BioPredict: Data-Driven Modeling of Serological Markers in Plasma Donors

Alex Roldan

June 30, 2025

Abstract

This project addresses the inefficient and costly screening of plasma donors by developing BioPredict, a predictive modeling system for Grifols. The primary objective is to build an interpretable machine learning model that identifies donors with a high probability of being positive for specific serological markers, using Hepatitis B as a proof-of-concept. An end-to-end pipeline was engineered using the NHANES dataset, culminating in an optimized XGBoost model. The final model demonstrates significant business value, concentrating 54% of all positive cases within the top 30% of highest-risk donors, an efficiency 1.8 times greater than random screening. Key contributions include a reproducible MLOps framework and the use of SHAP for model transparency, ensuring predictions translate into actionable strategies. This work provides a data-driven pathway to reduce operational costs and enhance the supply of valuable biological materials.

Keywords: Predictive Modeling, Hepatitis B, Donor Screening, XGBoost, Model Interpretability, Model Benchmarking, XAI, SHAP, Risk Stratification, Business Aligned

1 Introducction - Project and Business Context

Grifols, as a leading multinational healthcare company specializing in human plasma-derived products, faces ongoing logistical and operational challenges in securing a supply of high-quality biological materials. The Bio Supplies business unit is responsible for providing plasma and serum from both healthy donors and those with specific pathologies for research, pharmaceutical manufacturing, and in vitro diagnostic development.

A fundamental challenge in this process is the efficient identification of donors who present specific serological markers. The traditional screening method, which often involves performing indiscriminate laboratory tests on large donor populations, is a costly and time-consuming process with a low success rate. This inefficiency not only increases operational costs but also limits the availability of valuable and hard-to-source biological materials.

This thesis addresses this challenge by framing it as a binary classification problem solvable with supervised machine learning. The core objective is to develop BioPredict, a data-driven system designed to predict the serological status of plasma donors. The project hypothesizes that a model, trained on historical donor data, can effectively learn the complex, non-linear relationships between a donor's attributes and the presence of a specific biomarker. By transforming a diverse set of inputs—spanning demographic, clinical, and behavioral data—into a high-dimensional feature space, we can construct a model capable of generating a calibrated risk score for each donor.

The strategic value of BioPredict for Grifols is crucial. First, it enables the prioritization of testing on donors with the highest likelihood of being positive, thereby optimizing resource allocation. Second, it aims to significantly reduce the costs and processing times associated with screening. Finally, it seeks to improve the availability of specialized plasma, a critical resource for research and new product development.

In essence, this project represents a shift from a reactive screening model to a proactive, predictive strategy, applying data-driven modeling techniques to solve a real and tangible business challenge in the biopharmaceutical industry.

Beyond achieving a certain level of predictive accuracy, it is crucial for this system to be interpretable and its outputs translatable into actionable business strategies. For Grifols to trust and deploy this model, understanding *why* it makes

• Contact E-mail: alexroldanins@gmail.com
• Supervised by: Ernest Valveny
• Academic Year 2024/25
• NOTE: Used techniques are described on the appendix

certain predictions is as important as *what* it predicts.

The core business goals driving this project are to:

- **Reduce operational costs** by minimizing indiscriminate, low-yield laboratory tests.
- **Manage the donor screening process more efficiently**, concentrating resources on high-probability candidates.
- **Enable targeted donor campaigns** to increase the supply of valuable and hard-to-source biological materials.

So, the fundamental contribution of this TFG is the creation of a model that is not only predictive but also transparent. The system is designed to be deployable in a production environment, leveraging the types of data realistically available within Grifols and providing interpretable outputs that can directly inform screening priorities. While this academic project utilizes the NHANES public dataset as a proxy, the end-to-end pipeline, from data processing to model interpretation, serves as a robust proof-of-concept for Grifols' internal implementation. The selection of Hepatitis B as the target pathology was informed by a literature review which confirmed the viability of using machine learning for its prediction and highlighted a core set of predictive features available in the proxy data.

2 State of the Art and Literature Review

The development of the BioPredict model required a suitable proxy dataset, as direct access to sensitive internal donor data was not feasible for this academic project. After evaluating several large-scale public health surveys, the National Health and Nutrition Examination Survey (NHANES) was selected. This decision was primarily driven by its unique alignment with the project's requirements: its population is representative of the U.S. demographic, matching the target donor base for Grifols' U.S. operations.

Furthermore, NHANES provides a rich, granular combination of demographic, behavioral, and clinical variables alongside the necessary ground-truth laboratory results for serological markers, making it an ideal resource for training supervised learning models.

A review of existing literature confirmed the viability of using machine learning for predicting infectious diseases, with a notable precedent for Hepatitis B (HBV). Studies such as Kim et al. (2023) have successfully used NHANES data to develop predictive models for HBV, achieving high performance ($AUC \approx 0.80$) and demonstrating the predictive power of variables like age, ethnicity, and clinical history. Furthermore, reviews by Su & Kao (2024) highlight the efficacy of ensemble methods, such as Gradient Boosting and Random Forest, in handling the complexity of HBV-related data, reinforcing the choice of these algorithms for this project.

The literature consistently identifies a core set of predictive feature categories, including demographic factors (age, gender, country of birth), behavioral risks (drug use, sexual history), and clinical indicators (BMI, vaccination status). Based on this strong scientific precedent and the high predictive potential reported in multiple studies, Hepatitis B (specifically the HBsAg marker) was selected as the primary target for the BioPredict model. While other pathologies like Toxoplasmosis were initially considered, focusing on HBV allowed for the development of a robust, end-to-end pipeline, with the extension to other biomarkers identified as a clear direction for future work.

3 System Design and Methodology

The development of the BioPredict system was guided by principles from both Agile project management and modern Machine Learning Operations (MLOps). This dual focus ensured that the project not only maintained scientific rigor but was also executed in a manner that was transparent, reproducible, and aligned with the strategic objectives of the business. This section details the methodological framework, the data processing pipeline, and the modeling strategies employed.

3.1 Agile Framework and MLOps Principles

The project was executed using an **Agile Scrum methodology**, mirroring the professional environment at Grifols. The workflow was structured into iterative development cycles (*sprints*), each focused on delivering a specific, functional part of the system. This framework was reinforced by core Scrum rituals, including brief daily stand-up meetings to report progress and identify impediments, and regular sprint reviews to demonstrate completed work. This process facilitated a continuous feedback loop with the Grifols Bio Supplies team, ensuring that technical development remained consistently aligned with their practical needs and evolving requirements.

Beyond the project management framework, the technical architecture was engineered with a strong focus on MLOps principles to ensure robustness, scalability, and maintainability.

Modularity and Configurability The entire predictive pipeline is designed to be highly modular, separating responsibilities into logical components such as data ingestion, preprocessing, model training, evaluation, and interpretability. Crucially, the pipeline is controlled by a central configuration file (`config.yaml`), which acts as a control panel for all experiments. Key parameters—including the target variable, the feature set to be used, model hyperparameters, and data processing steps like normalization or resampling—can be modified in this file. This design decouples the experimental setup from the core logic, enabling rapid iteration and adaptation—such as switching the target biomarker or adjusting the feature set—without modifying the Python source code.

Reproducibility and Versioning Ensuring the scientific validity and traceability of results was a cornerstone of the

design. The codebase is managed using **Git** for version control. Furthermore, a systematic experiment tracking protocol was implemented. Each execution of the pipeline generates a unique, timestamped output directory (e.g., `run_{timestamp}`), creating a complete archive of the experiment. This archive includes the serialized model object, a comprehensive log file detailing every step of the process, all generated performance metrics, and key visualizations (e.g., evaluation plots and SHAP analyses). This systematic approach guarantees full reproducibility—allowing any result to be precisely replicated—and provides a clear audit trail for debugging and validation.

