University of Bahrain

College of Science

Department of Mathematics

PREDICTING FOOTBALL PLAYER SUCCESS USING MACHINE LEARNING ON FOOTBALL MANGER 2024 GAME SIMULATION

A report submitted to the Department of Mathematics
In partial fulfillment of the requirement for the degree of
B.Sc. in Statistics and Data science

EBRAHIM JUMA SHAKAK ALSAWAN
ID: 202009241

Under the Supervision of Ms. Aseel Masoud Ebrahim Alhermi
December 2025

Acknowledgment

I would like to express my sincere gratitude to my supervisor, Ms. Aseel Masoud Ebrahim Alhermi, for her continuous guidance, patience, and invaluable feedback throughout this project. Her support and direction were essential in shaping the study and ensuring its successful completion.

I am also grateful to the Department of Mathematics at the University of Bahrain for supporting this project and giving me the space to explore an unconventional topic. Their openness allowed me to work on something I truly enjoy, and I deeply appreciate their encouragement of academic creativity.

I would also like to extend my heartfelt thanks to my family and friends, whose support has been a constant source of motivation throughout my university journey. Their encouragement, understanding, and help played a significant role in enabling me to complete both this project and my degree.

Finally, I would like to acknowledge the developers of Football Manager 2024. Their work in creating such a detailed and sophisticated simulation environment not only made this research possible but also introduced me to a platform that I would not have discovered otherwise. I believe that modern games can serve as valuable alternative, artificial data sources, and this project would not have been possible without the rich and structured data provided by FM.

Abstract

This study develops a machine learning framework to predict long-term football player success using ten-year simulations in Football Manager 2024 (FM) [1]. The game's structured database and progression engine provide a controlled alternative to real-world longitudinal datasets, which are typically scarce. The analysis focuses on players aged 15–23 from clubs in the top 18 leagues worldwide and evaluates each player at two points: the start of the simulation (Year-0) and ten years later (Year-10). All experiments are repeated across three independent runs to reduce simulation randomness.

Two configurations are examined—With-Age and No-Age—to isolate the independent contribution of age as a predictive feature. Each configuration is evaluated in both a Realistic Mode (excluding hidden attributes such as CA and PA) and a Full Mode (including them), enabling a structured four-way comparison.

Success is defined using a dual benchmark. Real-world benchmarking is based on scraping Transfermarkt [6] market values identify the top 25% (Q3) of global players. Mapping this to the simulated distribution shows that the equivalent cutoff inside FM corresponds to the top 10% in Year-10. A player is labeled successful if they reach this threshold in at least two of the three simulations.

After preprocessing, feature refinement, and addressing extreme class imbalance with SMOTE, multiple models are trained, including Logistic Regression, Decision Trees, Random Forests, SVM, and XGBoost. Hyperparameters are optimized using Optuna, and interpretability is assessed through SHAP, coefficients, and permutation importance. Cluster analysis (K-Means) is also used to explore structural patterns among successful players.

Results show that the Full Mode consistently outperforms the Realistic Mode, with XGBoost achieving a balanced accuracy of 0.8999 in the With-Age Full configuration—the strongest overall performance. Even in Realistic Mode, ensemble methods demonstrated meaningful predictive capability (balanced accuracy ≈ 0.76–0.79), confirming that observable attributes carry genuine signals about long-term success. The study successfully validated FM as a complementary platform for talent evaluation, though simulation-based data cannot replace professional scouting.

List of figures

List of Tables

Table of Contents

1. Introduction

Long-term success prediction in football is one of the most complex challenges in sports analytics. Real-world player development is influenced by many interacting factors, such as physical growth, tactical adaptation, injuries, environmental conditions, and quality of coaching. While modern clubs increasingly use data-driven methods to support scouting and decision-making, comprehensive longitudinal datasets comprising thousands of young players tracked over many years are seldom available. Most public datasets focus on mature players and top-tier leagues, hence offering limited insight into early-career development or long-term progression. These limitations motivate an exploration of alternative data sources offering structure, consistency, and depth.

FM, developed by Sports Interactive, offers a huge simulation environment with tens of thousands of players described by detailed technical, mental, and physical attributes. FM is widely regarded as highly realistic, supported by one of the most extensive scouting networks. Several real-life stories illustrate how much the game reflects football logic, for example, the well-documented story of Will Still [7], whose early tactical learning came from FM before becoming a professional manager himself. Community reports and media coverage further illustrate how FM's structured attribute system and tactical environment often resonate with the intuitions of practitioners, analysts, and aspiring coaches.

Based on this foundation, the current study investigates whether machine learning models can predict which young players—who are between 15 and 23 years old at the start of the simulation—eventually become successful after ten simulated seasons in FM [1]. Success is defined using a dual-benchmark framework that calibrates simulated outcomes against real-world market valuations, as detailed below.

Defining Success through A Dual-Benchmark Framework

To ensure that the success definition is grounded in real-world standards rather than arbitrary in-game thresholds, this study adopts a dual-benchmark approach:

1. Real-world benchmark (Transfermarkt [6]):

Real-world success is defined using market value data scraped from Transfermarkt on October 12, 2025, for the top 500 players globally. A player is considered successful in real-world terms if they fall into the Top 25% (Q3) of this distribution, reflecting elite-level valuation within the global football market.

2. In-game benchmark (FM):

By comparing the real-world Transfermarkt distribution to simulated Year-10 FM market values, the real-world Top 25% threshold was found to correspond approximately to the Top 10% within FM's simulated market value distribution. Thus, in-game success is defined as attaining a market value ranking in the Top 10% after ten simulated years.

Majority-vote labeling procedure:

To account for FM's inherent randomness arising from injuries, transfers, and match outcomes, each player is simulated across three independent 10-year runs. A player is labeled as successful only if they reach the Top 10% threshold in at least two of the three simulations. Figure 1 illustrates this procedure.

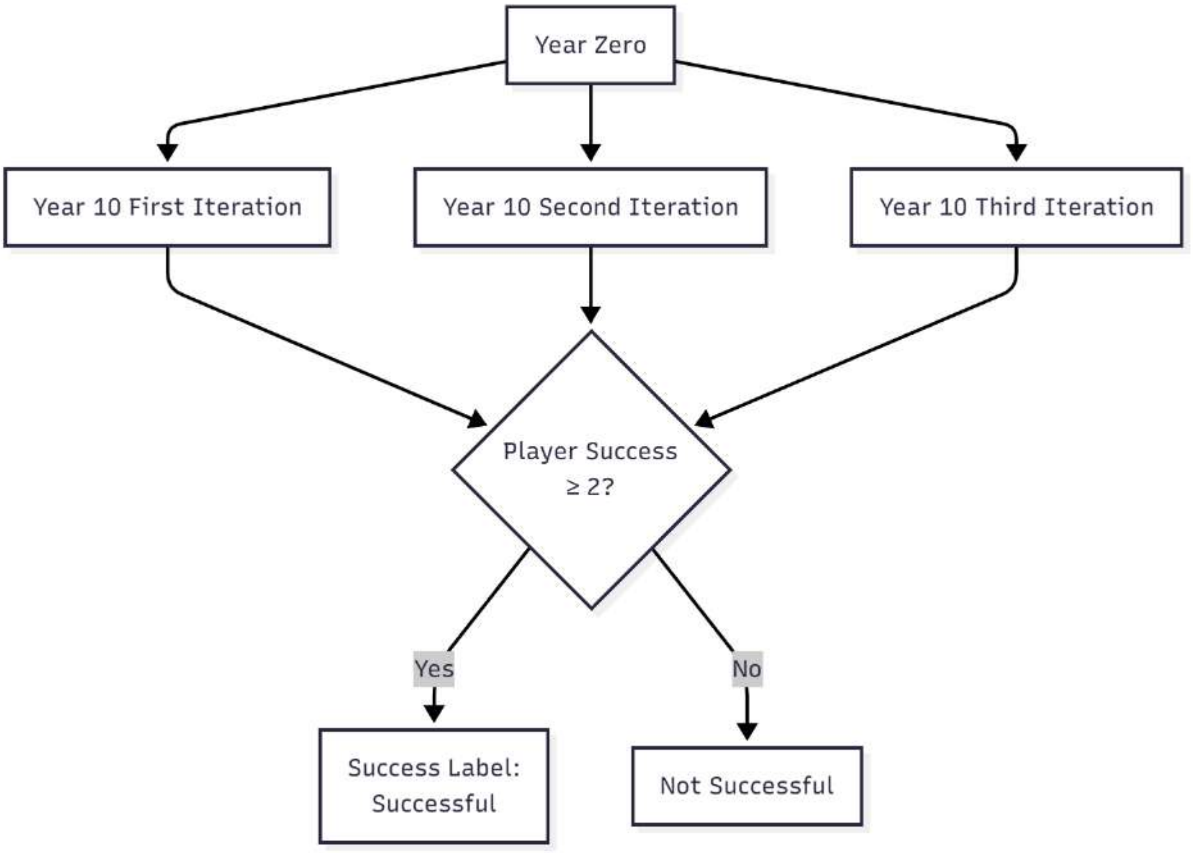


Figure 1: Success Label Determination

Why market value?

Market value is used as the core success metric because it is the only quantitative indicator that integrates all dimensions of a player's long-term career into a single measurable outcome. Unlike isolated performance statistics—such as goals, assists, or minutes played—which differ substantially by playing position, tactical role, and league strength, market value reflects a holistic evaluation that incorporates performance consistency, physical and technical development, injury history, competition level, contractual stability, and external demand from clubs. Additionally, football is one of the few global sports that operates in an open transfer market in where every player has a dynamic financial value. This makes market value uniquely suited for longitudinal analysis, unlike sports such as the NBA, NFL, or MLB where salary caps and draft systems prevent meaningful valuation of individual players. The availability of real-world Transfermarkt data also enables a direct calibration between simulated outcomes and real-world economic valuations.

For these reasons, market value provides a robust, realistic, and widely recognized proxy for long-term professional success in football.

As might be expected from such a rigid and consistency-based definition, only 3.3% of players achieve elite long-term outcomes across multiple simulations. This distribution is shown in Figure 2. The rarity of success is intuitive and aligned with real-world football.

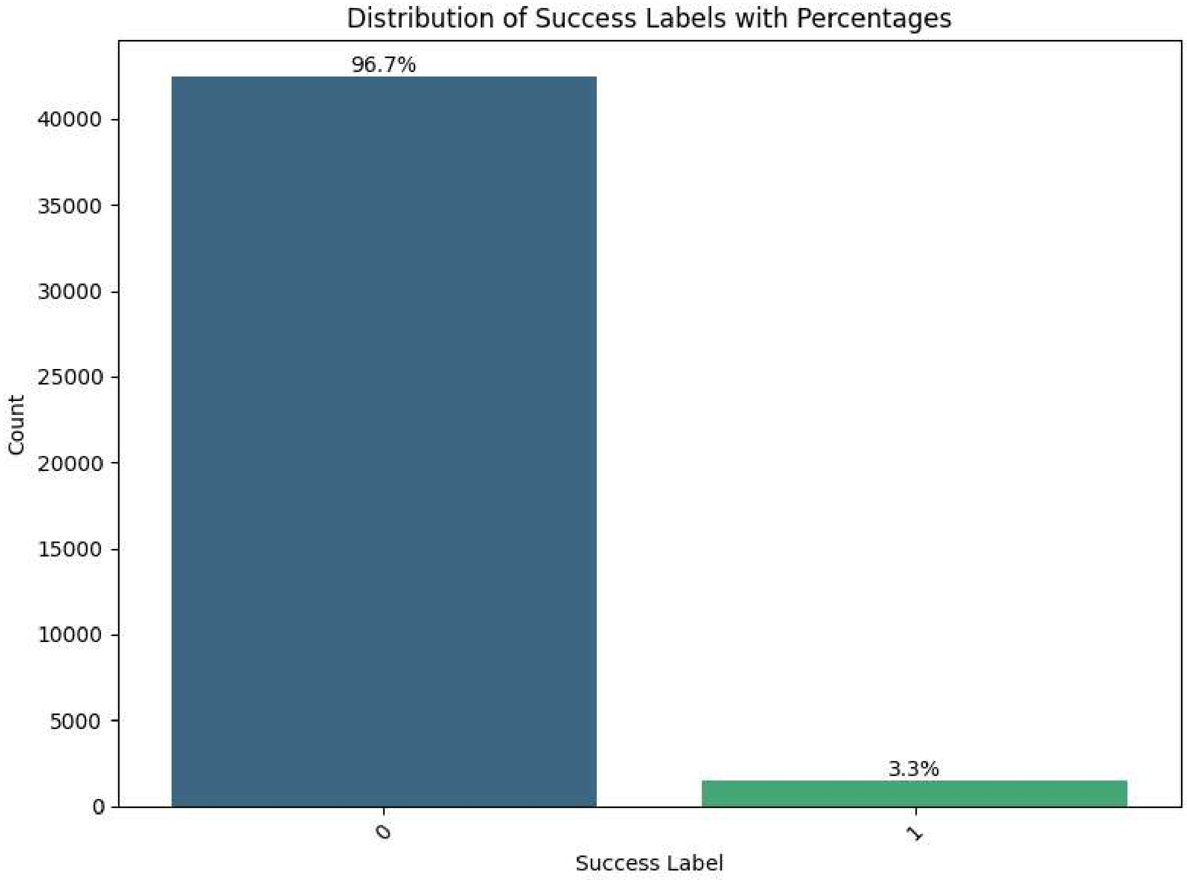


Figure 2: Distribution of Success Labels with Percentages

The dataset includes players aged 15-23 from the top 18 football leagues, extracted at Year-0 and Year-10. FM does not publish its internal algorithms for player development or match simulation. This study therefore treats FM as a black-box environment, with all simulations executed under identical conditions to ensure that outcome differences arise from inherent stochasticity rather than user input.

Model Configurations

This study employs a four-way experimental design that evaluates models under two configurations (With-Age vs. No-Age) and two information modes (Realistic vs. Full). The With-Age and No-Age configurations isolate the specific contribution of age as a predictive feature, while the Realistic and Full modes distinguish between information available to real-world scouts versus complete internal game data including hidden attributes such as Current Ability (CA) and Potential Ability (PA). This design allows the study to answer two key questions: (1) Does Age provide genuine predictive value or merely act as a proxy for contextual factors? (2) How much additional predictive power is gained from hidden attributes that scouts cannot access?

Study Overview

Multiple machine learning models are evaluated, including Logistic Regression, Decision Trees, Random Forests, SVM, Cluster analysis, and XGBoost. The goal is not to reach any specific accuracy threshold but to understand the intrinsic complexity of the elite success prediction task, how extreme class imbalance influences model behavior, and which early-career attributes meaningfully contribute to predictive performance. Class imbalance is addressed using SMOTE, and interpretability is supported by SHAP, Coefficient/permutation importance. The complete methodology is described in Section 3.

2. Related work

2.1 **FM in Academic Research**

FM has increasingly attracted academic interest as a structured, data-rich environment for studying player development and decision-making. One of the most relevant works is the master’s thesis by van Wijk (2022), which used the FM20 database to predict a player’s hidden Potential Ability (PA) using machine learning models such as Logistic Regression, Decision Trees, and Support Vector Machines. His study demonstrated that FM’s internal attribute system is sufficiently consistent for predictive modeling, and that certain early-career attributes—particularly balance—strongly correlate with long-term potential. Although his work focused on predicting PA rather than real or simulated career outcomes, it established FM as a credible platform for research on player evaluation and talent identification.

A second significant contribution comes from Rocha Lima et al. (2018), who examined FM as a developmental tracking tool for young Brazilian players. Their study simulated several seasons and compared the progression of 16- and 20-year-old players, concluding that FM’s progression patterns reasonably approximate real-world development trajectories. This work positioned FM as a viable proxy for longitudinal analysis when real datasets are limited or inaccessible. Collectively, these studies validate the use of FM as a controlled experimental environment and provide the methodological foundation upon which the present study builds—extending prior research by shifting from PA prediction to long-term success forecasting grounded in market value outcomes.

2.2 **Positioning This Study**

While van Wijk (2022) [3] provided valuable insights into predicting player potential using FM data, this study extends and differentiates from that work in several fundamental ways, addressing key methodological and practical limitations while introducing novel approaches to validation and evaluation.

2.2.1 **Prediction Target: Market Value vs. Potential Ability**

The most fundamental difference lies in the prediction target itself. van Wijk (2022) predicted Potential Ability (PA), a fixed hidden attribute in the FM database that represents the maximum theoretical level a player could reach. PA is a static value determined at the start of the game and never changes throughout a player's simulated career a game-internal rating with no direct connection to real-world football economics.

In contrast, this study predicts long-term market value after 10 simulated years, representing a player's actual value in the simulated football economy. This approach offers three key advantages:

1. Economic realism: Market value reflects the cumulative outcome of performance, development, injuries, transfers, and external demand—factors that matter to real clubs and scouts.
2. Real-world calibration: By anchoring the in-game success threshold to real-world Transfermarkt [6] data (mapping the real-world top 25% to

FM's top 10%), the study grounds its success definition in actual market dynamics rather than an arbitrary game-internal cutoff.
3. Practical relevance: Clubs invest in players based on expected value and impact, not abstract potential ratings. Market value serves as a holistic proxy for professional success.

2.2.2 Multi-Simulation Validation vs. Single Snapshot

van Wijk (2022) used a single dataset extracted from FM 2020—a cross-sectional snapshot where each player had one set of attributes and one PA value. However, FM introduces substantial randomness into player development, match performance, and career trajectories.

This study implements three independent 10-year simulations for each player, all starting from the same Year-0 state. A player is labeled as successful only if they reach the top 10% market value threshold in at least two out of three simulations (majority-vote rule). This multi-simulation validation strategy provides:

1. Reduced noise: Random events (e.g., a catastrophic injury, an unlikely transfer) are less likely to dominate the success label.
2. Robust outcomes: Success labels reflect consistent performance across multiple independent career paths, not one-off trajectories.
3. Longitudinal design: The prediction task becomes: given a player's Year-0 attributes, will they achieve elite success 10 years later? Mirroring real scouting decisions that evaluate young players based on projections of future success.

2.2.3 Model Configurations: Ablation Analysis and Information Modes

van Wijk (2022) evaluated models on a single feature set that included all public attributes, without systematically investigating the contribution of specific features or distinguishing between information available to real scouts versus hidden game internals.

This study introduces a four-way experimental design:

Configuration (Age Handling):

- No-Age: Excludes Age to test whether predictions depend on legitimate attribute-based signals or simply exploit age as a proxy.
- With-Age: Includes Age to measure its independent contribution.

Information Mode:

- Realistic Mode: Uses only publicly visible attributes that real-world scouts can access (technical, mental, physical attributes).
- Full Mode: Includes hidden attributes (CA, PA), providing an upper-bound benchmark when complete internal information is available.

This ablation approach answers two critical questions: (1) Does Age provide genuine predictive value? (2) How much predictive power is gained from hidden information scouts cannot access?

2.2.4 Additional Methodological Enhancements

Beyond these core differences, this study incorporates several other improvements:

- SMOTE sensitivity analysis: Systematically evaluates three oversampling ratios (0.2, 0.5, 1.0) rather than a single arbitrary setting, ensuring conclusions are robust across varying degrees of class balance.
- Enhanced evaluation framework: Following He and Garcia [9], adopts a comprehensive multi-metric approach including F1-score, Balanced Accuracy, MCC, Cohen's Kappa, ROC-AUC, and Precision-Recall curves, particularly important given the extreme class imbalance (~3.3% success rate).
- Automated data extraction: Developed a custom PyAutoGUI [5] pipeline for reproducible, scalable extraction across three simulations (vs. using a third-party static Kaggle dataset).
- Efficient hyperparameter optimization: Uses Optuna [8] with TPE sampler (100 trials) instead of grid search, providing better exploration of the hyperparameter space.
- Model-agnostic interpretability: Employs SHAP, Coefficient and permutation importance for unified feature importance analysis

Table 1 is a summary of the key differences between the two studies

Table 1: Summary of Key Differentiators

Aspect	van Wijk (2022)	This Study
Prediction Target	Potential Ability (PA) - fixed game attribute	Market value at Year-10 - dynamic economic outcome
Real-World Calibration	None	Dual-benchmark framework using Transfermarkt data
Temporal Scope	Cross-sectional snapshot	Longitudinal (10-year simulation)
Validation Strategy	Single dataset	Three independent simulations with majority-vote labeling
Model Configurations	Single feature set	Four configurations (With/No-Age × Realistic/Full modes)
SMOTE Sensitivity	Single ratio	Three ratios (0.2, 0.5, 1.0) with structured analysis
Evaluation Metrics	Accuracy, precision, recall, F1	Multi-metric framework including MCC, Kappa, ROC-AUC, PR curves
Hyperparameter Tuning	Grid search	Optuna with TPE sampler (100 trials)
Interpretability	Coefficient/permutation importance	SHAP/ Coefficient/permutation importance
Dataset Size	56,562 players	43,094 players (filtered from 88,000)
Success Rate	6% (PA ≥ 130)	3.3% (top 10% market value, 2/3 simulations)

3. Methods

To structure the methodological design clearly and systematically, the following subsections outline the study objectives, data collection process, model configurations, class-imbalance strategy, model-training procedures, and the evaluation metrics used to assess performance.

3.1 **Study Objectives**

The methodological design is guided by four core objectives that connect the simulation environment, machine learning framework, and analytical validation into a unified research structure:

1. Evaluate FM as a simulation environment by examining whether long-term in-game success aligns with real-world valuation patterns fromTransfermarkt.

2. Bridge sports analytics with modern data science methodology through the construction of a rigorous machine learning framework that incorporates proper experimental controls and comprehensive evaluation metrics.

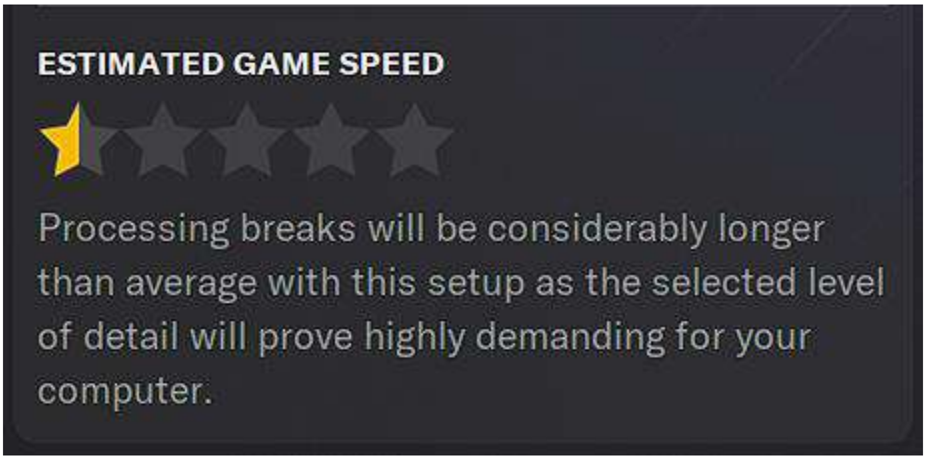
3. Extend and differentiate prior academic work especially van Wijk (2022) [3] by replacing PA prediction with a market-value–based success definition, implementing multi-simulation validation, and comparing Realistic versus Full modes alongside With-Age versus No-Age configurations.

4. Identify key early-career attributes that shape long-term success through multiple ML models, imbalance-handling techniques, and feature-importance analyses including SHAP, coefficient-based, and permutation-based methods.

3.2 **Data Collection**

3.2.1 **Computational Constraints**

The full FM database contains approximately 206,740 players across all leagues and divisions worldwide, covering more than 50 countries across Africa, Asia, Europe, North America, and South America. Running long-term simulations on this entire population would significantly slow down game performance to an estimated 0.5/5 simulation speed using my personal hardware as shown in Figure 3, making it computationally infeasible for this project.


ESTIMATED GAME SPEED

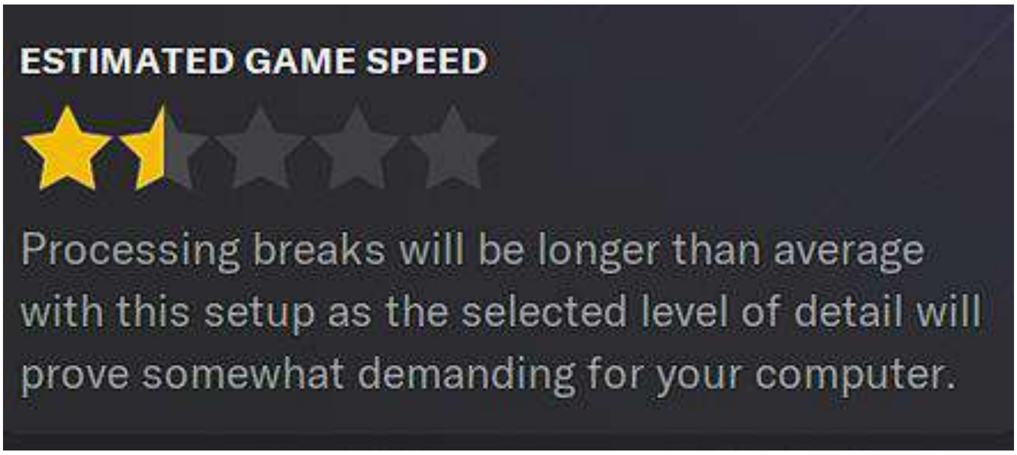
Processing breaks will be considerably longer than average with this setup as the selected level of detail will prove highly demanding for your computer.

Figure 3: Game Speed Using the Full Dataset

League Filtering to Improve Performance

To balance simulation time and data quality, the player pool was restricted to the top 18 leagues, selected based on the Global Football Rankings [4]. These leagues represent highly competitive football environments, ensuring the dataset remains realistic, relevant, and diverse while reducing its size.

This filtering reduced the player count to approximately 88,000 players, allowing the simulation to run at a manageable speed of 1.5/5 as shown in Figure 4 below.



ESTIMATED GAME SPEED

★½☆☆☆

Processing breaks will be longer than average with this setup as the selected level of detail will prove somewhat demanding for your computer.

Figure 4: Game Speed Using the Top 18 Leagues Dataset

Included Leagues (Top 18)

- Argentina: Argentine Premier Division
- Belgium: Jupiler Pro League
- Brazil: Brazilian National First Division
- Croatia: Croatian First League
- Denmark: 3F Superliga
- England: English Premier Division, Sky Bet Championship
- France: Ligue 1 Uber Eats
- Germany: Bundesliga, Bundesliga 2
- Italy: Italian Serie A
- Japan: J1 League

- Mexico: Mexican First Division
- Netherlands: Eredivisie
- Poland: PKO Bank Polski Ekstraklasa
- Portugal: Portuguese Premier League
- Spain: Spanish First Division
- United States: Major League Soccer (MLS)

This filtering strategy drastically reduces computational load while maintaining a high-quality player sample from the world's most competitive leagues.

3.2.2 Data Source, extraction pipeline

All in-game data used in this project was extracted directly from FM [1]. Since the game provides no built-in database export functionality and limits manual exporting to 200 players at a time, a fully automated extraction pipeline was developed using the PyAutoGUI library [5]. This tool enables scripted mouse-and-keyboard interactions, allowing the system to operate the interface exactly as a human user would.

The extraction process begins in Year-0. A custom script scans the player search interface, selects players in batches of 200, saves each batch as a new shortlist, and exports it is using the game's HTML export feature. Each file is automatically named to preserve batch order. Once all batches are exported, a second script converts every shortlist into FM's internal FMF format. These FMF files store the exact same set of players and ensure they can be reliably reloaded later in the simulation, even if players retire or are replaced by synthetic players.

After completing the Year-0 extraction, the game is simulated forward for ten in-game years with no further automation. At Year-10, a third script loads each FMF shortlist, displays the corresponding players, and exports updated HTML files containing their Year-10 attributes and market values. The game is then reloaded back to the original Year-0 save, and the entire ten-year simulation is repeated two more times to produce three independent outcomes. Each run begins from the same initial state, but FM’s stochastic development system introduces variability in transfers, injuries, and progression.

This process generates many files: each extraction point (Year-0 or Year-10) produces 224 HTML files containing roughly 200 rows and 73 columns. To avoid manual merging, a Python data-processing script automatically scans the export directory, identifies all HTML files using the glob module, reads their tables with pandas’ read_html(), and concatenates them into a unified dataset. Duplicate entries are removed using each player’s unique identifier (UID), and the final dataset is exported as a single CSV file. This automated consolidation reduces several hours of manual work to approximately thirty seconds per dataset while ensuring consistency, reliability, and scalability across all simulation iterations. Figure 5 summarizes the full data-extraction pipeline.

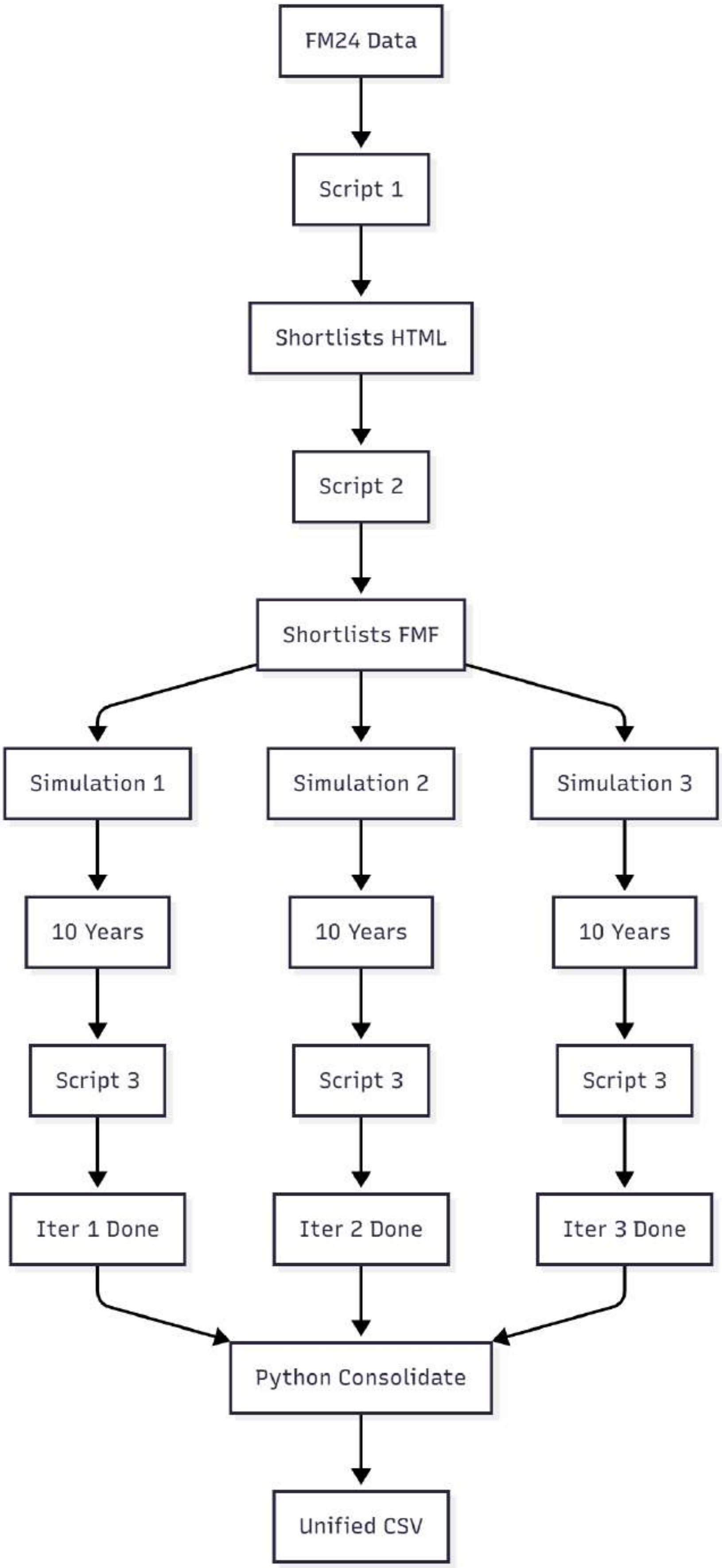


Figure 5: End-to-End Data Extraction Pipeline for FM

Details on data processing, attribute selection, and dataset structure are provided in appendix 1

3.3 Model Configurations

To examine the role of information visibility and feature design, the study evaluates four configurations combining two dimensions: age inclusion and attribute visibility. This experimental design allows for systematic investigation of how different types of information contribute to predictive performance.

Age-Based Configurations

The With-Age and No-Age configurations isolate the specific contribution of age as a predictive feature. Age raised methodological concerns during the early stages of this project regarding whether it could introduce unintended leakage or provide an unfair predictive advantage by merely acting as a proxy for contextual factors such as league quality or positional distributions. To address this uncertainty rigorously, the analysis was structured to evaluate models both with and without Age under identical conditions, following a standard ablation-style approach that allows the effect of Age to be measured independently of all other features.

Information Modes

Each age configuration is evaluated under two information modes that reflect different levels of access to player attributes:

- Realistic Mode: includes only attributes that are visible to a regular FM player and therefore represents the information available to real-world scouts. This mode excludes hidden internal attributes such as Current Ability (CA) and Potential Ability (PA), simulating realistic scouting conditions where decisions must be made based on observable player characteristics: technical skills, mental attributes, and physical measurements—rather than game-internal ratings.
- Full Mode: includes all attributes from Realistic Mode plus the hidden internal attributes CA and PA. This mode serves as an upper-bound benchmark, demonstrating how much predictive performance becomes available when the model has access to complete internal information that scouts cannot observe in real-world settings. By comparing Realistic and

Full modes, the study quantifies the predictive advantage gained from perfect knowledge of the game's internal player ratings.

Four-Way Experimental Design

The combination of these two dimensions yields four distinct configurations, as illustrated in Figure 6:

1. **No-Age + Realistic Mode:** Uses only publicly visible attributes, excluding Age
2. **No-Age + Full Mode:** Includes CA and PA, but excludes Age
3. **With-Age + Realistic Mode:** Uses publicly visible attributes, including Age
4. **With-Age + Full Mode:** Includes all attributes (visible + hidden + Age)

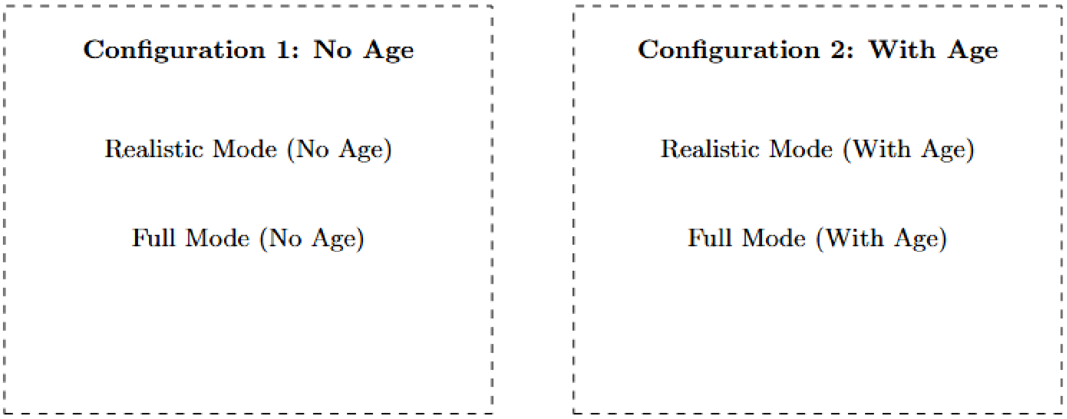


Figure 6: Overview of the two model configurations used in this study, comparing No-Age and With-Age setups, each containing both Realistic and Full prediction modes.

This four-way comparison enables the study to address two critical research questions:

Research Question 1: What is the independent effect of including or excluding Age?
By comparing No-Age versus With-Age configurations while holding the information mode constant, the study can isolate Age's contribution to predictive accuracy and determine whether it provides legitimate predictive value or merely reflects underlying league/position distributions.

Research Question 2: How much additional predictive power is gained from hidden attributes (CA/PA)?
By comparing Realistic versus Full modes while holding the age configuration constant, the study quantifies the value of internal game information that real scouts cannot access in practice.

SHAP values, coefficient and permutation importance are used to interpret model predictions under these configurations, identifying which early-career attributes contribute most strongly to long-term success and whether those patterns differ between information modes. This structured approach ensures that the evaluation framework remains consistent across all experiments while providing interpretable insights into feature contributions. In Section 4, the performance of the four configurations is compared to determine whether Age materially improves predictive accuracy.

3.4 **Handling Class Imbalance**

The dataset exhibits extreme class imbalance, with only 3.3% of players achieving long-term success. A classifier predicting all players as unsuccessful would achieve 96.7% accuracy while completely failing to identify successful players. To address this issue, the Synthetic Minority Over-sampling Technique (SMOTE) [10] was adopted. Unlike random oversampling, SMOTE generates synthetic minority-class samples by interpolating new points along the feature space of similar instances based on k-nearest-neighbor structure, thereby reducing overfitting.

Multi-Ratio Sensitivity Analysis

Rather than fixing a single oversampling ratio, three SMOTE configurations were evaluated to ensure conclusions remain robust across varying degrees of class balance:

- Small oversampling (ratio = 0.2): Minority class increased to 20% of majority
- Medium oversampling (ratio = 0.5): Minority class increased to 50% of majority
- Large oversampling (ratio = 1.0): Full class balance achieved

Different model families respond differently to class distribution changes. Tree-based models (Random Forest, XGBoost) often perform well under moderate imbalance, while linear models (Logistic Regression) typically require stronger balancing. SVM can be sensitive to the exact ratio and the geometry of the minority class. Evaluating multiple ratios prevents drawing conclusions from a single arbitrary SMOTE setting and reveals model stability under different resampling intensities.

Integration with Hyperparameter Optimization

SMOTE is not applied globally. Instead, each machine learning model is trained under each SMOTE ratio, and Optuna selects the best-performing configuration by jointly optimizing model hyperparameters, SMOTE-generated decision boundaries, and classification threshold tuning. This means different models may ultimately prefer different SMOTE ratios exactly what is observed in Section 4.

Implementation Details

- SMOTE is applied only to the training set after the train-validation-test split, ensuring no data leakage
- Validation and test sets maintain the original 3.3% success rate, producing realistic performance estimates
- Oversampling is incorporated directly inside the imbalanced-learn pipeline, ensuring consistent preprocessing across models

This multi-ratio framework provides three key advantages:

1. sensitivity analysis: ensures conclusions are not dependent on a single sampling assumption
2. model-specific optimization reveals which algorithms are robust to imbalance and which rely heavily on resampling
   practical guidance emerges for future talent identification systems facing similarly skewed datasets.

Model Training

Model training was organized around the four configurations described in Section 3.3 (No-Age vs. With-Age × Realistic vs. Full). For each configuration, a separate feature matrix X and target vector y were constructed and passed through the same supervised-learning pipeline, differing only in the set of available predictors (e.g., exclusion of CA/PA in Realistic Mode, exclusion of Age in the No-Age configuration). This structure ensures that any performance differences are attributable to information visibility and Age handling rather than changes in the training procedure.

3.4.1 **Algorithms**

The study evaluates five supervised machine learning algorithms and one unsupervised method; all applied consistently across the four model configurations described in Section 3.3. Each supervised model was embedded in an imbalanced-learn Pipeline consisting of:
(SimpleImputer → SMOTE → StandardScaler → Classifier).
This design ensures:

1. consistent preprocessing across models
2. correct application of SMOTE only on the training portion of each split,
3. fair comparability across Realistic vs. Full Modes and With-Age vs. No-Age configurations.

Supervised Algorithms

- **Logistic Regression (LR)** [11]
  Serves as a linear baseline model and provides interpretable coefficients under standardized features. Its simplicity allows examination of linear relationships between player attributes and long-term success.

