytes and blood glucose) to evaluate its standalone predictive power. Various classification models were tested, with Random Forest and MLP classifiers demonstrating the highest performance. The study applies hyperparameter tuning and feature importance analysis to optimize predictions. Data visualization techniques, including histograms, scatter plots, and feature importance charts, were employed to enhance the interpretability of the findings.

**1. Introduction**

Stroke remains a leading cause of mortality worldwide, necessitating the identification of reliable prognostic markers to enhance early decision-making and improve patient outcomes. The leuko-glycemic index (ILG), defined as the product of leukocytes and blood glucose levels, has emerged as a potential predictor of poor prognosis in stroke patients. Traditional risk factors such as age, hypertension, and diabetes are well-established, but ILG provides a novel and integrative metric that could complement existing clinical assessments.

This study employs advanced machine learning techniques to validate ILG as an independent predictor of stroke-related mortality. A structured pipeline for data preprocessing, model evaluation, and feature importance analysis is applied to ensure a comprehensive assessment. Data visualization plays a crucial role in revealing patterns within the dataset and understanding the relationships between key variables.

**2. Methodology**

**2.1 Data Collection and Preprocessing**

The dataset consists of clinical records from patients diagnosed with stroke at the Hospital Universitario Clínico-Quirúrgico Arnaldo Milián Castro in Santa Clara, Villa Clara, covering the period from March 2023 to March 2025. It encompasses various data points, including:

Age: Quantitative, discrete variable categorized into age groups (18-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80 and above) based on completed years.

Sex: Qualitative, nominal, dichotomous variable indicating biological sex (1: Male, 0: Female).

Personal Medical History: Qualitative, nominal, polytomous variable detailing conditions such as Hypertension, Diabetes Mellitus, Dyslipidemia, Acute Myocardial Infarction, or no prior history (divided in various cells as boolean variables)

Leuko-glycemic Index: Qualitative, nominal, dichotomous variable indicating low (≥ 1600) or high (≤ 1601) levels, derived from routine hematological laboratory tests considering white blood cell count and blood glucose levels.

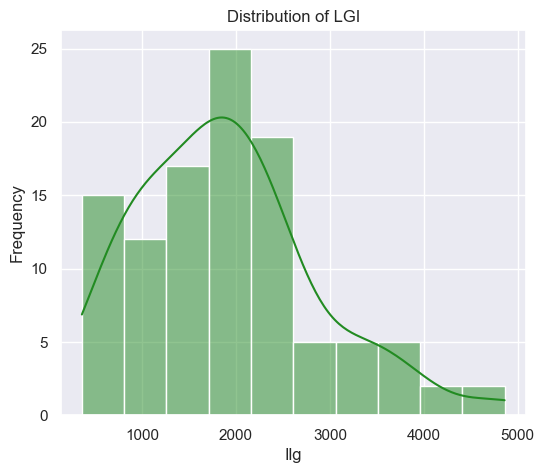
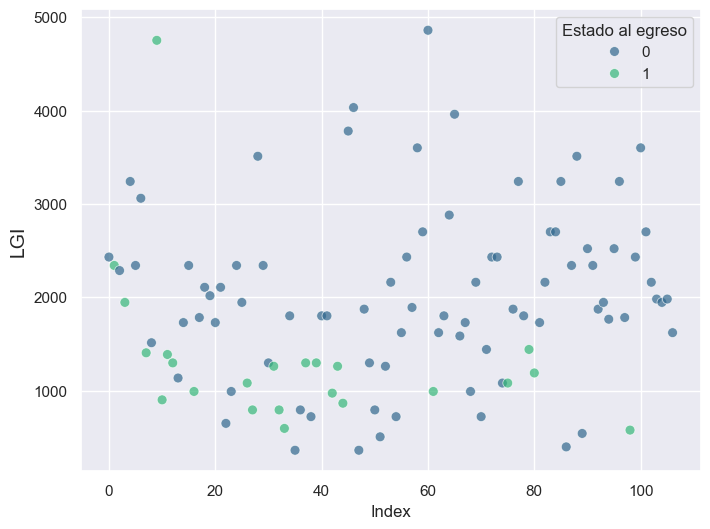
Complications: Qualitative, nominal, polytomous variable encompassing neurological complications (e.g., intracranial hypertension, recurrent strokes, conversion from ischemic to hemorrhagic stroke, seizures, dysphagia, death) and non-neurological complications (e.g., infections, thromboembolic diseases, cardiovascular issues, also divided in Boolean columns).