3.2 Data Processing Pipeline

A robust and automated data processing pipeline was developed to transform the raw source data into a clean, suitable dataset for model training. The pipeline consists of two main stages: data ingestion and transformation.

Data Ingestion and Unification The process began with the raw NHANES data files, provided in the survey-specific `.xpt` format. These files, corresponding to multiple survey years and different data categories (e.g., demographics, questionnaires, laboratory), were programmatically converted into the standard `.csv` format. Subsequently, the relevant files were merged and concatenated into a single, cohesive dataset using the unique participant identifier, `SEQN`, ensuring a complete record for each individual.

Automated Cleaning and Transformation Once unified, the dataset underwent a sequence of automated cleaning and feature engineering steps implemented within the pipeline:

1. **Target Variable Handling:** Records with a missing value for the target variable (HBsAg) were removed, as they offer no value for supervised training.

2. **Missing Value Imputation:** A predefined strategy was used to handle missing values in feature columns, typically imputing based on survey-specific codes (e.g., using '9' for "Don't Know").

3. **Feature Pruning:** To reduce noise and model complexity, features exhibiting excessively high rates of missingness (e.g., 90%) or near-zero variance (e.g., 0.01) were programmatically excluded from the analysis.

4. **Feature Scaling:** All numerical features were standardized using the Z-score method. This ensures that features on different scales contribute equitably to the model's learning process, which is critical for many algorithms.

5. **Handling Class Imbalance:** To address the significant class imbalance inherent in the dataset, the **SMOTE (Synthetic Minority Over-sampling Technique)** was applied. This resampling was performed exclusively on the training portion of the data within each cross-validation fold to prevent data leakage and overly optimistic performance estimates.

3.3 Modeling and Optimization

Following data preparation, the project proceeded to the modeling phase, which was structured into three key stages: systematic benchmarking, hyperparameter optimization, and a final, unbiased evaluation.

Model Benchmarking A comprehensive set of classification algorithms was systematically benchmarked to identify the most promising model architecture for the prediction task. The performance of each model was evaluated using **Stratified K-Fold Cross-Validation** (with k=10). This technique was specifically chosen to ensure that the class distribution of the target variable was preserved across all folds. Models were subsequently ranked based on a primary performance metric, typically the F1-Score or AUC, but configurable according to the scope, to guide the selection for the next stage.

Hyperparameter Optimization The best-performing model from the benchmarking phase was subjected to advanced hyperparameter optimization using the **Optuna** framework. Optuna employs a **Bayesian optimization** approach, guided by a **TPE (Tree-structured Parzen Estimator)** sampler. This allows for an efficient and intelligent search of the high-dimensional hyperparameter space, converging more quickly on the optimal configuration that maximizes the target evaluation metric compared to exhaustive methods like Grid Search.

Final Evaluation Upon completion of the optimization process, the final, tuned model was evaluated one last time on the unseen hold-out test set. This final evaluation provides an unbiased estimate of the model's generalization performance on new, independent data, serving as the definitive measure of its predictive capability in a real-world scenario.

4 Results and Analysis

The analysis begins with an Exploratory Data Analysis (EDA) to understand the fundamental characteristics of the dataset. Following this, the performance of the predictive models is evaluated, and the results are interpreted to extract actionable, business-relevant insights.

4.1 Exploratory Data Analysis (EDA)

The initial analysis was intentionally constrained to a limited set of features to establish a baseline model that simulates the data readily available within Grifols' Bio Supplies unit. This approach allows for an evaluation of the predictive power derived solely from core demographic variables: `Age`, `Gender`, and `Race_Ethnicity`. The target variable for this analysis is `Hepatitis_B`, a binary indicator representing the presence (1) or absence (0) of the target marker.

4.1.1 Target Variable Distribution

As illustrated in Figure 2, the dataset exhibits a significant class imbalance. The negative class (Hepatitis B absent, 0)

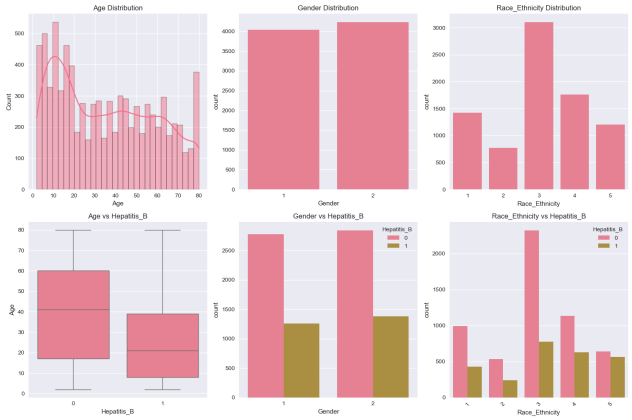

Fig. 1: Distributions of the predictor variables (`Age`, `Gender`, `Race_Ethnicity`) (top row) and their relationship with the `Hepatitis_B` target variable (bottom row). The analysis reveals varying distributions and potential predictive power across features.

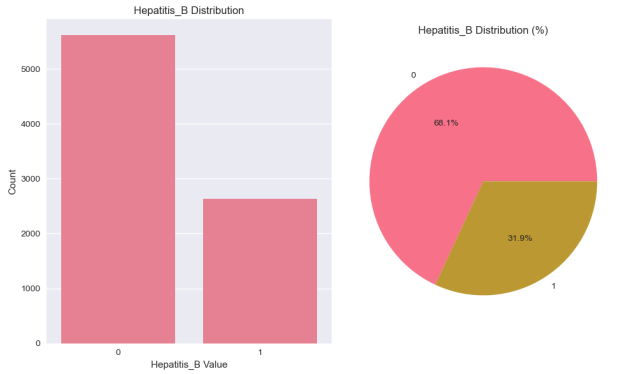

Fig. 2: Distribution of the target variable, `Hepatitis_B`, shown in absolute counts and as a percentage. The significant class imbalance (31.9% Positive) is a critical consideration for model training and evaluation.

constitutes 68.1% of the samples, while the positive class (Hepatitis B present, 1), which is the primary target of interest for Grifols' donor screening, represents only 31.9%. This imbalance has critical implications for the modeling process. Standard accuracy would be a misleading metric, as a naive model could achieve high accuracy by simply predicting the majority class. Consequently, this observation validates the methodological decision to employ techniques like **SMOTE** (Synthetic Minority Over-sampling Technique) for resampling the training data and to prioritize evaluation metrics that are robust to imbalance, such as the **F1-Score**, **Area Under the ROC Curve (AUC)**, and **Matthews Correlation Coefficient (MCC)**. For Grifols, accurately identifying this smaller positive cohort is paramount for efficient resource allocation.

4.1.2 Predictor Variable Analysis

An analysis of the core predictor variables (Figure 1) provides initial insights into their potential utility for the classification task.

Age The `Age` distribution in the dataset is right-skewed, indicating a larger proportion of younger individuals. When examining its relationship with Hepatitis B status, the boxplot reveals a notable difference: the median age of the `Hepatitis_B` positive group is visibly lower than that of the negative group. This suggests that `Age` is a potentially valuable predictor, with younger age potentially associated with a higher likelihood of being positive in this specific dataset.

Gender The `Gender` variable is almost perfectly balanced within the dataset. However, the proportion of positive `Hepatitis_B` cases appears nearly identical across both genders when viewed in isolation. This initial observation suggests that `Gender`, on its own, might offer limited direct predictive power for this specific serological marker, though it could interact with other features in a more complex model.