- **Decision Tree (DT)** [12]
  A non-linear model capable of capturing hierarchical interactions between technical, mental, and physical attributes. DT also provides transparent feature-split logic useful for interpretability.

- **Random Forest (RF**) [13]
  An ensemble of bootstrapped trees that reduces variance and improves stability over a single DT. RF is robust to noise, captures non-linear patterns, and handles high-dimensional data effectively.

- **XGBoost** [14]
  A gradient-boosted decision-tree model well-suited for imbalanced, structured datasets. XGBoost frequently achieves strong performance in

tabular prediction problems and was included as a high-capacity model to benchmark upper-bound predictive performance.

- Support Vector Machine (SVM, RBF kernel) [15]
  A margin-based classifier that models non-linear decision boundaries through a radial basis function kernel. SVM has been previously used in FM research and provides a useful contrast to tree-based methods, especially in high-dimensional spaces.

Unsupervised Algorithm (Exploratory Analysis)

- **K-Means Clustering** [16]
  Used solely for exploration analysis to identify latent player archetypes. Clustering was conducted primarily on hidden ability attributes (CA, PA) and market value, with optional inclusion of selected performance attributes. The optimal number of clusters was selected using elbow and silhouette criteria.
  To aid interpretation, the resulting clusters were visualized through:
  - **Principal Component Analysis (PCA)** [17] linear dimensionality reduction to 2D space
  - **t-SNE [18]** non-linear embedding that highlights local player-similarity structure

These visualizations were diagnostic only and were not used as inputs to supervise models.

3.4.2 Hyperparameter Optimization (Optuna, TPE, 100 trials)

Hyperparameter tuning for all supervised models (Logistic Regression, Decision Tree, Random Forest, XGBoost, and SVM) was performed using Optuna with the Tree-structured Parzen Estimator (TPE) sampler. Each model–configuration pair (Realistic vs. Full × With Age vs. No-Age) was optimized using 100 trials, following Optuna's documented recommendation [8] that at least ~100 trials are required for the sampler to explore the search space effectively.

Within each trial, Optuna sampled classifier-specific hyperparameters as well as the SMOTE oversampling level (small, medium, large), allowing each model to select the imbalance-handling strategy most suitable for its structure. The exact search space used for each classifier is documented in Appendix 2.

- **Logistic Regression** – Regularization strength *C*, penalty type (ℓ_1/ℓ_2 where appropriate), and solver
- **Decision Tree** – Maximum depth, minimum samples per split, minimum samples per leaf, and splitting criterion
- **Random Forest** – Number of trees, maximum depth, minimum samples per split/leaf, and maximum features
- **XGBoost** – Number of estimators, learning rate, maximum depth, subsample, and column sampling parameters
- **SVM (RBF)** – Regularization strength *C* and kernel width γ

Optimization Objective: Maximizing Balanced Accuracy

Data Partitioning Strategy (Train–Validation–Test Split)

All supervised experiments employed a three-way partitioning strategy to maintain clean separation between model fitting, hyperparameter selection, and final evaluation. After constructing the feature matrix *X* and target variable *y* for each configuration, the dataset was divided as follows:

1. Test Set (15%) – Held out completely until final evaluation. No model training or hyperparameter tuning accessed this split.
2. Training + Validation Pool (85%) – Further subdivided into:
    - *Training Set (~70%)* – Used to fit models during each Optuna trial
    - *Validation Set (~15%)* – Used exclusively to evaluate Balanced Accuracy as the Optuna optimization objective

All splits were stratified to preserve the natural class imbalance (~3.3% success rate) across subsets.

Rationale for Fixed Validation Instead of Cross-Validation

Cross-validation was intentionally avoided for several interconnected reasons:

1. Computational infeasibility. Each Optuna trial fits a complete pipeline (imputation → SMOTE → scaling → classifier). Introducing 5-fold CV

would multiply the cost of every trial by 5×. Across 5 algorithms, 4 configurations, and 100 trials per combination, this becomes prohibitively expensive.

2. SMOTE overhead and complexity. Proper CV implementation requires applying SMOTE separately within each training fold to prevent data leakage. Repeating this resampling across folds, trials, and models dramatically amplifies runtime and introduces additional instability.
3. Instability under extreme imbalance. With only ~3.3% positive cases, individual CV folds contain very few successful players, producing noisy and unreliable objective values that hinder effective hyperparameter optimization.

The train–validation–test split therefore strikes the optimal balance between computational feasibility, methodological rigor, and reproducibility for this study.

3.5 **Evaluation Metrics**

Evaluating predictive performance in this study requires metrics that remain reliable under severe class imbalance, where only ~3.3% of players achieve long-term success. Standard accuracy is not meaningful in this setting, as a trivial classifier predicting all players as unsuccessful would exceed 96% accuracy.

As emphasized by He and Garcia [9], "a singular evaluation metric, such as overall classification error rate, is not sufficient when handling imbalanced learning problems." They recommend "a combination of singular-based metrics (e.g., precision, recall, F-measure, and G-mean) together with curve-based assessment metrics [e.g., receiver operating characteristic (ROC) curve, precision–recall (PR) curve]" to provide comprehensive evaluation. Accordingly, this evaluation framework emphasizes metrics that balance performance across classes and highlight the model's ability to correctly identify rare successful players.

3.5.1 **Balanced Accuracy (Primary Metric)**

Balanced Accuracy [19] serves as the primary evaluation metric and the objective function for Optuna hyperparameter optimization. It is defined as:

Balanced Accuracy = ½(Sensitivity + Specificity)

This metric treats both classes equally, ensuring that the model cannot achieve a high score by simply predicting the majority class. Balanced Accuracy therefore provides a stable, unbiased measure under extreme imbalance and explicitly forces optimization toward detecting the minority class.

3.5.2 Precision and Recall

Precision and Recall [20] are critical for understanding the error profile of the minority "successful" class:

- Precision: Precision = TP/(TP + FP)
  Measures how many predicted successful players were successful.
- Recall (Sensitivity): Recall = TP/(TP + FN)
  Measures the model's ability to detect successful players.

Recall is especially important in this context because missing successful players (false negatives) represents a significant practical error.

3.5.3 F1-Score (Positive Class)

The F1-score [21] is the harmonic mean of Precision and Recall:

F1 = $2 \times (Precision \cdot Recall)/(Precision\ +\ Recall)$

Because detecting successful players is the primary focus, the F1-score of the minority class serves as a key evaluation metric. A second operating threshold is selected for each model by maximizing F1-score on the test set, reflecting the model's best-case retrieval performance.

3.5.4 Precision–Recall (PR) Curves

Following the recommendations of He and Garcia [9], PR curves [22] are emphasized over ROC curves because they better reflect minority-class performance under extreme imbalance. While ROC-AUC can inflate perceived performance when class distributions are severely skewed, PR curves directly illustrate the trade-off between Precision and Recall across all decision thresholds, enabling clearer visual comparison of each model's ability to retrieve successful players while limiting false positives.

3.5.5 Dual-Threshold Evaluation

Each model is evaluated under two decision thresholds:

1. **F1-optimized threshold** – Prioritizes recall of successful players
2. **Balanced-Accuracy-optimized threshold** – Provides a more conservative, balanced operating point

This dual-threshold approach illustrates how model behavior shifts depending on whether the goal is maximizing detection or balancing errors across classes.

3.5.6 Geometric Mean (G-Mean)

$$\text{G-Mean} = \sqrt{(\text{Sensitivity} \cdot \text{Specificity})}$$

G-Mean [23] penalizes poor minority-class recall more strongly than Balanced Accuracy and captures how well the model balances performance across both classes.

3.5.7 Matthews Correlation Coefficient (MCC)

$$\text{MCC} = (TP \cdot TN - FP \cdot FN) / \sqrt{[(TP+FP)(TP+FN)(TN+FP)(TN+FN)]}$$

MCC [24] is a robust correlation-based metric that incorporates all four confusion matrix components and remains reliable even when classes are extremely imbalanced.

3.5.8 Cohen's Kappa

$$\kappa = (p_0 - p_e) / (1 - p_e)$$

where:

- p_0 = observed agreement (accuracy)
- p_e = expected agreement by chance

Cohen's Kappa [25] measures agreement beyond chance and provides a complementary view of classifier reliability compared to raw accuracy or recall-based metrics. It ranges from -1 to 1, where 1 indicates perfect agreement, 0 indicates agreement equivalent to chance, and negative values indicate agreement worse than chance.

4. Results

4.1 Configuration Comparison

All detailed visualizations—including confusion matrices, threshold–metric curves, configuration comparison plots, and the Precision–Recall and ROC curves—have been moved to Appendix 3 for readability. The full Performance Summary tables are provided in Appendix 4, and all Feature-Importance plots (SHAP, coefficient weights, and permutation importance) are moved to Appendix 5. This section reports only the key comparative findings and aggregated results.

4.1.1 Statistical Relationship Between Age and Success

Before evaluating the impact of Age within the predictive models, we first determine whether age is statistically related to long-term success in the dataset. To do this, several complementary tests are applied:

1. **Point-biserial correlation [26]** – quantifies the linear association between Age and the binary success_label.
2. **Mann–Whitney U test [27]** – compares the age distributions of successful and unsuccessful players.
3. **Chi-square test of independence [28]** – evaluates whether categorical age groups (≤20 vs 21–23) differ in success rates.
4. **Descriptive subgroup comparison** – examines observed success rates within each age group.
5. **Visual analysis** – boxplot, histogram, and bar chart illustrating how Age varies across outcomes.

The results of these tests, along with the corresponding figures, are reported as follows:

Point-Biserial Correlation

A point-biserial correlation was computed between Age and success_label:

- r = 0.1213

- $p < 1.44 \times 10^{-143}$

A correlation heatmap summarizing the top attributes associated with Age is shown in .

Mann–Whitney U Test

A Mann–Whitney U [27] test compared the age distributions of successful and unsuccessful players:

- $U = 42{,}866{,}537$
- $p < 1.64 \times 10^{-149}$

The distributional differences are visualized using a boxplot of Age by success_label in Figure 8.

Chi-Square Test of Independence

Age was grouped into two categories (≤20 vs 21–23)the chi-square test [28] returned:

- $\chi^2 = 1237.84$
- $p < 3.64 \times 10^{-271}$

The contingency table revealed that the Peak age group (21–23) had 13,998 unsuccessful and 1,122 successful players, while the Young age group (≤20) had 28,462 unsuccessful and 321 successful players.

Descriptive Subgroup Comparison

Observed success rates by age category:

- Young (≤20): N = 28,783, success rate = 1.12%
- Peak (21–23): N = 15,120, success rate = 7.42%

This difference is visualized in Figure 9, which presents a bar chart comparing success rates across the two age categories.

To further summarize distributional patterns, Figure 10 presents a histogram of Age for successful and unsuccessful players..

These statistical outputs describe the empirical relationship between Age and long-term success and provide the basis for evaluating predictive performance under the With-Age and No-Age model configurations.

4.1.2 With-Age vs No-Age Comparison

While these statistical tests confirm a strong empirical association between Age and long-term success, the direction and cause of this relationship remain uncertain. The observed effect may partly reflect underlying league-level differences in age structure rather than Age itself. Therefore, instead of assuming Age is valid or invalid a priori, the study evaluates its practical impact directly through model performance. The next section compares the With-Age and No-Age configurations to determine how much predictive value Age contributes within the controlled modelling framework.

4.2 Overall Model Performance

This section presents the predictive performance of all machine learning models across the four configurations defined in Section 3 (With-Age vs. No-Age × Realistic vs. Full). All models were evaluated on the held-out test set using the evaluation framework established in Section 3.6 (Evaluation Metrics), under two operating thresholds:

1. F1-optimized threshold
2. Balanced-Accuracy-optimized threshold

4.2.1 Logistic Regression Performance

Performance Summary.

Across all configurations, the Full mode achieved higher balanced accuracy and recall compared to the Realistic mode. In addition, With-Age models outperformed No-Age models under both F1-optimized and balanced-accuracy–optimized thresholds.

For Logistic Regression, the Realistic With-Age models used SMOTE level = small, as shown in Table 3, while the Full With-Age models also used SMOTE level = small, as reported in Table 4.

For the No-Age configurations, Realistic mode used SMOTE level = small, as summarized in Table 5, and Full mode likewise used SMOTE level = small, as presented in Table 6.

4.2.2 Random Forest

Performance Summary.

Across all configurations, the Full mode achieved higher balanced accuracy compared to the Realistic mode. Recall was also higher in the Full mode under the balanced-accuracy–optimized threshold, although the Realistic mode produced slightly higher recall under the F1-optimized threshold in the Without-Age configuration. In addition, With-Age models outperformed No-Age models under both F1-optimized and balanced-accuracy–optimized thresholds.

For Random Forest, the Realistic With-Age models used SMOTE level = small, as shown in Table 7, while the Full With-Age models used SMOTE level = medium, as reported in Table 8.
For the No-Age configurations, the Realistic mode again used SMOTE level = small, as summarized in Table 9, whereas the Full mode used SMOTE level = high, as presented in Table 10.

4.2.3 Decision Tree

Performance Summary.

Across the balanced-accuracy–optimized thresholds, the Full mode achieved higher balanced accuracy and recall compared to the Realistic mode for both the With-Age and Without-Age configurations. Under the F1-optimized threshold, the Full mode outperformed the Realistic mode in the With-Age configuration, while the Realistic mode achieved slightly higher balanced accuracy and recall in the Without-Age configuration.

For Decision Tree, the With-Age configurations used SMOTE level = small for both the Realistic and Full modes, as shown in Table 11 and Table 12. For the No-Age configurations, both the Realistic and Full modes used SMOTE level = medium, as presented in Table 13 and Table 14.

4.2.4 SVC

Performance Summary

Across the balanced-accuracy–optimized thresholds, the Full mode achieved higher balanced accuracy and recall compared to the Realistic mode for the Without-Age configuration. For the With-Age configuration, the Realistic mode produced higher balanced accuracy and F1 scores, while the Full mode achieved perfect recall at the cost of substantially lower balanced accuracy. Under both optimization strategies, the With-Age models generally performed better than the No-Age models in terms of recall and F1 score.

For SVC, the With-Age configurations used SMOTE level = small for both the Realistic and Full modes, as shown in Tables 15 and 16. The No-Age configurations also used SMOTE level = small for both modes, as reported in Tables 17 and 18.

4.2.5 XGBoost

Performance Summary

Across the balanced-accuracy–optimized thresholds, the Full mode achieved higher balanced accuracy and recall compared to the Realistic mode for both the With-Age and Without-Age configurations. Under the F1-optimized threshold, the Realistic mode obtained slightly higher balanced accuracy for both configurations, while the Full mode produced higher recall values. Across both threshold settings, the With-Age models generally performed better than the No-Age models, particularly in terms of balanced accuracy and recall.

For XGBoost, all configurations used SMOTE level = high, as shown in Tables 19 through 22, covering the Realistic and Full modes for both the With-Age and Without-Age conditions.

4.3 Feature Importance Analysis

This section presents the feature-importance results for all models under the With-Age and Without-Age configurations and across both the Realistic and Full modes. Each model reports feature importance using the method appropriate to its underlying structure: coefficient-based importance for Logistic

Regression, impurity-based importance for Decision Tree and Random Forest, SHAP values for XGBoost, and permutation importance for SVC.

4.3.1 Logistic Regression (LR) — Coefficient-Based Importance

Feature importance for Logistic Regression was derived from the absolute values of the learned coefficients. Figures 69 and 70 display the top 10 features for each configuration.

- **With-Age – Realistic Mode (Figure 69):**

  The most influential predictors were Age_Group_Young, Age, Injury Proneness, Anticipation, and Bravery, followed by technical and mental attributes such as Strength, Determination, Concentration, Acceleration, and Finishing.

- **With-Age – Full Mode (Figure 69):**

  When hidden attributes were included, Potential Ability (PA) and Current Ability (CA) became the dominant predictors. Age_Group_Young and Injury Proneness remained among the top contributors, with smaller effects observed from Determination, Finishing, First Touch, and Marking.

- **Without-Age – Realistic Mode (Figure 70):**

  In the absence of age-related features, the ranking shifted toward football-specific attributes. Injury Proneness, Anticipation, Strength, Bravery, Determination, Finishing, Acceleration, and Composure formed the top set of predictors.

- **Without-Age – Full Mode (Figure 70):**

  As expected, PA and CA again dominated the coefficient magnitudes. Other influential predictors included Injury Proneness, Growth Ratio, Determination, Finishing, Anticipation, Decisions, and Marking**.**

4.3.2 Random Forest (RF) — Feature Importance (Impurity-Based)

Feature importance for the Random Forest models was computed using impurity-based importances, derived from reductions in Gini impurity across all trees. Figures 71 and 72 display the top 10 most important features in each configuration.

- **With-Age – Realistic Mode (Figure 71)**

The most influential predictors under the Realistic mode included Age_Group_Young and Age, followed by a group of technical and mental attributes such as Anticipation, Bravery, Teamwork, Balance, Strength, Composure, Determination, and Stamina.

- **With-Age – Full Mode (Figure 71)**

  When hidden attributes were available, Current Ability (CA) and Potential Ability (PA) became dominant. Age-related variables (Age_Group_Young, Age) continued to contribute, alongside additional attributes such as Strength, Bravery, Teamwork, Balance, Composure, and Anticipation.

- **Without-Age – Realistic Mode (Figure 72)**

  Without age features, the model emphasized football-specific attributes. The highest importances were assigned to Bravery, Anticipation, Teamwork, Concentration, Balance, Strength, Composure, Determination, Stamina, and Injury Proneness.

- **Without-Age – Full Mode (Figure 72)**

  With hidden attributes included, CA and PA were the strongest predictors. Other influential attributes included Teamwork, Strength, Bravery, Anticipation, Concentration, Composure, Injury Proneness, and Balance.

4.3.3 Decision Tree (DT) — Feature Importance (Impurity-Based)

Feature importance for the Decision Tree models was computed using impurity-based importances, derived from reductions in Gini impurity along the tree splits. Figures 73 and 74 present the top 10 contributing features for each configuration.

- **With-Age – Realistic Mode (Figure 73)**

The strongest predictor was Age_Group_Young, followed by Age, with additional contributions from Composure, Bravery, Concentration, Balance, Injury Proneness, 1v1, and minor influence from Aggression and Finishing.

- **With-Age – Full Mode (Figure 73)**

When hidden attributes were included, Current Ability (CA) became the dominant feature, with Age_Group_Young and Potential Ability (PA) also ranking highly. Other contributing attributes included Age, Composure, Injury Proneness, First Touch, Command of Area, Professionalism, and Leadership.

- **Without-Age – Realistic Mode (Figure 74)**

Without age variables, the model placed the highest importance on Teamwork, followed by Anticipation, Determination, Injury Proneness, Bravery, Dribbling, Free Kicks, Tackling, Command of Area, and Concentration.

- **Without-Age – Full Mode (Figure 74)**

With hidden attributes available, CA overwhelmingly dominated the importance distribution, with PA as the next strongest contributor. Lower-ranked features included Injury Proneness, Aerial Reach, First Touch, Composure, Decisions, Vision, Acceleration, and Growth Ratio.

4.3.4 Support Vector Classifier (SVC) — Feature Importance (Permutation-Based)

Feature importance for the SVC models was computed using permutation importance, measured as the reduction in balanced accuracy when each feature was randomly permuted. Figures 75 through 78 present the top 10 features for each configuration.

- **With-Age – Realistic Mode (Figure 75)**

The most influential features were Age_Group_Young and Age_Group_Peak, followed by Injury Proneness, Determination, Dirtiness, Natural Fitness, Important Matches, Leadership, Anticipation, and Concentration. Age-group variables contributed the largest balanced-accuracy drops when permuted.

With-Age – Full Mode (Figure76)

When hidden attributes were enabled, permutation importance became highly dispersed, with no feature producing a consistent decrease in balanced accuracy. Among the top-ranked variables were Age_Group_Peak, Age_Group_Young, Growth_Room, Growth_Ratio, Height, Weight, Work Rate, Jumping Reach, Natural Fitness, and Versatility, though their importance values were near zero and did not form a clear hierarchy.

This unusual dispersion will be revisited and discussed in Section 5

- **Without-Age – Realistic Mode (Figure 77)**

In the absence of age-related variables, the strongest predictors were Injury Proneness, Anticipation, Consistency, Determination, Aggression, Strength, Temperament, Leadership, Concentration, and Natural Fitness. These features showed the largest accuracy drops under permutation, although overall magnitudes remained small.

- **Without-Age – Full Mode (Figure 78)**

With hidden attributes available, Potential Ability (PA) and Current Ability (CA) dominated the permutation-importance ranking, producing the largest reductions in balanced accuracy. Additional contributors included Growth Ratio, Injury Proneness, First Touch, Marking, Heading, Finishing, Growth Room, and Tackling.

4.3.5 XGBoost — Feature Importance & SHAP Analysis

XGBoost provides two complementary forms of feature-importance reporting:

1. Built-in tree-based feature importance (gain-based importance), and

2. SHAP (SHapley Additive exPlanations), which quantifies each feature's average impact on predictions and enables detailed distribution analysis.

Figures 79 through 86 present the top features for all four configurations, using both mean(|SHAP|) bar plots and SHAP summary (distribution) plots.

With-Age – Realistic Mode (Figures 79 & 80)

The top predictors were Age_Group_Young, Age, and a group of football-specific attributes such as Anticipation, Bravery, Determination, Concentration, Injury Proneness, Strength, and Age_Group_Peak.
The SHAP summary plot shows that younger age categories generally had the strongest positive contribution to predicted success, while technical and mental attributes produced more moderate SHAP effects.

- **With-Age – Full Mode (Figures 81 & 82)**

When hidden attributes were introduced, Current Ability (CA) became the dominant predictor, followed by Age_Group_Young, Potential Ability (PA), and Age.
SHAP distributions show extremely large positive SHAP effects for CA and PA, confirming their overwhelming influence in the Full mode. Other attributes such as Determination, Injury Proneness, Composure, and Teamwork (TRO) showed smaller but consistent contributions.

- **Without-Age – Realistic Mode (Figures 83 & 84)**

Removing Age shifted the model's focus toward technical and mental attributes. The most influential features were Anticipation, Strength, Concentration, Bravery, Injury Proneness, Determination, Balance, Aerial Ability, and Consistency.
SHAP summary distributions illustrate a more balanced spread of effects compared to the Full mode, with no single feature dominating the predictions.

- **Without-Age – Full Mode (Figures 85 & 86)**

In the Full mode without age, **CA and PA again dominated**, followed by Growth Ratio, Injury Proneness, First Touch, Marking, Heading, Finishing, Growth Room, and Tackling.
The SHAP summary plot shows that CA and PA produce the largest SHAP values by a substantial margin, while all other attributes contribute comparatively small but directionally consistent effects.

Interpretation of these patterns, including why CA/PA dominate and why age-driven features have strong SHAP values, is deferred to the Discussion section, where these effects will be analyzed in depth.

4.4 **Best Performing Models**

This subsection identifies the strongest models under the two evaluation thresholds used in this study:

1. F1-optimized threshold – highlights minority-class performance.
2. Balanced-Accuracy-optimized threshold – the primary metric of the study.

A model is considered “best performing” when it achieves high Balanced Accuracy, strong recall of successful players, competitive F1, and favorable MCC.

4.4.1 **Best Models Under the F1-Optimized Threshold**

Across all configurations, the top three F1-optimized models are:

1st – XGBoost (Full – With Age)

- F1 Score: 0.4649 *(highest overall)*
- Balanced Accuracy: 0.7542
- Precision: 0.4113
- Recall: 0.5346
- MCC: 0.4484 (highest among F1 configurations)

2nd – Random Forest (Full – With Age)

- F1 Score: 0.45
- Balanced Accuracy: 0.7916
- Precision: 0.352
- Recall: 0.6221
- MCC: 0.4448

3rd – XGBoost (Full – Without Age)

- F1 Score: 0.4186
- Balanced Accuracy: 0.6616
- Precision: 0.5669
- MCC: 0.4195

4.4.2 Best Models Under the Balanced-Accuracy-Optimized Threshold

Ranking the models by Balanced Accuracy yields:

1st – XGBoost (Full – With Age)

- Balanced Accuracy: 0.8999
- F1 Score: 0.2807
- Precision: 0.1642
- MCC: 0.3609

2nd – Random Forest (Full – With Age)

- Balanced Accuracy: 0.8872
- F1 Score: 0.272
- Precision: 0.1589
- MCC: 0.3483

3rd – Logistic Regression (Full – With Age)

- Balanced Accuracy: 0.8810
- F1 Score: 0.2481
- Precision: 0.1425
- MCC: 0.3274

4.4.3 Overall Best-Performing Model

Considering both thresholds, and comparing performance across Balanced Accuracy, F1 Score, and MCC, the XGBoost (Full – With Age) model emerges as the best overall performer. It ranked first simultaneously in F1-optimized and Balanced-Accuracy-optimized evaluations and consistently outperforms all other models in terms of reliability and predictive strength.

Tables present all models evaluated under both thresholds, sorted by their respective primary metric for each threshold are found in Appendix 4.

5. Discussion

5.1 Key Findings

This subsection summarizes the most important insights from the entire study.

5.1.1 Predictive Performance – What did we learn about predicting success?

The results show that predicting long-term success in FM is feasible but requires models capable of capturing complex, non-linear patterns. The XGBoost Full With-Age configuration achieves the highest balanced accuracy (0.8999), demonstrating that combining hidden attributes (CA/PA), age information, and a strong ensemble architecture yields the most reliable predictions. Even in the absence of hidden attributes, the Realistic-Mode Random Forest performs strongly, indicating that meaningful predictive signals exist within the publicly visible attributes available to real scouts.

A consistent pattern across all experiments is the clear separation in performance between ensemble models (XGBoost, Random Forest) and simpler methods (Logistic Regression, Decision Tree). Ensemble methods handle the high-dimensional interaction structure of FM attributes effectively, while linear and single-tree models struggle to capture the game's multi-factor development dynamics.

SVC Failure Under Extreme Class Imbalance.

One notable finding is the collapse of the Support Vector Classifier in the Full Mode With-Age configuration. In this case, the model degenerates into predicting only a single class for all players. This behavior arises from the combination of extreme class imbalance, SMOTE oversampling, and the internal probability formulation of SVC. Because the classifier outputs the same class for every instance, it effectively does not use any features for discrimination; accordingly, permutation importance results were entirely flat, with all feature importances ≈ 0. This confirms that SVC provides no useful predictive value in this setup and cannot form a meaningful decision boundary under these conditions. As this issue is consistent, reproducible, and isolated to SVC, the model is excluded from the final comparative evaluation.

Finally, model performance varies noticeably depending on the decision-threshold strategy. Optimizing for F1 emphasizes detecting more successful players (higher recall), whereas optimizing for balanced accuracy yields more stable overall discrimination between classes. This threshold sensitivity highlights the importance of matching evaluation criteria to the project’s objectives—especially in settings with severe class imbalance such as elite-talent identification.

5.1.2 Which Attributes Were Most Important?

Feature-importance analyses across all models reveal a consistent hierarchical structure of attributes that drive long-term success in FM. The following synthesis integrates findings from LR coefficients, RF/DT impurity-based importance, SVC permutation importance, and XGBoost SHAP values (Figures 69–86).

Hierarchy of Predictive Attributes:

1. Hidden Attributes (Full Mode Only)

In every Full-Mode model, Current Ability (CA) and Potential Ability (PA) were the dominant predictors, producing the largest SHAP values and feature importances by substantial margins. This confirms that FM's internal progression logic is heavily governed by these two ratings, which represent current skill level and maximum achievable potential.

2. Age and Developmental Timing (Realistic Mode)

When CA and PA are unavailable, age-related variables become highly influential:

- Age (Discrete)
- Age_Group_Young (≤20 years)
- Age_Group_Peak (21–23 years)

These variables consistently appear in the top features for all Realistic-Mode models, indicating that FM strongly favors players who are younger and still within peak development windows.

3. Core Football Intelligence Attributes (Mental)

The following mental attributes emerged as top predictors across all Realistic-Mode models, reflecting decision-making, reading of the game, resilience, and tactical awareness:

- Anticipation: Ability to predict ball and player movement
- Decisions: Quality of tactical choices under pressure
- Determination: Resolve to succeed and mental resilience
- Composure: Calmness in high-pressure situations
- Concentration: Focus and attention over 90 minutes
- Bravery: Willingness to take risks in dangerous situations
- Teamwork: Following tactical instructions and supporting teammates

These attributes consistently ranked in the top 10 across LR, RF, DT, and XGBoost

Realistic Mode configurations, aligning with real-world scouting frameworks that prioritize tactical intelligence and mental strength.

4. Physical Foundation Attributes

Physical qualities play a meaningful secondary role, particularly for durability

and athleticism over a 10-year career:

- Strength: Power in physical duels
- Balance: Stability when moving or being challenged
- Pace: Top running speed
- Acceleration: Speed in reaching maximum pace
- Stamina: Ability to maintain exertion throughout matches
- Natural Fitness: Physical condition maintenance and recovery

These attributes appeared consistently in RF, XGBoost, and SVC models, reflecting FM's emphasis on physical sustainability.

5. Technical Skill Indicators

Key technical attributes that reflect execution quality:

- First Touch: Quality of ball control upon receiving
- Technique: Overall technical skill and clean execution
- Passing: Accuracy of distribution

- Finishing: Conversion of goal-scoring chances (for attackers)

6. Composite Measures

Total Score—the sum of all technical, mental, and physical attributes—emerged as a strong holistic indicator, especially in ensemble models. Cluster analysis (Figures 97–98) showed sharp transitions in market value and success rates at the P90 threshold, confirming that aggregated ability captures player quality more reliably than isolated attributes.

Summary

Across all models and feature-importance methods, the strongest early-career predictors form a consistent profile: mental intelligence (Anticipation, Decisions, Determination, Composure), physical foundation (Strength, Balance, Pace, Stamina), technical skill (First Touch, Technique), developmental timing (Age), and injury resistance. This profile aligns with professional scouting principles and fulfills the study's fourth objective of identifying key attributes that shape long-term success.

6.1.1 Does Age Provide Genuine Predictive Value?

The results show that Age provides genuine predictive value, supported by both statistical testing and model behavior. The statistical analyses in Section 4.1.1 correlation, Mann–Whitney U, chi-square, and subgroup success rates consistently indicate that younger players, especially those in the 21–23 “Peak” group, succeed at much higher rates. This pattern is mirrored in the model results. In every Realistic-Mode model, adding Age improves all major evaluation metrics, including balanced accuracy, recall, F1 score, MCC, and G-Mean, typically yielding a 4–7 percentage point increase in balanced accuracy and similar gains across the remaining metrics. Age and Age_Group_Young also appear in the top 10 most influential features for LR, RF, DT, and XGBoost, confirming that the models actively rely on developmental age information. Even in Full Mode (where CA and PA dominate) age-related features still retain measurable importance, suggesting that Age captures aspects of player growth and timing that are not fully explained by ability ratings alone. These findings align with FM’s internal logic: younger players benefit from more development time, and the simulation engine favors those with higher PA-driven growth trajectories.

6.1.2 Exploratory Cluster Analysis (Supplementary Insight)

To investigate whether players form natural groups based on early-career attributes and whether these groups relate to long-term success several exploratory K-Means analyses were conducted. Although clustering does not contribute directly to prediction, it provides descriptive insight into the internal structure of FM’s attribute space and helps contextualize the supervised results

All clustering figures, including scatterplots, PCA projections, and t-SNE visualizations, are provided in Appendix 3.

1. **Clustering Visible Attributes**

   K-Means applied across all visible technical, mental, and physical attributes produced two broad clusters (≈4,600 vs. 39,000 players). Descriptive statistics showed only mild differences across attribute means, and success rates were similarly low (4.73% vs. 3.12%). The PCA and t-SNE projections (Figures 87 and 88) illustrate clearly the two clusters overlap substantially

These results validate that visible attributes alone do not form strong structural divisions within the player population. The lack of separability in both PCA and t-SNE indicates that early-career visible attributes do not naturally group players into meaningful categories related to long-term success. This reinforces the finding that predicting success requires supervised models capable of capturing non-linear interactions, rather than relying on unsupervised clusters or linear separations in the raw attribute space.

The success-colored PCA and t-SNE visualization (Figure 89 and 90) shows successful and unsuccessful players scattered throughout both clusters.

The PCA success-colored shows that successful players are spread throughout the visible-attribute space rather than concentrated in any region. This indicates that visible attributes do not form a linear structure that separates successful from unsuccessful players.

The t-SNE success-colored plot confirms the same pattern: even under nonlinear embedding, successful players remain intermixed with unsuccessful ones. No natural success clusters appear, reinforcing that visible attributes alone do not separate success outcomes.

2. **Clustering on Hidden Attributes (CA–PA Space)**

To explore whether FM’s hidden ability attributes exhibit meaningful structural patterns, K-Means clustering was applied to the CA–PA space (43,903 players × 2 features). Silhouette analysis indicated k = 2 as the optimal solution, producing two nearly balanced groups: Cluster 0 (22,365 players) and Cluster 1 (21,538 players). These clusters differ substantially in their ability profiles: Cluster 0 shows lower CA (≈63) and PA (≈89), while Cluster 1 displays significantly higher CA (≈92) and PA (≈120). This structural difference directly translates into outcomes—only 0.21% of Cluster 0 players achieved long-term success, compared with 6.48% in Cluster 1, representing an approximate 30× advantage.

The separation is clearly visible in the PCA and t-SNE projections (Figures 91 and 92), where the clusters form two distinct, non-overlapping regions.

When the same projections are recolored by success_label instead of cluster assignment (Figures 93 and 94), nearly all successful players appear inside Cluster 1, with very few exceptions. This alignment confirms that FM’s longterm development engine is strongly governed by the hidden CA–PA attributes the clustering structure formed purely from CA and PA almost perfectly explains the distribution of successful players.

3. CA–PA Space Colored by Success Label

K-Means clustering applied to the CA–PA space (43,903 players) reveals a clear structural division in FM’s hidden ability system. As shown in Figure 95, the model separates players into two distinct groups: Cluster 0, which occupies the lower-left region of the CA–PA space and consists of players with both low Current Ability (CA) and low Potential Ability (PA), and Cluster 1, which dominates the upper-right region and contains players with substantially higher CA and PA values. This diagonal split reflects FM’s inherent talent hierarchy: higher ability levels naturally form a dense, coherent high-talent cluster.

When success outcomes are overlaid in Figure 96, the relationship becomes Certain nearly all successful players (success_label = 1) fall within Cluster 1, while Cluster 0 contains almost no successful cases. The combined plot shows that high-CA/high-PA players not only cluster together but also overwhelmingly achieve long-term success, whereas low-ability players rarely do. This confirms that CA and PA alone encode nearly the entire structure of FM’s long-term player development, and that success is fundamentally tied to a player's initial CA–PA profile.

4. Analysis of Ability Aggregation and Market Valuation Patterns

The results from Figures 97 and 98 clearly demonstrate that Total Score is one of the strongest predictors of both market value and long-term success in FM. Because Total Score aggregates all physical, mental, and technical attributes, it behaves as a global measure of overall player quality. In the gradient plot Figure 97, Transfer Value increases smoothly and consistently as total Score rises, with a sharp transition occurring around the P90 threshold, beyond which nearly all high-value and successful players appear. Players below the mean or median thresholds almost never accumulate meaningful market value, indicating that low aggregated ability cannot compensate for long-term development.

The clustering plot Figure 98 reinforces this structure: the K-Means algorithm naturally separates players into ability-valuation segments. Cluster 0 captures low-score, low-value players; Cluster 1 and 2 occupy the mid-to-high value range; and Cluster 3 isolates elite outliers with exceptional ability and valuation. Crucially, successful players (success_label = 1) are overwhelmingly concentrated in the high-score clusters, especially those exceeding the P90 cutoff. This alignment shows that total Score not only reflects a player's current profile but also encapsulates the developmental ceiling that governs FM’s long term simulation outcomes.

In summary, the results confirm that total Score is a powerful and reliable summary metric for differentiating player ability, predicting transfer value, and identifying long-term success trajectories.

6.2 **Practical Implications**

The results of both the predictive modeling and the cluster analysis yield several practical implications for football scouting, data-driven talent identification, and the use of FM as an analytical environment.

1. Early-career attributes provide meaningful—but incomplete—predictive signal

The Realistic Mode models demonstrated that visible technical, mental, and physical attributes carry measurable information about long-term success. This confirms that real-world scouts, who rely on similar observable traits, can extract useful signals from players aged 15–23. However, the cluster analysis shows that these visible attributes do not naturally form distinct tiers of talent: K-Means clusters overlapped heavily, and successful players were spread across multiple attribute profiles. Practical implication: scouts cannot rely on simple threshold-based filtering (e.g., "Anticipation ≥ 15")—success arises from complex interactions across attributes.

2. Age is a genuinely informative predictor
Across all models, younger players exhibited systematically higher success rates due to having more developmental runway. The clear improvement in model performance in the With-Age configuration supports longstanding scouting logic: identifying talent early yields higher long-term returns.

3. Hidden attributes (CA/PA) illustrate the limits of real-world prediction

The dramatic improvement in model performance when CA/PA were included highlights how much FM's development engine is driven by unobservable qualities. Cluster analysis reinforces this: in CA–PA space, players form clean, separated clusters where nearly all future elite players fall into the high-CA/high-PA region. Practical implication: a large portion of a player's eventual peak depends on factors that scouts cannot directly observe—potential ceiling, learning capacity, natural growth. This mirrors real-world uncertainty and explains why forecasting football careers is inherently probabilistic.

4. Extreme class imbalance reflects real-world rarity of elite outcomes.
Only ~3.3% of players reach elite status. This severe imbalance affects both model behavior and practical scouting decision-making.Clubs must accept that high failure rates in youth development are not errors—they reflect the true distribution of football outcomes.

5. Ensemble models outperform linear models due to non-linear interaction effects

The strong performance of Random Forest and XGBoost indicates that player success is shaped by non-linear combinations of attributes rather than by single standout traits. This agrees with the cluster finding that visible attributes do not separate cleanly into talent groups—only models capable of capturing interactions can extract meaningful structure.

6. Total Score emerges as a useful composite measure

The Total Score–Market Value patterns (sharp jump at the P90 threshold) show that aggregated measures capture player quality more reliably than individual attributes. This matches real-world scouting, where evaluators often rely on holistic impressions rather than isolated metrics. Practical implication: clubs may benefit from internal composite indices rather than relying solely on raw attribute tables.

7. FM as a research and strategic testing environment

While FM cannot replace real-world scouting, the study shows it can act as a safe, controlled sandbox to:

- test scouting heuristics,
- evaluate the importance of specific attributes,
- examine long-term development pathways, and
- prototype data-driven decision systems.
- It provides structure and repeatability that real-world environments cannot.

Overall, the practical implications confirm that while early-career data can meaningfully guide scouting decisions, success prediction remains inherently uncertain driven by nonlinear interactions, hidden factors, and the extreme rarity of elite outcomes. FM offers a valuable experimental platform to explore these patterns, but it complements rather than replaces real-world scouting expertise.

6.2.1 How This Study Extends and Improves Upon van Wijk (2022)

While Section 2 outlines the methodological differences, their practical impact becomes clear when interpreting the results. Each extension beyond van Wijk (2022) contributed directly to a deeper and more realistic understanding of long-term player success:

1. **Replacing PA prediction with a market-value–based success definition**

Unlike PA—which is static and internally defined by the game—market value incorporates dynamic elements such as injuries, playing time, transfers, development environment, and competition level. As a result, this study evaluates a more realistic and more difficult target, revealing that FM's long-term progression is influenced by many interacting factors, not only PA.

2. **Multi-simulation validation (3 independent runs)**

   Running each career three times exposed the inherent randomness in FM (injuries, loan spells, team context).