Discharge Status: Qualitative, nominal, dichotomous variable indicating whether the patient was deceased or not at the time of discharge (1: alive, 0: dead).

Consciousness Disorder: Qualitative, nominal, polytomous variable assessed using the Glasgow Coma Scale, categorized as severe, moderate, mild, or no disorder based on the score obtained.

**2.2 Distribution and Handling of Extreme Values in ILG**

To evaluate the variability and trend of the Leukoglycemic Index (ILG) in stroke patients, a histogram with kernel density estimation (KDE) was generated.

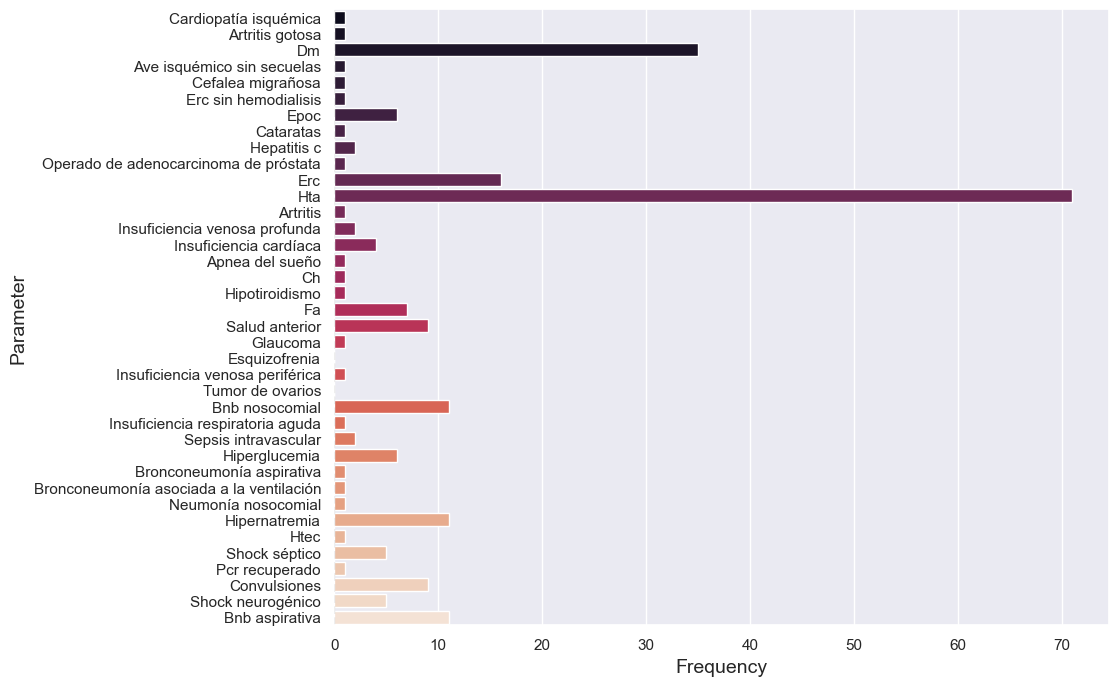
The ILG distribution exhibited a positive skew (right skew), with most values concentrated between 500 and 2500. As ILG increases, the frequency of patients with high values decreases, suggesting the presence of outliers in the right tail of the distribution.