Race/Ethnicity In contrast, the `Race_Ethnicity` feature demonstrates significant potential. The distribution across its five categories is unequal, with some categories being more prevalent than others. More importantly, the prevalence of positive `Hepatitis_B` cases varies noticeably across these ethnic groups. The visual differences in the proportion of positive instances (represented by the brown segment of the bars in Figure 1) indicate that this feature is likely to be a strong predictor and a key driver of the model's decisions. This aligns with known epidemiological patterns where Hepatitis B prevalence can differ among various ethnic populations due to historical exposure, vaccination rates, and other socio-cultural factors.

4.1.3 Feature Correlation Analysis

To further understand the linear relationships between the baseline features and the target variable, `Hepatitis_B`, a Pearson correlation matrix was generated, as shown in Figure 3.

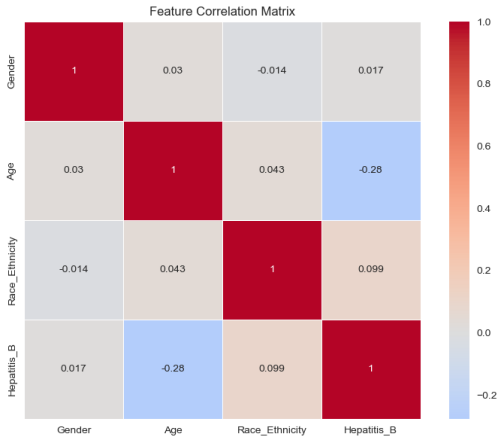

Fig. 3: Pearson Correlation Matrix for the baseline features (`Age`, `Gender`, `Race_Ethnicity`) and the target variable (`Hepatitis_B`).

The correlation matrix reveals the following pairwise linear relationships:

- **`Age` vs. `Hepatitis_B` (-0.28):** This shows a moderate negative linear correlation. It quantitatively supports the observation from the boxplots that lower age

is associated with a higher likelihood of testing positive for Hepatitis B in this dataset. For Grifols, this suggests that age might be a useful initial filter, although the reasons for this trend (e.g., cohort effects related to vaccination programs in the US) would require further investigation beyond the scope of this model.

- **`Race_Ethnicity` vs. `Hepatitis_B` (0.099):** The linear correlation is weak positive. This low value indicates that a simple linear model might not capture the full predictive power of `Race_Ethnicity`. However, as noted in the visual EDA (Figure 1), specific categories within `Race_Ethnicity` show distinct prevalence rates, implying a more complex, non-linear relationship that tree-based models like XGBoost are adept at identifying.

- **`Gender` vs. `Hepatitis_B` (0.017):** This indicates a very weak, almost negligible linear correlation. This reinforces the earlier observation that `Gender`, when considered independently, has limited linear predictive power for Hepatitis B in this dataset.

- **Inter-feature Correlations:** The correlations between the predictor variables themselves are also very low (e.g., `Gender` vs. `Age`: 0.03; `Gender` vs. `Race_Ethnicity`: -0.014; `Age` vs. `Race_Ethnicity`: 0.043). This lack of strong multicollinearity is generally beneficial for model stability and interpretability, as it means the features provide relatively independent pieces of information.

In summary, the EDA, including the correlation analysis, confirms that even with a limited, business-aligned feature set, there are discernible patterns (`Age` showing a moderate linear relationship, and `Race_Ethnicity` hinting at non-linear predictive value) that a machine learning model can potentially leverage to distinguish between positive and negative cases for Hepatitis B.

4.2 Model Performance Evaluation

Following the exploratory analysis, the core modeling phase was executed to quantify the predictive power of the selected demographic features. This involved systematically benchmarking a suite of algorithms to identify the most suitable architecture, followed by a rigorous optimization and evaluation process to determine the final model's performance on unseen data. The insights gained from this evaluation are critical for understanding the model's capabilities and its potential value to Grifols.

4.2.1 Benchmarking Results: Identifying Promising Candidates

To ensure a data-driven approach to model selection, a comprehensive suite of classification algorithms was initially benchmarked. This process involved training each algorithm on the training dataset—balanced using the SMOTE technique to address the inherent class imbalance—and evaluating its performance using Stratified 10-Fold Cross-Validation. The primary metrics for comparison were ROC-AUC, reflecting overall discriminative power, and F1-Score, representing the balance between precision and recall. Table 1 summarizes these cross-validation performances.

TABLE 1: Cross-Validation Performance of Benchmarked Models on the Training Set (Ranked by Accuracy). This initial screening identified Gradient Boosting as the most promising architecture.

Model	Acc.	AUC	Recall	Prec.	F1	Kappa	MCC
Gradient Boosting Classifier	**0.7205**	**0.7259**	**0.3992**	**0.5935**	**0.4763**	**0.2958**	**0.3071**
Light Gradient Boosting Machine	0.7191	0.7094	0.3944	0.5901	0.4720	0.2912	0.3025
Ada Boost Classifier	0.7158	0.7237	0.3759	0.5853	0.4561	0.2769	0.2898
Extra Trees Classifier	0.6927	0.6661	0.3266	0.5305	0.4034	0.2124	0.2240
Decision Tree Classifier	0.6921	0.6650	0.3266	0.5292	0.4030	0.2114	0.2229
Quadratic Discriminant Analysis	0.6880	0.6843	0.1446	0.5555	0.2270	0.1086	0.1489
Ridge Classifier	0.6857	0.6794	0.1614	0.5230	0.2460	0.1137	0.1449
Random Forest Classifier	0.6857	0.6708	0.3748	0.5106	0.4316	0.2218	0.2271
Naive Bayes	0.6849	0.6761	0.1565	0.5208	0.2403	0.1094	0.1409
Logistic Regression	0.6823	0.6793	0.1906	0.5062	0.2765	0.1239	0.1482
Linear Discriminant Analysis	0.6821	0.6794	0.1923	0.5061	0.2783	0.1246	0.1487
K Neighbors Classifier	0.6818	0.6586	0.3797	0.5024	0.4318	0.2170	0.2214
SVM - Linear Kernel	0.6818	0.6440	0.0130	0.0571	0.0211	0.0106	0.0151
Dummy Classifier	0.6807	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000

The benchmarking results in Table 1 show that the **Gradient Boosting Classifier** is the undisputed top-performing model across all key metrics, including Accuracy (0.7205), AUC (0.7259), Precision (0.5935), Recall (0.3992), and F1-Score (0.4763). This provides a clear and data-driven direction for the project.

From a business perspective, these results are promising but also highlight areas for improvement. The Precision of 0.59 suggests that when the model identifies a donor as positive, it is correct nearly 60% of the time, significantly reducing wasted tests compared to random screening. However, the Recall of 0.40 is a critical point of concern for Grifols. This metric indicates that the model, in its default configuration, is only able to identify 40% of the total true positive donors in the population. While efficient, this would leave 60% of valuable, hard-to-find donors undiscovered. The F1-Score of 0.4763 reflects this challenging trade-off between finding positive donors and avoiding false alarms.

Therefore, while the Gradient Boosting algorithm is clearly the most suitable architecture, its out-of-the-box performance is not yet optimal for Grifols' dual objectives of maximizing yield (Recall) and controlling costs (Precision). This motivated the decision to select **Extreme Gradient Boosting (XGBoost)**, a highly optimized and powerful implementation of the gradient boosting algorithm, for an intensive hyperparameter optimization phase. The goal of this next step is to leverage XGBoost's extensive tunability to improve upon these benchmark results, specifically aiming to increase the Recall and F1-Score to develop a model that is not only predictive but also delivers maximum strategic value.