This produced a more stable and credible success label and demonstrated that high-PA players do not always succeed—an insight not captured in single-simulation studies.

3. **Realistic vs Full Mode analysis**

Splitting models into Realistic (no CA/PA) and Full (with CA/PA) provided a clear quantification of how much predictive power comes from observable attributes versus hidden internal ratings. This distinction highlights why real-world scouting is inherently uncertain, a nuance not explored in prior work.

4. **With-Age vs No-Age configurations**

Including and excluding age revealed that development time is a statistically meaningful factor affecting success probabilities. Van Wijk concluded that age was unimportant when predicting PA; this study shows that age is important when predicting actual long-term outcomes.

Overall Impact:

These methodological extensions do not merely replicate earlier work but produce a more realistic, more generalizable, and more actionable understanding of long-term success. They demonstrate how FM behaves under more realistic scouting constraints and reveal structural patterns that earlier PA-based studies could not detect.

6.3 **Limitations**

This study has several limitations that should be considered when interpreting the results:

1. FM as a controlled simulation environment

Player outcomes within FM are generated by internal algorithms—particularly Current Ability (CA) and Potential Ability (PA)—which necessarily simplify the multifaceted complexity of real football careers. While the simulation incorporates injuries, transfers, playing time, and competitive context, it cannot fully replicate psychological factors, cultural adaptation, coaching relationships, or systemic inequalities that shape real-world career trajectories. Consequently, direct generalization from FM-based findings to actual football environments must be approached with appropriate caution.

2. Market value is only a proxy for success.

Value captures performance and potential but does not fully reflect tactical fit, injuries, or psychological factors. Some aspects of "success" cannot be measured within this definition.

3. The dataset is highly imbalanced.

Elite outcomes are rare (~3.3%), which constrains model performance even with SMOTE and threshold optimization. Predicting extremely rare events is inherently difficult.

4. Hidden attributes create two different predictive realities.

Full-mode models use CA/PA, which real scouts cannot observe. This makes Full-mode results a theoretical upper bound rather than a practical prediction scenario.

5. Only Year-0 and Year-10 snapshots were extracted.

Intermediate development patterns (e.g., Year-1, Year-5) were not captured, preventing analysis of growth curves or mid-career fluctuations.

6. The study focuses on top leagues and ages 15–23.

Results may not extend to lower divisions, later-blooming players, or different competitive environments.

7. Hyperparameter optimization was limited by computation time.

Although Optuna was used, search depth was bounded by available resources and may not reflect the absolute global optimum.

6.4 **Future Work**

Several directions can extend and strengthen this study:

1. Develop a time-series modeling framework.

Future work can extract multiple intermediate snapshots to build true developmental time series. This would allow the use of temporal models capture how player attributes evolve over time rather than relying solely on initial and final states.

2. Explore sequence-based or temporal deep-learning models.

With time-series data available, neural architecture could model non-linear development paths and detect early signals of rapid or stalled progression.

3. Expand the data set to include lower leagues and broader age ranges.

This would test whether the observed patterns generalize beyond top-tier environments and early-career players.

4. Investigate alternative definitions of success.

Combining market value with playing time, league level, or international appearances could provide a multi-dimensional measure of career achievement.

5. Integrate contextual or environmental variables.

Factors such as injuries, coaching quality, and loan spells (if accessible) could improve the realism and explanatory power of predictive models.

6. Improve computational depth for model optimization.

Larger Optuna search spaces or Bayesian hyperparameter strategies could yield further performance gains.

Conclusion

This study developed a complete machine-learning framework for predicting long-term football player success using Football Manager 2024 (FM) as a controlled simulation environment. By tracking 43,094 players aged 15–23 across three independent ten-year simulations, the study achieved all four research objectives: validating FM as a structured data source, applying rigorous data-science methodology, extending prior academic work, and identifying early-career attributes that influence long-term success.

The dual-benchmark approach—mapping real-world Transfermarkt valuations to in-game market values—confirmed that FM produces stable and reproducible long-term outcomes despite inherent randomness. The four-way experimental design (Realistic vs Full × With-Age vs No-Age) enabled a systematic investigation of model configurations under varying levels of information visibility. Across all models, age and key mental attributes (Anticipation, Decisions, Determination, Composure) consistently emerged as strong predictors of future success. In Full Mode, the internal hidden attributes CA and PA dominated feature importance, revealing how heavily FM's progression engine depends on unobservable internal ratings.

Model performance further reinforced these insights. The XGBoost Full With-Age model achieved the strongest results (balanced accuracy ≈ 0.90), establishing the upper bound of what is possible when complete game information is available. Ensemble models outperformed linear and single-tree models, demonstrating that FM career trajectories require models capable of capturing complex non-linear interactions. Cluster analysis also showed that visible attributes alone do not form natural tiers of talent, whereas clustering on CA–PA space produced clear separation between successful and unsuccessful players.

A key result is the substantial performance gap between Full Mode and Realistic Mode. This gap highlights a fundamental limitation: FM cannot replace real-world scouting, because its most predictive information comes from hidden values designed by FM's creators. Importantly, the fact that Realistic Mode performs substantially worse is not a weakness of the study; it reflects the true difficulty of predicting elite success when relying only on observable attributes—a difficulty shared by real scouts and real-world analytics systems. When CA and PA are removed, the models operate under the same constraints as human evaluators, and the reduced accuracy becomes an informative and realistic outcome rather than a methodological flaw.

Even so, FM proves extremely valuable as a research tool. It offers a low-cost, high-scale, repeatable environment for studying player development—something that real-world datasets rarely allow. While FM should not be viewed as a substitute for professional scouting, it is a powerful complementary platform for hypothesis testing, feature analysis, and early-stage model prototyping.

Several limitations must be acknowledged. FM simplifies psychological, environmental, and contextual factors that shape real careers. Market value, although broad, is only a proxy for success. The dataset excludes lower leagues and late-blooming players, and the study uses only Year-0 and Year-10 snapshots, preventing analysis of mid-career dynamics. Future work should incorporate time-series modeling, expand player coverage, include contextual variables such as injuries and coaching quality, and explore cross-version comparisons to strengthen external validity.

In conclusion, this study demonstrates both the promise and the boundaries of predicting long-term football success using FM simulations. Machine-learning models can extract meaningful signals from early-career data, but long-term success remains inherently probabilistic and shaped by complexity, randomness, and hidden variables. FM stands not as a replacement for human scouting, but as a robust experimental ground that offers unique opportunities for analytics research, talent-development studies, and the advancement of data-driven methodologies in football.

References

1. Sports Interactive, "FM24," SEGA Europe Ltd., London, UK, 2023.
2. E. M. R. Lima, I. W. Tertuliano, A. L. Aroni, A. A. Machado, and C. N. Fischer, "Saga FM na gestão esportiva: uma ferramenta tecnológica para monitorar jogadores promissores," Lecturas: Educación Física y Deportes, vol. 23, no. 239, pp. 1-15, Apr. 2018.
3. W. van Wijk, "Predicting the potential ability of football players in the FM game," M.S. thesis, School of Humanities and Digital Sciences, Tilburg University, Tilburg, Netherlands, 2022.
4. “Football League Rankings | Analysis of Top Soccer Leagues,” *Globalfootballrankings.com*, 2024. https://globalfootballrankings.com/
5. A. Sweigart, “Welcome to PyAutoGUI’s documentation! — PyAutoGUI 1.0.0 documentation,” *Readthedocs.io*, 2014. https://pyautogui.readthedocs.io/en/latest/
6. “Most valuable players,” _. https://www.transfermarkt.com/spieler-statistik/wertvollstespieler/marktwertetop
7. K. Smith, “Meet Will Still: From FM Addict to Managerial Powerhouse,” *GiveMeSport*, Mar. 11, 2023. https://www.givemesport.com/who-is-will-still-football-manager/
8. Readthedocs.io. (2018). *optuna.samplers — Optuna 4.3.0 documentation*. [online] Available at: https://optuna.readthedocs.io/en/stable/reference/samplers/index.html.
9. H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1284, 2009. Available: https://cyberenlightener.com/wp-content/uploads/2025/03/Imbalanced-Learning_-Foundations-Algorithms-and-Applications-2013-Wiley-IEEE-Press-10.1002_9781118646106-libgen.li_.pdf#page=193
10. imbalanced-learn.org. (n.d.). *SMOTE — Version 0.9.0*. [online] Available at: https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html
11. scikit-learn (2014). *LogisticRegression*. [online] Scikit-learn.org. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

12. scikit-learn (2025). *1.10. Decision Trees — scikit-learn 0.22 documentation*. [online] Scikit-learn.org. Available at:https://scikit-learn.org/stable/modules/tree.html
13. Scikit-Learn (2025). *sklearn.ensemble.RandomForestClassifier — scikit-learn 0.20.3 Documentation*. [online] Scikit-learn.org. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
14. xgboost developers (2022). *XGBoost Documentation — xgboost 1.5.1 documentation*. [online] xgboost.readthedocs.io. Available at: https://xgboost.readthedocs.io/en/stable/
15. scikit-learn (2018). *1.4. Support Vector Machines*. [online] Scikit-learn.org. Available at: https://scikit-learn.org/stable/modules/svm.html
16. scikit-learn (2019). *sklearn.cluster.KMeans — scikit-learn 0.21.3 documentation*. [online] Scikit-learn.org. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html.
17. scikit-learn (2009). *sklearn.decomposition.PCA — scikit-learn 0.20.3 documentation*. [online] Scikit-learn.org. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html.
18. scikit-learn developers (2014). *sklearn.manifold.TSNE — scikit-learn 0.21.3 documentation*. [online] Scikit-learn.org. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html.
19. Scikit-learn.org. (2010). *sklearn.metrics.balanced_accuracy_score — scikit-learn 0.21.3 documentation*. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.balanced_accuracy_score.html
20. Scikit-Learn (2019). *Precision-Recall — scikit-learn 0.21.3 documentation*. [online] Scikit-learn.org. Available at: https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html
21. scikit-learn (2019). *sklearn.metrics.f1_score — scikit-learn 0.21.2 documentation*. [online] Scikit-learn.org. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html
22. scikit-learn. (n.d.). *sklearn.metrics.precision_recall_curve*. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_curve.html
23. Readthedocs.io. (2021). *G-Mean Score (GMS) — Permetrics 2.0.0 documentation*. [online] Available at: https://permetrics.readthedocs.io/en/latest/pages/classification/GMS.html

24. Readthedocs.io. (2021). *Matthews Correlation Coefficient (MCC) — Permetrics 2.0.0 documentation*. [online] Available at: https://permetrics.readthedocs.io/en/latest/pages/classification/MCC.html.

25. Readthedocs.io. (2021). *Cohen Kappa Score (CKS) — Permetrics 2.0.0 documentation*. [online] Available at: https://permetrics.readthedocs.io/en/latest/pages/classification/CKS.html.

26. “NCSS Statistical Software Point-Biserial and Biserial Correlations.” Available: https://www.ncss.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Point-Biserial_and_Biserial_Correlations.pdf

27. R. Shier, “Statistics: 2.3 The Mann-Whitney U Test,” 2004. Available: https://www.statstutor.ac.uk/resources/uploaded/mannwhitney.pdf

28. M. McHugh, “(PDF) The Chi-square test of independence,” *ResearchGate*, 2013. https://www.researchgate.net/publication/253336860_The_Chi-square_test_of_independence

Data and Code Availability Statement

All code used for data extraction, preprocessing, model training, evaluation, and visualization in this study is publicly available on GitHub at https://github.com/E2J1/STAT-499----Senior-project

The repository includes:

- PyAutoGUI automation scripts for FM data extraction
- Complete data preprocessing pipeline
- Optuna hyperparameter optimization implementations for all models
- SMOTE configuration and class imbalance handling procedures
- Evaluation metric calculations and visualization code
- SHAP analysis and feature importance extraction scripts
- Cluster analysis and dimensionality reduction implementations

The final processed dataset (Year-0 and Year-10 attributes with success labels) is also provided to enable full reproducibility of results. Note that raw FM database files cannot be shared due to licensing restrictions, but the extraction pipeline can be replicated by any user with a legitimate copy of Football Manager 2024.

Appendix 1:

Original Time-Point Plan (Year-0, Year 1, Year 5, Year-10)

The initial design of the study included extracting player data at Year-0, Year 1, Year 5, and Year-10. The purpose of including Year 1 and Year 5 was to:

- Evaluate early and mid-career development patterns within the game.
- Demonstrate that FM produces realistic, gradual progression rather than sudden jumps.
- Compare development curves between successful and unsuccessful players.
- Strengthen the research analysis with longitudinal evidence beyond the model inputs.

These time-points were intended strictly for analytical validation, not for training the machine learning models (which only use Year-0 as input and Year-10 as the target).

Reason for Excluding Year 1 and Year 5

Due to the large player pool and the manual extraction method using PyAutoGUI, each extraction point requires a long time. Including Year 1 and Year 5, three simulations would require significantly extending the total project duration.

Since Year 1 and Year 5 are not essential for predictive modelling, and their contribution is limited to supporting analysis-not model training-they were excluded to meet project time constraints. The final study therefore focuses exclusively on Year-0 and Year-10

Data Content

The dataset underwent several preprocessing steps to ensure suitability for both modeling modes used in this study: the Realistic Mode and the Full Mode. In the Realistic Mode, attributes that are not visible to real-world analysts or scouts-specifically the hidden attributes Current Ability (CA) and Potential Ability (PA)-were excluded to simulate realistic scouting conditions. In contrast, Full Mode retained these hidden attributes to evaluate the maximum theoretical predictive performance when all in-game information is available.

Several categories of attributes were removed to maintain consistency and prevent data leakage. Seasonal performance statistics such as goals, assists, expected goals (xG), and clean sheets were excluded because they reset every season and therefore do not represent stable long-term indicators. Club-dependent attributes-including wages, contract details, reputation values, and loan information-were removed due to their variability across clubs and their limited relevance to intrinsic player ability.

After applying the league and age filters selecting only players from the top 18 leagues and restricting ages to 15-23 at the start of the simulation-the targeted dataset consisted of 43,094 players. Players in this age range were chosen because they represent the developmental phase of a football career, where growth, progression, and long-term potential can be meaningfully observed. Older players (above 23) generally exhibit limited room for further development, while younger players often lack stable data or are not yet registered in professional leagues. Importantly, this filtering does not make the simulation run faster; FM still processes the full database of approximately 88,000 players internally. The filtering only affects which players appear in the exported dataset, not the number of players the game engine simulates in the background.

To reduce randomness inherent in FM's simulation system, three independent 10-year simulations were performed. Initially, the extracted data from each simulation was stored separately, creating parallel fields such as UID_k1, UID_k2, and UID_k3 (player identifiers across simulations), TransferValue_k1, Value_k2, Value_k3, and success_k1, success_k2, success_k3 where k is the number of the simulation. These were later merged into a single integrated dataset using the unique player UID as a foreign key, ensuring that each row represented the same player across all three simulations. The final success_label was derived by aggregating outcomes from all simulation runs.

During the Year-10 extraction, a subset of players appeared with a transfer value marked as "Not for Sale." This occurs in FM when a club refuses to assign market value to a player due to strategic reasons-such as the player being a core part of the team, being considered indispensable, having a long-term contract, or the club simply choosing not to engage in transfer negotiations. In some cases, "Not for Sale" also appears because of internal game logic unrelated to a player's actual CA or PA.

Since "Not for Sale" is non-numeric, the hidden attribute AP (Asking Price) was used solely to assign a numerical value to these cases. FM typically assigns a default AP of 350 million, representing the upper valuation ceiling within the game. Crucially, AP was never used as an input feature in any machine learning model, neither in the Realistic Mode nor in the Full Mode. Including AP would cause data leakage, as it directly reflects a club's refusal to sell and therefore conveys unavailable information through normal scouting channels. AP was used strictly as a technical solution to convert non-numeric values into consistent numeric entries, ensuring dataset completeness without compromising model integrity.

Table 2 Below shows the final contents of the dataset; it includes only stable and intrinsic player attributes suitable for long-term prediction. These consist of all technical, mental, and physical attributes, along with general demographic and positional information such as height, weight, nationality, and playing position. Market values at both the beginning of the simulation (Year-0) and after ten seasons (Year-10) were retained to measure long-term progression. Player IDs and Simulation IDs were preserved to maintain structural integrity and support traceability across the three simulation runs.

Table 2: Contents of the Final Dataset.

Field	Description	Range	Attribute type
UID	Unique Identifier. A distinct number or assigned to a player.	-	General Information
Name	The full name of the player.	-	General Information
Club	The current football club the player is signed to.	-	General Information
Age	The player's current age in years.	-	General Information
Nat	Nationality. The player's country of citizenship.	-	General Information
Position	The player's primary position(s) (e.g., ST, MC, DC).	-	General Information

Transfer Value	The estimated market value or transfer price of the player.	0 – 350,000,000	General Information
CA	**Current Ability**. A numerical rating representing the player's skill level now	1 - 200	Hidden Player Attribute
PA	**Potential Ability**. A numerical rating represents the player's maximum potential skill level.	1 - 200	Hidden Player Attribute
Acc	**Acceleration**. The player's speed in reaching full pace from a standstill.	1 - 20	Physical Attribute
Ada	**Adaptability**. How well a player settles in a new country or environment.	1 - 20	Hidden Personal Attributes
Agg	**Aggression**. A player's competitive and determined nature (not necessarily dirty play).	1 - 20	Mental Attributes
Aer	**Aerial Reach.** How effectively a goalkeeper can reach high balls, especially during crosses, corners, and aerial duels.	1 - 20	Goalkeeping Attributes
Amb	**Ambition**. A player's drive to play at the highest level and for better clubs.	1 - 20	Hidden Personal Attributes
Agi	**Agility**. How quickly a player can change direction and movement.	1 - 20	Physical Attributes

Ant	**Anticipation**. How well a player predicts the movement of the ball or other players.	1 - 20	Mental Attributes
Bal	**Balance**. A player's ability to stay steady when moving or being tackled.	1 - 20	Physical Attributes
Bra	**Bravery**. Willingness to put themselves in dangerous situations (e.g., blocking shots).	1 - 20	Mental Attributes
Cmp	**Composure**. A player's ability to remain calm under pressure, especially when finishing or passing.	1 - 20	Mental Attributes
Com	**Communication.** Measures how effectively the goalkeeper communicates with and organizes the defensive line.	1 - 20	Goalkeeping Attributes
Cnt	**Controversy**. A hidden attribute related to a player's tendency to cause trouble.	1 - 20	Hidden Player Attributes
Cons	**Consistency.** Indicates how reliably a player performs at a stable level from match to match.	1 - 20	Hidden Player Attributes
Cont	**Controversy.** Indicates a player’s tendency to create off-field issues or disciplinary problems.	1 - 20	Hidden Personal Attributes

Cro	**Crossing.** Measures how accurately a player delivers crosses into dangerous attacking areas.	1 - 20	Technical Attributes
Cmd	**Command of Area.** Shows how confidently a goalkeeper controls the penalty area, especially when dealing with crosses and aerial balls	1 - 20	Goalkeeping Attributes
Cor	**Corners.** Measures how accurately a player delivers corner kicks.	1 - 20	Technical Attributes
Dec	**Decisions**. A player's ability to make the correct choice of action quickly.	1 - 20	Mental Attributes
Det	**Determination**. The resolve to succeed and keep fighting, even when things are going poorly.	1 - 20	Mental Attributes
Dirt	**Dirtiness.** Reflects a player’s tendency to commit fouls or engage in unsporting behavior.	1 - 20	Hidden Player Attributes
Dri	**Dribbling.** Measures how well a player can run with the ball and maintain close control.	1 - 20	Technical Attributes
Ecc	**Eccentricity.** It indicates how likely a goalkeeper is to behave unpredictably or take risky,	1 - 20	Goalkeeping Attributes

	unconventional actions.		
Fin	**Finishing.** Measures how effectively a player converts goal-scoring chances.	1 - 20	Technical Attributes
Fla	**Flair**. A player's natural talent for the unexpected and creative.	1 - 20	Mental Attributes
Fir	**First Touch.** Shows how well a player controls the ball upon receiving it.	1 - 20	Technical Attributes
Fre	**Free Kick Taking.** Measures a player's accuracy and effectiveness when taking free kicks.	1 - 20	Technical Attributes
Inj Pr	**Injury Proneness**. How likely a player is to pick up injuries.	1 - 20	Hidden Player Attributes
Imp M	**Important Matches**. A player's ability to perform well in high-stakes, important games.	1 - 20	Hidden Player Attributes
Hea	**Heading.** Indicates how well a player wins and directs the ball in aerial duels.	1 - 20	Technical Attributes
Han	**Handling.** Shows how securely a goalkeeper catches or holds the ball when making saves.	1 - 20	Goalkeeping Attributes
Kic	**Kicking.** Measures a goalkeeper’s ability to kick the ball accurately and over distance.	1 - 20	Goalkeeping Attributes

Ldr	**Leadership**. The ability to influence and inspire teammates (often for a captain).	1 - 20	Mental Attributes
Lon	**Long Shots.** Indicates how accurately a player shoots from outside the penalty area.	1 - 20	Technical Attributes
L Th	**Long Throws.** Measures how far and accurately a player can perform long throw-ins.	1 - 20	Technical Attributes
Loy	**Loyalty**. How likely a player is to stay at the club.	1 - 20	Hidden Personal Attributes
Mar	**Marking.** Indicates how well a player tracks, follows, and defends against an assigned opponent.	1 - 20	Technical Attributes
OtB	**Off the Ball**. A player's movement when not in possession of the ball (finding space).	1 - 20	Mental Attributes
1v1	**One-on-Ones.** Shows how well a goalkeeper performs in direct one-on-one situations against attackers.	1 - 20	Goalkeeping Attributes
Pas	**Passing.** Measures how accurately a player can pass the ball to teammates.	1 - 20	Technical Attributes
Pac	**Pace**. The player's **top speed** while running.	1 - 20	Physical Attributes
Pen	**Penalty Taking.** Measures how	1 - 20	Technical Attributes

	effectively a player converts penalty kicks		
Pos	**Positioning**. A player's ability to take up the best defensive position.	1 - 20	Mental Attributes
Pres	**Pressure.** Shows how well a player handles high-pressure or high-stakes situations.	1 - 20	Hidden Personal Attributes
Prof	**Professionalism.** Indicates a player’s work ethic, discipline, and commitment to training and long-term development.	1 - 20	Hidden Personal Attributes
Pun	**Punching (Tendency).** Shows how likely a goalkeeper is to punch the ball clear instead of catching it.	1 - 20	Goalkeeping Attributes
Ref	**Reflexes.** Measures how quickly a goalkeeper reacts to unexpected shots or events.	1 - 20	Goalkeeping Attributes
TRO	**Throwing.** Indicates how accurately and effectively a goalkeeper distributes the ball using throws, often to start counterattacks.	1 - 20	Goalkeeping Attributes
Spor	**Sportsmanship.** Reflects how fairly and respectfully a player behaves toward opponents,	1 - 20	Hidden Personal Attributes

	teammates, and officials.		
Str	**Strength**. The player's power in physical challenges and holding off opponents.	1 - 20	Physical Attributes
Sta	**Stamina**. The player's ability to maintain physical exertion throughout a match.	1 - 20	Physical Attributes
Tck	**Tackling.** Measures how well a player wins the ball cleanly without committing fouls.	1 - 20	Technical Attributes
Tea	**Teamwork.** Indicates how well a player follows tactical instructions and supports teammates.	1 - 20	Mental Attributes
Tec	**Technique.** Shows the player’s overall technical skill and ability to execute difficult actions cleanly.	1 - 20	Technical Attributes
Thr	**Throwing (Goalkeepers).** Measures how accurately and effectively a goalkeeper distributes the ball with throws.	1 - 20	Goalkeeping Attributes
Temp	**Temperament.** Reflects how well a player maintains self-control and avoids	1 - 20	Hidden Personal Attributes

	emotional overreaction.		
Vis	**Vision.** Indicates how well a player sees potential passing or attacking opportunities.	1 - 20	Mental Attributes
Vers	**Versatility.** Shows how well a player performs outside their natural position and adapts to new roles.	1 - 20	Hidden Personal Attributes
Nat.1	**Natural Fitness.** Indicates how well a player maintains physical condition, recovers from fatigue, and sustains athletic ability over time.	1 - 20	Physical Attributes
Jum	**Jumping Reach**. How high a player can jump to head the ball.	1 - 20	Physical Attributes
Wor	**Work Rate.** It shows how hard a player is willing to work and how consistently they put effort into off-ball movement and defensive duties.	1 - 20	
Weight	The player's weight.	-	General Information
Height	The player's Height.	-	General Information
Source_File	Indicates the original exported HTML file from which the player’s data was extracted.	-	Dataset Metadata

UID_k1	Unique player identifier for Simulation 1.	-	Simulation Metadata
Transfer Value_k1	Player’s market value at Year-10 in Simulation 1.	0 – 350,000,000	Simulation Output
UID_k2	Unique player identifier for Simulation 2.	-	Simulation Metadata
Transfer Value_k2	Player’s market value at Year-10 in Simulation 2.	0 – 350,000,000	Simulation Output
UID_k3	Unique player identifier for Simulation 3.	-	Simulation Metadata
Transfer Value_k3	Player’s market value at Year-10 in Simulation 3.	0 – 350,000,000	Simulation Output
success_k1	Binary success outcome for Simulation 1 (1 = successful, 0 = unsuccessful).	0 - 1	Simulation Output
success_k2	Binary success outcome for Simulation 2.	0 - 1	Simulation Output
success_k3	Binary success outcome for Simulation 3.	0 - 1	Simulation Output
success_label	Final success label based on majority rule across the three simulations.	0 - 1	Final Derived Label

Appendix 2

To ensure reproducibility and consistent comparison across configurations, all models in this study were optimized using Optuna with 100 trials and the TPE sampler. The same SMOTE oversampling levels were used across all models:

- small = 0.2, medium = 0.5, high = 1.0

The following is the exact search spaces used for each classifier, as implemented in the model training code.

1. Logistic Regression (LR)

Hyperparameters were tuned with conditional logic depending on the selected solver.

Search Space:

- solver: {lbfgs, saga}
- penalty:
  - If lbfgs → "l2"
  - If saga → {l1, l2, elasticnet}
- l1_ratio: float (0, 1), only when penalty = elasticnet
- C: log-uniform $(10^{-3}, 10^{3})$
- max_iter: integer (500, 2000)
- SMOTE level: {0.2, 0.5, 1.0}

2. Decision Tree (DT)

Search Space:

- max_depth: integer (3, 15)
- min_samples_split: integer (5, 50)
- min_samples_leaf: integer (2, 20)
- criterion: {gini, entropy, log_loss}
- max_features: {sqrt, log2, None}

- SMOTE level: {0.2, 0.5, 1.0}

3. Random Forest (RF)

Search Space:

- n_estimators: integer (100, 1000), step = 50
- max_depth: integer (3, 30)
- min_samples_split: integer (2, 10)
- min_samples_leaf: integer (1, 10)
- max_features: {sqrt, log2, None}
- bootstrap: {True, False}
- class_weight: {None, balanced} selected through:
- use_class_weight {True, False}
- SMOTE level: {0.2, 0.5, 1.0}

4. Support Vector Machine (SVC)

Search Space:

- C: log-uniform ($10^{-3}, 10^{3}$)
- kernel: {linear, rbf}
- gamma: {scale, auto}
- max_iter: integer (1000, 5000)
- SMOTE level: {0.2, 0.5, 1.0}

5. XGBoost (XGBClassifier)

Search Space:

- n_estimators: integer (100, 500)
- max_depth: integer (3, 10)
- learning_rate: log-uniform (0.01, 0.3)

- subsample: float (0.6, 1.0)
- colsample_bytree: float (0.6, 1.0)
- gamma: float (0, 5)
- reg_lambda: log-uniform (0.1, 10)
- reg_alpha: float (0, 5)
- tree_method: "gpu_hist"
- eval_metric: "logloss"
- SMOTE level: {0.2, 0.5, 1.0}

Rationale for a Shared Search Space

Using a uniform search space across both configurations ensures:

- Fair comparability between With-Age vs. No-Age and Realistic vs. Full modes.
- Controlled experimentation, eliminating variation caused by differing search ranges.
- Transparent reproducibility, since all settings are explicitly documented and tied to the executable code.

Appendix 3: Plots

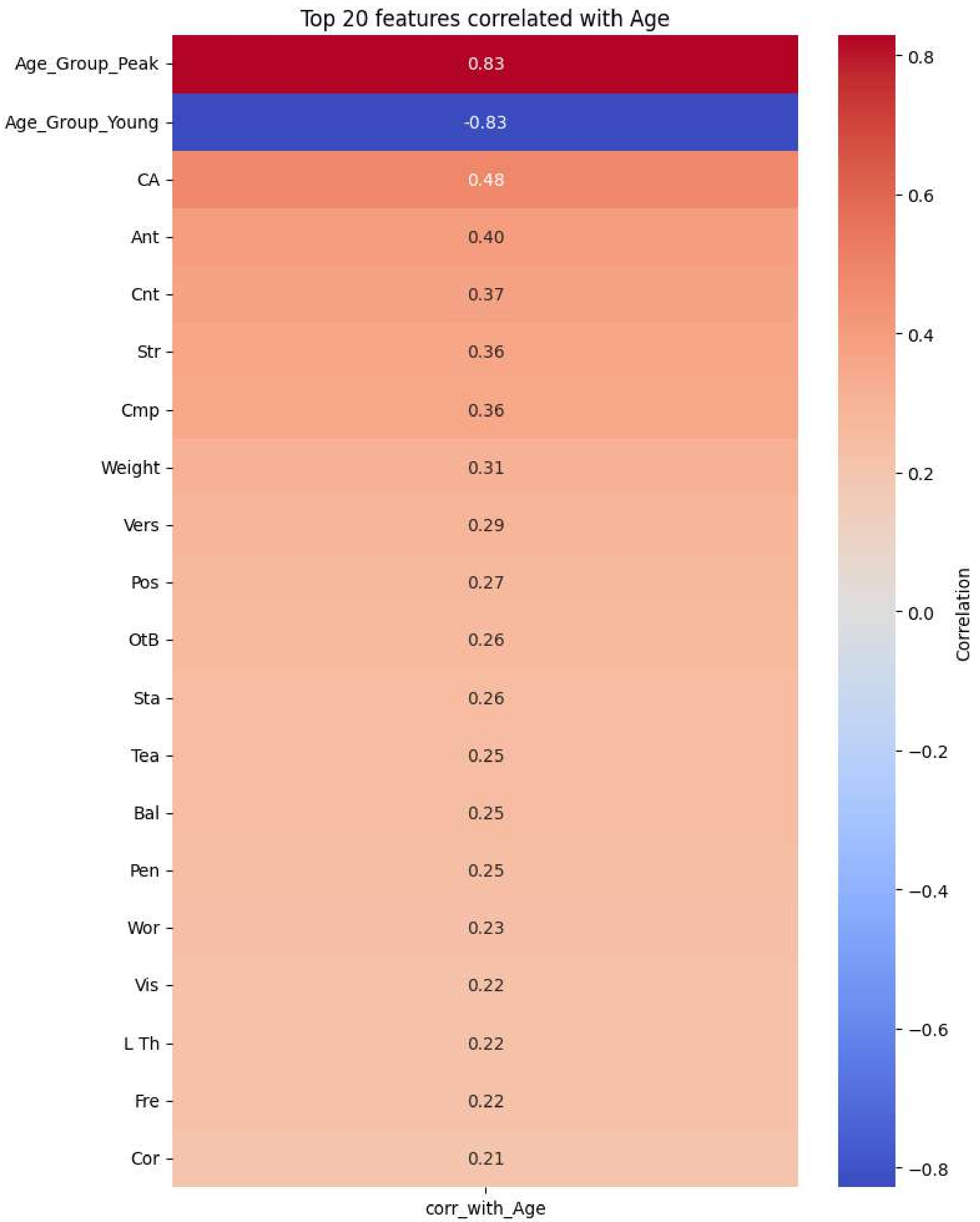


Figure 10: Heatmap of Top 20 Features Correlated with Age

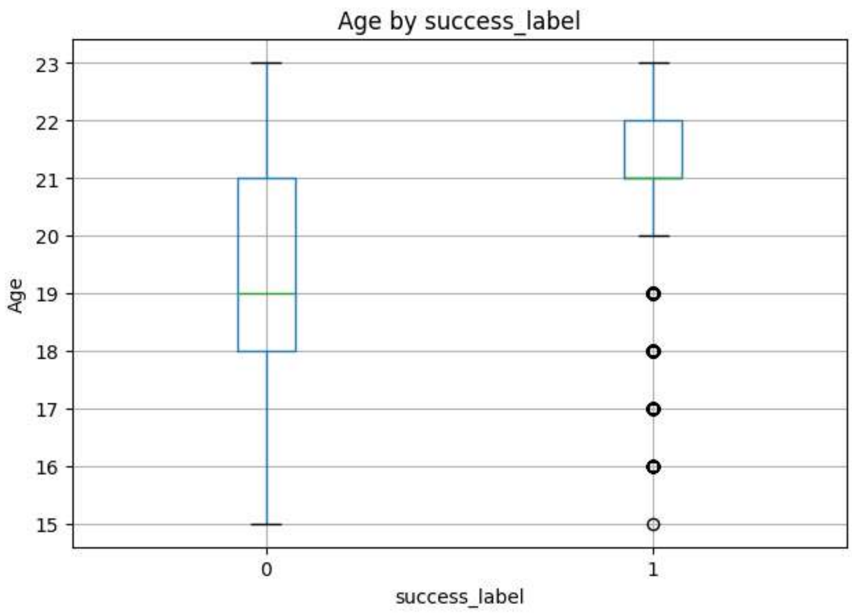


Figure 12: Boxplot of Age by Success Label

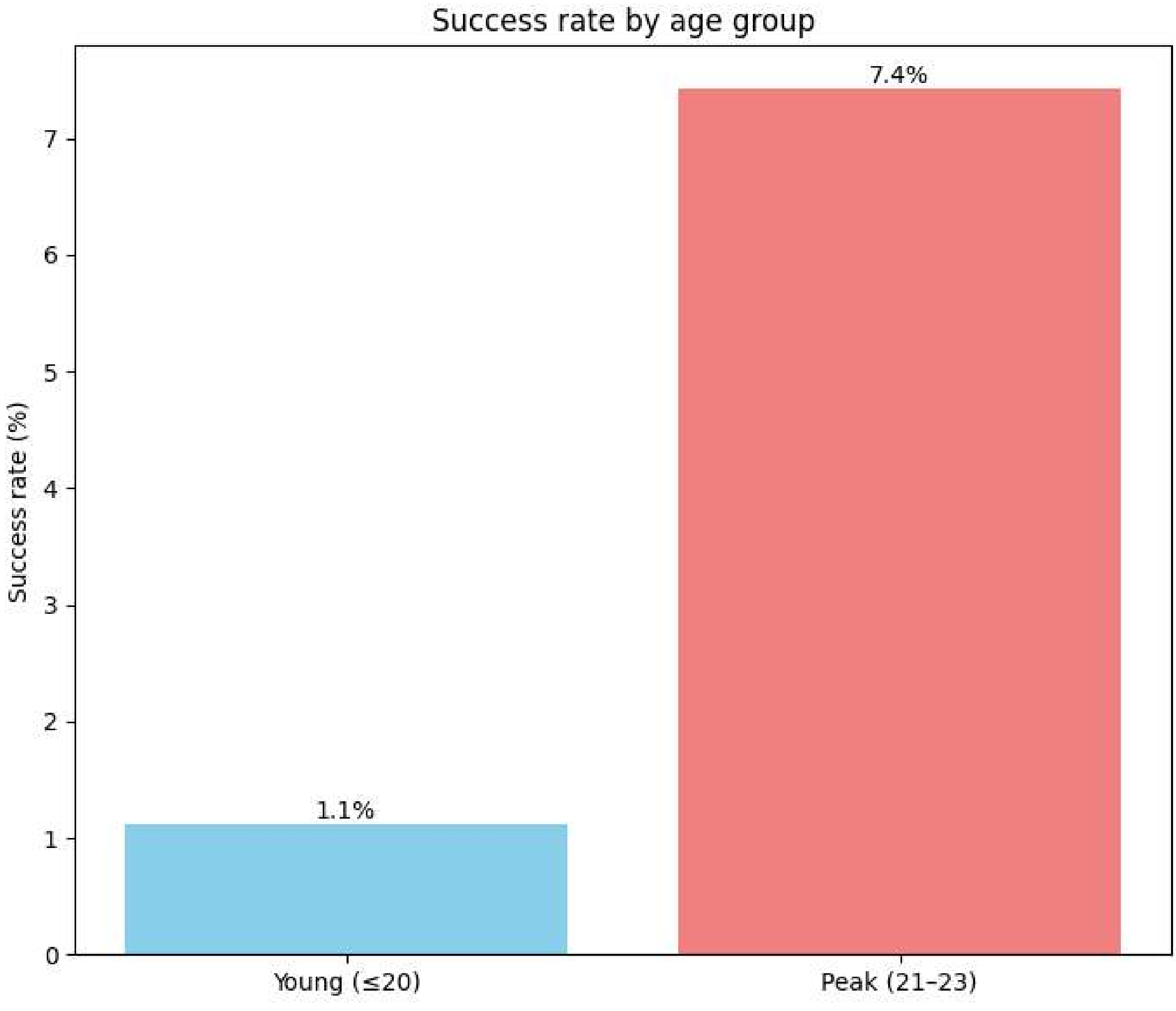


Figure 11: Bar chart of success rates by age group (Young ≤20 vs Peak 21–23)

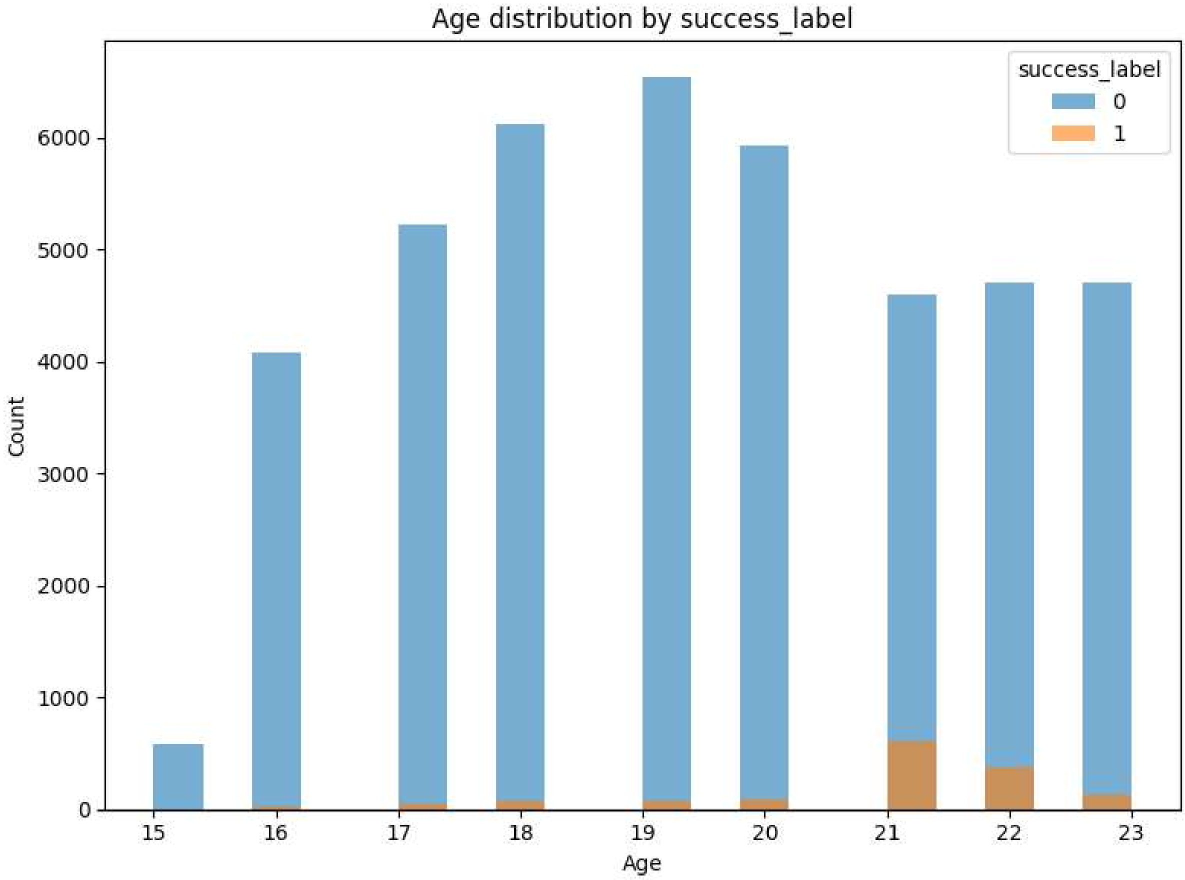


Figure 13: Histogram of Age distribution by success_label

Logistic Regression

Confusion Matrices.