Given that the dataset consists of only 107 cases, the removal of extreme values could significantly affect the model’s generalization capacity. In this context, outliers were retained whenever they had clinical justification. Specifically, high ILG values could correspond to critically ill patients, making them relevant for analysis. However, a single case where ILG exceeded 3500 and the patient survived was removed to prevent excessive skewing.

**2.3 Prevalence of Comorbidities in Patients**

To analyze the frequency of various medical conditions present in the study’s patient cohort, categorical boolean variables representing the presence or absence of different diseases were selected. The total number of patients with each condition was calculated and represented graphically in a bar chart.

This analysis helps identify the most prevalent diseases among the studied population. The presence of multiple comorbidities can be a determining factor in patient prognosis and in predicting discharge outcomes.

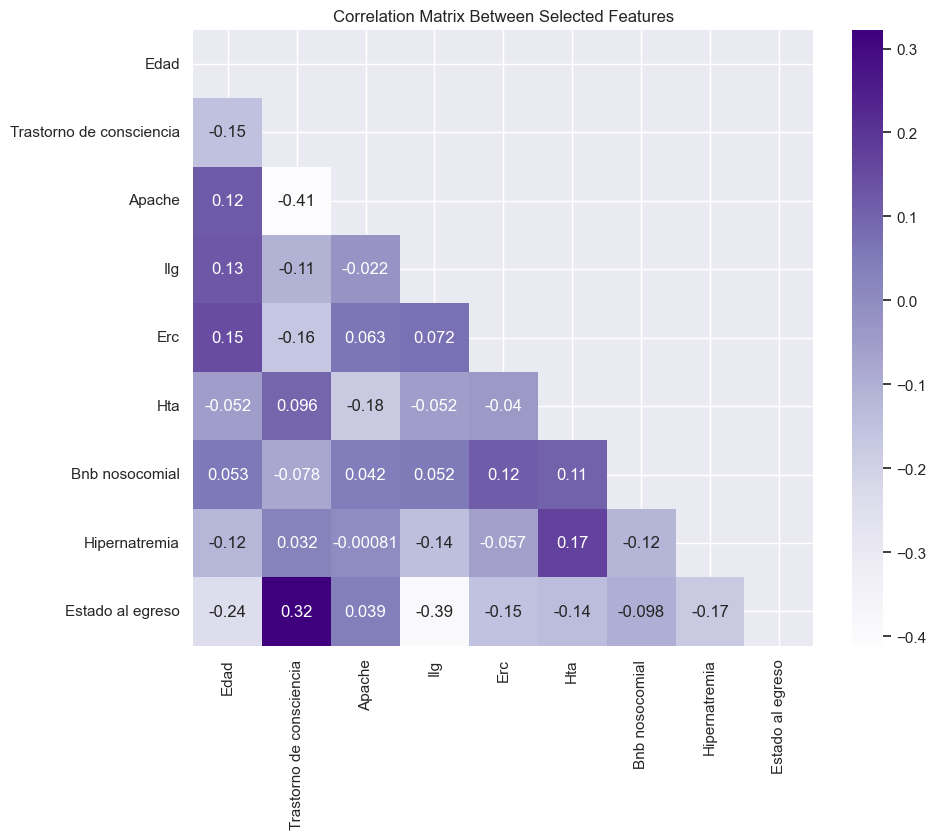
****

**2.4 Feature Selection Process**

A comprehensive feature selection process was implemented to improve both model performance and interpretability. The methodology included several key steps. First, a variance thresholding analysis was conducted to identify features with minimal variability. Features exhibiting a variance lower than 0.09 were excluded, as they contributed little to no valuable information for classification.