The optimization process successfully improved the model's overall discriminative power, as evidenced by the increase in ROC-AUC from the benchmark's best of 0.7259 to a final 0.7495. From a business perspective, the most significant change is the shift in the Precision-Recall balance. The optimized model demonstrates a notable increase in Precision for the positive class (from 0.59 to 0.64). This is a significant win from a cost-saving perspective, as it means that for every 100 donors the model flags as positive, 64 will be correct, reducing the number of wasted follow-up tests.

However, this gain in precision comes at a cost to Recall,

TABLE 2: Final Performance Metrics of the Optuna-Optimized XGBoost Model on the Hold-Out Test Set (N=1653). These metrics reflect the model's generalization capability on unseen donor data.

Metric	Value
Accuracy	0.7320
ROC-AUC	0.7495
F1-Score (Positive Class 1)	0.4600
Precision (Positive Class 1)	0.6400
Recall (Positive Class 1)	0.3674
F1-Score (Weighted Avg)	0.7079
Precision (Weighted Avg)	0.7167
Recall (Weighted Avg)	0.7320
Matthews Correlation Coeff. (MCC)	0.2885
Cohen's Kappa	0.2878

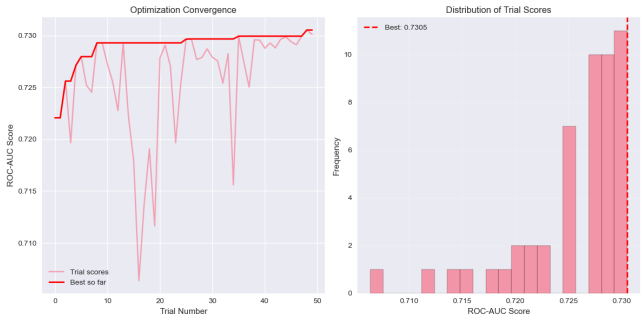

Fig. 4: Optuna Optimization Iterations

which decreased from the benchmark's 0.40 to 0.37. This indicates the model has become more conservative, identifying a smaller fraction of the total positive donors. The F1-Score, which balances these two metrics, saw a slight decrease from 0.4763 to 0.4600. This trade-off is critical: the optimized model is more cost-effective for the donors it identifies, but it misses more potential positive cases. This highlights a key strategic decision for Grifols: depending on the relative cost of a missed opportunity versus a wasted test, the model's prediction threshold can be adjusted to favor either higher Recall or higher Precision to align with specific business goals, since the parameter focus optimization on Optuna can be modified accordingly to the model expectations.

Confusion Matrix: Interpreting Prediction Outcomes for Grifols The confusion matrix for the Optuna-optimized XGBoost model on the test set is presented in Figure 5. This matrix is crucial for understanding the practical implications of the model's predictions in the context of Grifols' operations.

The matrix reveals:

- **True Positives (TP = 199):** The model correctly identified 199 donors who genuinely have Hepatitis B. For Grifols, these are successful identifications, enabling targeted confirmatory testing and efficient sourcing of valuable plasma.
- **True Negatives (TN = 996):** The model correctly identified 996 donors who do not have Hepatitis B. This translates directly into cost savings by avoiding unnecessary laboratory tests on a large segment of the donor population.
- **False Positives (FP = 129):** The model incorrectly flagged 129 Hepatitis B-negative donors as positive. Each False Positive represents an avoidable cost for Grifols, as these donors would undergo confirmatory testing only to be found negative.
- **False Negatives (FN = 329):** The model incorrectly classified 329 Hepatitis B-positive donors as negative. These are missed opportunities and constitute the most critical error type from a resource acquisition standpoint, as these valuable donors would not be prioritized for screening.

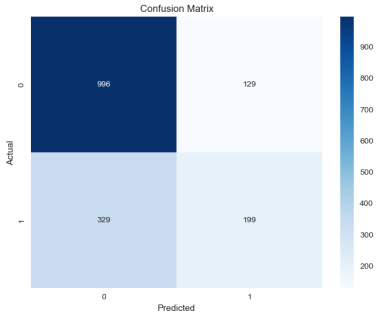

Fig. 5: Confusion Matrix for the Optuna-Optimized XGBoost Model on the Test Set (N=1653). Class 0: No Hepatitis B, Class 1: Hepatitis B. TN=996, FP=129, FN=329, TP=199.

5 Model Interpretation and Business Impact

Beyond achieving a certain level of predictive accuracy, it is crucial for this system to be interpretable and its outputs translatable into actionable business strategies. For Grifols to trust and deploy this model, understanding *why* it makes certain predictions is as important as *what* it predicts. This section leverages the SHAP (SHapley Additive exPlanations) framework to interpret the optimized XGBoost model trained on the baseline features, followed by an analysis of its direct business value using gain and decile charts.

5.1 Feature Importance and Interpretation (SHAP)

SHAP values provide a robust, theoretically grounded method for explaining the output of machine learning models. They quantify the contribution of each feature to the prediction for each individual instance, allowing for both global and local interpretability.

Global Feature Importance: Identifying the Key Drivers To understand which features have the most influence on the model's predictions overall, we analyze the mean absolute SHAP value for each feature. This metric, shown in Figure 6, ranks features by their average impact on the model's output magnitude, irrespective of direction.

The ranking confirms the insights from the initial EDA:

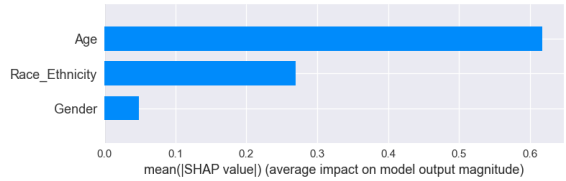

Fig. 6: Global Feature Importance for the Baseline Model, ranked by mean absolute SHAP value. `Age` is clearly the most impactful feature.

1. **`Age`:** Is unambiguously the most influential feature, with a mean SHAP value magnitude of approximately 0.6. This indicates that, on average, a donor's age has the largest impact on shifting their predicted risk score.
2. **`Race_Ethnicity`:** Ranks as the second most important feature, with roughly half the impact of `Age`. Its contribution is still substantial and critical to the model's performance.
3. **`Gender`:** Has a significantly lower impact compared to the other two features, confirming its role as a minor predictor in this specific model configuration.

Detailed Feature Impact: The SHAP Summary Plot The SHAP Summary Plot, often called a beeswarm plot (Figure 7), provides a much richer understanding than a simple bar chart. It shows not only the magnitude of each feature's impact (its position on the x-axis) but also its direction, while the color indicates the feature's original value for each individual donor (Red = High, Blue = Low).

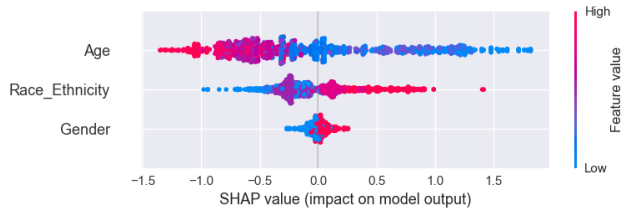

Fig. 7: SHAP Summary Plot for the Baseline Model. Each dot is a donor. The color represents the feature's value (High/Low), and its x-position shows its impact on the prediction (positive SHAP values increase the predicted risk of Hepatitis B).