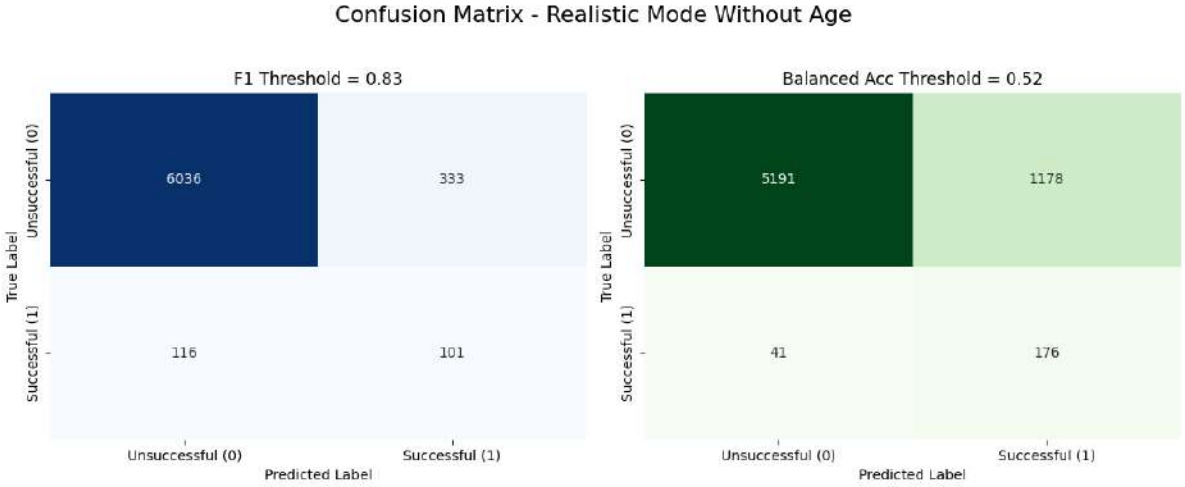


Figure 11: LR Without Age– Realistic Mode

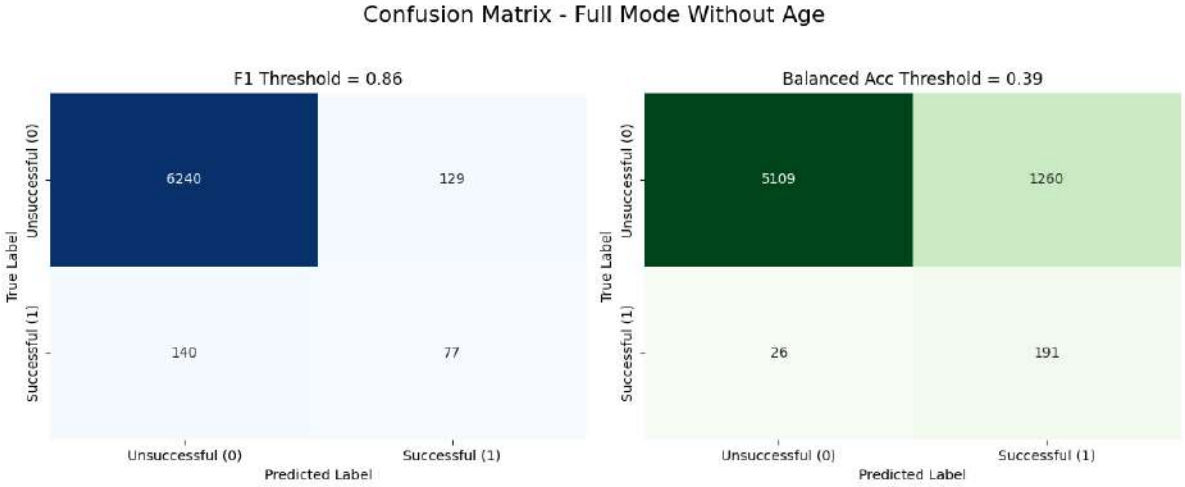


Figure 12: LR Without Age – Full Mode

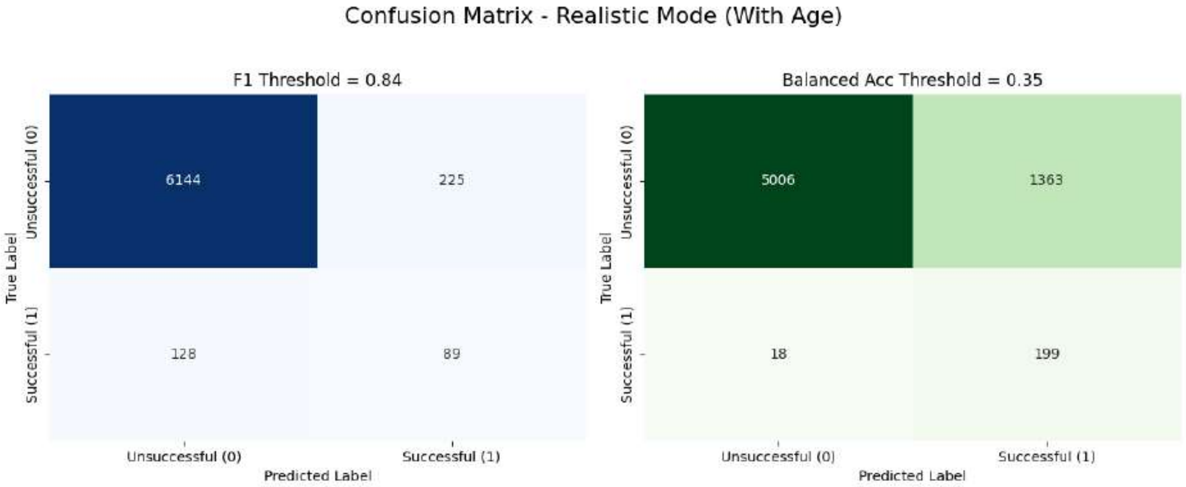


Figure 13: LR With Age – Realistic Mode

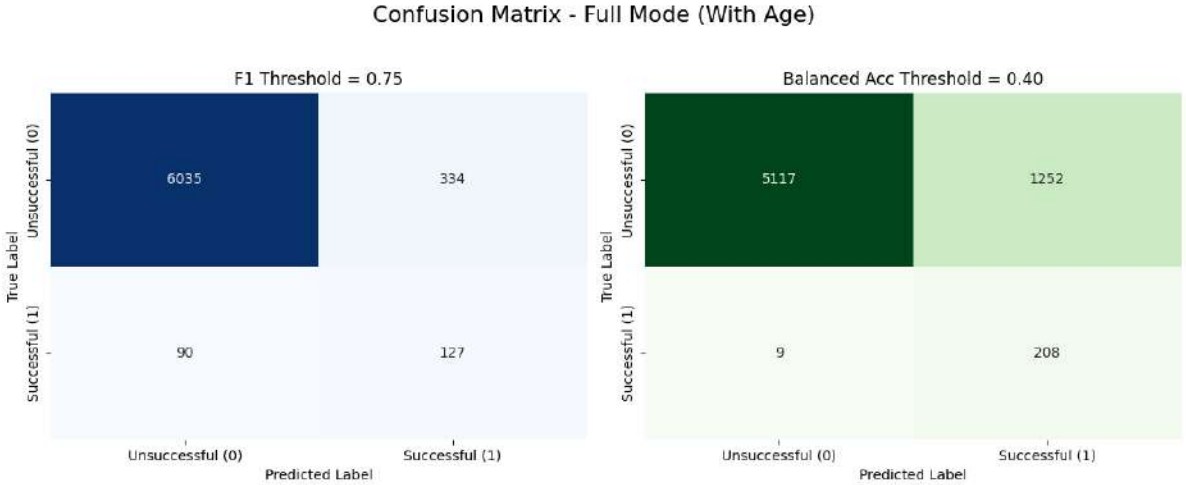


Figure 14: LR With Age – Full Mode

Threshold–Metric Profiles.

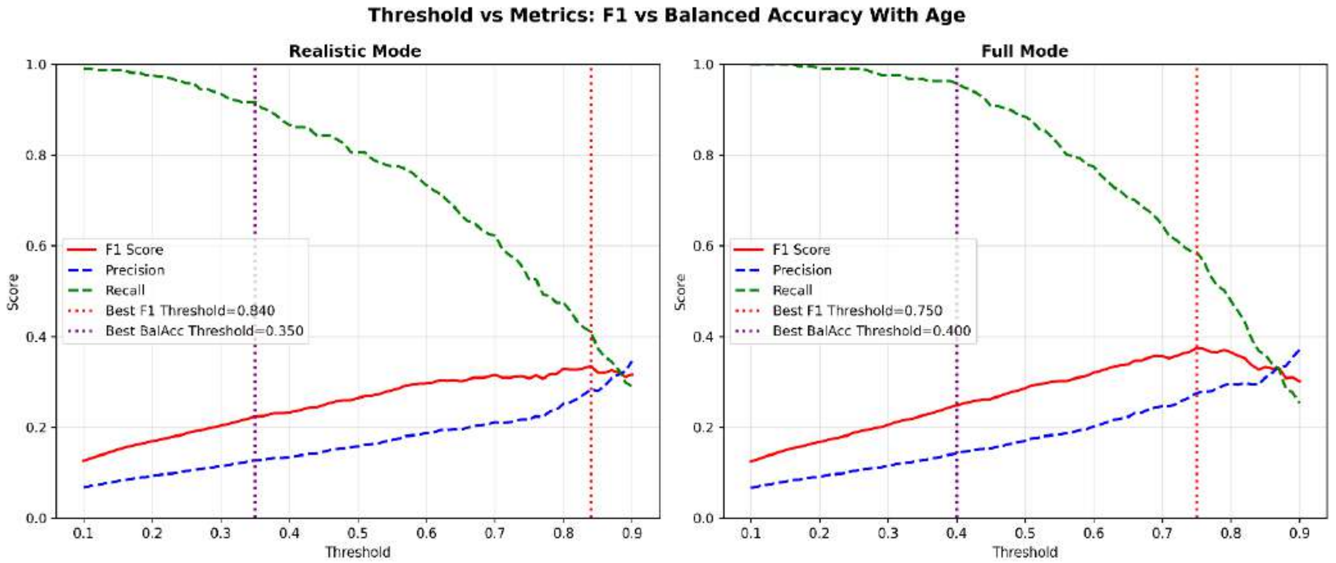


Figure 15: LR With Age Threshold–Metric

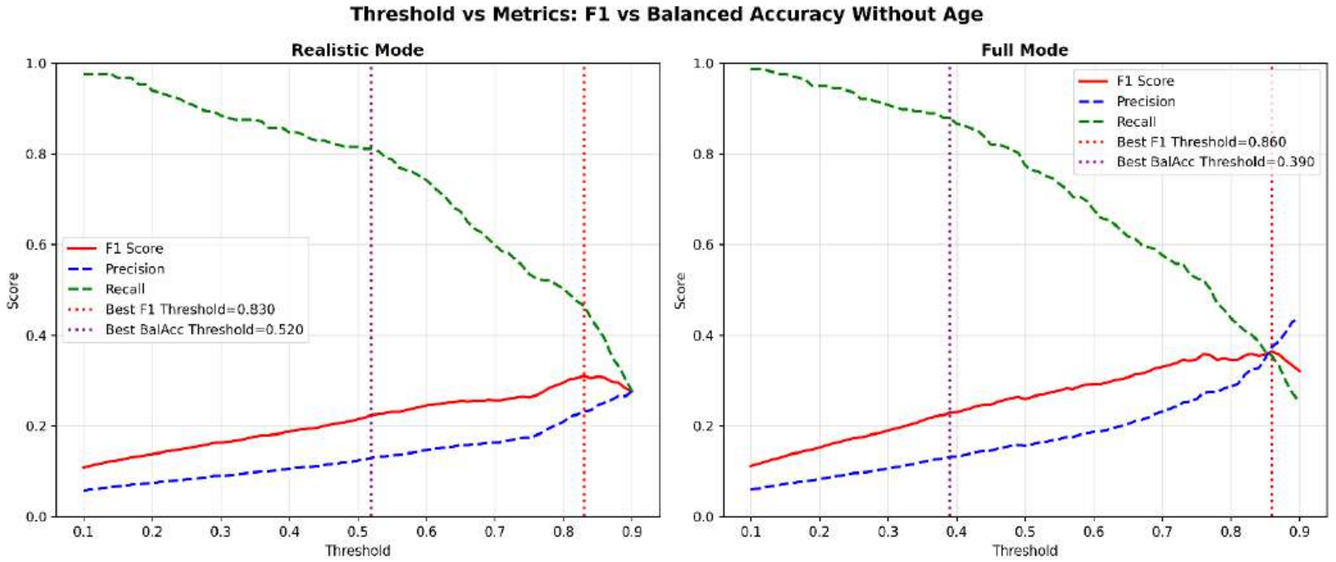
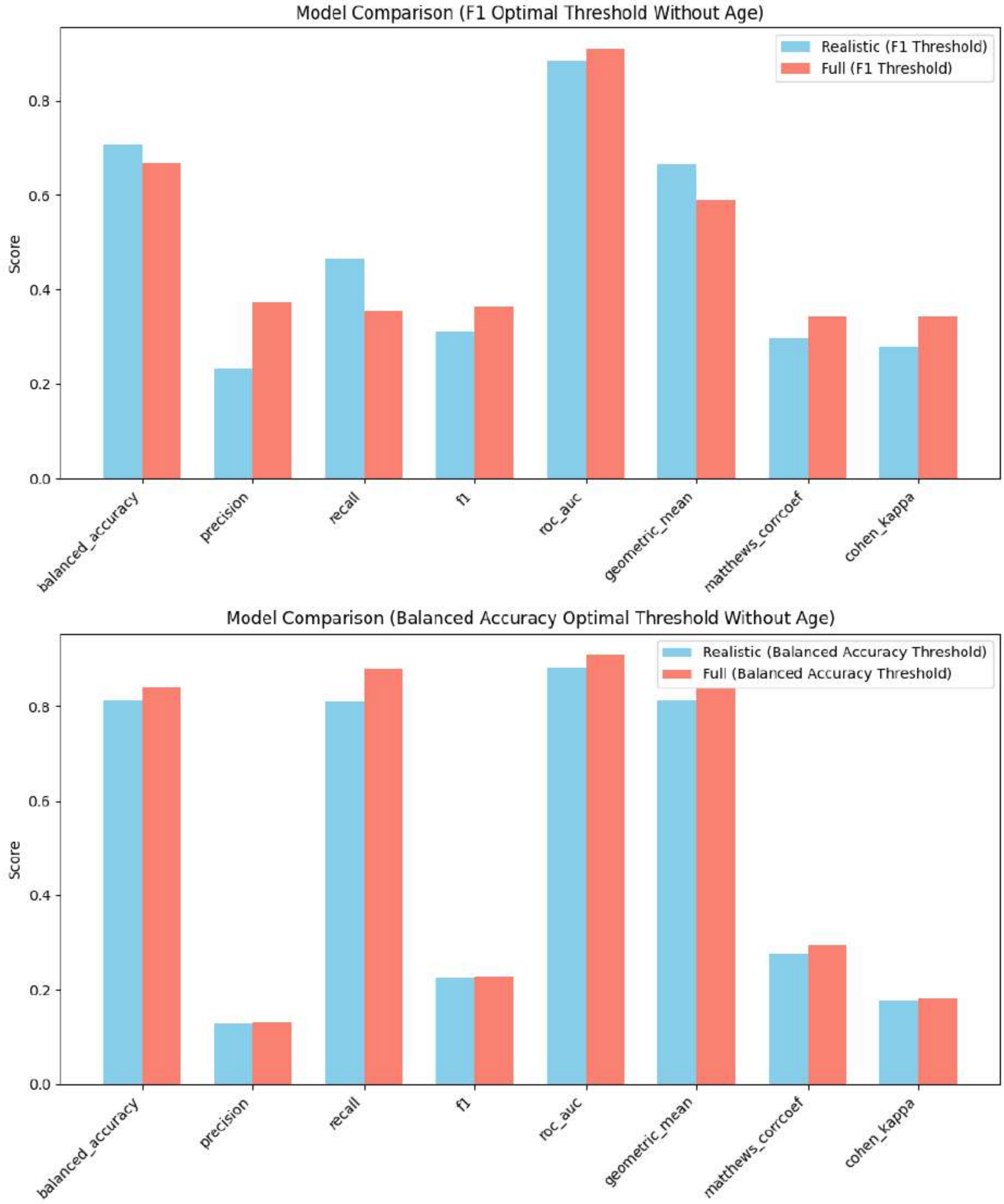


Figure 16: LR Without Age Threshold–Metric



Model Comparison (F1 Optimal Threshold Without Age)

Model Comparison (Balanced Accuracy Optimal Threshold Without Age)

Figure 15: LR model comparison without Age

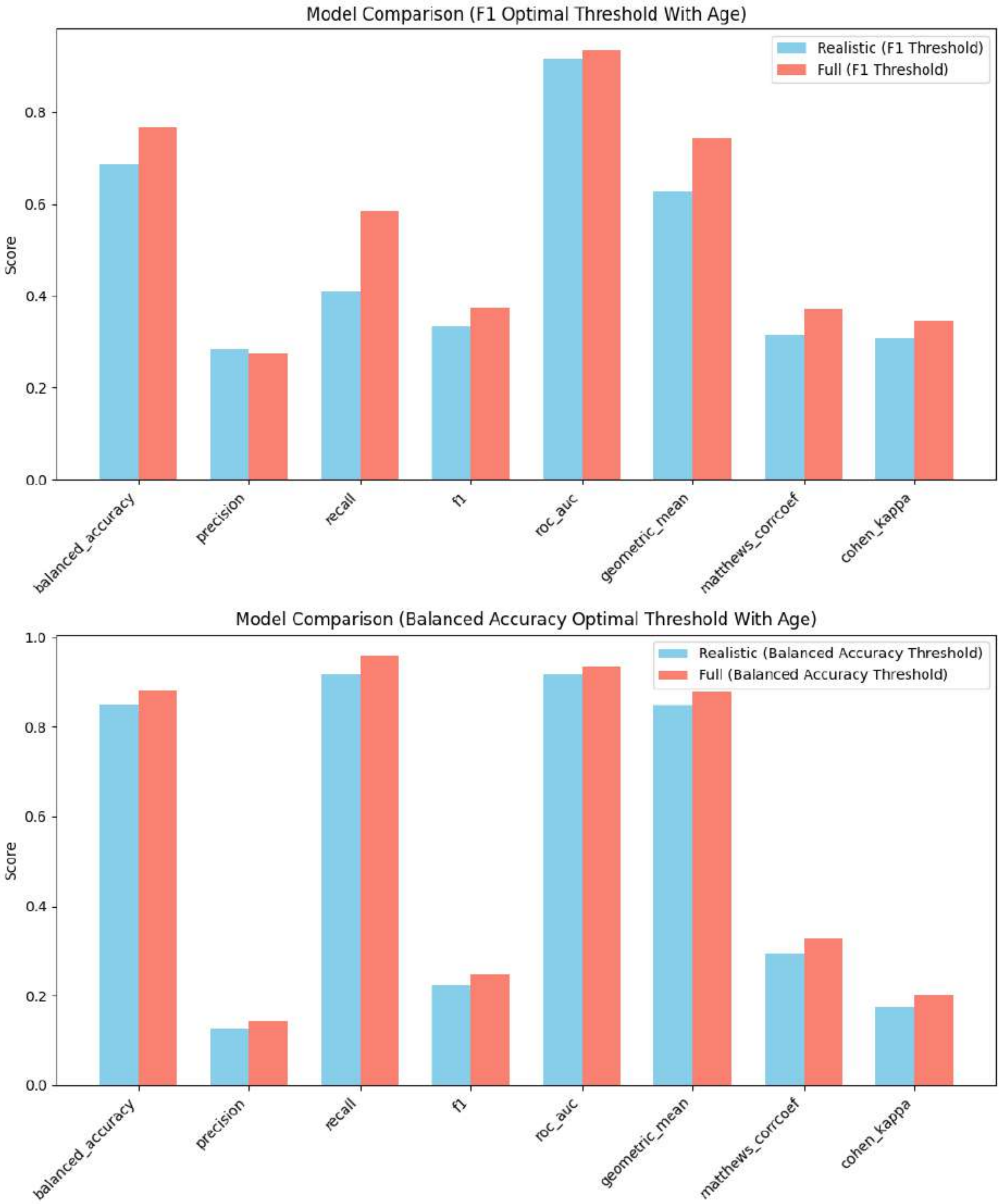


Figure 16: LR model comparison with Age

Precision–Recall and ROC Curves

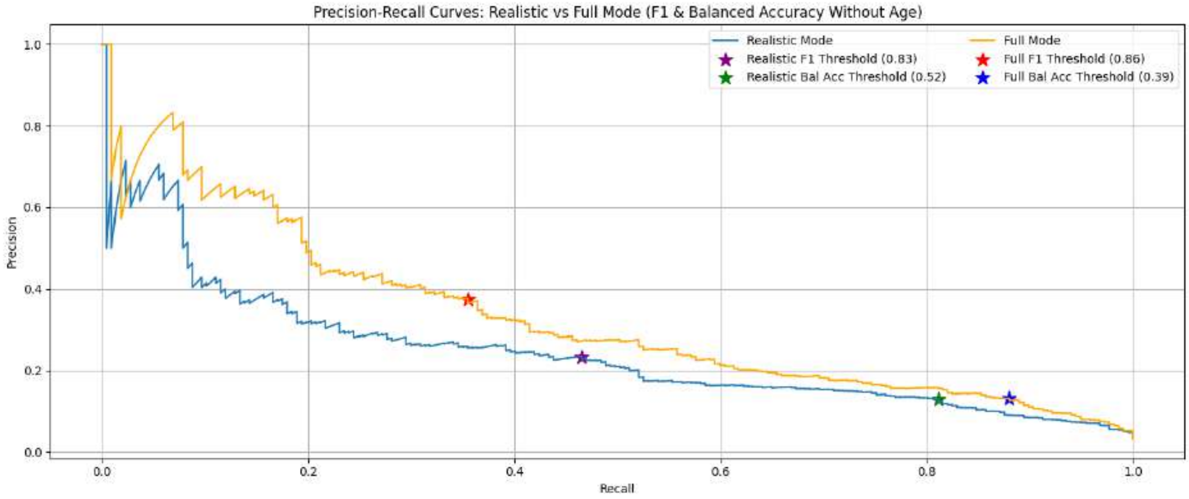


Figure 17: Precision Recall Curve Without Age

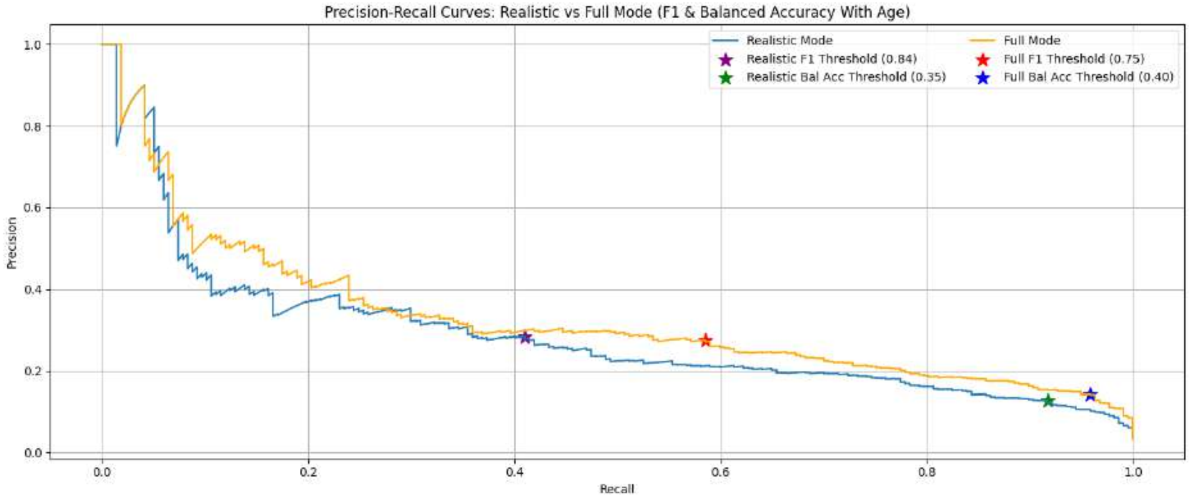


Figure 18: LR Precision Recall Curve With Age

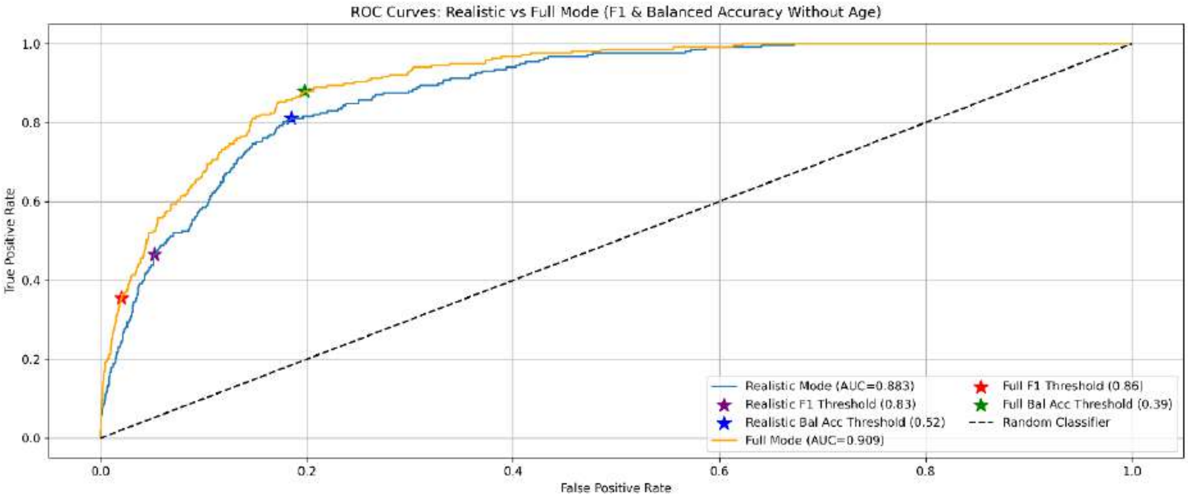


Figure 19: LR ROC-Curve Without Age

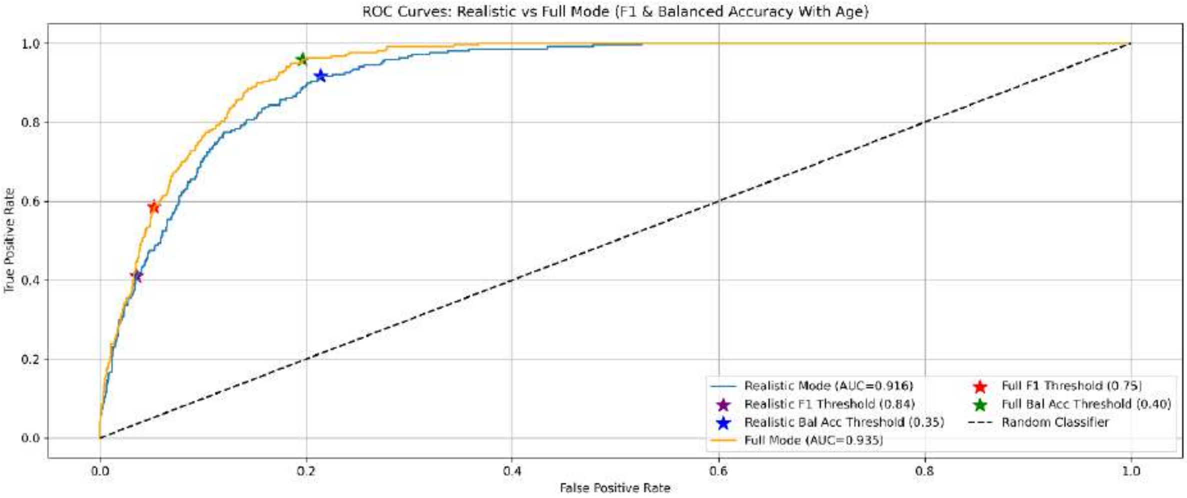


Figure 20: LR ROC-Curve With Age

RANDOM FOREST

Confusion Matrices.

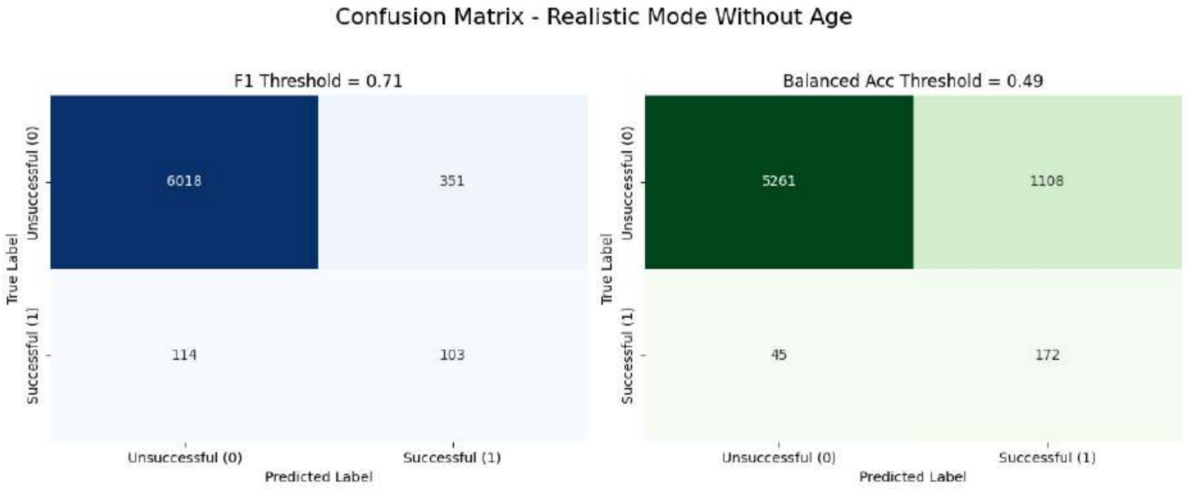


Figure 21: RF Without Age – Realistic Mode

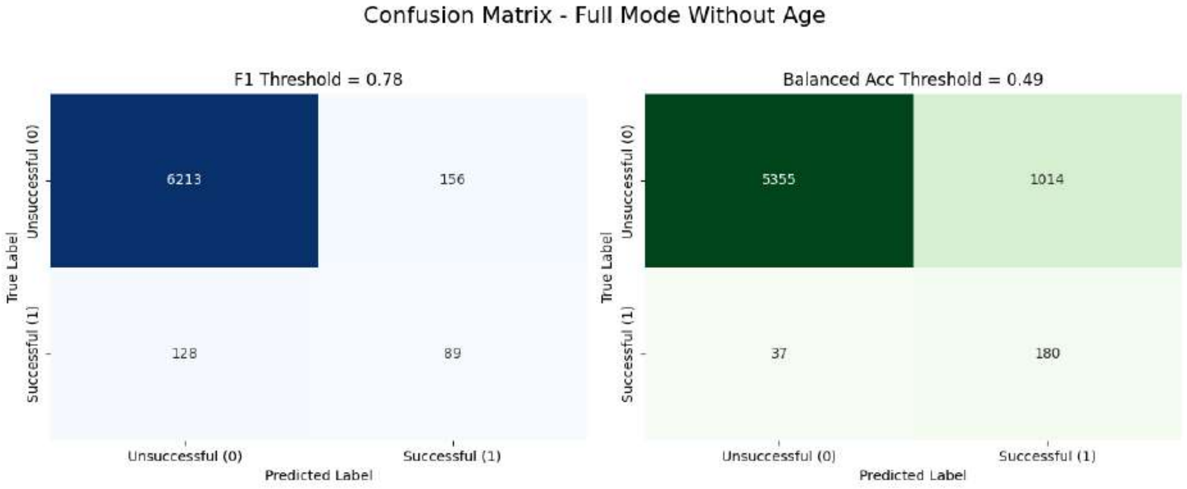


Figure 22: RF Without Age – Full Mode

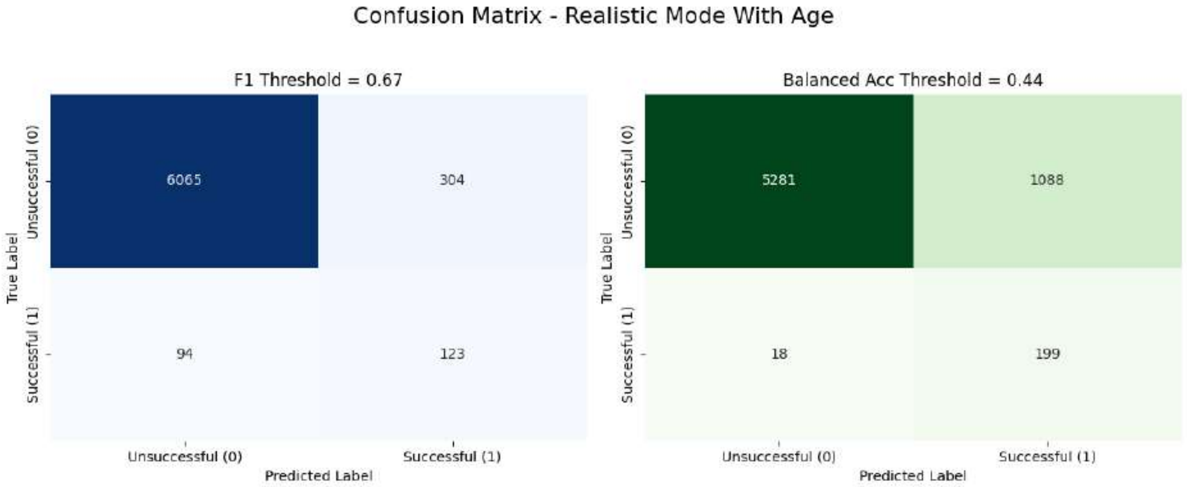


Figure 23: RF With Age – Realistic Mode

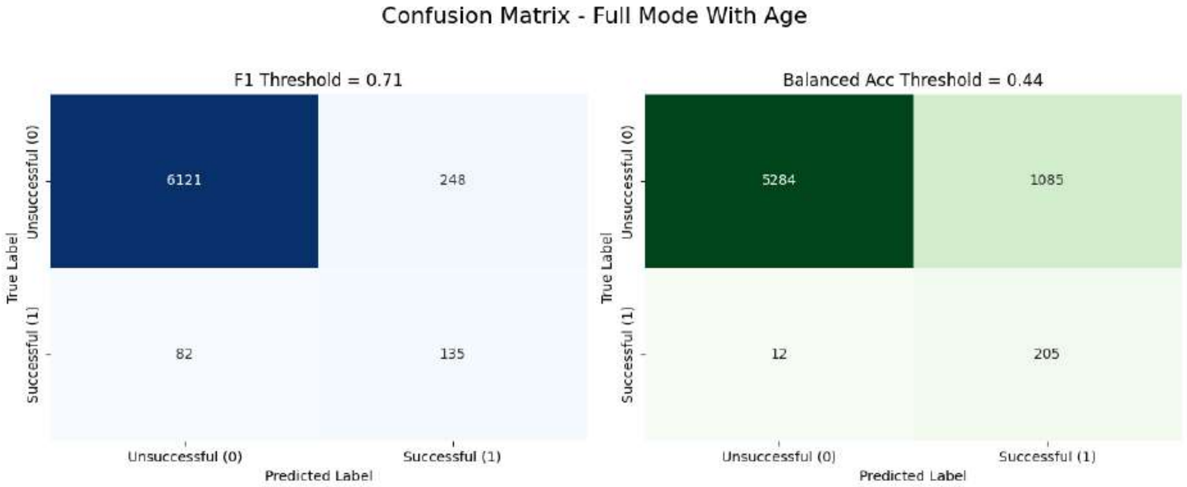


Figure 24: RF With Age – Full Mode

Threshold–Metric Profiles.