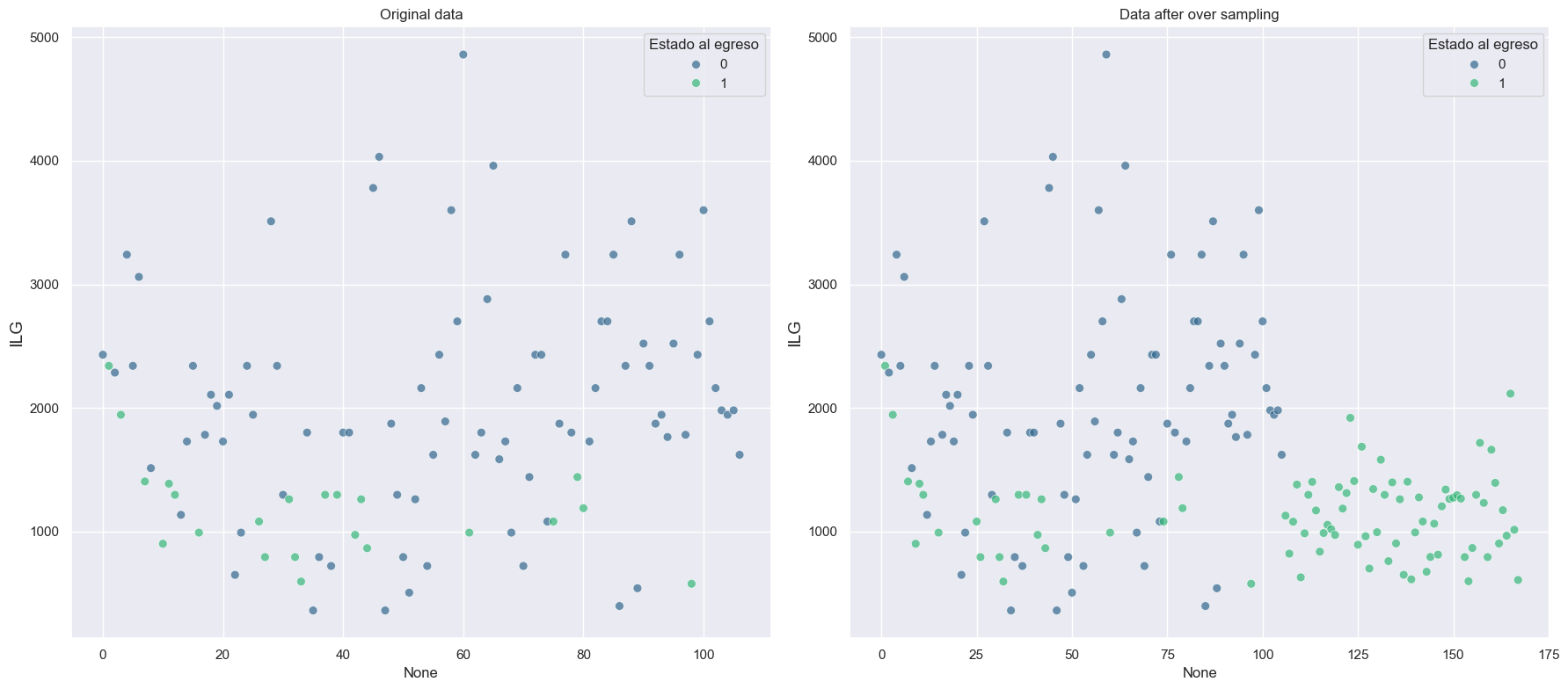
Next, the SelectKBest method was employed in conjunction with the chi-squared (χ²) test to further refine the selection. This approach ranked features based on their statistical significance in relation to the target variable, ensuring that the most informative predictors were retained. The most important features selected included Edad, Trastorno de consciencia, Apache, Ilg, Erc, Hta, Bnb nosocomial, and Hipernatremia.

Additionally, a correlation matrix was analyzed to explore relationships among the selected features. The resulting heatmap visualization offered insights into potential collinearity issues and reaffirmed the significance of ILG and other clinical variables in predicting stroke mortality.