Interpreting the plot reveals the nuanced logic learned by the model:

- **`Age`:** The plot shows a clear and consistent trend. Blue dots (lower age values) are predominantly on the right side of the zero line, contributing positive SHAP values. This means the model has learned that **lower age increases the predicted likelihood of having Hepatitis B**. Conversely, red/purple dots (higher age values) are on the left, contributing negative SHAP values, thus decreasing the predicted risk. This aligns with the negative correlation found in the EDA and is a clinically plausible finding in the context of the US population, where older cohorts may have higher rates of resolved infection or different historical risk exposures, while some younger populations might have lower vaccination uptake or different risk behaviors.
- **`Race_Ethnicity`:** The feature is encoded as: 1 (Mexican American), 2 (Other Hispanic), 3 (Non-Hispanic White), 4 (Non-Hispanic Black), and 5 (Other Race/Multi-Racial).
  - The plot shows a large cluster of purple-red dots (corresponding to values 4 and 5: **Non-Hispanic Black** and **Other Race**) on the right side of the zero line. This indicates that the model has learned that individuals from these groups have a higher predicted likelihood of Hepatitis B, as these features contribute positive SHAP values.
  - Conversely, there is a dense cluster of light-blue dots (corresponding to value 3: **Non-Hispanic White**) primarily on the left side, contributing negative SHAP values. This means being in this group significantly lowers the predicted risk.
  - The bluest dots (value 1: **Mexican American**) are spread more centrally but also show a tendency to be on the left, suggesting they also contribute to a lower predicted risk, though perhaps less strongly than for Non-Hispanic Whites.

  This demonstrates that the model is not using a simple linear trend but has learned that **specific ethnic categories are associated with different levels of risk**. This aligns with epidemiological data from the CDC, which historically shows a higher prevalence of chronic Hepatitis B among Non-Hispanic Black and some Asian populations (often categorized under "Other Race") compared to Non-Hispanic White populations in the United States. The model has successfully captured these well-documented demographic risk disparities from the data.
- **`Gender`:** The plot for `Gender` confirms its minor role. The dots are tightly clustered around zero, indicating a small impact on the final prediction for most individuals. There is a slight tendency for one gender (e.g., blue dots, perhaps representing males if encoded as a lower number) to have slightly positive SHAP values, but the effect is weak and not as clear-cut as with Age or Race/Ethnicity.

Local Interpretability: Explaining a Single Prediction Beyond global trends, SHAP can explain individual predictions, making the model's logic transparent on a case-by-case basis. The waterfall plot in Figure 8 deconstructs the prediction for a single donor.

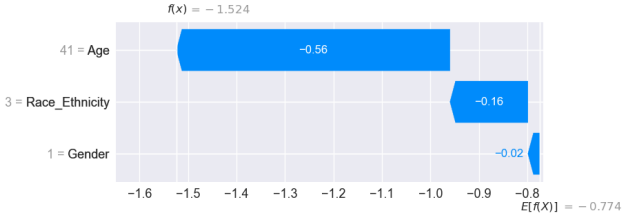

Fig. 8: SHAP Waterfall Plot explaining a single prediction. It shows how the base prediction ($E[f(X)] = -0.774$) is modified by each feature's contribution to arrive at the final log-odds output ($f(x) = -1.524$).

For this specific donor, who was likely predicted as negative for Hepatitis B (as $f(x) = -1.524$ is a low log-odds value):

- The prediction starts at the base value ($E[f(X)] = -0.774$), which is the average prediction across the entire dataset.
- The donor's **`Age` of 41** contributes a large negative push of **-0.56**, significantly decreasing their risk score. This is consistent with the global trend where higher age lowers risk.
- Their **`Race_Ethnicity` of 3** contributes a smaller negative push of **-0.16**.
- Their **`Gender` of 1** has a negligible negative impact of **-0.02**.

These individual feature contributions are summed up, moving the prediction from the average baseline to the final, specific output for this donor. This level of transparency is invaluable for Grifols, as it would allow an analyst or clinician to review why a particular donor was flagged (or not flagged) by the model, building trust and facilitating a human-in-the-loop validation process.

In conclusion, the SHAP analysis confirms that the model's decision-making is logical and based on the patterns identified in the EDA. It relies heavily on Age and Race/Ethnicity, treating them with a nuance that aligns with epidemiological expectations. This interpretability is a cornerstone for transitioning the BioPredict model from an academic exercise to a trusted operational tool within Grifols.

5.2 Business-Facing Analysis and Impact

The ojective of BioPredict's success is its ability to deliver tangible, operational value to Grifols, Cumulative Gains and Decile charts are indispensable for translating predictive performance into actionable strategies for cost savings and resource optimization. These analyses directly answer the critical business question: "How much more efficiently can we find our target donors by using this model?"

5.2.1 Cumulative Gains: Quantifying Efficiency

The Cumulative Gains chart, presented in Figure 9, is a powerful visualization of the model's efficiency compared to a random screening approach. It plots the percentage of total positive cases found (the "gain") against the percentage of the donor population that would need to be screened, after ranking all donors by the model's predicted risk score.

The significant gap between the model's performance (blue line) and the random baseline (grey dashed line) illustrates the immense operational advantage offered by BioPredict:

- **High-Efficiency Targeting:** By screening just the **top 30%** of donors whom the model deems most likely to be positive, Grifols can expect to identify **54%** of all available Hepatitis B positive individuals in the population. In contrast, a random screening of 30% of the population would, on average, only find 30% of the positives. This represents a **lift of 1.8x** (54% / 30%), meaning the model is 80% more efficient at finding positive cases within this segment than random chance.
- **Strategic Resource Allocation:** The chart allows for strategic planning. For example, to find approximately 83% of all positive donors, the model indicates that Grifols would need to screen the top 60% of the population. This allows decision-makers to balance the desired yield of positive plasma units against the available budget for confirmatory testing.

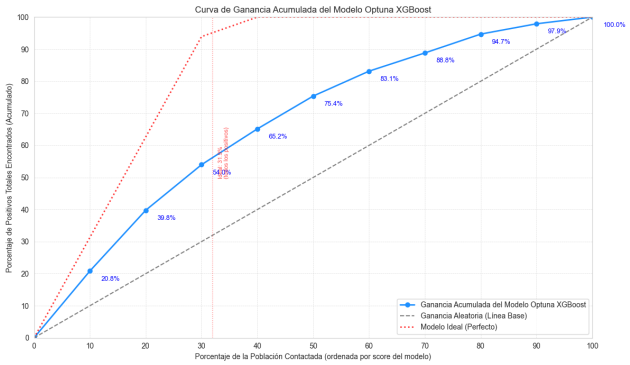

Fig. 9: Cumulative Gains Chart for the Optimized XGBoost Model. The blue line shows the percentage of total positive donors identified by screening a given percentage of the highest-risk population, as ranked by the model. The grey dashed line represents random screening. The red dashed line represents the ideal perfect model.

This analysis demonstrates that the model acts as a powerful focusing lens, allowing Grifols to concentrate its expensive and time-consuming laboratory testing on a much smaller, higher-probability subset of donors, thereby maximizing the return on its screening investment.

5.2.2 Decile Analysis: An Actionable Roadmap for Screening

While the gains chart shows cumulative benefit, the Decile Chart (Figure 10) provides an even more granular and operationally intuitive breakdown. It segments the donor population, ranked by the model's risk score, into ten equal groups (deciles) and shows the number of actual positive cases found within each. This provides a clear, step-by-step roadmap for a prioritized screening strategy.

The decile chart starkly illustrates the model's ability to concentrate positive cases in the top-ranking groups:

- **Decile 1 (Top 10%):** By testing only this highest-risk group (approx. 166 donors in the test set), the model successfully identifies **110 positive cases**. The precision or "hit rate" within this decile is an exceptional **66.3%** (110 out of 166). This single decile captures 110 out of the 528 total positives, or **20.8%** of the entire positive population.
- **Deciles 2 and 3:** The second decile adds another 100 positive cases (60.6% precision), and the third adds 75 more (45.5% precision). By screening just the top three deciles (30% of donors), Grifols would identify a total of $110 + 100 + 75 = 285$ positive cases, confirming the 54% gain shown in the previous chart.