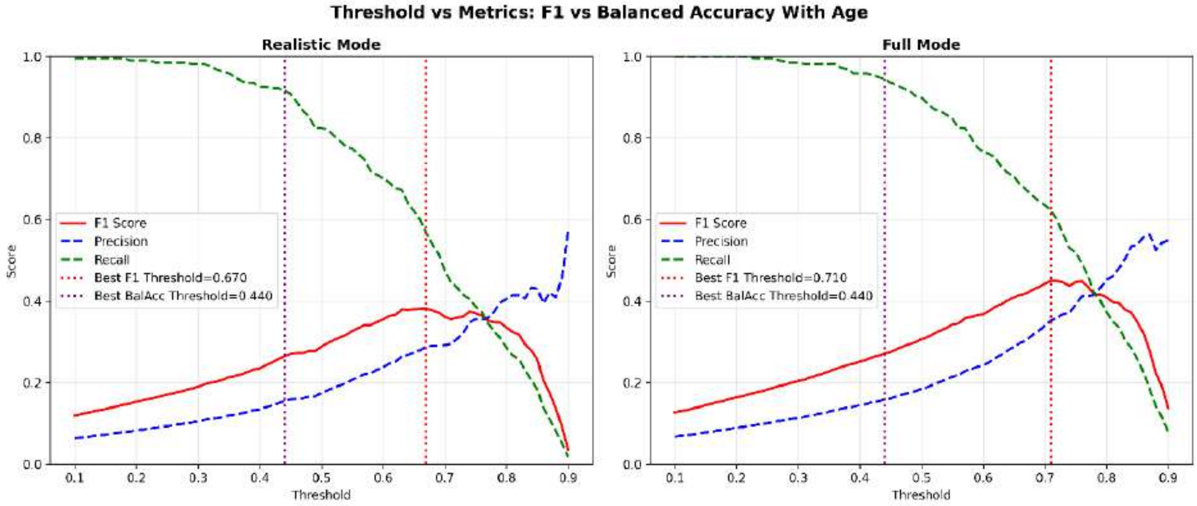


Figure 25: RF With Age Threshold–Metric

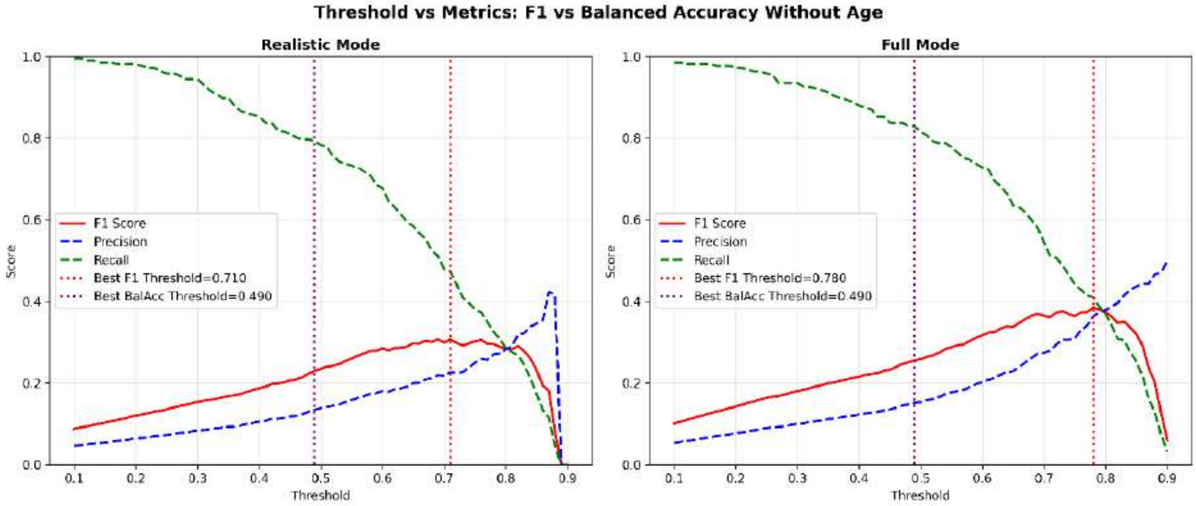


Figure 26: RF Without Age Threshold–Metric

Configuration Comparison Plots

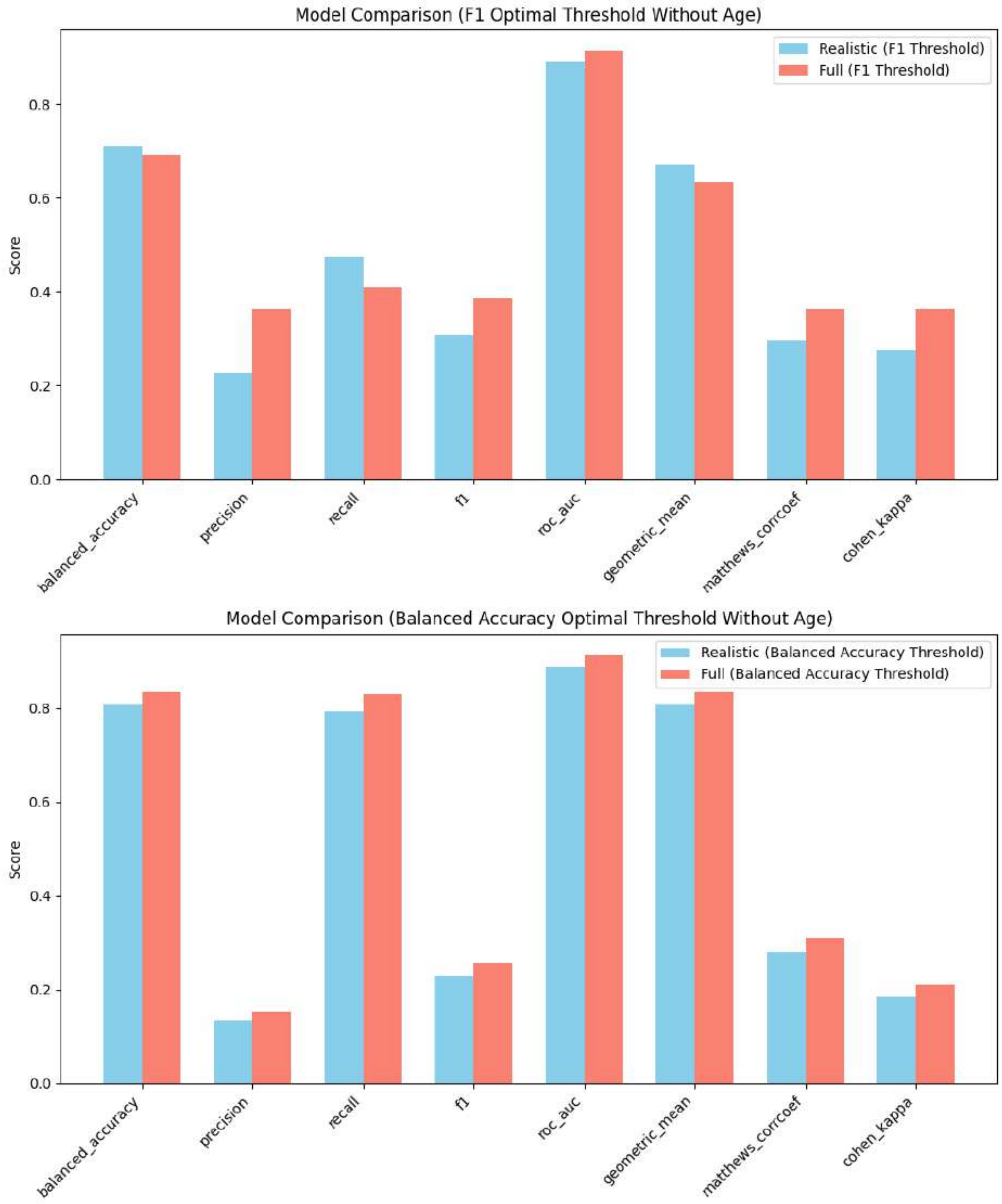


Figure 27: RF model comparison without Age

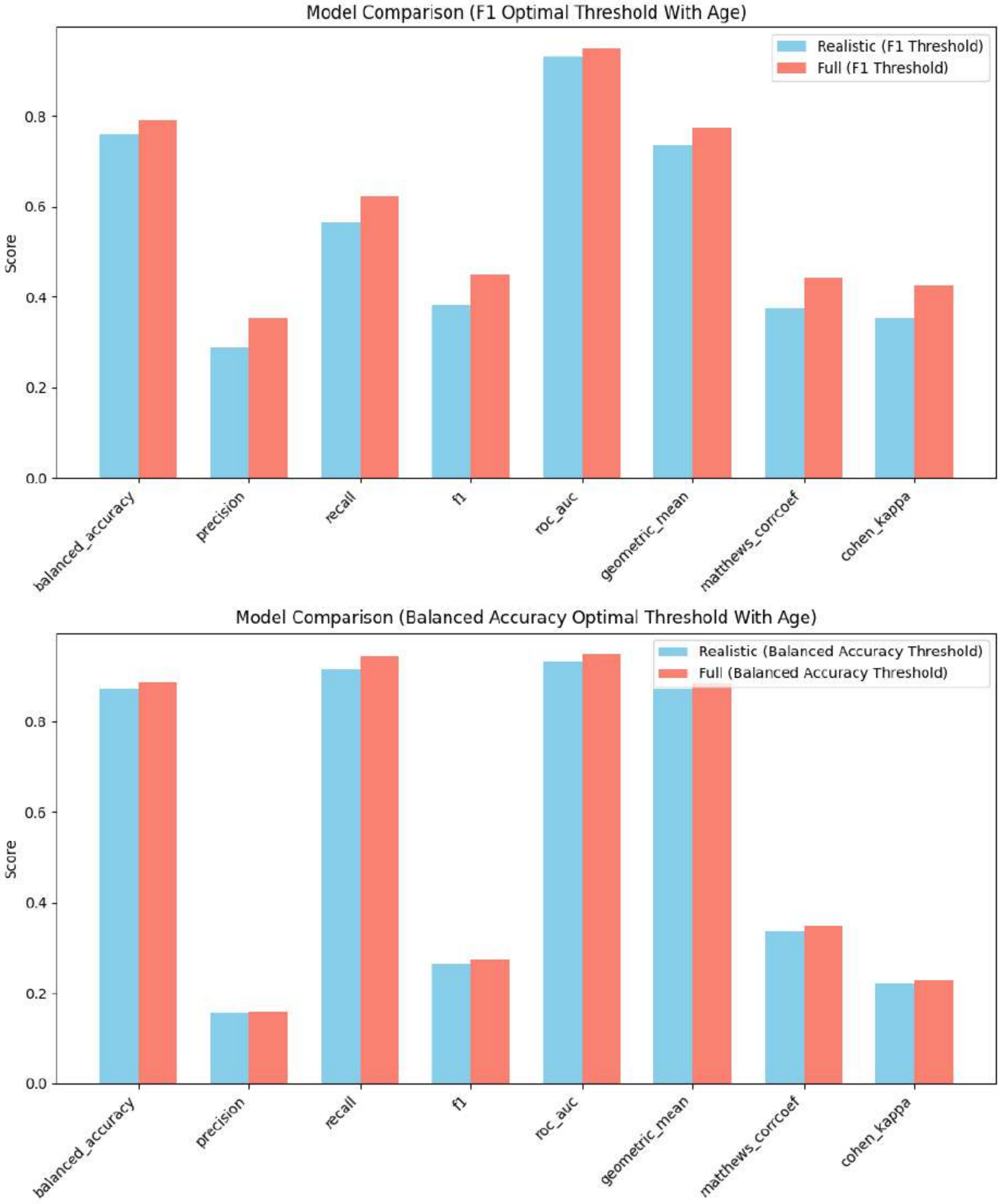


Figure 28: RF model comparison with Age

Precision–Recall and ROC Curves

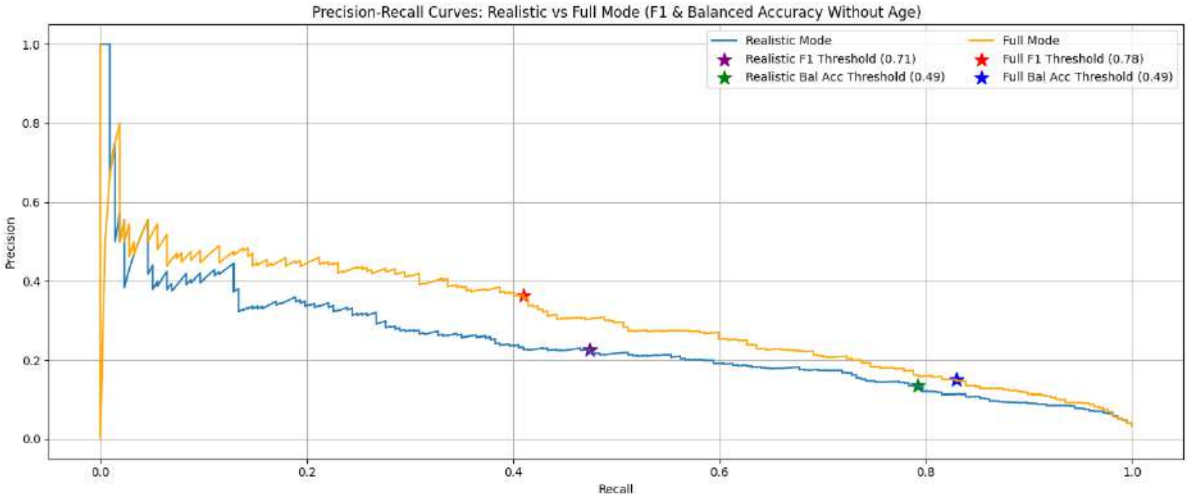


Figure 29: RF Precision Recall Curve Without Age

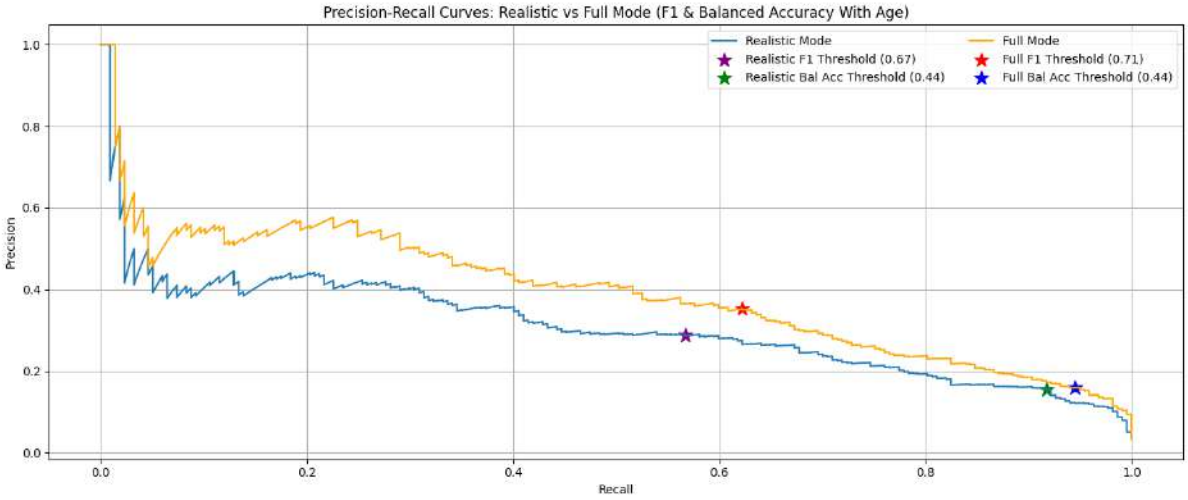


Figure 30: RF Precision Recall Curve With Age

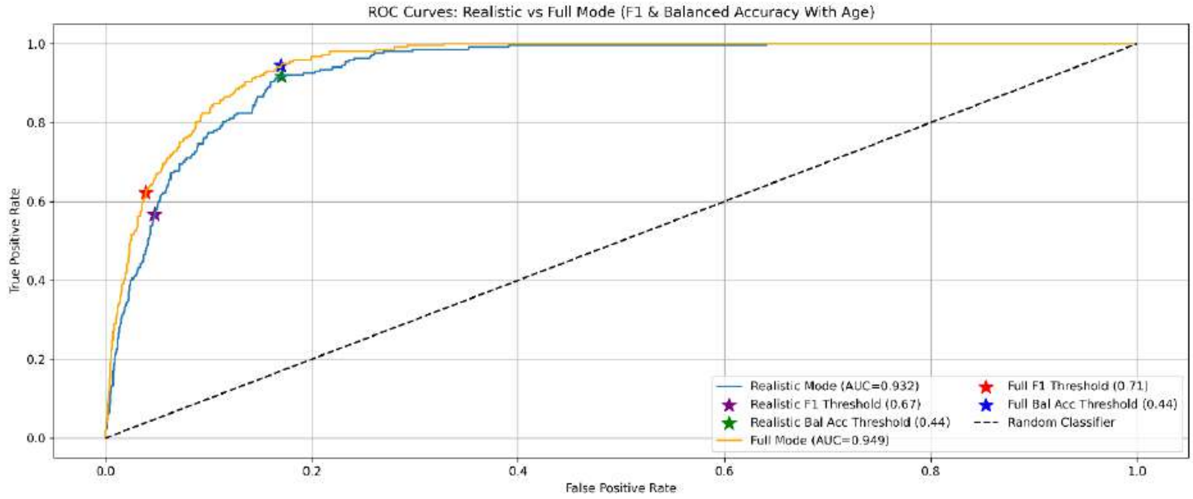


Figure 31: RF ROC-Curve With Age

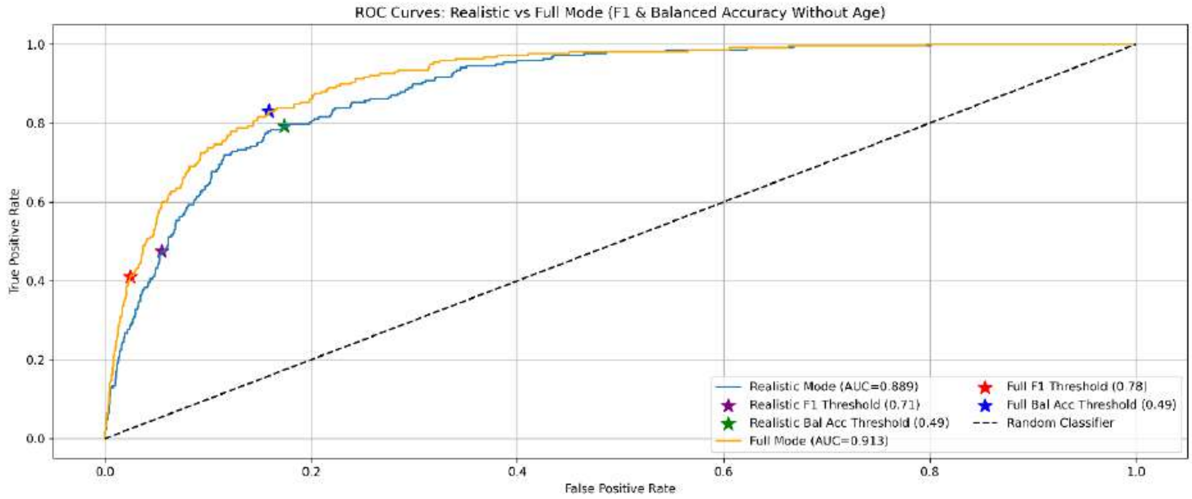


Figure 32: RF ROC-Curve Without Age

Decision Tree

Confusion Matrices.

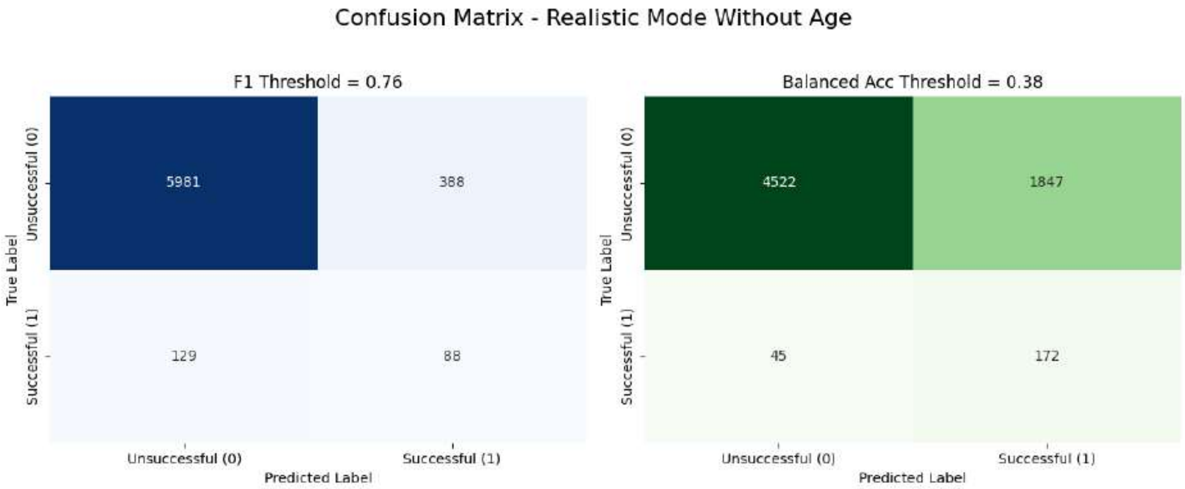


Figure 33: DT Without Age – Realistic Mode

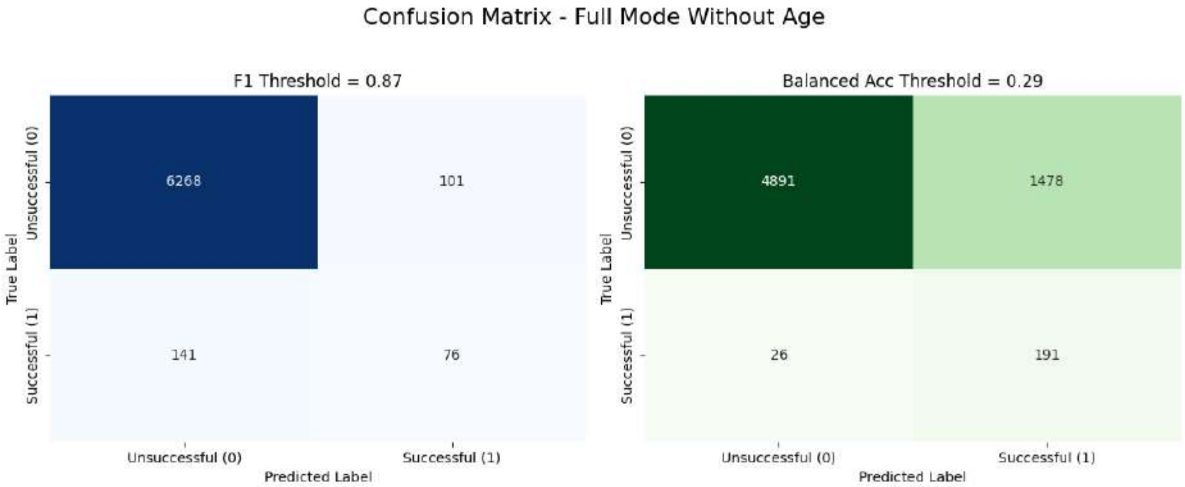


Figure 34: DT Without Age – Full Mode

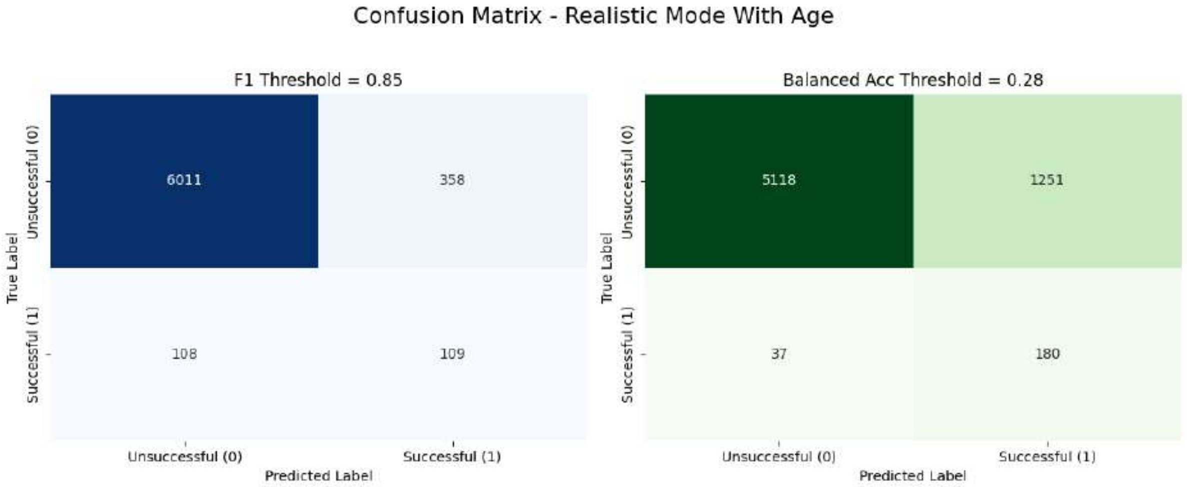


Figure 35: DT With Age – Realistic Mode

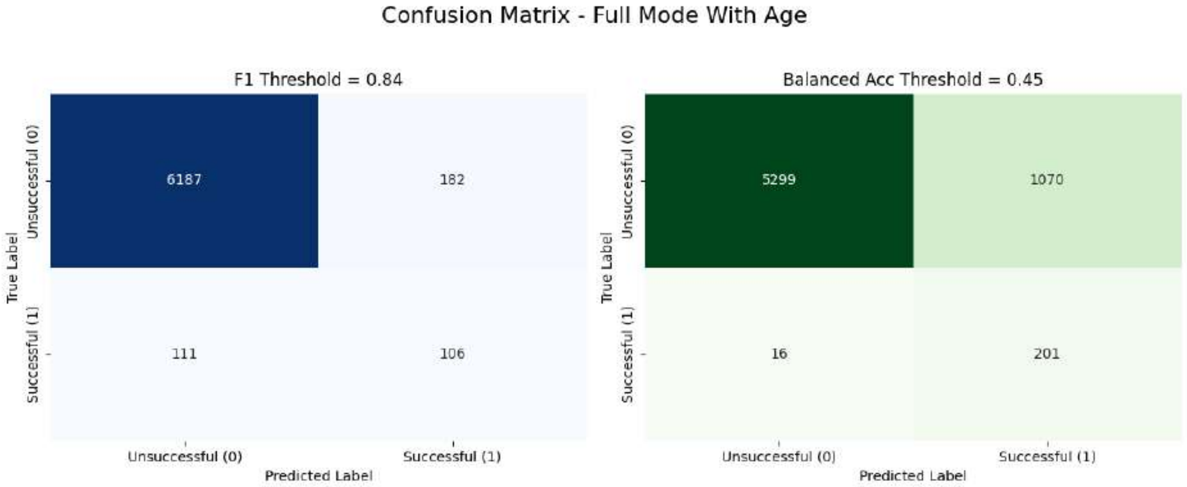


Figure 36: DT With Age – Full Mode

Threshold–Metric Profiles.

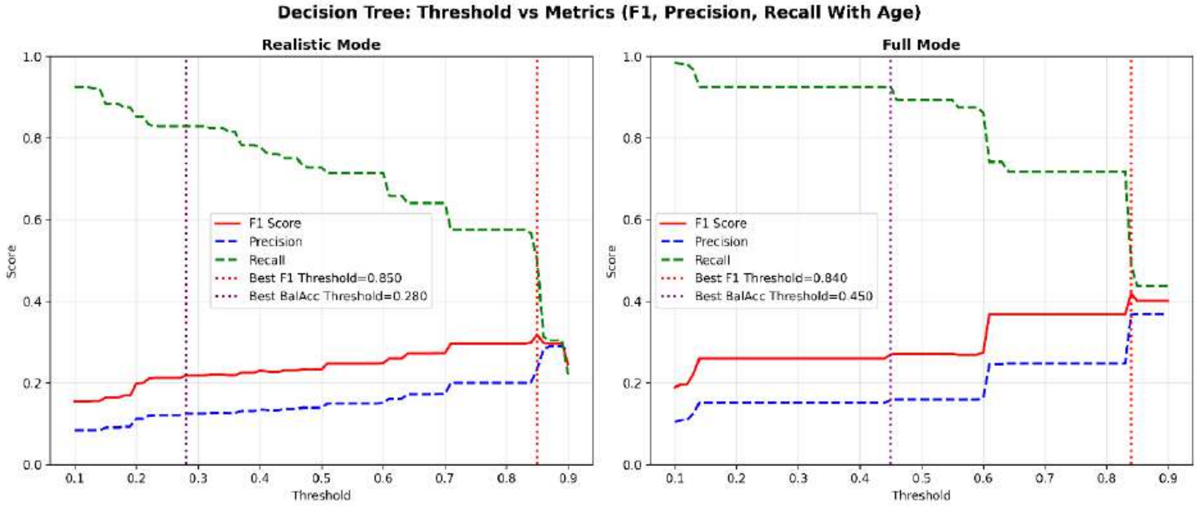


Figure 37: DT With Age Threshold–Metric

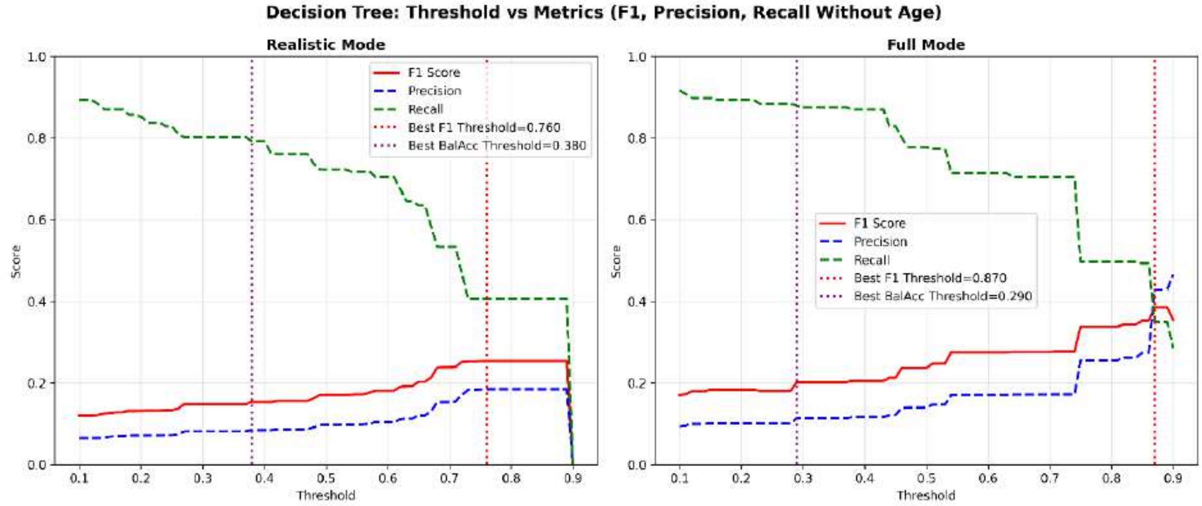


Figure 38: DT Without Age Threshold–Metric

Configuration Comparison Plots

Figures 39 and 40 summarize the performance differences between the Realistic and Full modes for Decision Tree under both threshold settings, with separate visualizations for the No-Age and With-Age configurations.

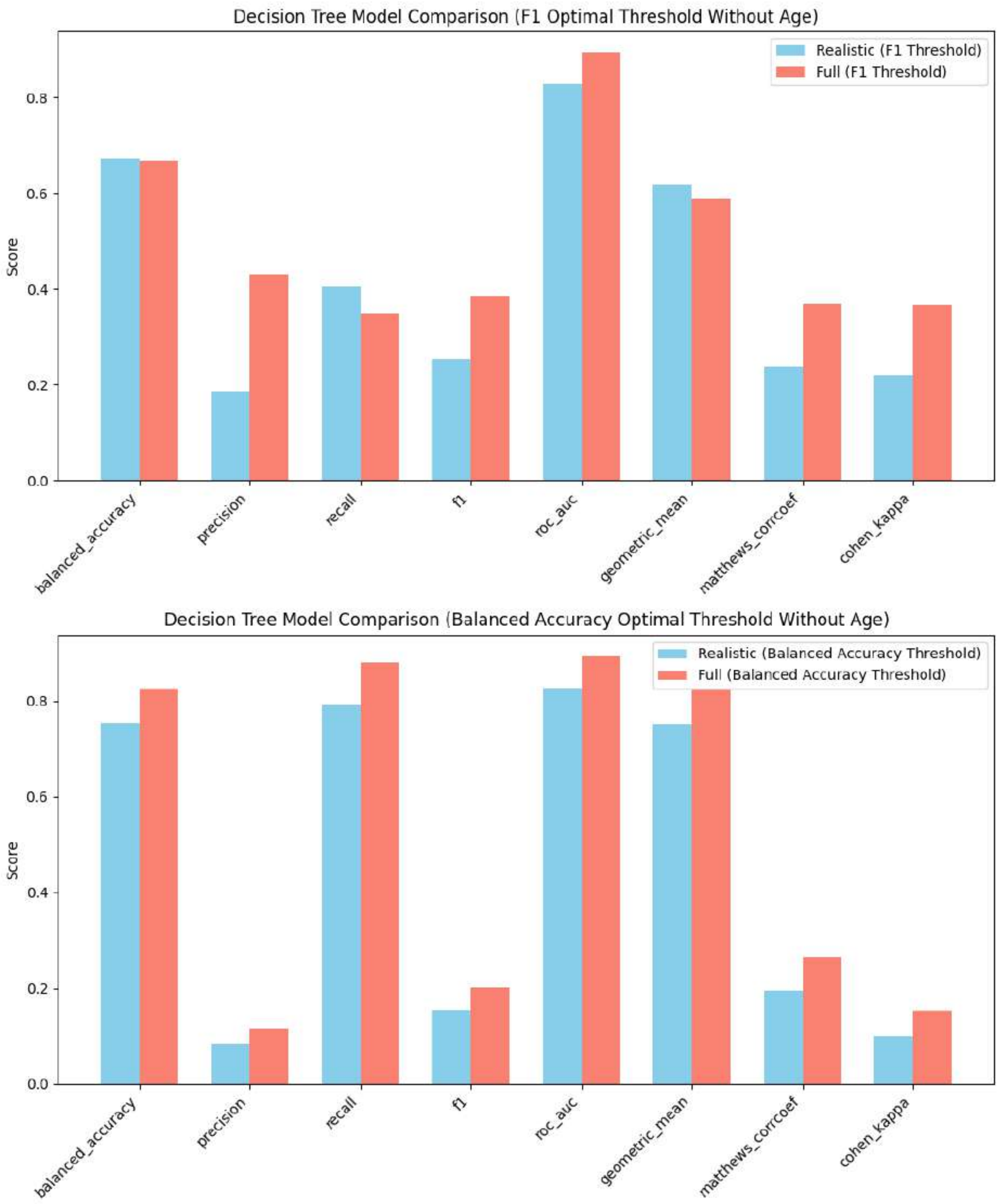


Figure 39: DT model comparison without Age

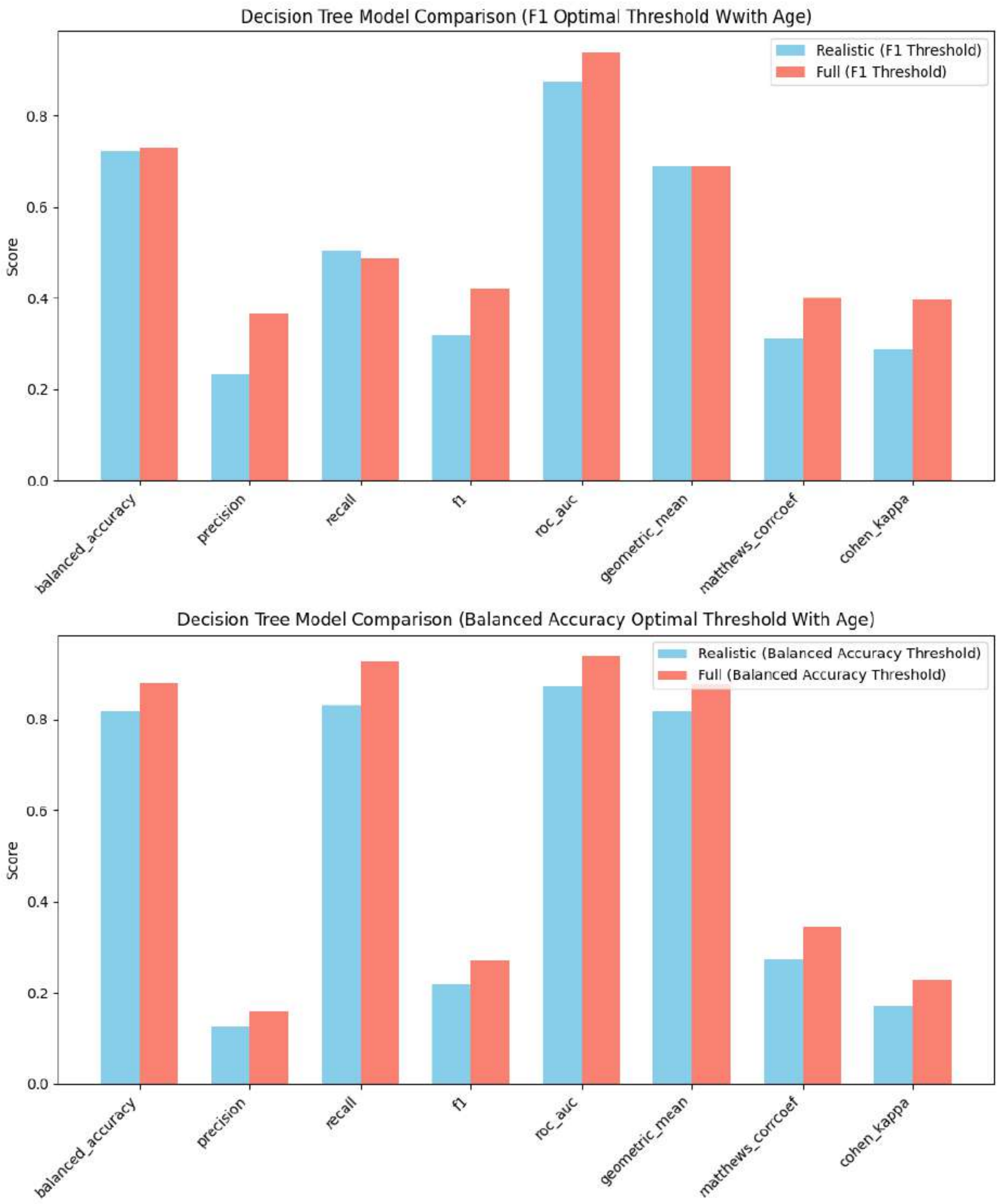


Figure 40: DT model comparison with Age

Precision–Recall and ROC Curves

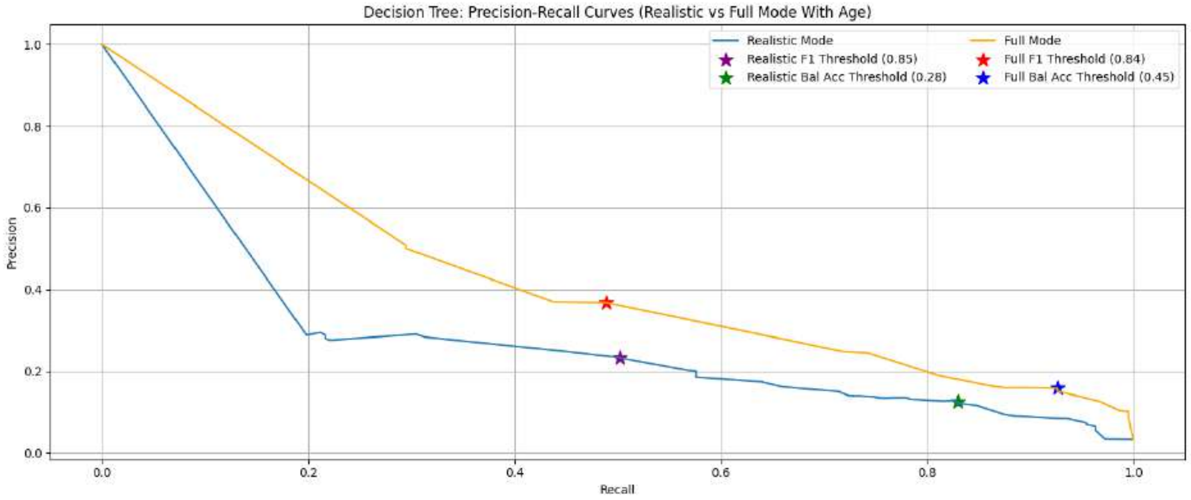


Figure 41: DT Precision Recall Curve With Age

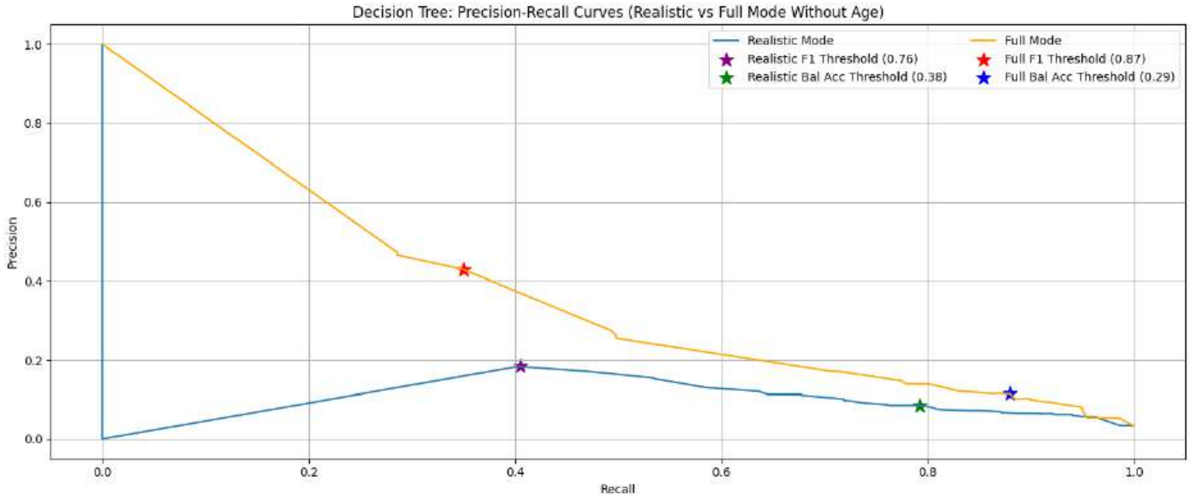
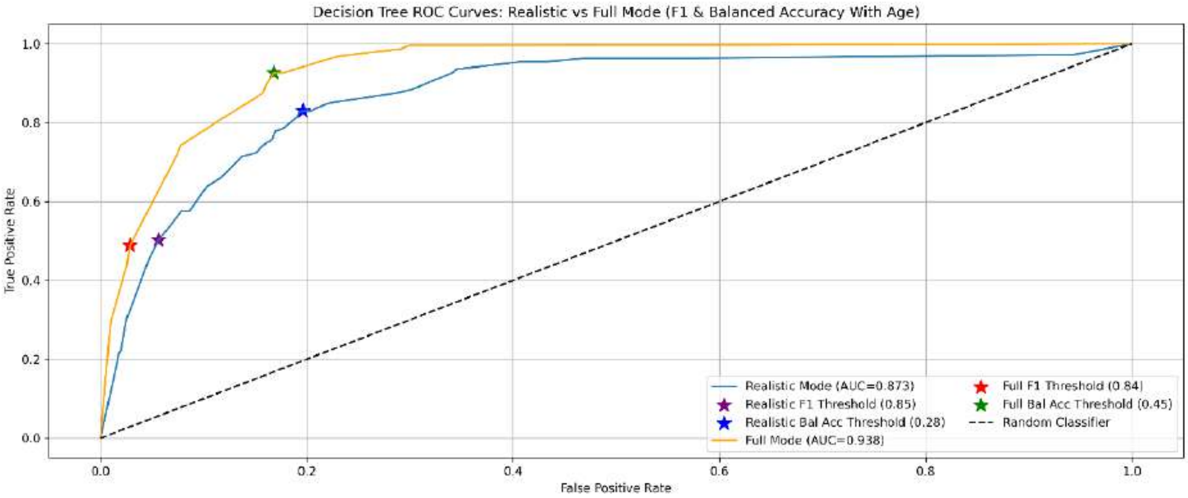


Figure 42: DT Precision Recall Curve Without Age

Figure 43: DT ROC-Curve With Age



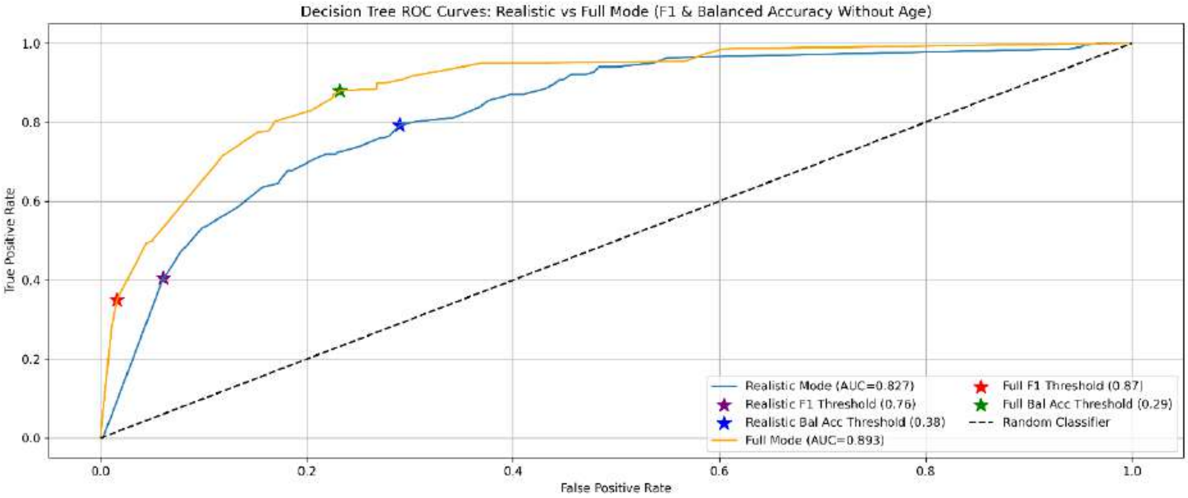


Figure 44: DT ROC-Curve Without Age

SVC

Confusion Matrices.