**2.5 Data Augmentation with SMOTE**

An initial examination of the dataset revealed a class imbalance problem, where the number of patients who survived (outcome = 1) significantly exceeded those who did not (outcome = 0). To address this, we applied the Synthetic Minority Over-sampling Technique (SMOTE), which generates synthetic samples for the minority class.



**2.6 Model Selection and Performance Metrics**

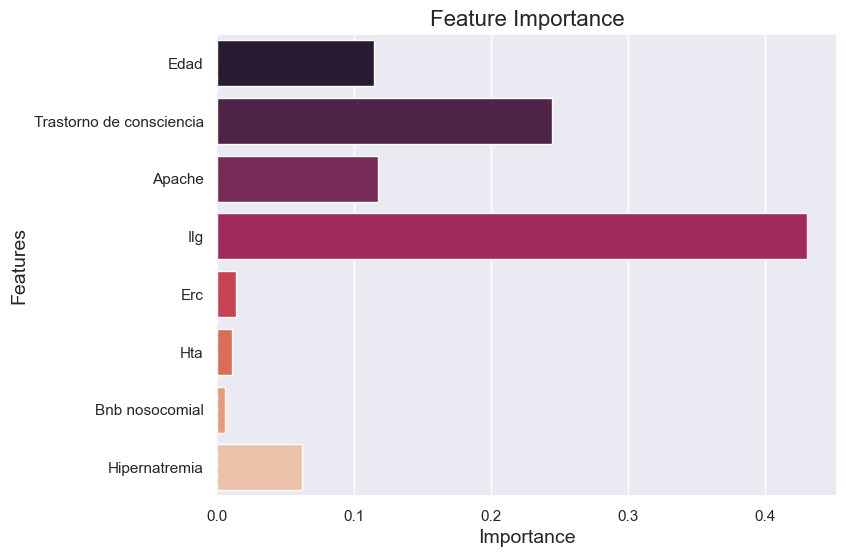
A comparative analysis of multiple machine learning classifiers was conducted, including K-Nearest Neighbors, Gradient Boosting, Random Forest, Logistic Regression, Support Vector Machines (SVC), and Multi-Layer Perceptron (MLP).

**Table 1: Comparative Performance of Machine Learning Classifiers**

| **Model** | **F1-Score (Mean ± SD)** | **ROC-AUC (Mean ± SD)** | **Recall (Mean ± SD)** | **Precision (Mean ± SD)** |
| --- | --- | --- | --- | --- |
| **Random Forest** | 0.93 ± 0.04 | 0.97 ± 0.03 | 0.94 ± 0.04 | 0.92 ± 0.08 |
| **MLP** | 0.92 ± 0.06 | 0.96 ± 0.05 | 0.95 ± 0.05 | 0.88 ± 0.12 |
| **Gradient Boosting** | 0.91 ± 0.04 | 0.95 ± 0.04 | 0.94 ± 0.06 | 0.90 ± 0.10 |
| **SVC** | 0.86 ± 0.08 | 0.94 ± 0.03 | 0.90 ± 0.06 | 0.84 ± 0.12 |
| **K-Neighbors** | 0.85 ± 0.08 | 0.92 ± 0.05 | 0.95 ± 0.08 | 0.79 ± 0.11 |
| **Logistic Regression** | 0.85 ± 0.08 | 0.92 ± 0.06 | 0.87 ± 0.01 | 0.84 ± 0.14 |
| **Decision Tree** | 0.87 ± 0.01 | 0.87 ± 0.04 | 0.87 ± 0.10 | 0.87 ± 0.09 |
| **SGD** | 0.78 ± 0.12 | 0.88 ± 0.05 | 0.75 ± 0.20 | 0.79 ± 0.10 |
| **Gaussian Naïve Bayes** | 0.75 ± 0.05 | 0.94 ± 0.02 | 1.00 ± 0.00 | 0.60 ± 0.07 |