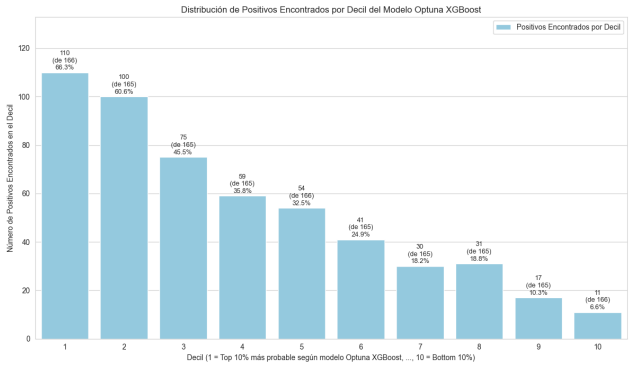

Fig. 10: Distribution of Positive Hepatitis B Cases Found per Decile. Donors are ranked by the model's prediction score (Decile 1 = Top 10% highest risk). The bars show the absolute number of positives found, while the percentages indicate the "hit rate" (precision) within each decile.

- **Diminishing Returns:** The number of positives found drops off sharply in the lower deciles. Decile 10, representing the donors the model deems least likely to be positive, contains only 11 positive cases, for a low hit rate of 6.6%. Screening this group would be highly inefficient.

Translating Insights into Cost Savings: An Example Let's illustrate the direct financial impact. Assume the goal is to find approximately **285 positive donors**.

- **Without the Model (Random Screening):** The overall prevalence of positive cases in the test set is $528/1653 \approx 31.9\%$. To find 285 positives through random screening, Grifols would need to test approximately $285/0.319 \approx$ **893** donors.
- **With the BioPredict Model:** According to the decile chart, to find 285 positives, Grifols only needs to test the donors in the top three deciles. The total number of donors in these three deciles is approximately $166 \times 3 =$ **498** donors.

In this scenario, deploying the BioPredict model would allow Grifols to achieve its goal by performing **395 fewer tests** $(893 - 498)$, a reduction of over 44% in the required screening volume. If each confirmatory test costs a significant amount in reagents, equipment time, and labor, this reduction translates directly into substantial and recurring operational cost savings. Grifols can use this decile-based framework to define a dynamic screening strategy based on their weekly or monthly targets for positive plasma units, ensuring maximum efficiency and a more predictable supply chain for this valuable resource.

5.2.3 Impact of Feature Set: From Baseline to Enhanced Model

To test the hypothesis that a richer donor profile improves predictive accuracy, an "enhanced" model was trained using an expanded set of features. These variables were selected based on established literature and known epidemiological risk factors for Hepatitis B, representing non-laboratory data that Grifols could potentially collect. For instance, `Injected_Drugs_Ever` was included due to the high scientific correlation between intravenous drug use and Hepatitis B transmission through shared needles. Similarly, `Dental_Visit_Reason` was chosen as a proxy for potential exposure to non-sterilized equipment in certain settings, a known, albeit less common, transmission route. The full enhanced set also included socio-demographic and health markers like `Country_of_Birth`, `Education_Level`, `Income_to_Poverty_Ratio`, and `Waist_Circumference`.

Feature Importance in the Enhanced Model SHAP (SHapley Additive exPlanations) analysis was used to evaluate the contribution of each feature in the optimized enhanced model. The global feature importances are ranked in Figure 11.

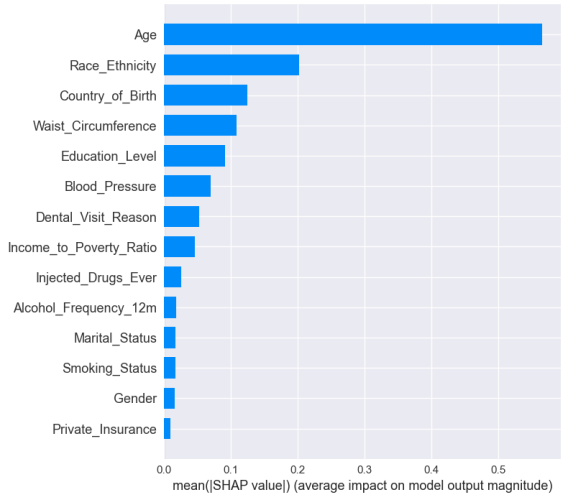

Fig. 11: Global Feature Importances (Mean Absolute SHAP Value) for the optimized model trained with the enhanced feature set. Features are ranked by their average impact on model output.

The SHAP analysis reveals that `Age` and `Race_Ethnicity` remain the most dominant predictors, highlighting the foundational importance of core demographics. Among the new features, `Country_of_Birth` and `Waist_Circumference` contribute meaningfully, suggesting that nativity (and potential exposure in endemic regions) and general health status are valuable signals. Interestingly, established risk factors like `Injected_Drugs_Ever` rank lower in this model. This could be due to low prevalence of this behavior in the dataset sample or because its predictive signal is already partially captured by other correlated socioeconomic features. The very low ranking of `Gender` is consistent across both baseline and enhanced models, confirming its limited direct impact in this context.

Performance Uplift and Discussion The primary goal of expanding the feature set was to increase predictive performance. Table 3 compares the final metrics of the optimized 3-feature baseline model against the new optimized enhanced model on the test set.

The results indicate that while incorporating a richer feature set does provide a performance uplift, the gains

TABLE 3: Performance Comparison: Optimized Baseline Model (XGBoost, 3 Features) vs. Optimized Enhanced Model (All Features) on the Test Set.

Model Configuration	Acc.	AUC	Recall (Cls 1)	Prec. (Cls 1)	F1 (Cls 1)
Baseline (XGBoost, 3 Feat.)	0.7320	0.7495	0.3674	0.6400	0.4666
Enhanced (All Features)	0.7344	0.7555	0.4100	0.6300	0.5000
Uplift / Change	+0.0024	+0.0060	+0.0426	-0.0100	+0.0334
Rel. Uplift (%)	+0.33%	+0.80%	+11.6%	-1.56%	+7.16%

Cls 1 refers to the Hepatitis B positive class. Its F1-Score, Recall, and Precision are the most critical metrics for Grifols' business case.

are modest. The overall accuracy improved marginally to 0.7344. More importantly for the business objective, the F1-Score for the positive class (Class 1) improved from 0.4666 to 0.5000, a relative increase of over 7%. This improvement was driven by a notable 11.6% relative increase in Recall (from 37% to 41%), meaning the enhanced model is able to identify a larger fraction of the true positive donors.

In summary, the enhanced feature set provides a slightly better model. The improvement in identifying valuable positive donors (higher recall and F1-score for Class 1) validates the inclusion of additional donor attributes. However, the modest nature of the overall performance gain suggests that the core demographic variables carry the majority of the predictive weight in this dataset. For Grifols, this implies that even a simple model with basic demographic data can be surprisingly effective, and while collecting more data is beneficial, it may yield diminishing returns.

6 Demo Deployment and Application

To translate the model's predictive power into a tangible business tool, a web-based demonstration was developed using the Streamlit framework. This interactive application serves as a proof-of-concept, showcasing how the BioPredict system can be operationalized for stakeholders at Grifols.

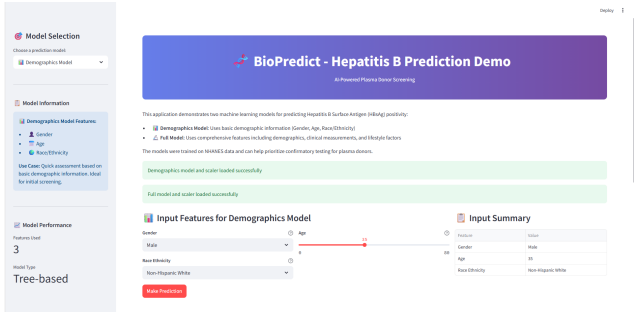

Fig. 12: The BioPredict Streamlit demo, showing the interactive input form and the resulting prediction analysis with a calibrated risk score.