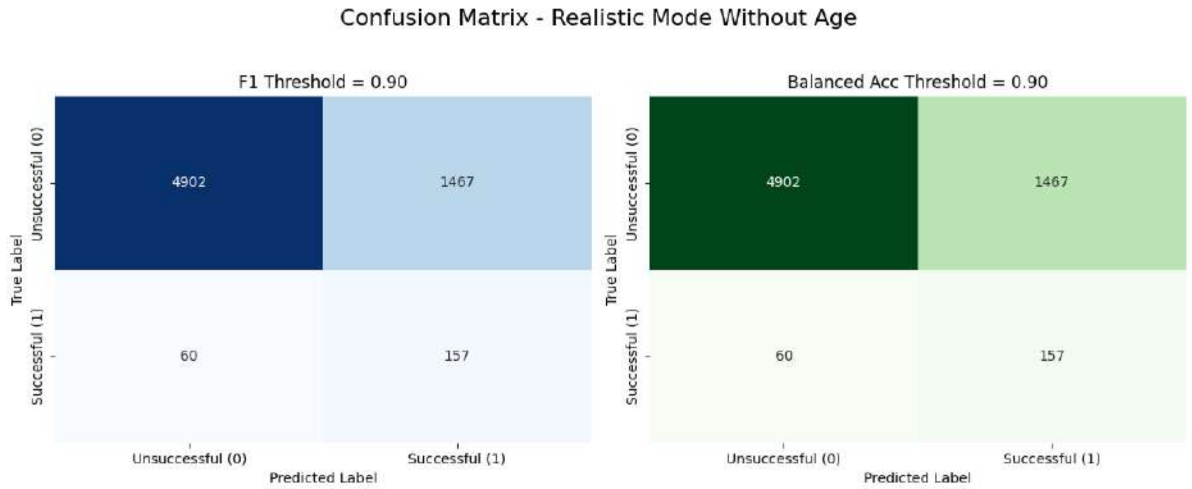


Figure 45: SVC Without Age – Realistic Mode

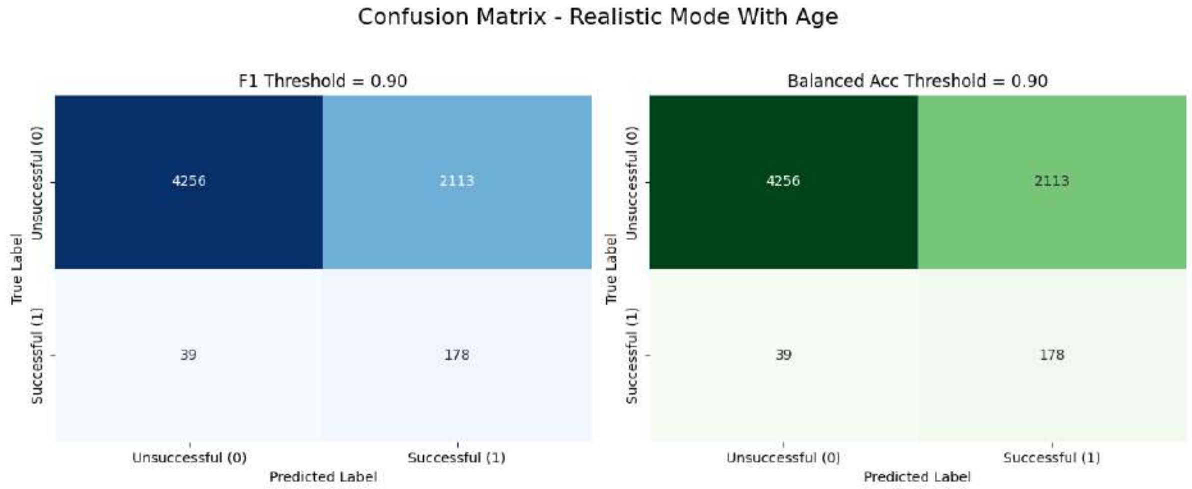


Figure 46: SVC Without Age – Full Mode

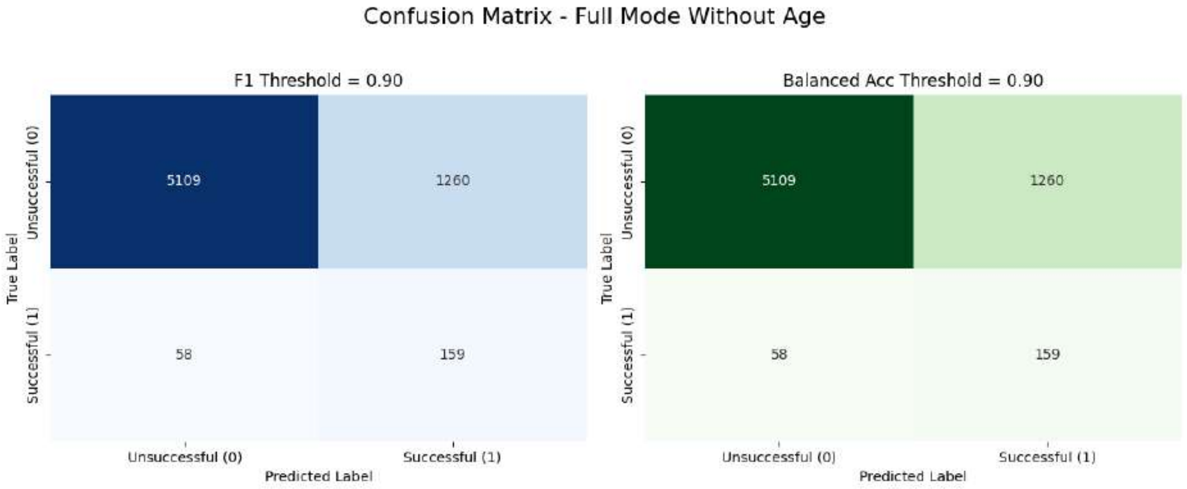


Figure 47: SVC With Age – Realistic Mode

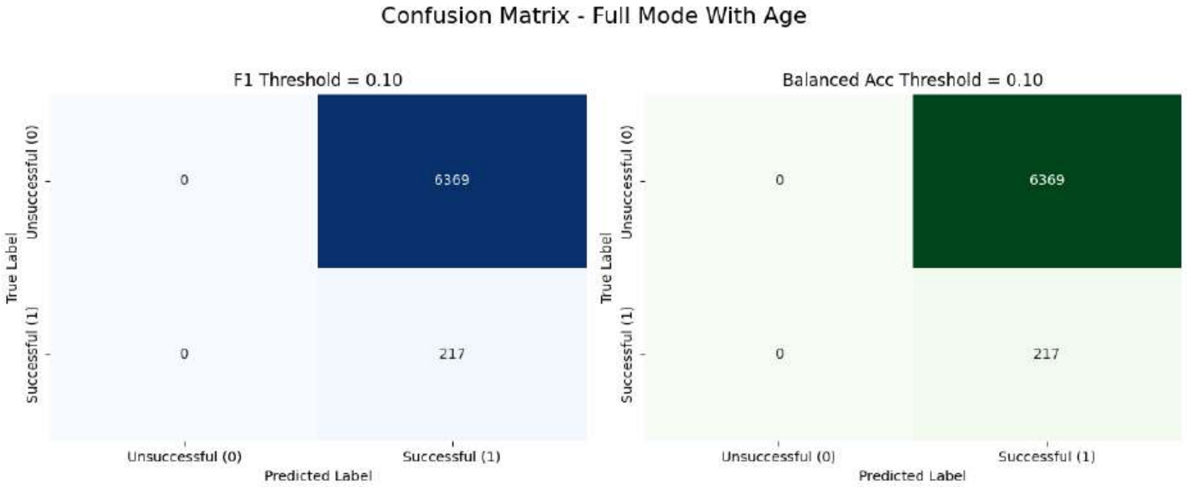


Figure 48: SVC With Age – Full Mode

Threshold–Metric Profiles.

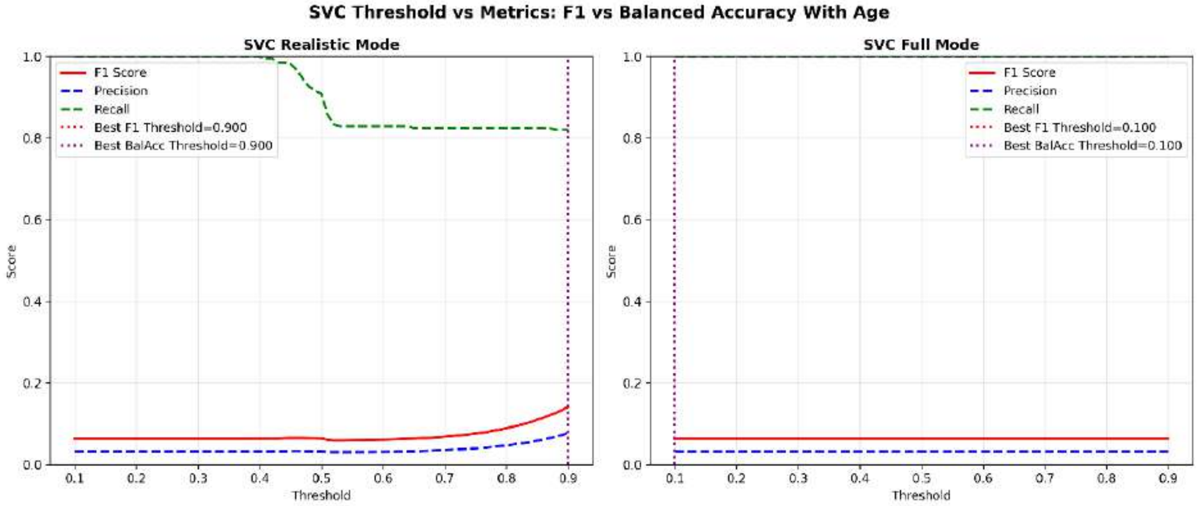


Figure 49: SVC With Age Threshold–Metric

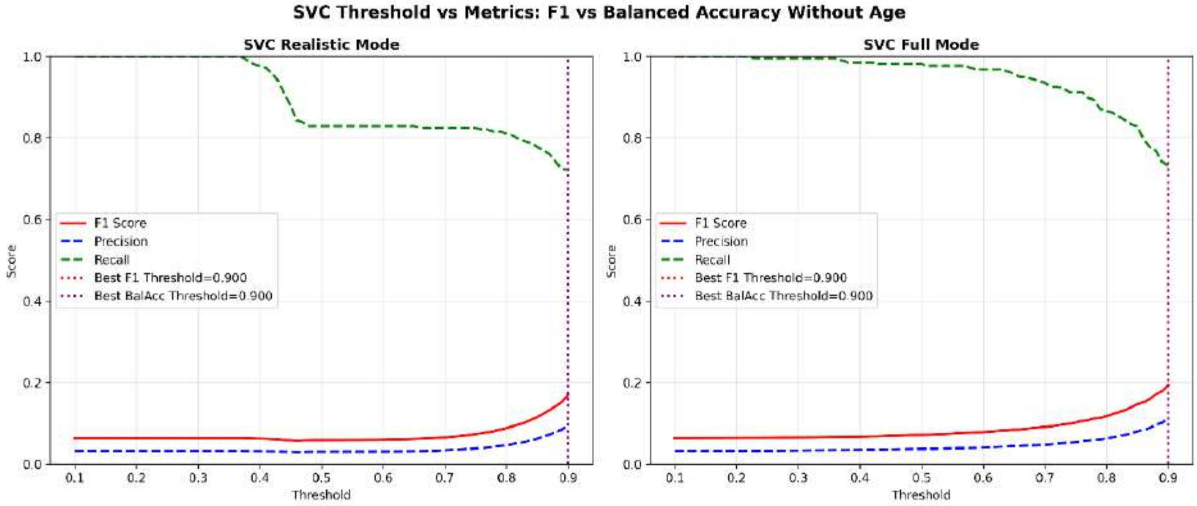
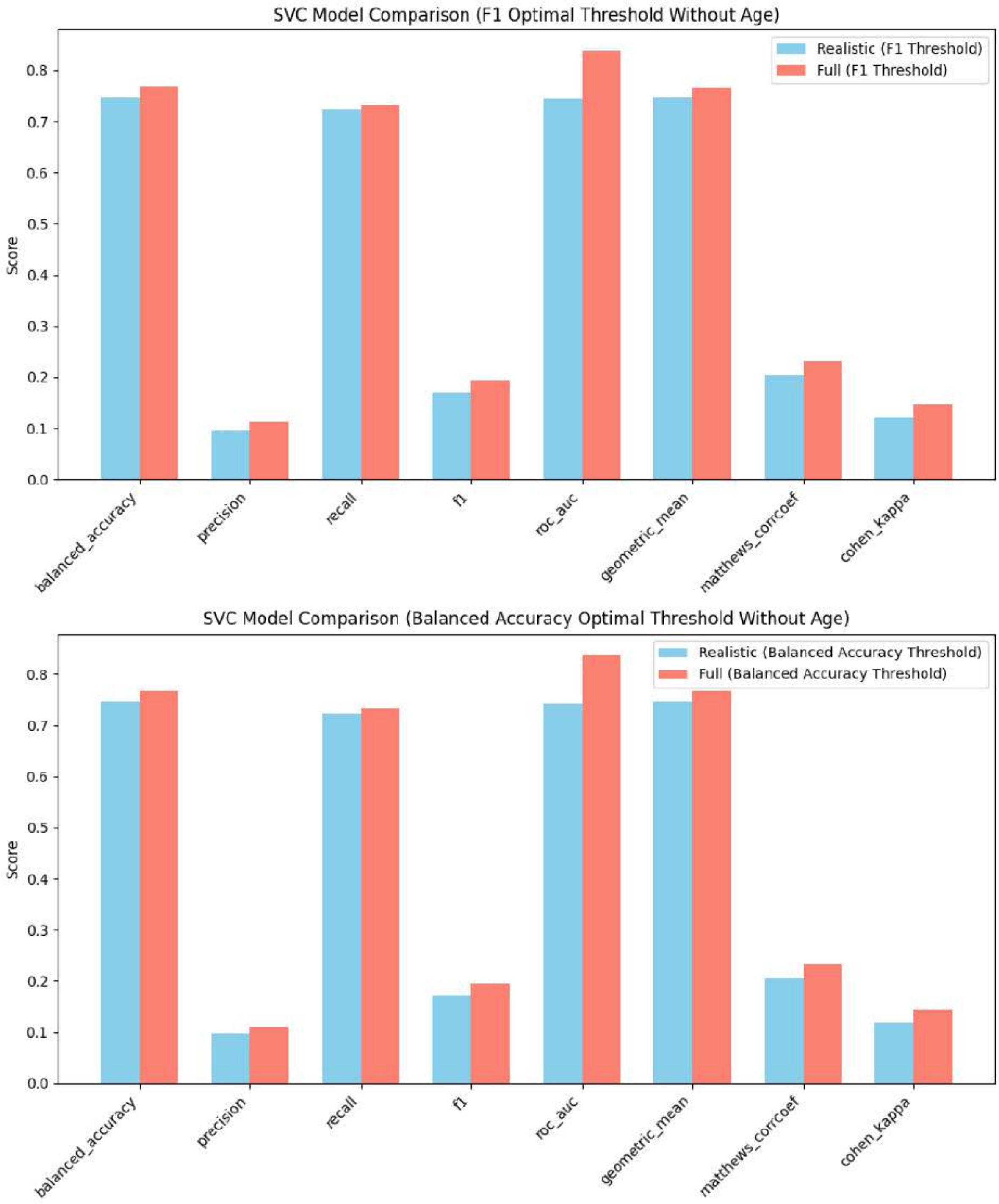


Figure 50: SVC Without Age Threshold–Metric

Configuration Comparison Plots



SVC Model Comparison (F1 Optimal Threshold Without Age)

SVC Model Comparison (Balanced Accuracy Optimal Threshold Without Age)

Figure 51: SVC model comparison without Age

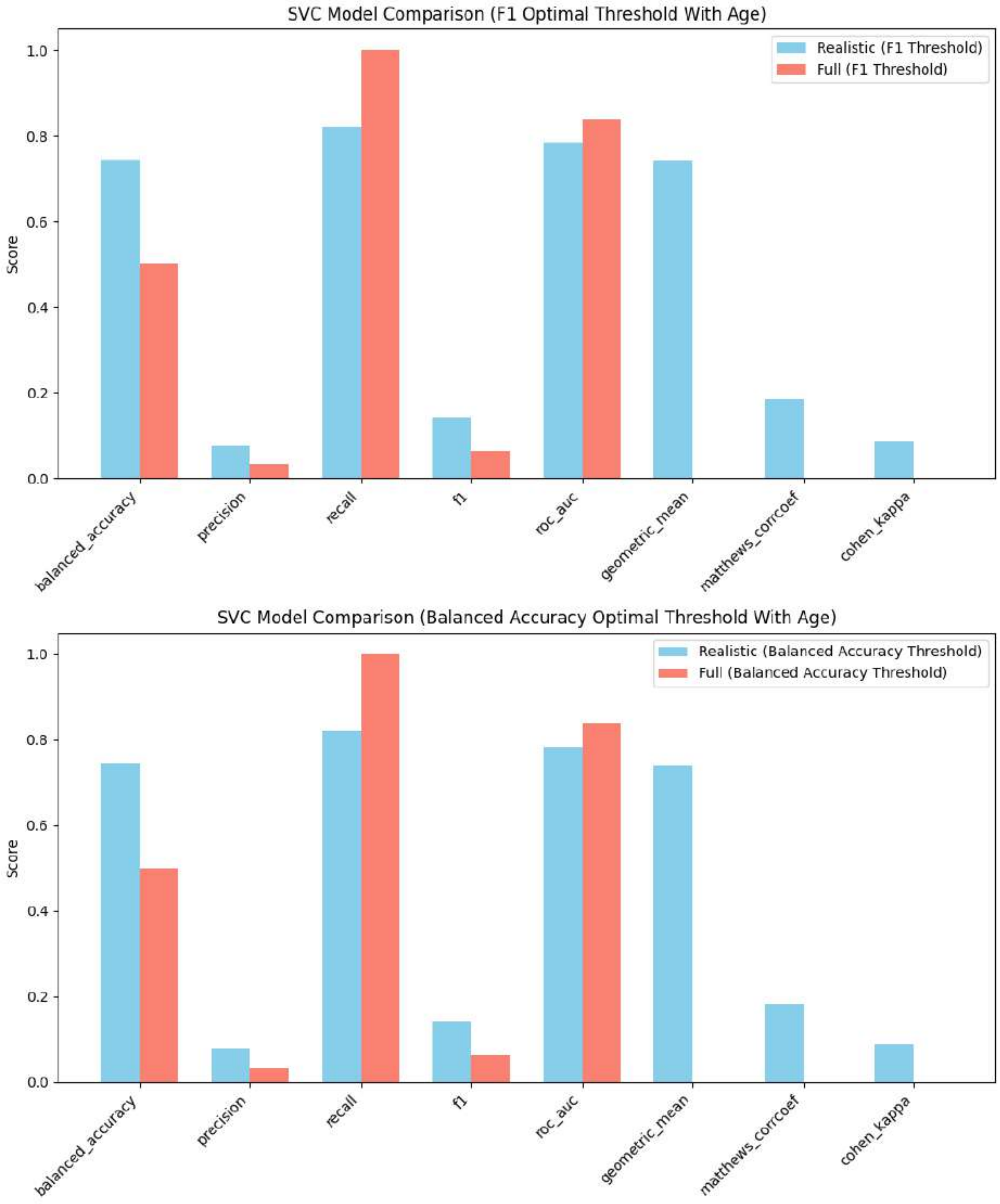


Figure 52: SVC model comparison with Age

Precision–Recall and ROC Curves

Four diagnostic plots were generated to summarize the threshold-independent performance of SVM. The Precision–Recall curves for the With-Age and Without-Age configurations are presented in Figures 53 and 54, respectively, each comparing the Realistic and Full modes under both threshold optimization strategies. The corresponding ROC curves for the With-Age and Without-Age configurations are shown in Figures 55 and 56

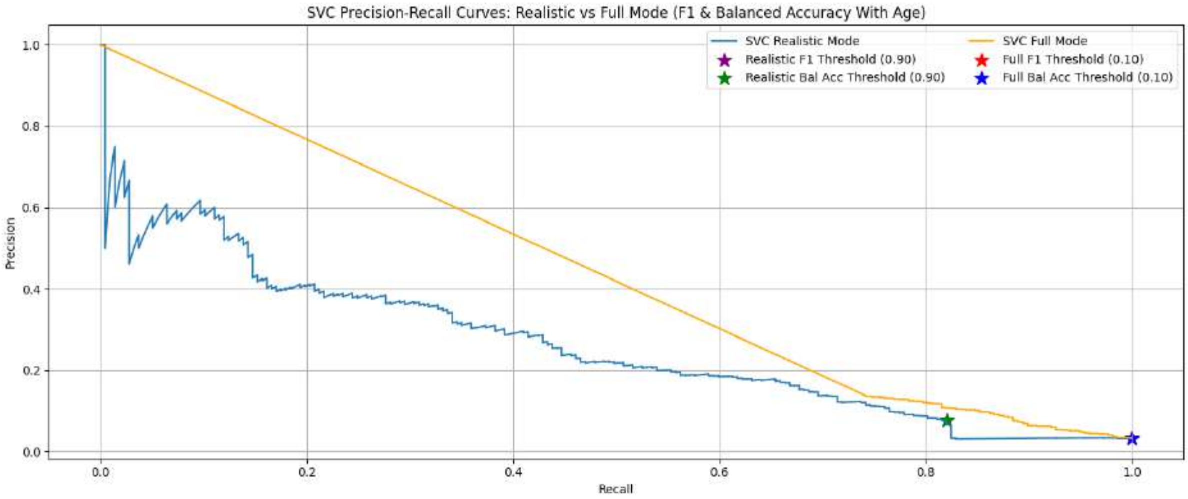


Figure 53: SVC Precision Recall Curve With Age

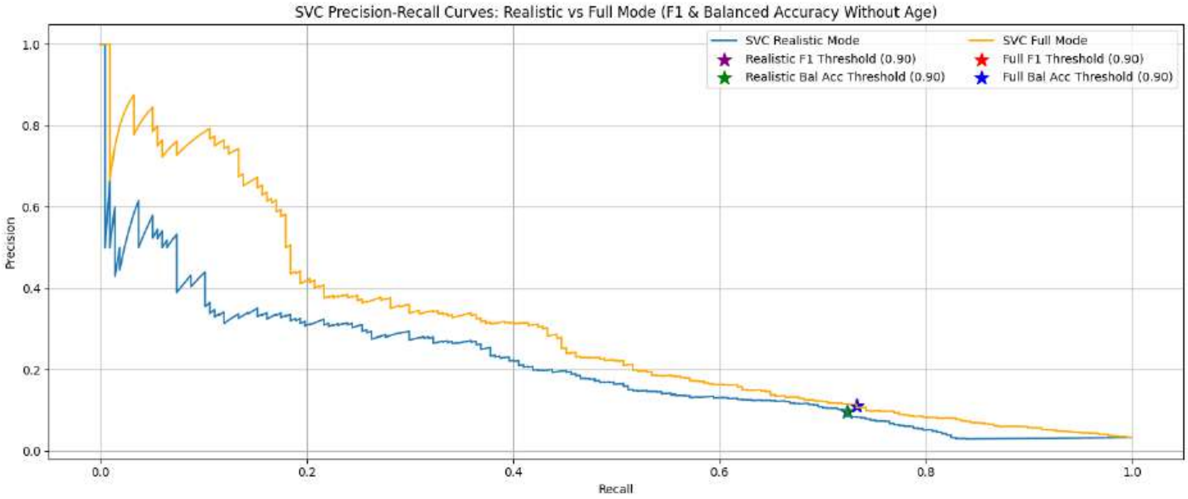


Figure 54: SVC Precision Recall Curve Without Age

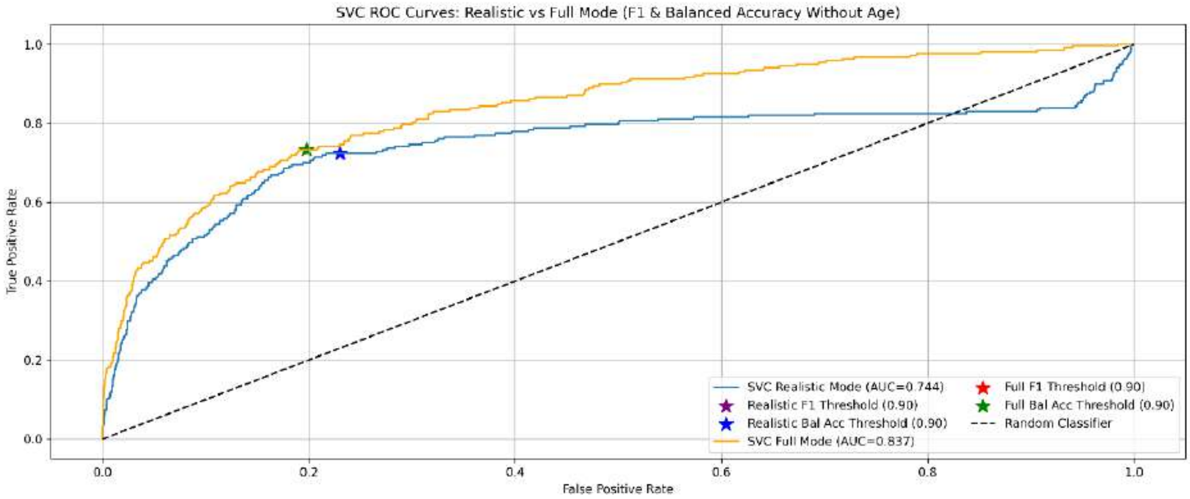


Figure 55: SVC ROC-Curve Without Age

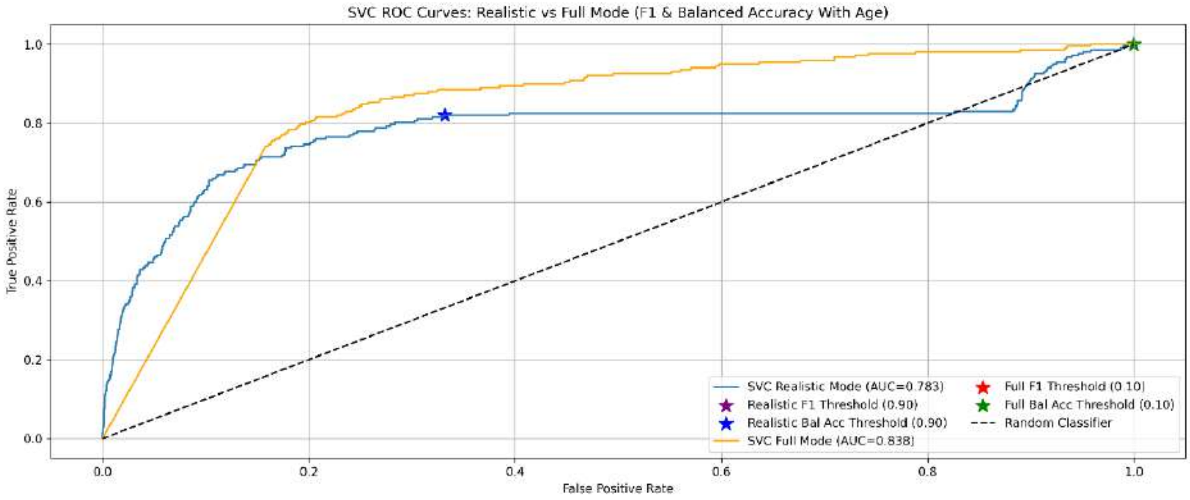


Figure 56: SVC ROC-Curve With Age

XGBoost

Confusion Matrices.

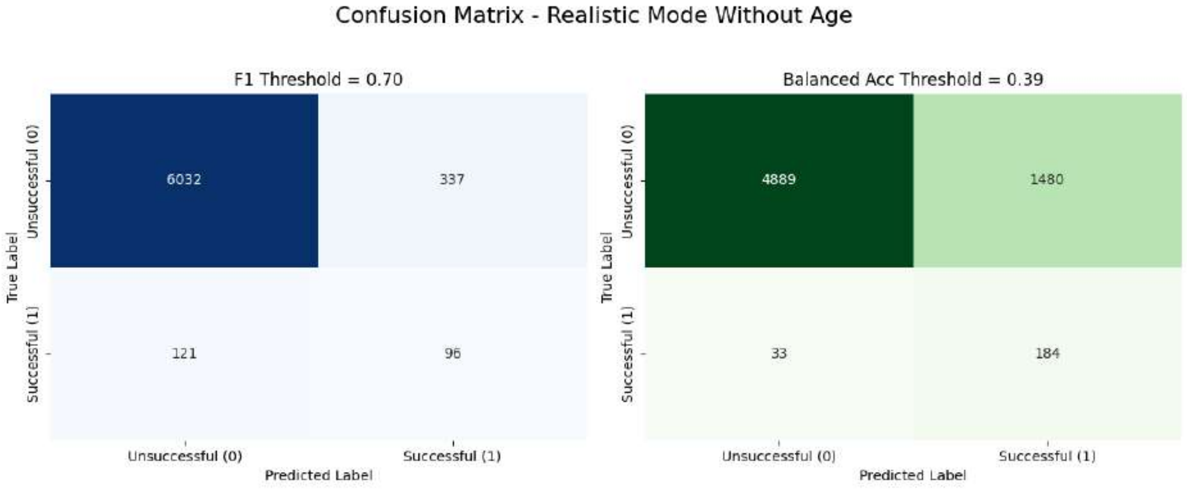


Figure 57: XGBoost Without Age – Realistic Mode

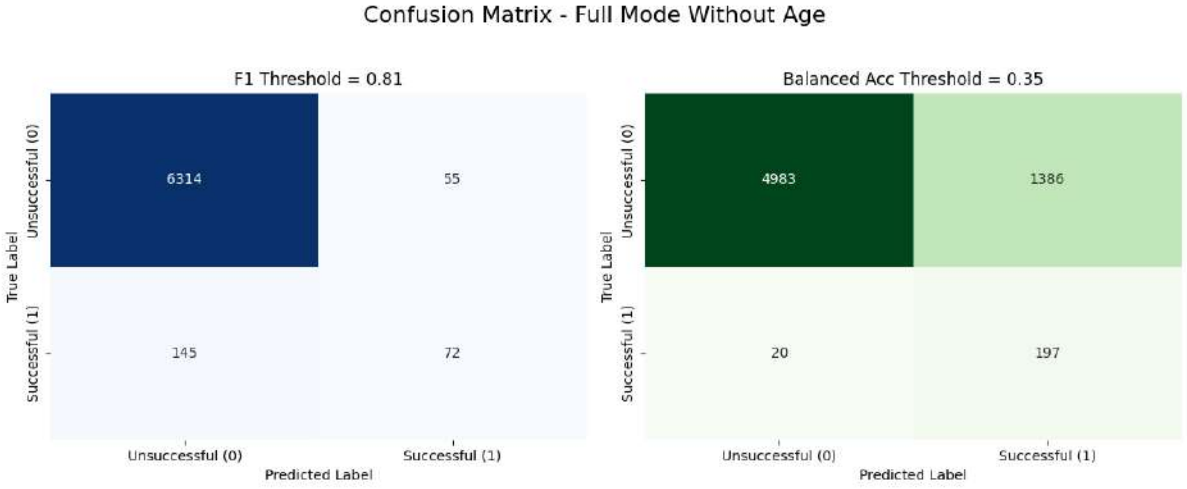


Figure 58: XGBoost Without Age – Full Mode

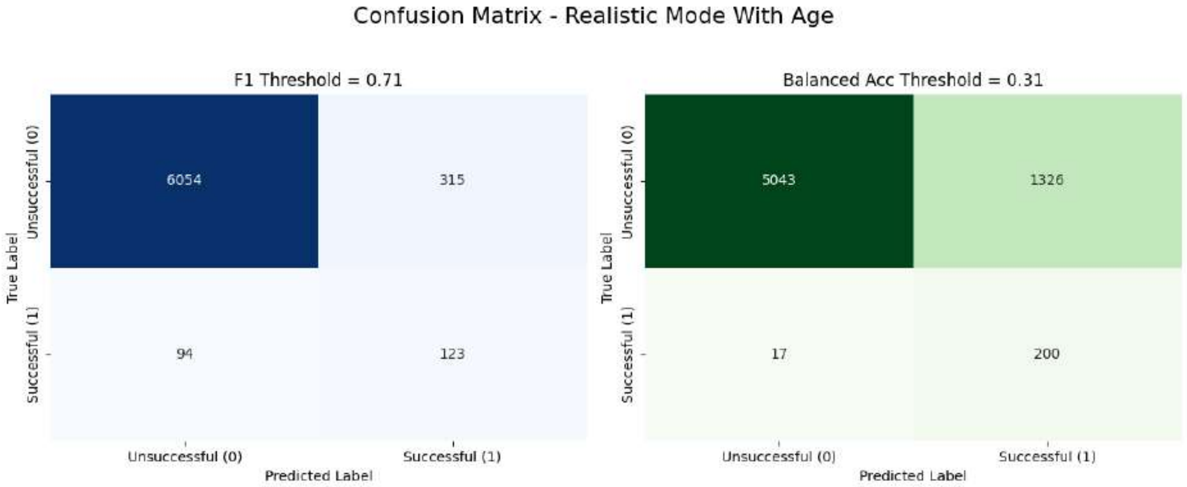


Figure 59: XGBoost With Age – Realistic Mode

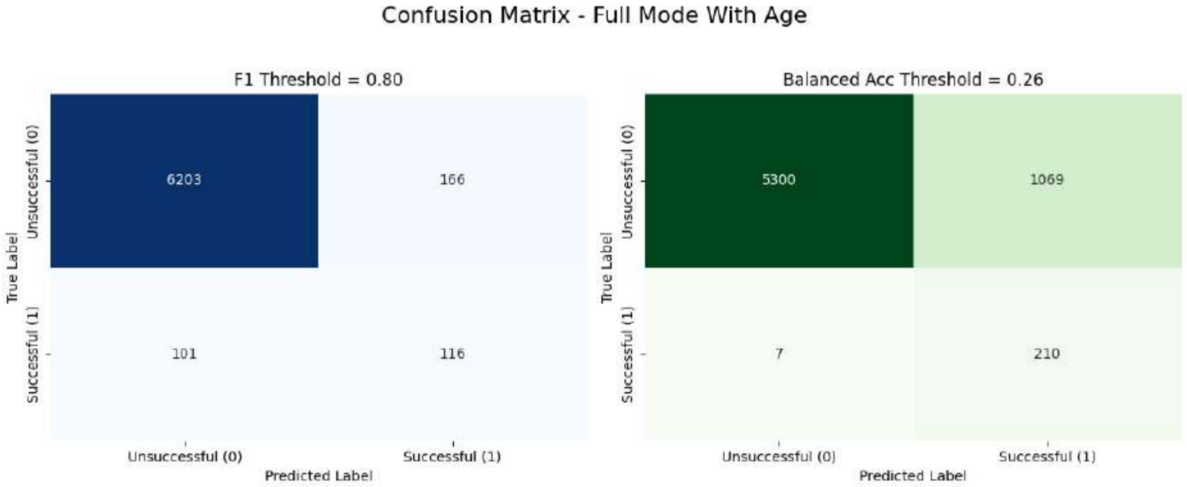


Figure 60: XGBoost With Age – Full Mode

Threshold–Metric Profiles.