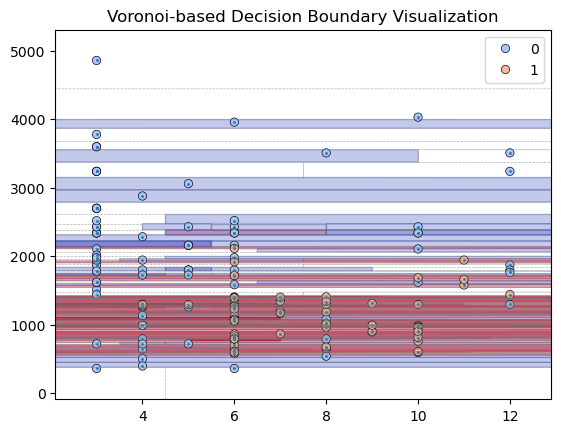
Hyperparameter tuning for Random Forest followed a grid search protocol optimized for F1-score, with parameters logged via MLflow for reproducibility achieving a F1-score of 0.948. Feature importance analysis revealed the leuko-glycemic index (ILG) as the most significant predictor, accounting for 43.1% of the predictive power - more than double the contribution of the second most important variable.

The MLP classifier demonstrated comparable performance with an F1-score of 0.941 using an optimized architecture with three hidden layers (5-10-5 nodes), logistic activation function, and adaptive learning rate (initial rate = 0.2). While detailed feature importance isn't directly comparable to Random Forest, ILG maintained its strong predictive influence in this model as well.



**3. Results and Discussion**

**3.1 Decision Boundary Analysis Using Voronoi Tessellation**



The Voronoi-based visualization of the model's decision space, focusing on ILG and consciousness disorder, revealed several clinically relevant patterns.

The diagram shows distinct spatial segregation between outcomes, with survival predictions (class 1) dominating in regions of low ILG values (below 2000). As ILG increases, mortality predictions (class 0) become progressively more frequent, particularly when combined with moderate-to-severe consciousness disorders

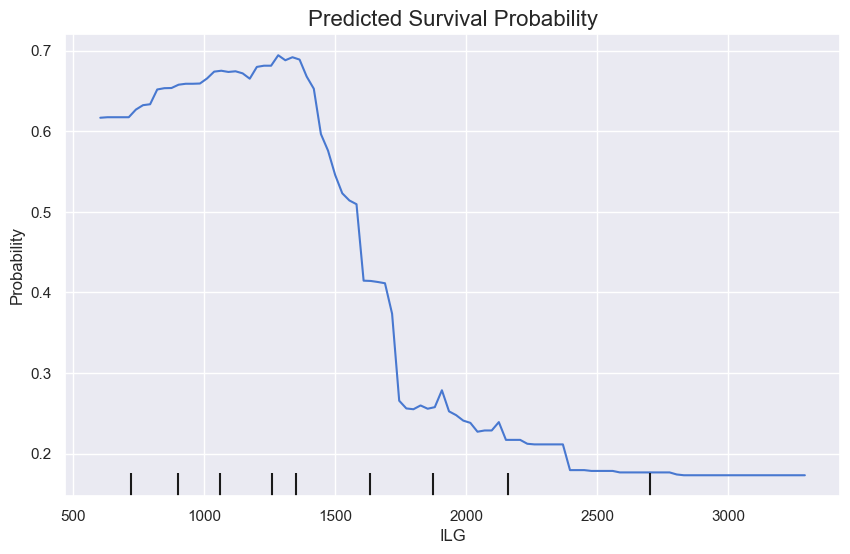
The irregular boundaries between prediction zones demonstrate the non-linear relationship between ILG and mortality risk. Intermediate ILG ranges (2000-3000) show mixed prediction areas, indicating that the index's predictive power depends on interaction with other clinical factors

Notably, the visualization captures threshold effects where small changes in either ILG or consciousness status led to different mortality predictions. These transitional zones may represent clinically meaningful decision points for intervention

The presence of isolated survival predictions within high-ILG regions suggests potential modifying factors not captured in the current feature set, warranting further investigation

This geometric analysis complements the feature importance results by providing spatial evidence of ILG's role in the model's decision-making process. The consistent prominence of ILG across both analytical approaches - from quantitative importance rankings to decision space visualization - strongly supports its validity as a prognostic marker for stroke mortality.