Interactive Interface As shown in Figure 14, the interface allows non-technical users to input hypothetical donor data using intuitive controls like sliders and dropdowns. A key feature is the ability to switch between the baseline (demographics-only) and the enhanced (full-feature) models, enabling a direct comparison of their predictive outputs and demonstrating the value of additional data collection.

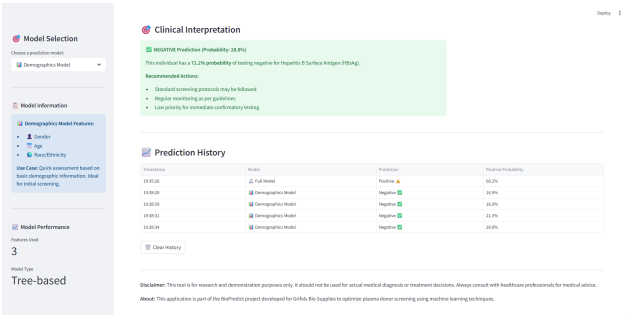

Fig. 13: The BioPredict Streamlit demo visualization 2.

Actionable and Interpretable Results Upon submitting the data, the demo provides more than a simple binary outcome. The output is designed to be directly actionable:

- **Prediction:** A clear "Positive" or "Negative" classification.
- **Calibrated Risk Score:** The application displays a precise probability of the donor being positive. This is a calibrated score, meaning a predicted 70% probability corresponds to a true, real-world likelihood of 70%. This is crucial for risk stratification, allowing Grifols to prioritize a donor with a 90% predicted risk over one with a 60% risk. The method for achieving this calibration is detailed in Section 6.1.
- **Clinical Interpretation:** A brief text summary translates the prediction into a recommended business action, such as "Prioritize for confirmatory testing."

This interactive tool effectively transforms the complex model into an operational prototype, enabling stakeholders to explore diverse donor profiles and understand the model's value in a tangible, accessible format.

6.1 Model Calibration for Reliable Risk Scores

Powerful classifiers like XGBoost are optimized for discrimination (correctly separating classes) but their raw probability outputs are often not well-calibrated. For instance, the model might assign a 90% probability to a group of donors where, in reality, only 75% are positive. Model calibration is a crucial post-processing step to adjust these scores so they accurately reflect the true likelihood of an outcome, specially when data is augmented in unbalanced classes.

The process involves training the primary XGBoost model and then using a separate, unseen validation set to train a second, simpler "calibrator" model. This calibrator learns to map the XGBoost model's potentially skewed probabilities to new, reliable probabilities. The initial calibration error is assessed using a reliability diagram, which plots predicted probabilities against the actual fraction of positives.

Two primary calibration methods were considered:

- **Platt Scaling:** Uses a parametric logistic regression model. It is effective for correcting simple, sigmoidal (S-shaped) distortions and is robust on smaller datasets.

- **Isotonic Regression:** A more powerful, non-parametric method that fits a non-decreasing function. It can correct more complex distortions but requires more data to avoid overfitting.

For this project, Isotonic Regression was applied to the model's outputs. The final predictive system used in the demo therefore consists of the primary XGBoost model followed by the trained Isotonic Regression calibrator, ensuring the probability scores are trustworthy and directly usable for business risk assessment.

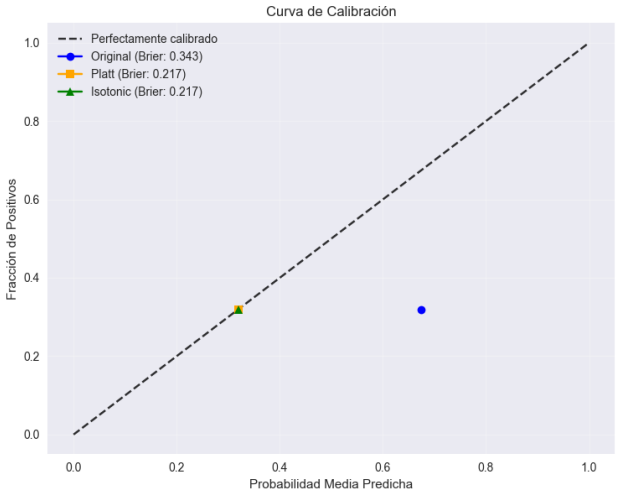

Fig. 14: Calibration Curve

7 Conclusions

This project was successfully developed accordingly to what Grifols initially expected, the work culminated in an optimized and interpretable machine learning model that demonstrates significant potential to reduce operational costs and enhance the supply of valuable biological materials for the unit.

A central requirement of this project was to develop a model that aligns with the data currently and realistically available within Grifols' operational systems, specifically the core demographic variables of age, gender, and race/ethnicity. This constraint was intentionally embraced to create a solution with a high and immediate return on investment (ROI), without requiring investment in new, costly data collection initiatives. The final model, built on just these three features, proved to be a highly effective, ready-to-use tool. While an enhanced model with additional variables showed a modest performance uplift, the success of the baseline model validates the core strategy: leveraging existing data assets to their fullest potential.

The final model's performance represents a dramatic improvement over indiscriminate or random screening. As demonstrated by the business impact analysis, the model acts as a powerful focusing lens, enabling Grifols to concentrate its resources on a small, high-probability segment of the donor population. This targeted approach translates directly into substantial cost savings from avoided laboratory tests and a more predictable and efficient donor management process.

While the use of the NHANES public dataset served as an effective proxy, the primary limitation of this work is the absence of validation on Grifols' proprietary donor data. Therefore, the immediate future work involves deploying this end-to-end pipeline to train and validate the model on internal data, which would likely uncover new patterns and further improve performance. Subsequent steps should include expanding the system to predict other valuable biomarkers and integrating the developed dashboard into Grifols' production environment to facilitate daily operational use.

In conclusion, BioPredict serves as a successful proof-of-concept, demonstrating that a well-designed, interpretable machine learning model can transform a reactive, costly screening process into a proactive, efficient, and data-driven strategy. The project delivers a clear and actionable pathway for Grifols to optimize resources, reduce costs, and ultimately strengthen its supply chain for critical biopharmaceutical products.

Acknowledgements

I would like to extend my sincere gratitude to my colleagues at Grifols for their proactive support throughout this project. I am especially grateful to my company tutor, Lucas Pastur, for his excellent supervision. His guidance was invaluable in steering the project in the right direction, and I have learned a great deal from his insights and expertise. I also wish to thank my academic tutor from the UAB, Ernest Valveny, for his constant willingness to help and for providing highly relevant feedback that was crucial for ensuring the quality of this work.