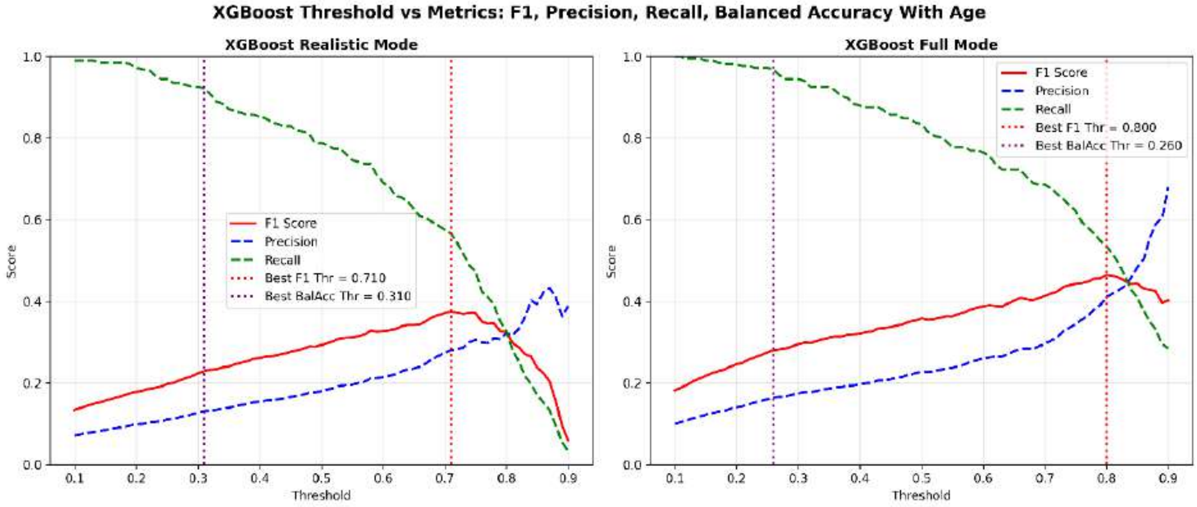


Figure 61: XGBoost With Age Threshold–Metric

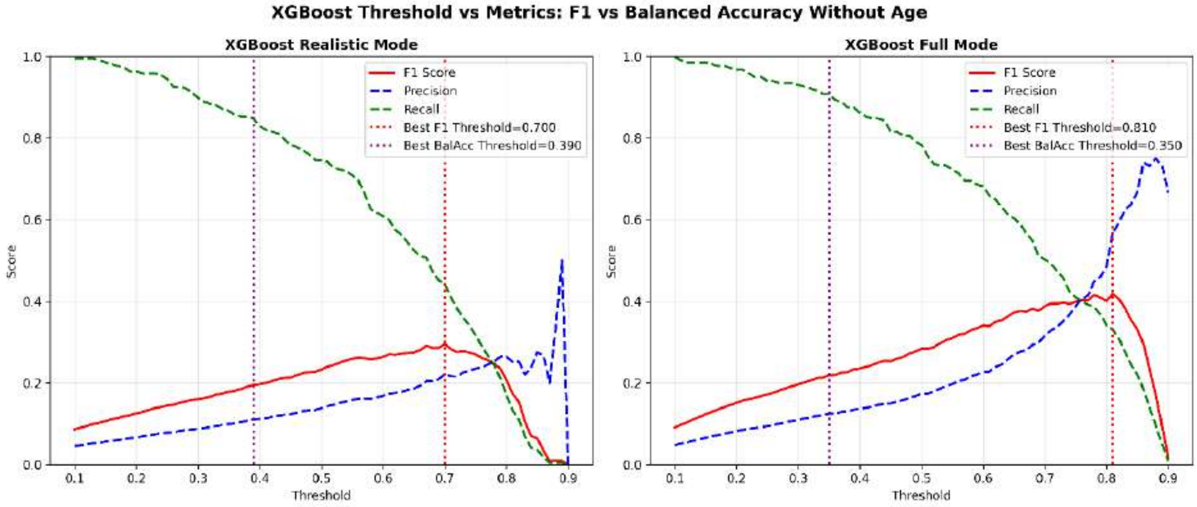


Figure 62: XGBoost Without Age Threshold–Metric

Configuration Comparison Plots

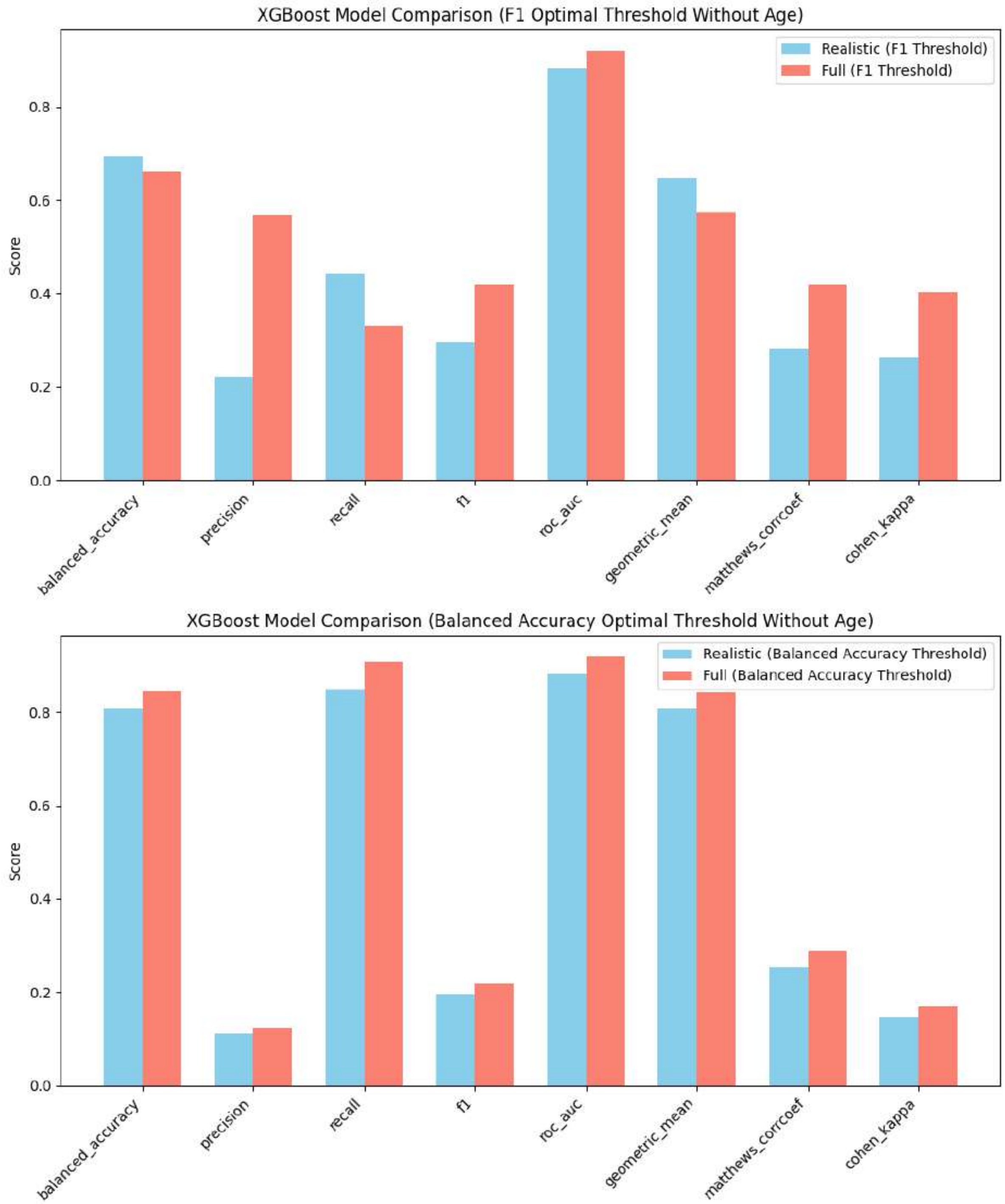


Figure 63: XGBoost model comparison without Age

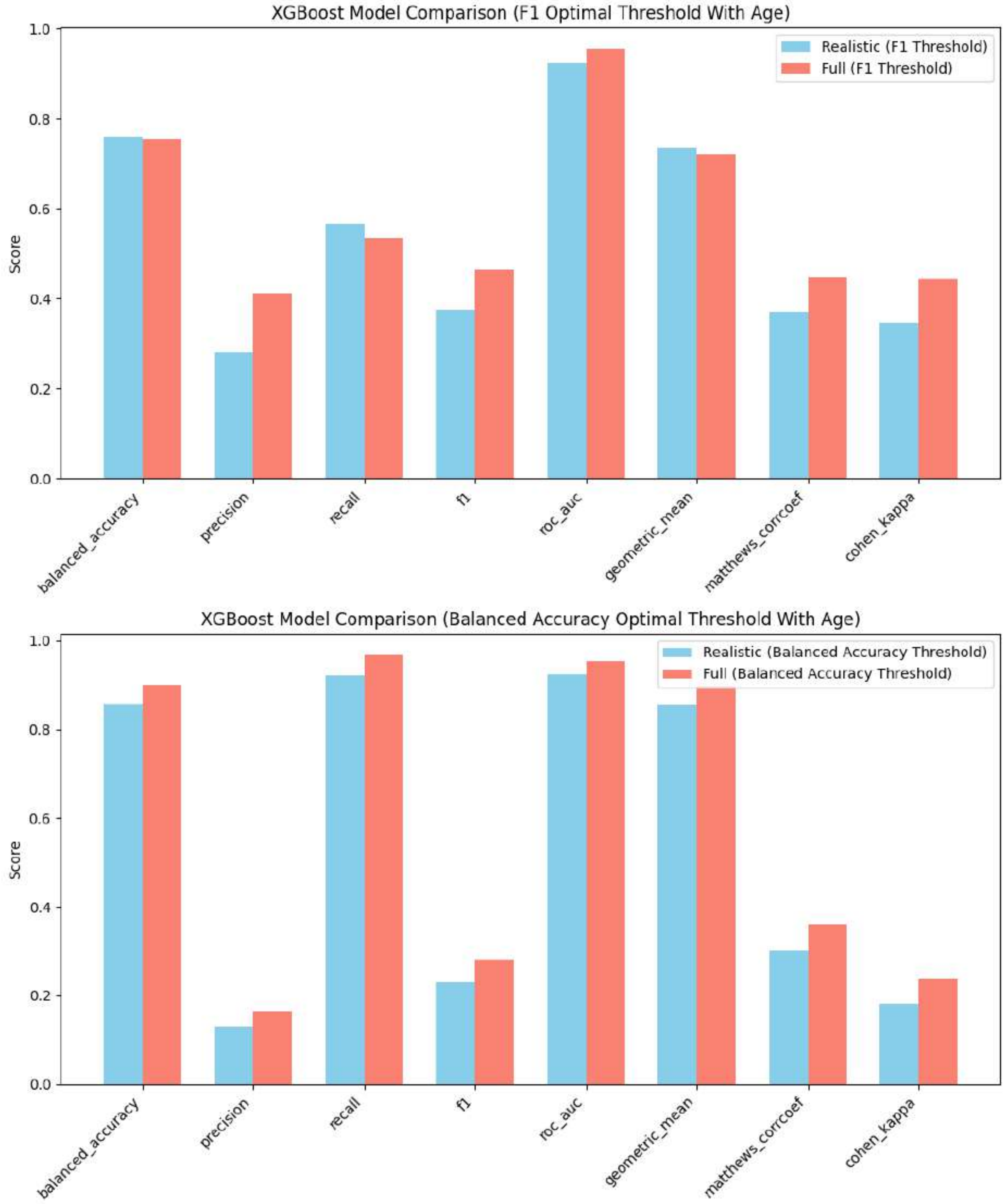


Figure 64: XGBoost model comparison with Age

Precision–Recall and ROC Curves

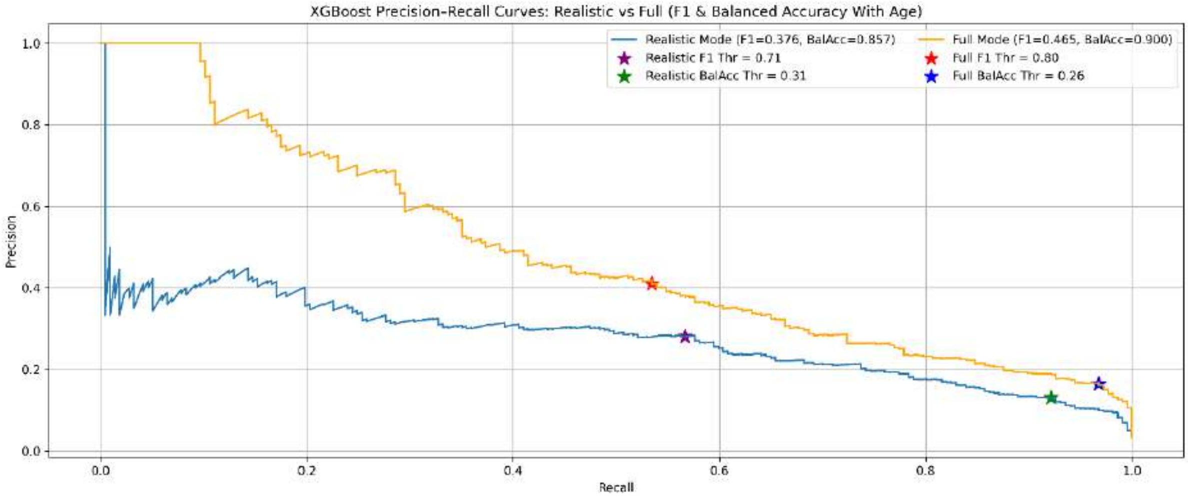


Figure 65: XGBoost Precision Recall Curve With Age

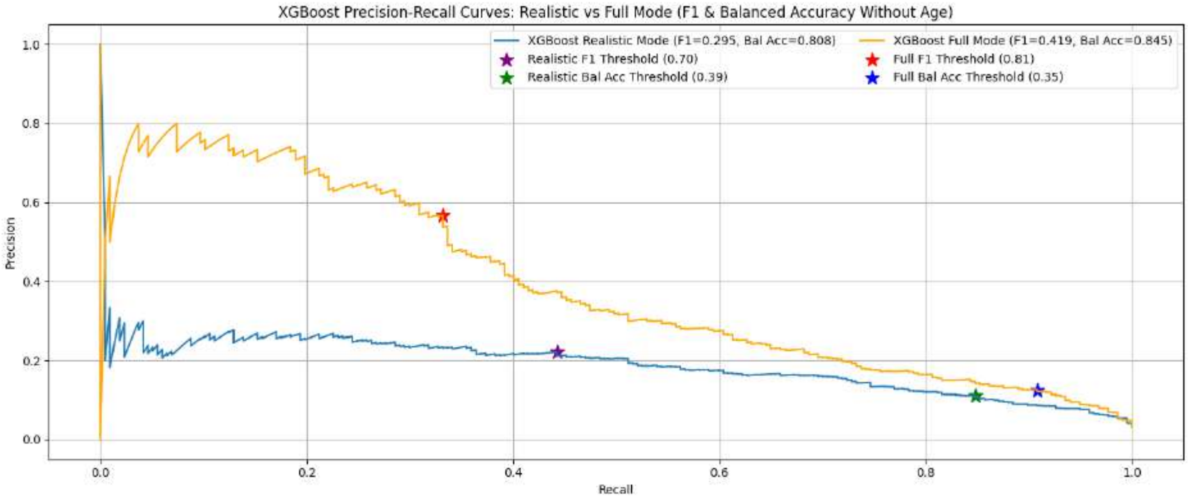


Figure 66: XGBoost Precision Recall Curve Without Age

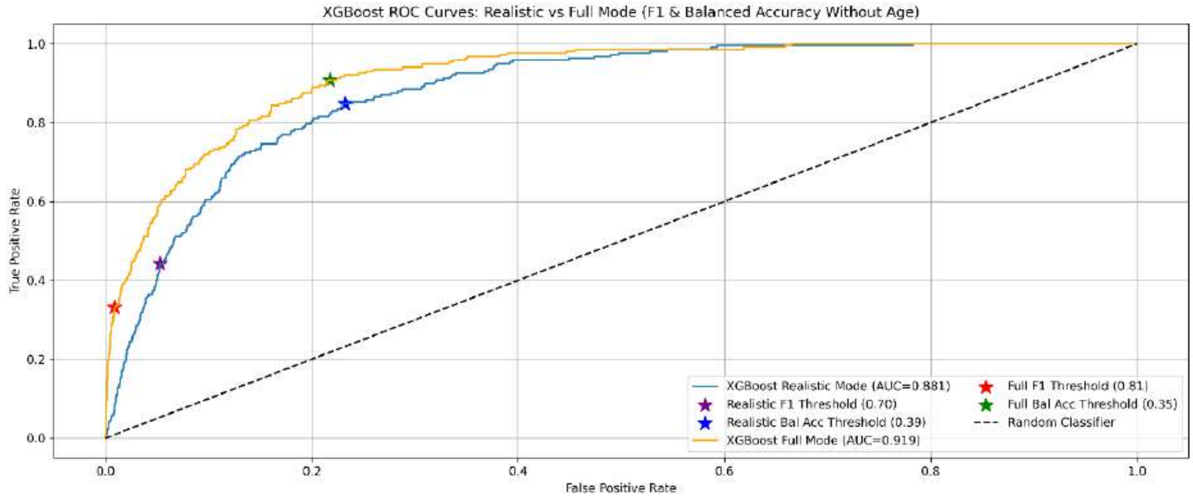


Figure 67: XGBoost ROC-Curve Without Age

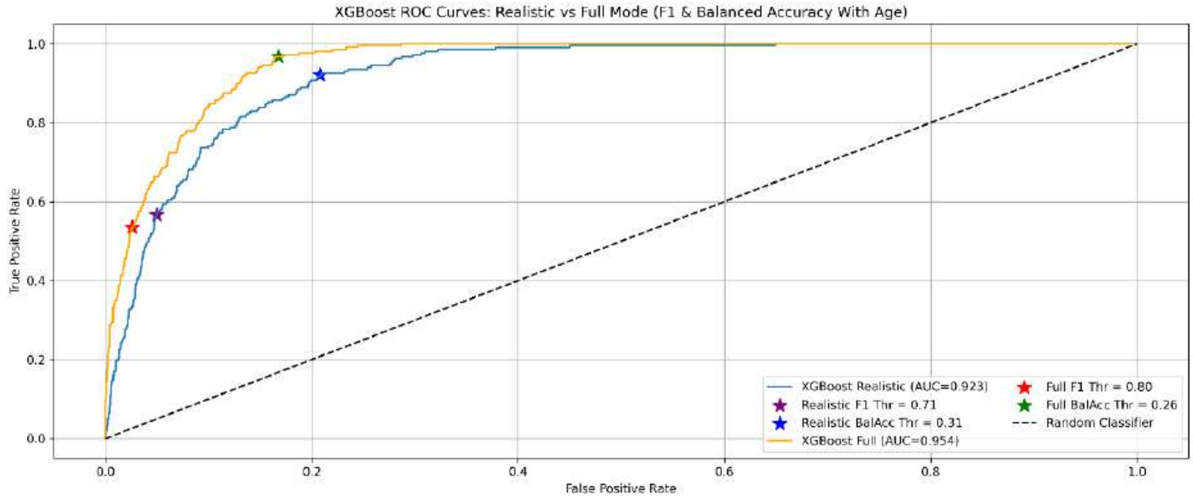


Figure 68: XGBoost ROC-Curve With Age

Exploratory Cluster Analysis (Supplementary Insight)

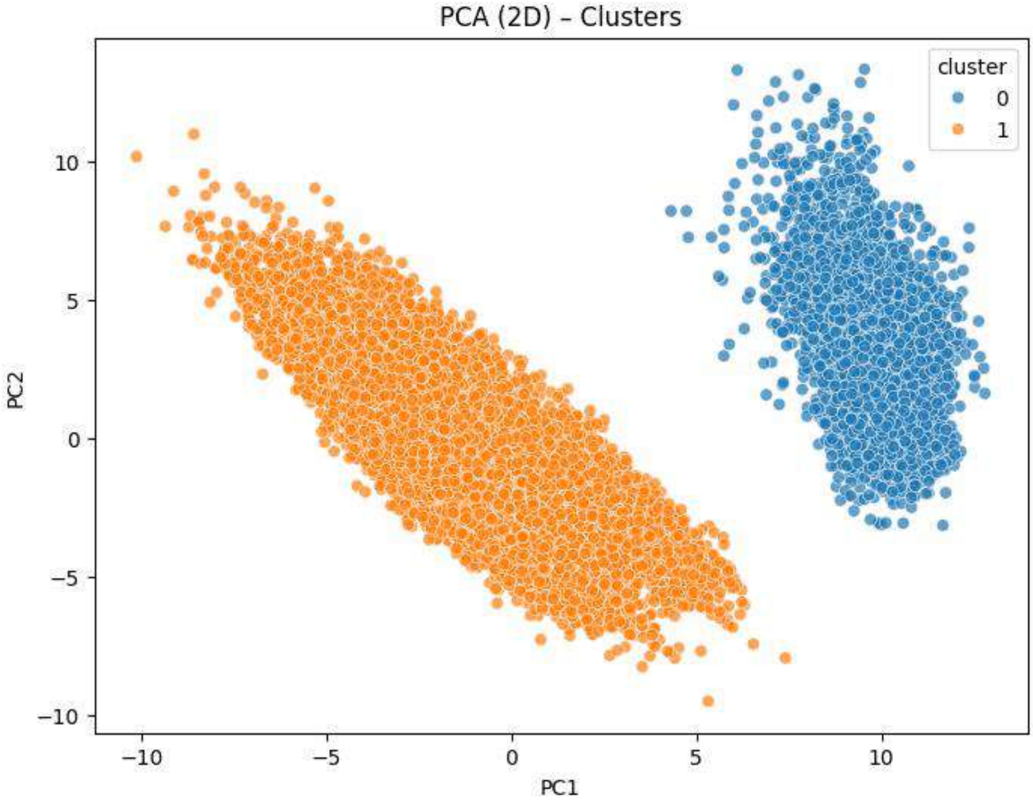


Figure 87: PCA Projection of Visible-Attribute Clusters.

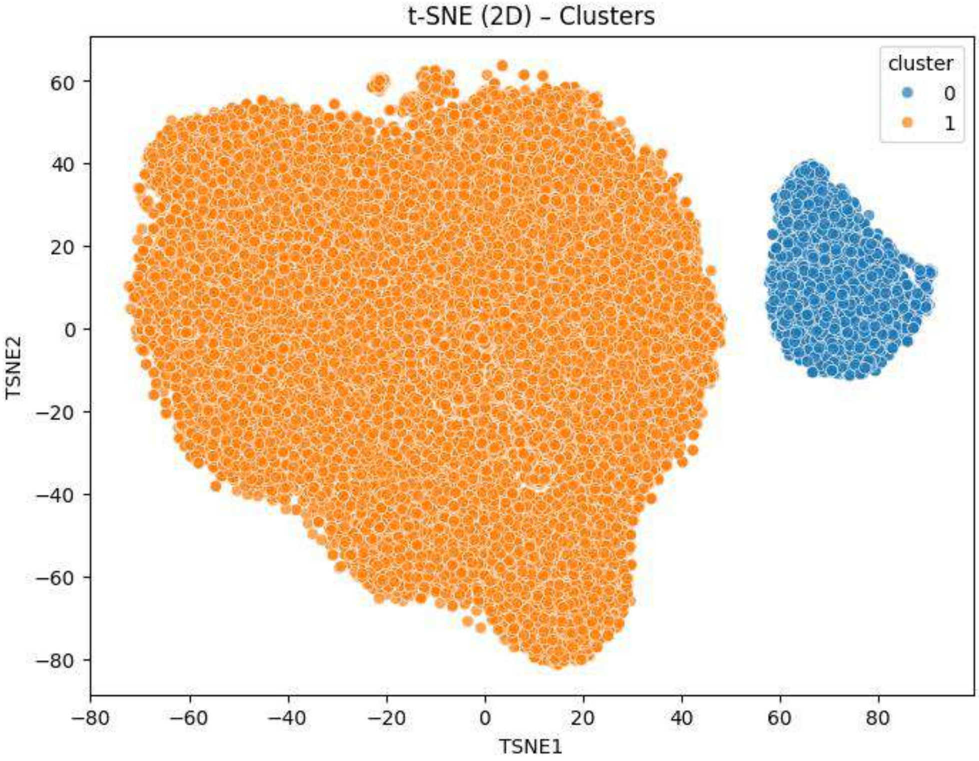


Figure 88: t-SNE Visualization of Visible-Attribute Clusters.

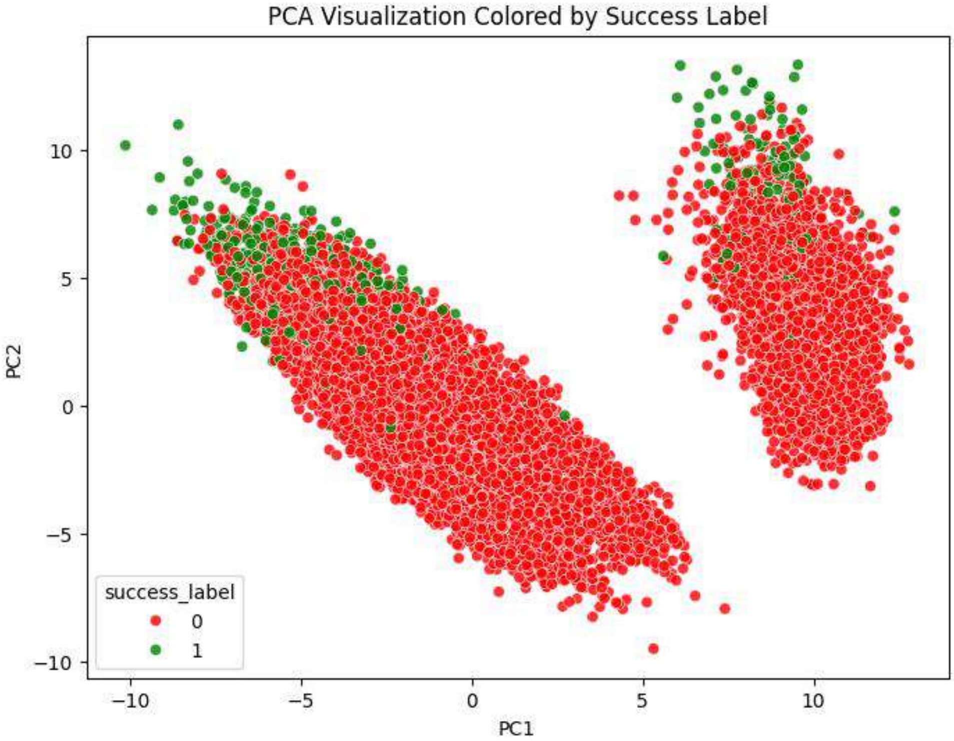


Figure 89: PCA Projection Colored by Success Label.

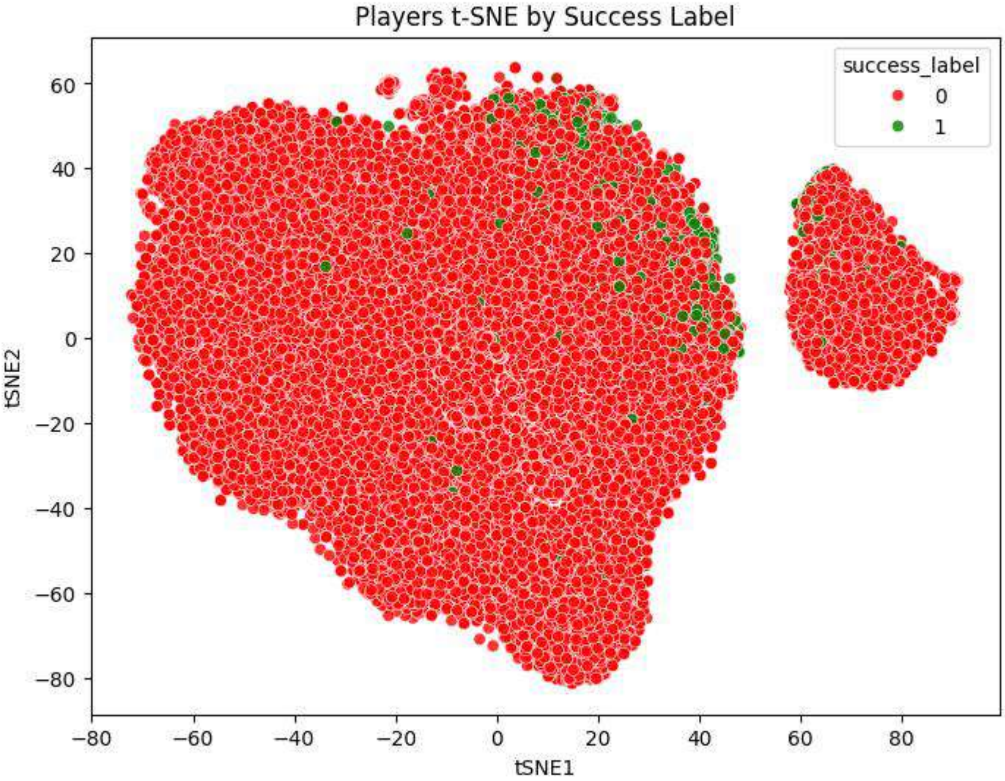


Figure 90: t-SNE Projection Colored by Success Label.

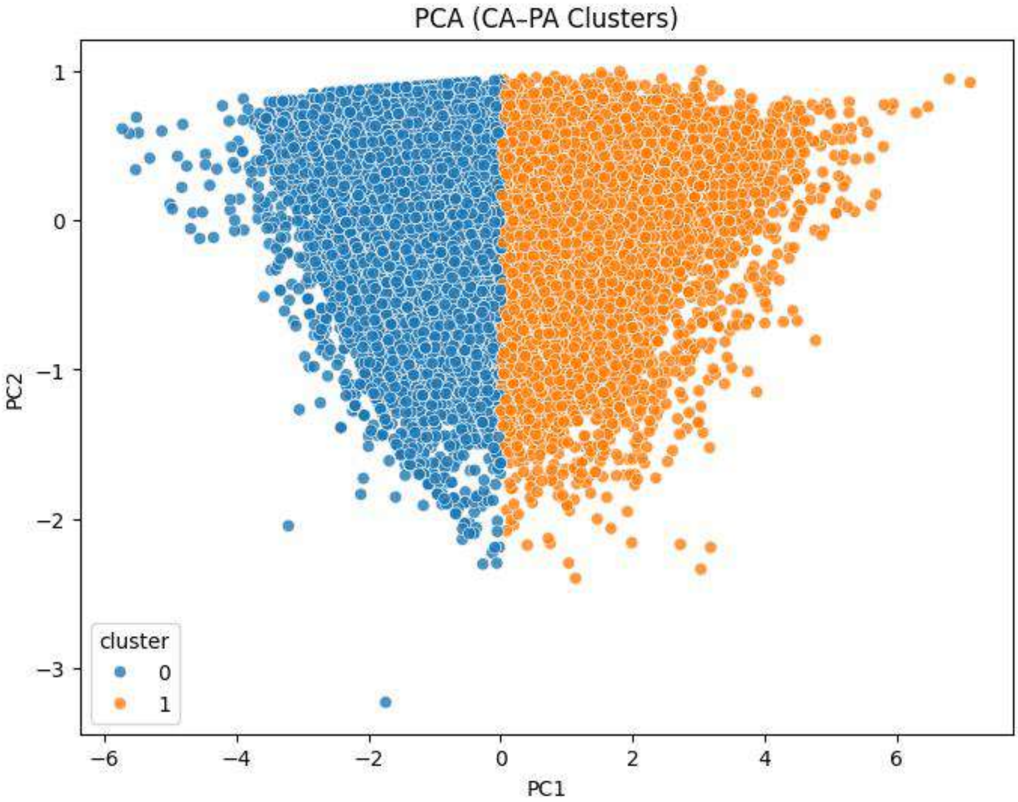


Figure 91: PCA Projection Colored by Cluster Assignment (Cluster 0 vs Cluster 1)

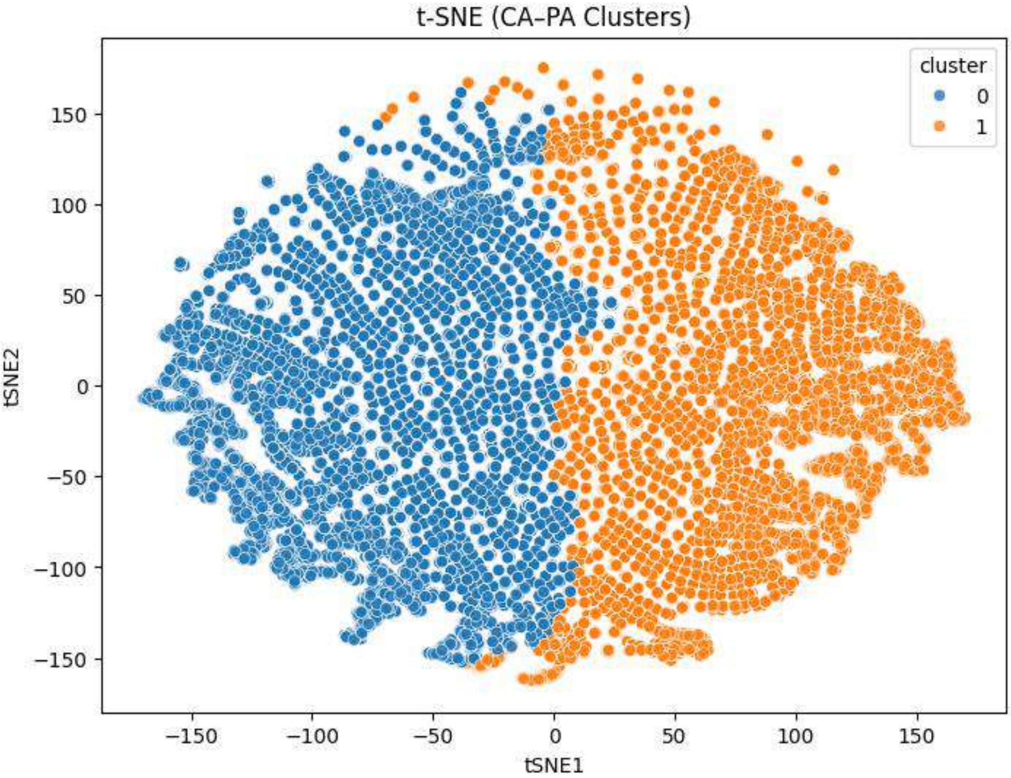


Figure 92: t-SNE Projection Colored by Cluster Assignment (Cluster 0 vs Cluster 1)

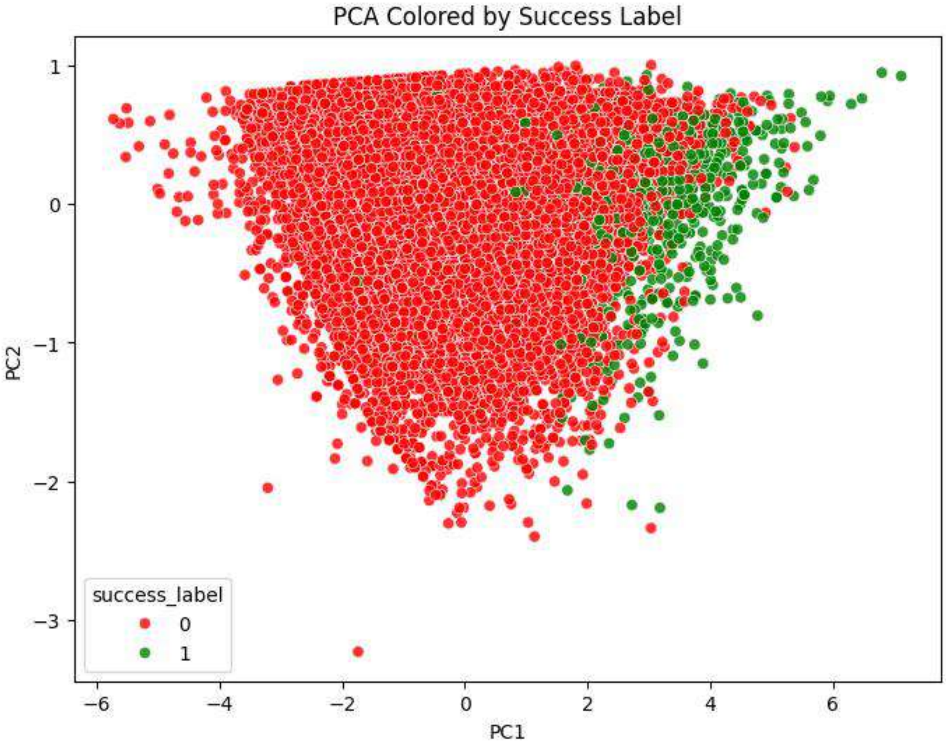


Figure 93: PCA Projection Colored by Success Label

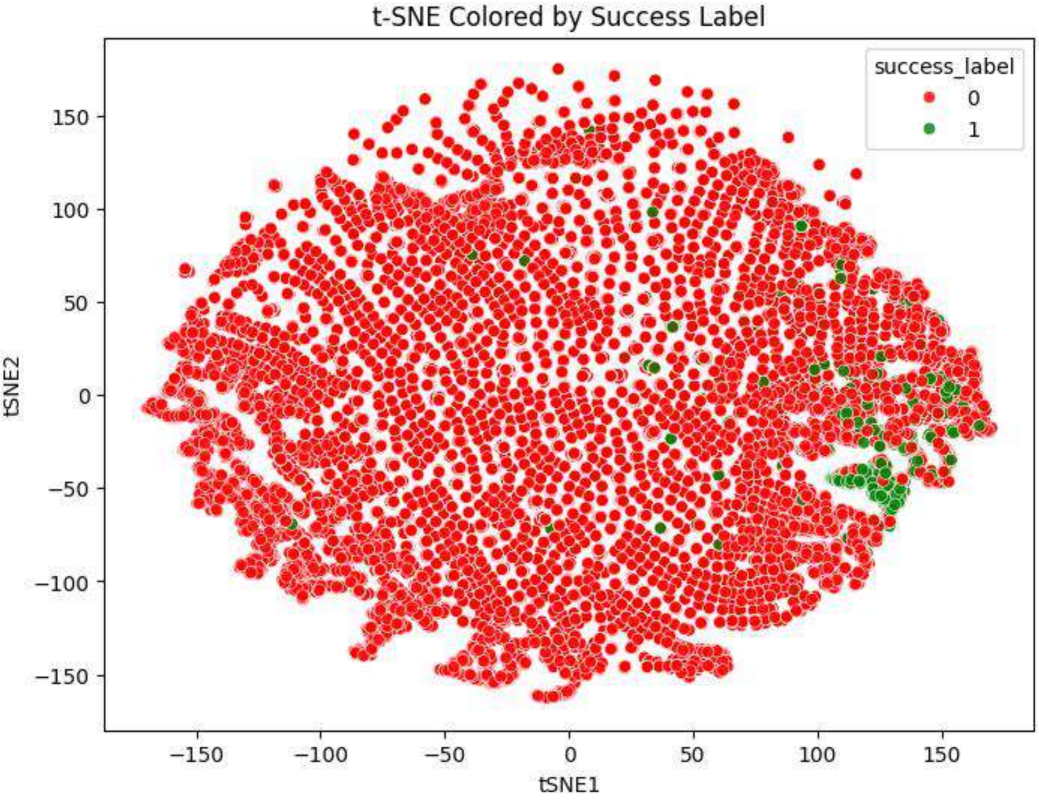


Figure 94: t-SNE Projection Colored by Success Label

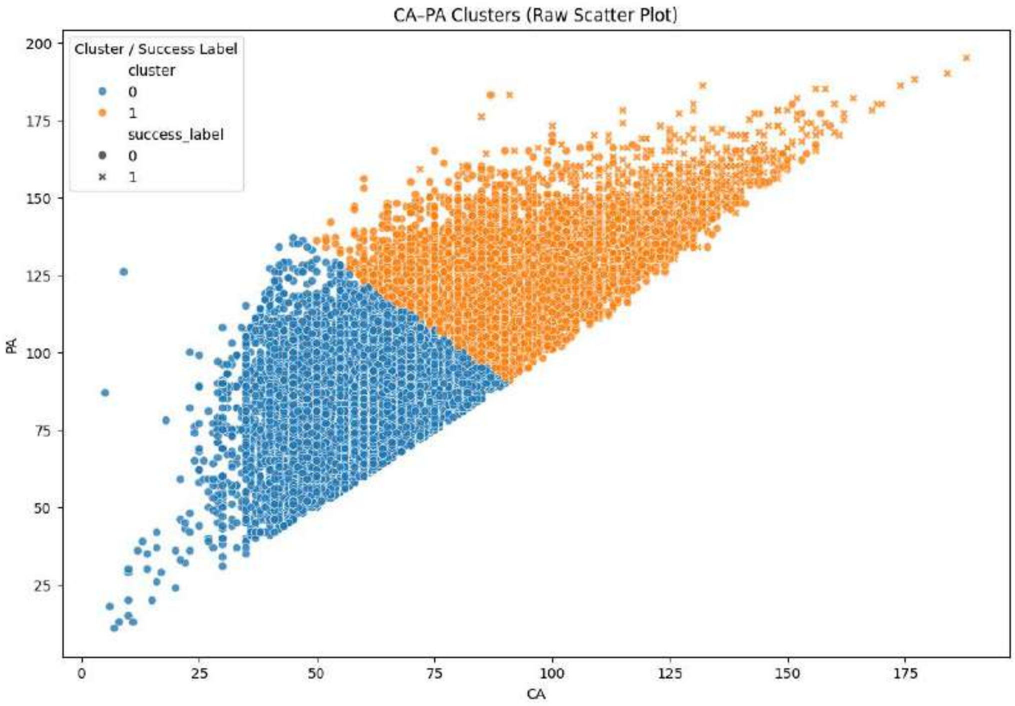


Figure 95: CA–PA Clusters (Raw Scatter Plot)