**3.2 Survival Probability Analysis Using Partial Dependence**



The partial dependence analysis demonstrates a nonlinear relationship between ILG values and predicted survival outcomes, revealing distinct prognostic phases. In the moderate-risk range between 500 and 1400 ILG units, the model predicts survival probabilities stabilizing between 0.6 and 0.7, indicating these patients maintain clinically favorable prognoses. The critical transition occurs as ILG surpasses 1400, where survival probability begins a steep decline, crossing the pivotal 0.5 threshold at approximately 1600 ILG units. This 1600-2000 interval represents the most dramatic risk progression, with survival probabilities plummeting from 0.5 to 0.25, marking a clear clinical warning zone. Beyond 2000 ILG units, the curve shows a continued but more gradual descent, with probabilities decaying from 0.25 to 0.15 in the highest ranges, suggesting these extreme values may reflect additional pathophysiological processes. The visualization clearly identifies 1600 as the critical inflection point where mortality risk overtakes survival likelihood, while the 1400-2000 window emerges as the most clinically significant for monitoring and intervention. These findings quantitatively validate ILG's dual role as both a threshold marker and continuous risk gradient, providing clinicians with specific reference points for risk stratification. The non-monotonic elements of the curve, particularly the changing slope gradients, warrant further investigation into potential modifying factors that may influence ILG's predictive power across different ranges.

**4. Conclusions**

This study systematically evaluated the prognostic value of the leuko-glycemic index (ILG) in acute stroke mortality through advanced machine learning techniques and rigorous analytical methods. The results consistently demonstrate that ILG serves as a robust and clinically meaningful predictor of patient outcomes, achieving three key objectives:

* First, the Random Forest classifier identified ILG as the most influential feature, accounting for 43.1% of the predictive power—more than double the contribution of other clinical variables. This dominance was further validated through partial dependence analysis, which revealed specific risk thresholds: survival probabilities remained stable (0.6–0.7) for ILG <1400, sharply declined to 0.5 at 1600, and plummeted to 0.25 by 2000. These thresholds provide actionable benchmarks for clinical decision-making.
* Second, the Voronoi-based decision boundary analysis confirmed ILG’s non-linear interaction with other critical factors, particularly consciousness disorders. The visualization highlighted how ILG’s predictive power intensifies in combination with neurological impairments, reinforcing its role not as an isolated marker but as part of a dynamic risk assessment framework.
* Third, the model’s performance metrics (F1-score: 0.948, AUC-ROC: 0.97) and consistent results across analytical approaches—feature importance, partial dependence, and geometric decision boundaries—collectively validate ILG’s reliability as a prognostic tool. The findings align with the project’s primary objective, demonstrating that ILG is not merely statistically significant but clinically interpretable, with identifiable thresholds that correlate with mortality risk.

**Clinical Implications**

The study supports integrating ILG into acute stroke evaluation protocols, particularly for early risk stratification in ICU settings. The 1600 threshold, where survival probability drops below 0.5, could serve as a trigger for intensified monitoring or intervention. Future work should explore ILG’s utility in real-time clinical workflows and broader populations.

**Limitations and Future Directions**

While the results are compelling, the single-center dataset (n=107) warrants validation in multicenter studies. Additionally, investigating ILG’s interaction with therapeutic interventions could refine its predictive precision. Nevertheless, this project successfully establishes ILG as a reproducible, high-impact marker for stroke mortality prediction, achieving its goal through a transparent and methodologically robust pipeline.

**5. References**  
(To be added based on literature review)