References

[1] V. Harabor *et al.*, "Machine learning approaches for the prediction of hepatitis b and c seropositivity," *International Journal of Environmental Research and Public Health*, vol. 20, no. 3, p. 2380, 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/36767747/

[2] S.-H. Kim *et al.*, "Machine learning for predicting hepatitis B or C virus infection in diabetic patients," *Scientific Reports*, vol. 13, p. 21518, 2023. [Online]. Available: https://www.nature.com/articles/s41598-023-49046-9

[3] W. Krueger *et al.*, "Drinking water source and human Toxoplasma gondii infection in the United States: a cross-sectional analysis of NHANES data," *BMC Public Health*, vol. 14, p. 711, 2014. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/25012250/

[4] T.-H. Su and J.-H. Kao, "Role of artificial intelligence in the management of chronic hepatitis B infection," *Clinical Liver Disease*, vol. 23, p. e0164, 2024. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11068129/

[5] A. Rostami *et al.*, "Global prevalence of latent toxoplasmosis in pregnant women: a systematic

review and meta-analysis," *Clinical microbiology and infection*, vol. 26, no. 6, pp. 698–710, 2020. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/31972316/

[6] J. Woodring *et al.*, "HIV infection in the united states household population aged 18–49 years: Results from 2007–2012 NHANES data," National Center for Health Statistics, National Health Statistics Reports 83, 2015. [Online]. Available: https://www.cdc.gov/nchs/data/nhsr/nhsr083.pdf

[7] G. McQuillan *et al.*, "Prevalence of hepatitis B virus infection in the united states: 1999–2006 and 1988–1994," National Center for Health Statistics, Tech. Rep., 2005, note: The user-provided URL points to the HIV report (nhsr083), not a specific HBV report from this author/year. A placeholder is used. [Online]. Available: https://www.cdc.gov/nchs/products/index.htm

[8] J. Jones *et al.*, "Toxoplasma gondii infection in the United States: seroprevalence and risk factors," *American journal of epidemiology*, vol. 154, no. 4, pp. 357–365, 2001. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/11495859/

[9] R. Lebo *et al.*, "Rubella immunity in the US population, NHANES 2010–2011," *The Journal of infectious diseases*, vol. 212, no. 10, pp. 1530–1537, 2015, the user-provided URL pointed to a different paper. Corrected based on title.

[10] S. Gottlieb *et al.*, "Prevalence of syphilis seroreactivity in the United States: NHANES 1999–2004," *Sexually Transmitted Diseases*, vol. 35, no. 5, pp. 497–501, 2008. [Online]. Available: https://journals.lww.com/stdjournal/abstract/2008/05000/prevalence_of_syphilis_seroreactivity_in_the.11.aspx

[11] National Center for Health Statistics (NCHS/CDC), "National health statistics reports, number 83," September 2015.

[12] "UK Biobank," https://www.ukbiobank.ac.uk/, accessed: [Date of Access].

[13] Institute for Health Metrics and Evaluation (IHME), "Global Health Data Exchange (GHDx)," https://ghdx.healthdata.org/, accessed: [Date of Access].

[14] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority oversampling technique," in *Journal of artificial intelligence research*, vol. 16, 2002, pp. 321–357.

[15] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[16] B. W. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," *Biochimica et Biophysica Acta (BBA)-Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975.

[17] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[18] J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Advances in neural information processing systems*, vol. 24, 2011.

[19] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in large margin classifiers*. MIT press, 1999, pp. 61–74.

[20] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002, pp. 694–699.

[21] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly weather review*, vol. 78, no. 1, pp. 1–3, 1950.

[22] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," *Ijcai*, vol. 14, no. 2, pp. 1137–1145, 1995.

ANNEX

The complete source code for this project, including all data processing pipelines, model training scripts, and the Streamlit demonstration, is publicly available in the official GitHub repository:

Link to Project GitHub Repository

Additionally, this project was supported by a comprehensive preliminary study on the early prediction of Hepatitis B and other pathologies. The full research notes, compiled in Spanish, can be accessed via the following link:

Link to Supplementary Research Document

TECHNICAL APPENDIX

This appendix provides a brief technical overview of key concepts and techniques employed in this project.

A Data Preprocessing and Resampling

Stratified K-Fold Cross-Validation Cross-validation is a resampling technique used to evaluate a model's performance on unseen data by partitioning the data into complementary subsets. In K-Fold cross-validation, the dataset is split into K folds (subsets). The model is trained on K-1 folds and validated on the remaining fold, a process repeated K times. **Stratified** K-Fold is a crucial variation used for imbalanced datasets, as in this project. It ensures that each fold maintains the same percentage of samples for each class as the complete dataset [22]. This prevents scenarios where a fold might, by chance, contain very few or no samples of the minority class, which would lead to an unreliable evaluation of the model's performance.

SMOTE (Synthetic Minority Over-sampling Technique) Class imbalance, where one class significantly outnumbers the other, can bias a model towards the majority class. SMOTE is an advanced over-sampling technique designed to mitigate this issue [14]. Instead of simply duplicating minority class instances, SMOTE generates new, *synthetic* samples. For each minority class instance, it identifies its k-nearest minority class neighbors and creates a synthetic sample along the line segment joining the instance and one of its chosen neighbors. This results in a more diverse and robust training set, helping the model learn the decision boundary for the minority class more effectively.

B Hyperparameter Optimization

Bayesian Optimization Finding the optimal hyperparameters for a model can be computationally expensive. Bayesian Optimization is an intelligent optimization strategy that is more efficient than exhaustive methods like Grid Search. It builds a probabilistic "surrogate model" of the objective function (e.g., the model's F1-score as a function of its hyperparameters). This surrogate model is then used to intelligently select the next set of hyperparameters to evaluate, focusing on regions of the parameter space that are most likely to yield improvement.

TPE (Tree-structured Parzen Estimator) TPE is a specific algorithm used for Bayesian Optimization and is the default sampler in the Optuna framework [18]. Instead of modeling the objective function directly, TPE models two separate probability distributions: one for hyperparameters associated with "good" scores (e.g., high F1) and another for those with "bad" scores. At each iteration, it samples candidate hyperparameters that are more likely under the "good" distribution and less likely under the "bad" one, allowing it to efficiently converge towards an optimal configuration.

C Performance Evaluation Metrics for Imbalanced Data

Matthews Correlation Coefficient (MCC) The MCC is a highly robust metric for binary classification, particularly on imbalanced datasets [16]. It is a correlation coefficient between the observed and predicted classifications and takes into account all four values in the confusion matrix (TP, TN, FP, FN). Its value ranges from -1 (total disagreement) to +1 (perfect prediction), with 0 indicating a performance no better than random guessing. A high MCC score is only achieved if the classifier obtains good results in all four confusion matrix categories, making it a balanced and reliable measure.

Cohen's Kappa Cohen's Kappa (κ) measures the agreement between two raters, or in this case, between the model's predictions and the ground truth. Crucially, it accounts for the possibility of agreement occurring by chance [17]. A Kappa score of 1 represents perfect agreement, 0 represents agreement equivalent to random chance, and negative values indicate agreement worse than chance. It provides a more robust measure than simple accuracy on imbalanced datasets because it penalizes models that simply predict the majority class.

D Model Interpretability

SHAP (SHapley Additive exPlanations) Understanding *why* a model makes a certain prediction is crucial for building trust and deploying it in a business context. SHAP is a game theory-based framework for explaining the output of any machine learning model [15]. It calculates Shapley values, which represent the contribution of each feature to pushing the model's output from a baseline value (the average prediction) to the final prediction for a specific instance. This allows for both global interpretability (which features are most important overall) and local interpretability (why a specific donor received a high-risk score).

E Model Calibration

Brier Score The Brier score is a proper score function that measures the accuracy of probabilistic predictions [21]. It is the mean squared error between the predicted probabilities and the actual outcomes (0 or 1). A lower Brier score indicates better calibration, meaning the model's predicted probabilities are more reliable. A score of 0 represents a perfect model.

Platt Scaling and Isotonic Regression Powerful models like XGBoost often produce poorly calibrated probabilities. Calibration is a post-processing step to correct this.

- **Platt Scaling** fits a simple logistic regression model to the classifier's outputs. It is effective for correcting sigmoidal (S-shaped) distortions in probabilities [19].
- **Isotonic Regression** is a more powerful, non-parametric method that fits a non-decreasing, piecewise-constant function [20]. It can correct more complex calibration errors but is more prone to overfitting on small datasets. For this project, Isotonic Regression was chosen to ensure the risk scores presented in the demo were as reliable as possible.