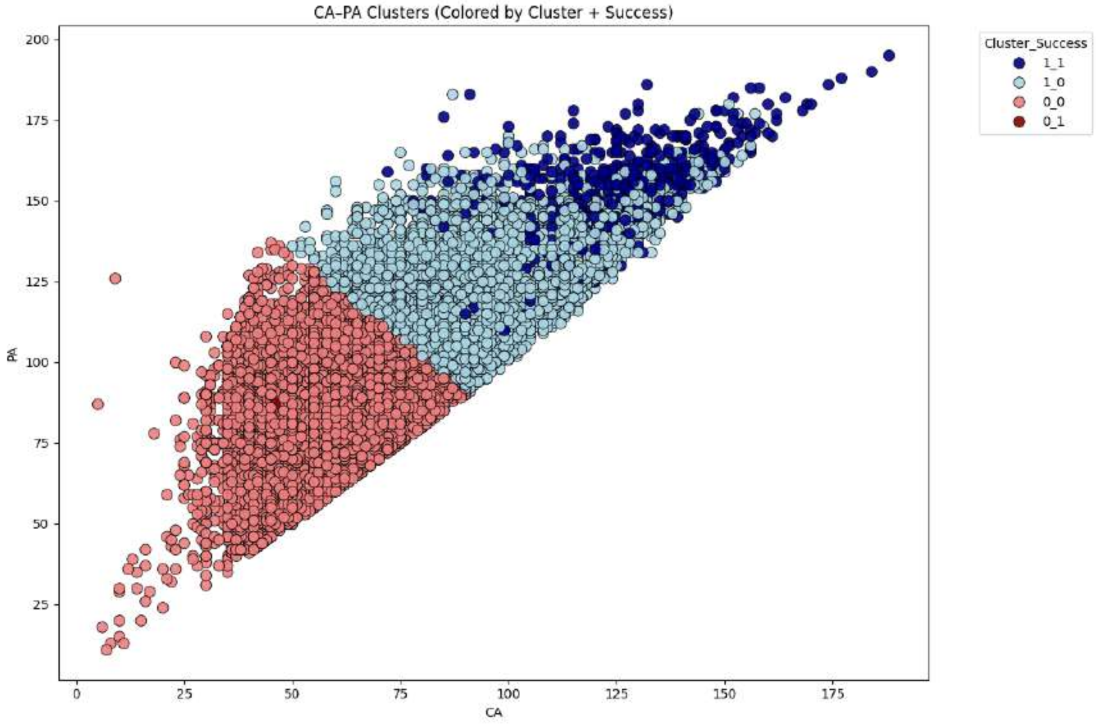


Figure 96: CA–PA Clusters Colored by Cluster + Success Label

- 1_1 (dark blue) = Cluster 1 & Successful
- 1_0 (light blue) = Cluster 1 & Not Successful
- 0_0 (light red) = Cluster 0 & Not Successful
- 0_1 (dark red) = Cluster 0 & Successful

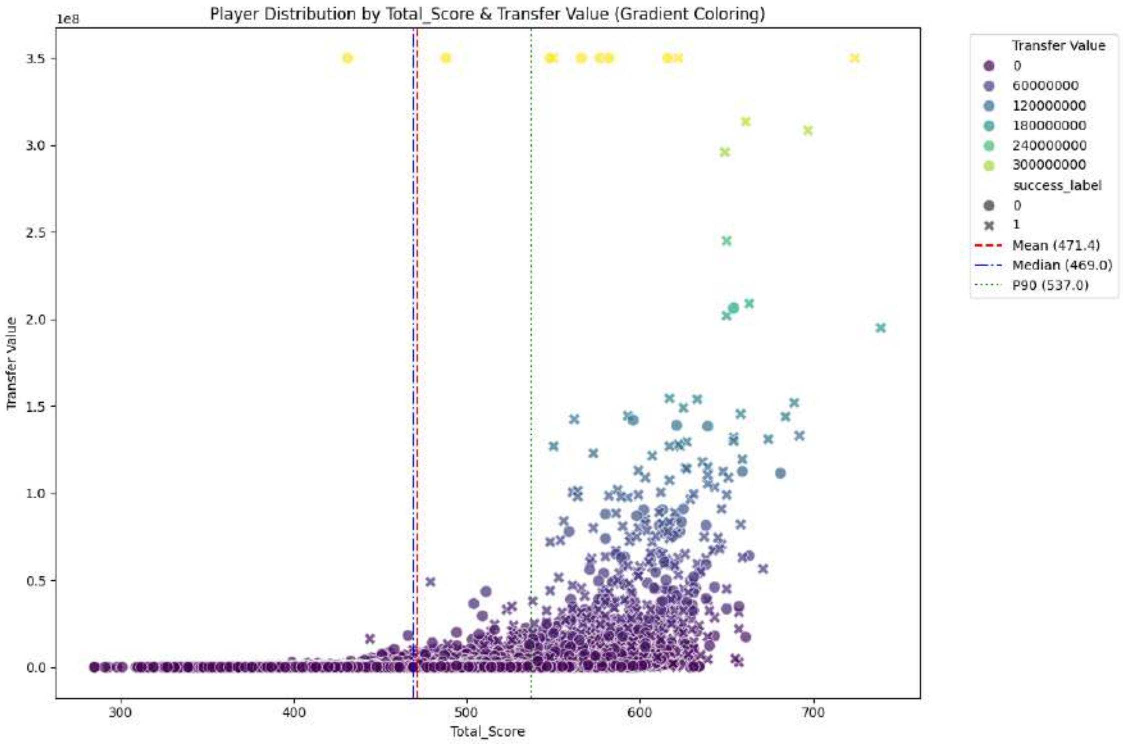


Figure 97: Player Distribution by Total Score and Transfer Value

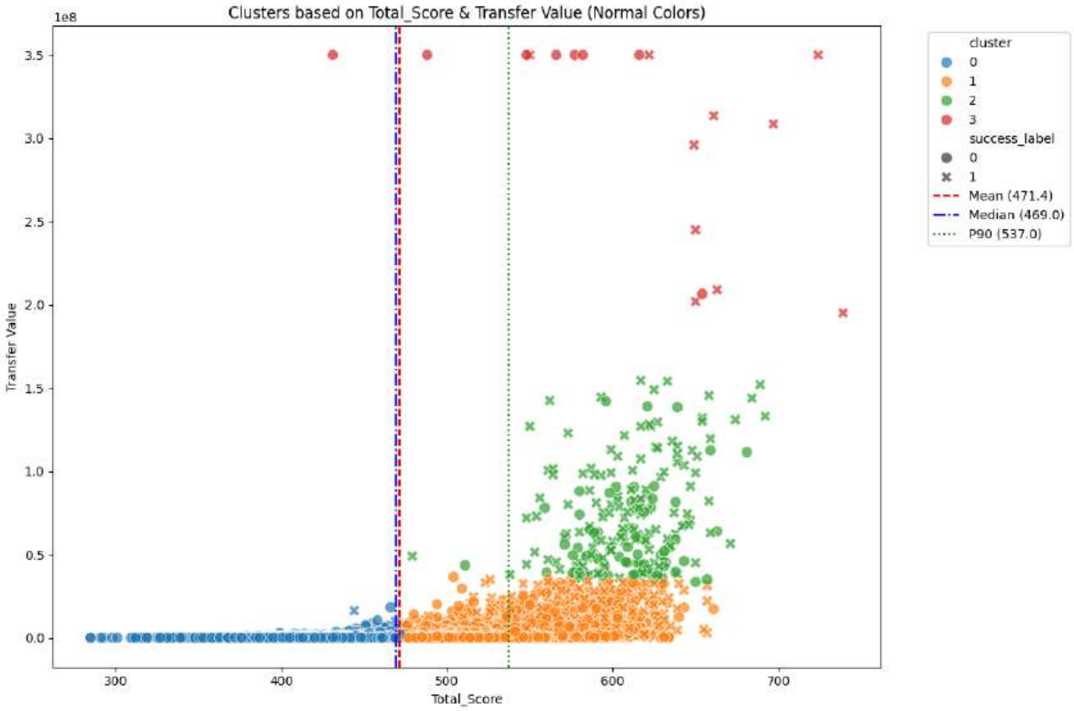
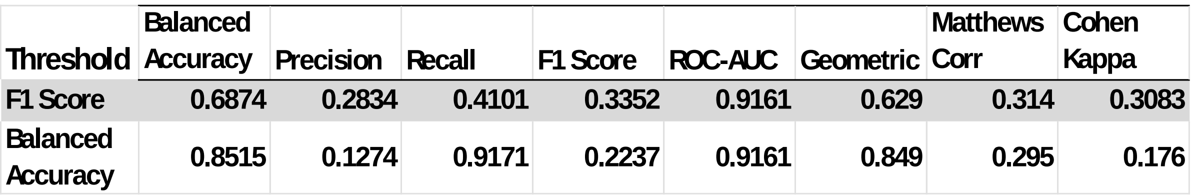


Figure 98: Clusters Based on Total Score and Transfer Value (Normal Colors)

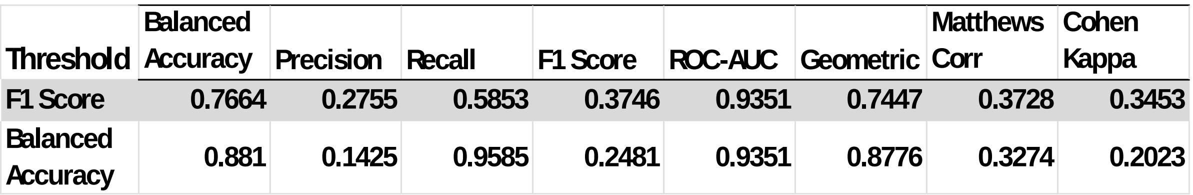
Appendix 4: Tables

Logistic Regression



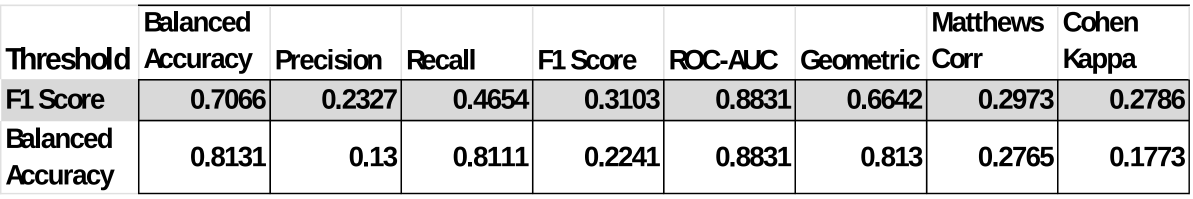
Threshold	Balanced Accuracy	Precision	Recall	F1 Score	ROC-AUC	Geometric	Matthews Corr	Cohen Kappa
F1 Score	0.6874	0.2834	0.4101	0.3352	0.9161	0.629	0.314	0.3083
Balanced Accuracy	0.8515	0.1274	0.9171	0.2237	0.9161	0.849	0.295	0.176

Table 3: LR With Age – Realistic Mode Evaluation Metric



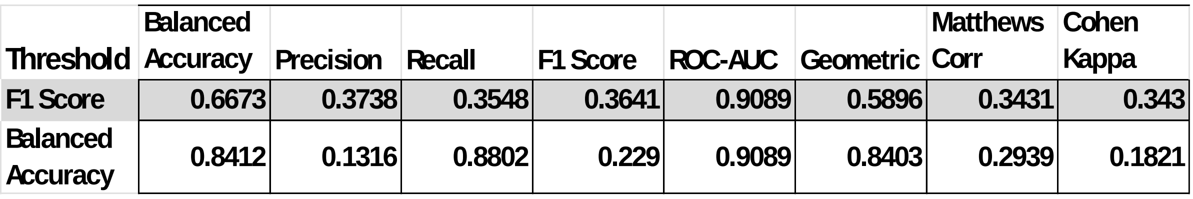
Threshold	Balanced Accuracy	Precision	Recall	F1 Score	ROC-AUC	Geometric	Matthews Corr	Cohen Kappa
F1 Score	0.7664	0.2755	0.5853	0.3746	0.9351	0.7447	0.3728	0.3453
Balanced Accuracy	0.881	0.1425	0.9585	0.2481	0.9351	0.8776	0.3274	0.2023

Table 4: LR With Age – Full Mode Evaluation Metric



Threshold	Balanced Accuracy	Precision	Recall	F1 Score	ROC-AUC	Geometric	Matthews Corr	Cohen Kappa
F1 Score	0.7066	0.2327	0.4654	0.3103	0.8831	0.6642	0.2973	0.2786
Balanced Accuracy	0.8131	0.13	0.8111	0.2241	0.8831	0.813	0.2765	0.1773

Table 5: LR Without Age – Realistic Mode Evaluation Metric



Threshold	Balanced Accuracy	Precision	Recall	F1 Score	ROC-AUC	Geometric	Matthews Corr	Cohen Kappa
F1 Score	0.6673	0.3738	0.3548	0.3641	0.9089	0.5896	0.3431	0.343
Balanced Accuracy	0.8412	0.1316	0.8802	0.229	0.9089	0.8403	0.2939	0.1821

Table 6: LR Without Age – Full Mode Evaluation Metric

RANODM FORSET

Threshold	Balanced Accuracy	Precision	Recall	F1 Score	ROC-AUC	Geometric	Matthews Corr	Cohen Kappa
F1 Score	0.7595	0.2881	0.5668	0.382	0.9323	0.7347	0.3763	0.3538
Balanced Accuracy	0.8731	0.1546	0.9171	0.2646	0.9323	0.872	0.3359	0.2207

Table 7: RF With Age – Realistic Mode Evaluation Metric

Threshold	Balanced Accuracy	Precision	Recall	F1 Score	ROC-AUC	Geometric	Matthews Corr	Cohen Kappa
F1 Score	0.7916	0.3525	0.6221	0.45	0.9495	0.7732	0.4448	0.4258
Balanced Accuracy	0.8872	0.1589	0.9447	0.2721	0.9495	0.8853	0.3483	0.2285

Table 8: RF With Age – Full Mode Evaluation Metric

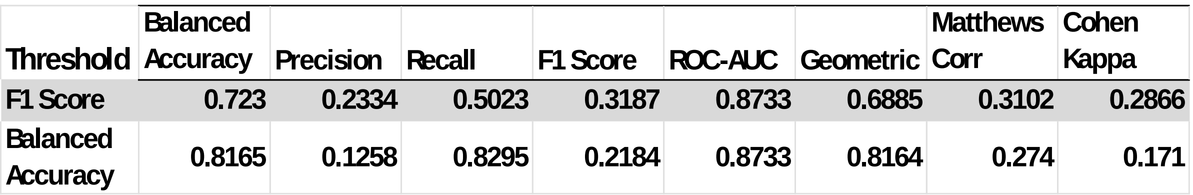
Threshold	Balanced Accuracy	Precision	Recall	F1 Score	ROC-AUC	Geometric	Matthews Corr	Cohen Kappa
F1 Score	0.7098	0.2269	0.4747	0.307	0.8889	0.6697	0.2956	0.2747
Balanced Accuracy	0.8093	0.1344	0.7926	0.2298	0.8889	0.8092	0.2791	0.1838

Table 9: RF Without Age – Realistic Mode Evaluation Metric

Threshold	Balanced Accuracy	Precision	Recall	F1 Score	ROC-AUC	Geometric	Matthews Corr	Cohen Kappa
F1 Score	0.6928	0.3633	0.4101	0.3853	0.9132	0.6325	0.3637	0.363
Balanced Accuracy	0.8351	0.1508	0.8295	0.2551	0.9132	0.8351	0.3106	0.2111

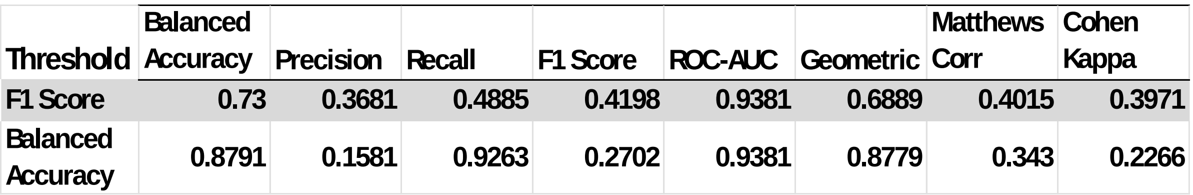
Table 10: RF Without Age – Full Mode Evaluation Metric

Decision Tree



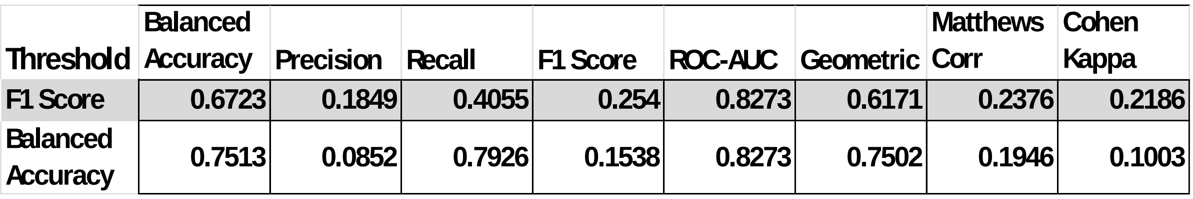
Threshold	Balanced Accuracy	Precision	Recall	F1 Score	ROC-AUC	Geometric	Matthews Corr	Cohen Kappa
F1 Score	0.723	0.2334	0.5023	0.3187	0.8733	0.6885	0.3102	0.2866
Balanced Accuracy	0.8165	0.1258	0.8295	0.2184	0.8733	0.8164	0.274	0.171

Table 11: DT With Age – Realistic Mode Evaluation Metric



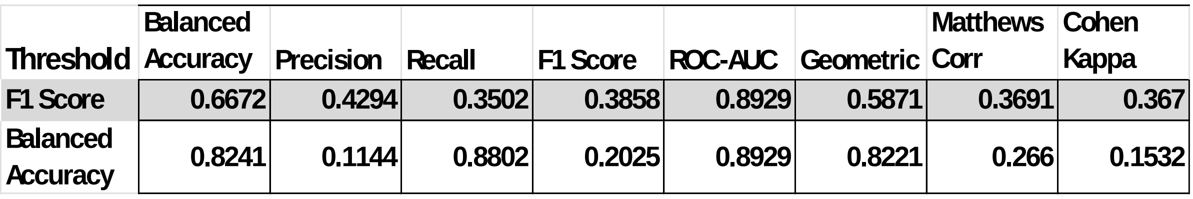
Threshold	Balanced Accuracy	Precision	Recall	F1 Score	ROC-AUC	Geometric	Matthews Corr	Cohen Kappa
F1 Score	0.73	0.3681	0.4885	0.4198	0.9381	0.6889	0.4015	0.3971
Balanced Accuracy	0.8791	0.1581	0.9263	0.2702	0.9381	0.8779	0.343	0.2266

Table 12: DT With Age – Full Mode Evaluation Metric



Threshold	Balanced Accuracy	Precision	Recall	F1 Score	ROC-AUC	Geometric	Matthews Corr	Cohen Kappa
F1 Score	0.6723	0.1849	0.4055	0.254	0.8273	0.6171	0.2376	0.2186
Balanced Accuracy	0.7513	0.0852	0.7926	0.1538	0.8273	0.7502	0.1946	0.1003

Table 13: DT Without Age – Realistic Mode Evaluation Metric



Threshold	Balanced Accuracy	Precision	Recall	F1 Score	ROC-AUC	Geometric	Matthews Corr	Cohen Kappa
F1 Score	0.6672	0.4294	0.3502	0.3858	0.8929	0.5871	0.3691	0.367
Balanced Accuracy	0.8241	0.1144	0.8802	0.2025	0.8929	0.8221	0.266	0.1532

Table 14: DT Without Age – Full Mode Evaluation Metric

SVC

Threshold	Balanced Accuracy	Precision	Recall	F1 Score	ROC-AUC	Geometric	Matthews Corr	Cohen Kappa
F1 Score	0.7443	0.0777	0.8203	0.1419	0.7833	0.7404	0.1831	0.087
Balanced Accuracy	0.7443	0.0777	0.8203	0.1419	0.7833	0.7404	0.1831	0.087

Table 15: SVC With Age – Realistic Mode Evaluation Metric

Threshold	Balanced Accuracy	Precision	Recall	F1 Score	ROC-AUC	Geometric	Matthews Corr	Cohen Kappa
F1 Score	0.5	0.0329	1	0.0638	0.8378	0	0	0
Balanced Accuracy	0.5	0.0329	1	0.0638	0.8378	0	0	0

Table 16: SVC With Age – Full Mode Evaluation Metric

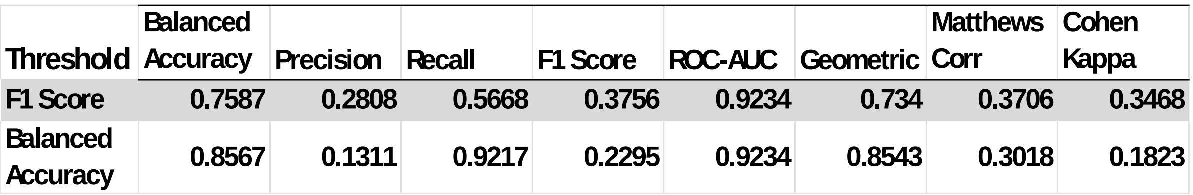
Threshold	Balanced Accuracy	Precision	Recall	F1 Score	ROC-AUC	Geometric	Matthews Corr	Cohen Kappa
F1 Score	0.7466	0.0967	0.7235	0.1706	0.7437	0.7462	0.2042	0.1194
Balanced Accuracy	0.7466	0.0967	0.7235	0.1706	0.7437	0.7462	0.2042	0.1194

Table 17: SVC Without Age – Realistic Mode Evaluation Metric

Threshold	Balanced Accuracy	Precision	Recall	F1 Score	ROC-AUC	Geometric	Matthews Corr	Cohen Kappa
F1 Score	0.7674	0.1121	0.7327	0.1944	0.8368	0.7667	0.2322	0.1455
Balanced Accuracy	0.7674	0.1121	0.7327	0.1944	0.8368	0.7667	0.2322	0.1455

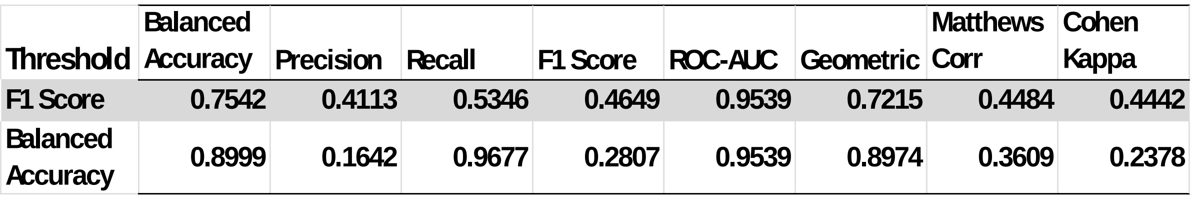
Table 18: SVC Without Age – Full Mode Evaluation Metric

XGBoost



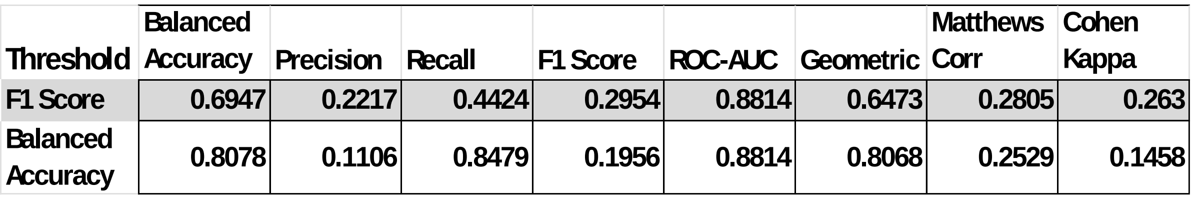
Threshold	Balanced Accuracy	Precision	Recall	F1 Score	ROC-AUC	Geometric	Matthews Corr	Cohen Kappa
F1 Score	0.7587	0.2808	0.5668	0.3756	0.9234	0.734	0.3706	0.3468
Balanced Accuracy	0.8567	0.1311	0.9217	0.2295	0.9234	0.8543	0.3018	0.1823

Table 19: With Age – Realistic Mode Evaluation Metric



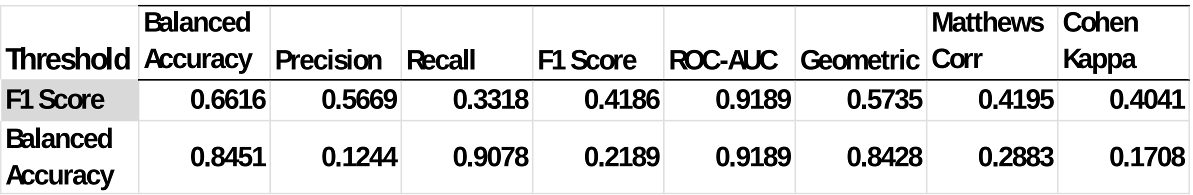
Threshold	Balanced Accuracy	Precision	Recall	F1 Score	ROC-AUC	Geometric	Matthews Corr	Cohen Kappa
F1 Score	0.7542	0.4113	0.5346	0.4649	0.9539	0.7215	0.4484	0.4442
Balanced Accuracy	0.8999	0.1642	0.9677	0.2807	0.9539	0.8974	0.3609	0.2378

Table 20: XGBoost With Age – Full Mode Evaluation Metric



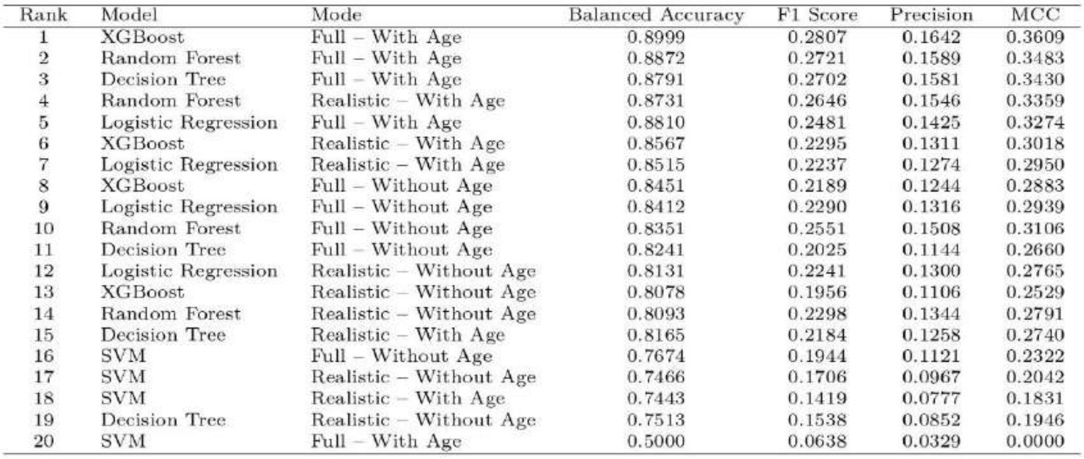
Threshold	Balanced Accuracy	Precision	Recall	F1 Score	ROC-AUC	Geometric	Matthews Corr	Cohen Kappa
F1 Score	0.6947	0.2217	0.4424	0.2954	0.8814	0.6473	0.2805	0.263
Balanced Accuracy	0.8078	0.1106	0.8479	0.1956	0.8814	0.8068	0.2529	0.1458

Table 21: XGBoost Without Age – Realistic Mode Evaluation Metric



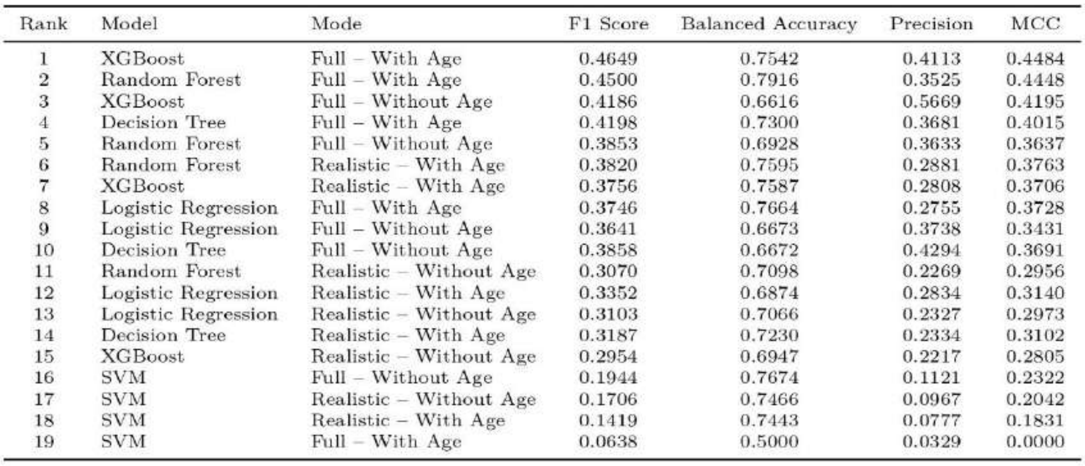
Threshold	Balanced Accuracy	Precision	Recall	F1 Score	ROC-AUC	Geometric	Matthews Corr	Cohen Kappa
F1 Score	0.6616	0.5669	0.3318	0.4186	0.9189	0.5735	0.4195	0.4041
Balanced Accuracy	0.8451	0.1244	0.9078	0.2189	0.9189	0.8428	0.2883	0.1708

Table 22: XGBoost Without Age – Full Mode Evaluation Metric



Rank	Model	Mode	Balanced Accuracy	F1 Score	Precision	MCC
1	XGBoost	Full – With Age	0.8999	0.2807	0.1642	0.3609
2	Random Forest	Full – With Age	0.8872	0.2721	0.1589	0.3483
3	Decision Tree	Full – With Age	0.8791	0.2702	0.1581	0.3430
4	Random Forest	Realistic – With Age	0.8731	0.2646	0.1546	0.3359
5	Logistic Regression	Full – With Age	0.8810	0.2481	0.1425	0.3274
6	XGBoost	Realistic – With Age	0.8567	0.2295	0.1311	0.3018
7	Logistic Regression	Realistic – With Age	0.8515	0.2237	0.1274	0.2950
8	XGBoost	Full – Without Age	0.8451	0.2189	0.1244	0.2883
9	Logistic Regression	Full – Without Age	0.8412	0.2290	0.1316	0.2939
10	Random Forest	Full – Without Age	0.8351	0.2551	0.1508	0.3106
11	Decision Tree	Full – Without Age	0.8241	0.2025	0.1144	0.2660
12	Logistic Regression	Realistic – Without Age	0.8131	0.2241	0.1300	0.2765
13	XGBoost	Realistic – Without Age	0.8078	0.1956	0.1106	0.2529
14	Random Forest	Realistic – Without Age	0.8093	0.2298	0.1344	0.2791
15	Decision Tree	Realistic – With Age	0.8165	0.2184	0.1258	0.2740
16	SVM	Full – Without Age	0.7674	0.1944	0.1121	0.2322
17	SVM	Realistic – Without Age	0.7466	0.1706	0.0967	0.2042
18	SVM	Realistic – With Age	0.7443	0.1419	0.0777	0.1831
19	Decision Tree	Realistic – Without Age	0.7513	0.1538	0.0852	0.1946
20	SVM	Full – With Age	0.5000	0.0638	0.0329	0.0000

Table 23: Full Ranking of All Models — Balanced-Accuracy-Optimized Threshold
Sorted by Balanced Accuracy (highest to lowest)



Rank	Model	Mode	F1 Score	Balanced Accuracy	Precision	MCC
1	XGBoost	Full – With Age	0.4649	0.7542	0.4113	0.4484
2	Random Forest	Full – With Age	0.4500	0.7916	0.3525	0.4448
3	XGBoost	Full – Without Age	0.4186	0.6616	0.5669	0.4195
4	Decision Tree	Full – With Age	0.4198	0.7300	0.3681	0.4015
5	Random Forest	Full – Without Age	0.3853	0.6928	0.3633	0.3637
6	Random Forest	Realistic – With Age	0.3820	0.7595	0.2881	0.3763
7	XGBoost	Realistic – With Age	0.3756	0.7587	0.2808	0.3706
8	Logistic Regression	Full – With Age	0.3746	0.7664	0.2755	0.3728
9	Logistic Regression	Full – Without Age	0.3641	0.6673	0.3738	0.3431
10	Decision Tree	Full – Without Age	0.3858	0.6672	0.4294	0.3691
11	Random Forest	Realistic – Without Age	0.3070	0.7098	0.2269	0.2956
12	Logistic Regression	Realistic – With Age	0.3352	0.6874	0.2834	0.3140
13	Logistic Regression	Realistic – Without Age	0.3103	0.7066	0.2327	0.2973
14	Decision Tree	Realistic – With Age	0.3187	0.7230	0.2334	0.3102
15	XGBoost	Realistic – Without Age	0.2954	0.6947	0.2217	0.2805
16	SVM	Full – Without Age	0.1944	0.7674	0.1121	0.2322
17	SVM	Realistic – Without Age	0.1706	0.7466	0.0967	0.2042
18	SVM	Realistic – With Age	0.1419	0.7443	0.0777	0.1831
19	SVM	Full – With Age	0.0638	0.5000	0.0329	0.0000

Table 24: Full Ranking of All Models — F1-Optimized Threshold
Sorted by F1 Score (highest to lowest)

Appendix 5:Feature Importance

Logistic Regression (LR) — Coefficient-Based Importance

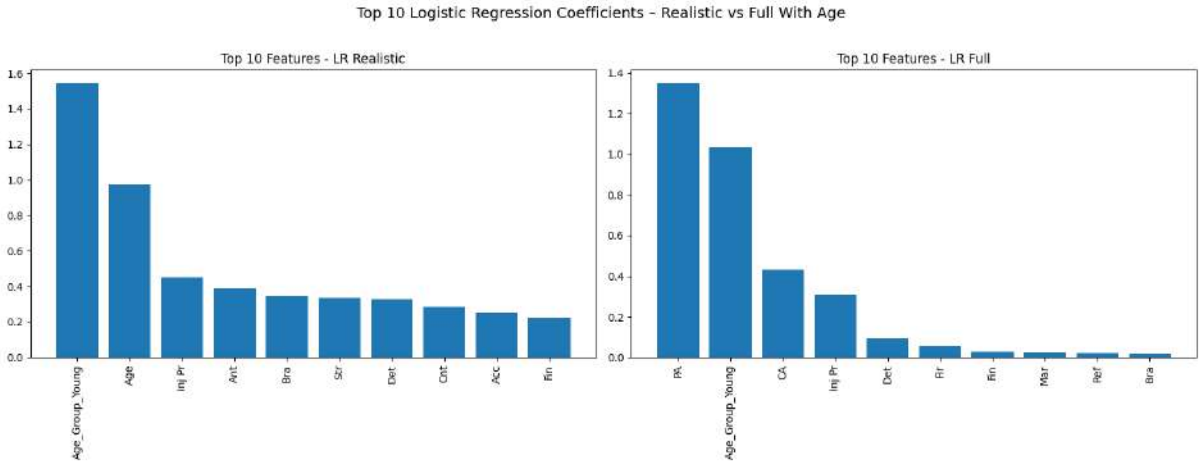


Figure 69: Top 10 Logistic Regression Feature Coefficients (Realistic Mode, With Age)

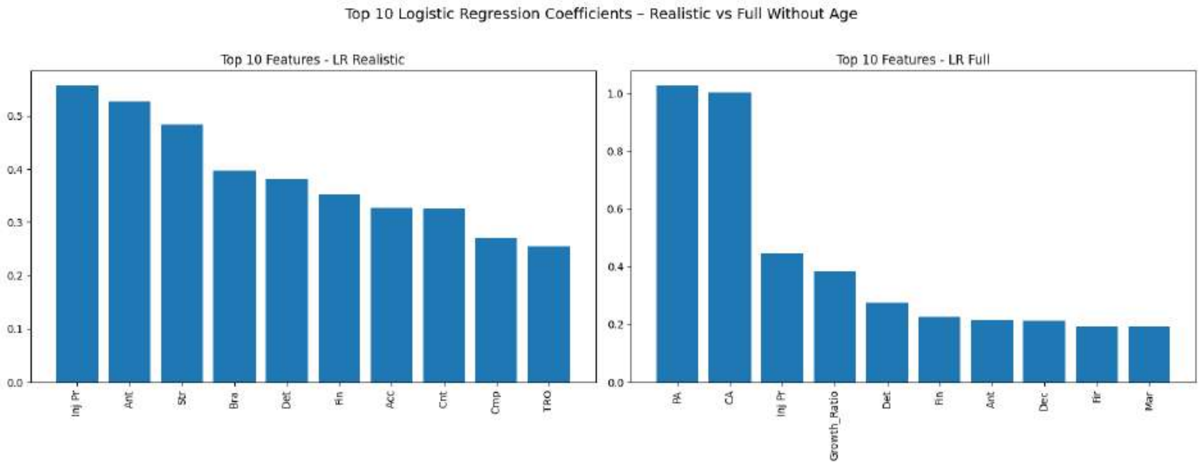


Figure 70: Top 10 Logistic Regression Feature Coefficients (Realistic Mode, Without Age)

Random Forest (RF) — Feature Importance (Impurity-Based)

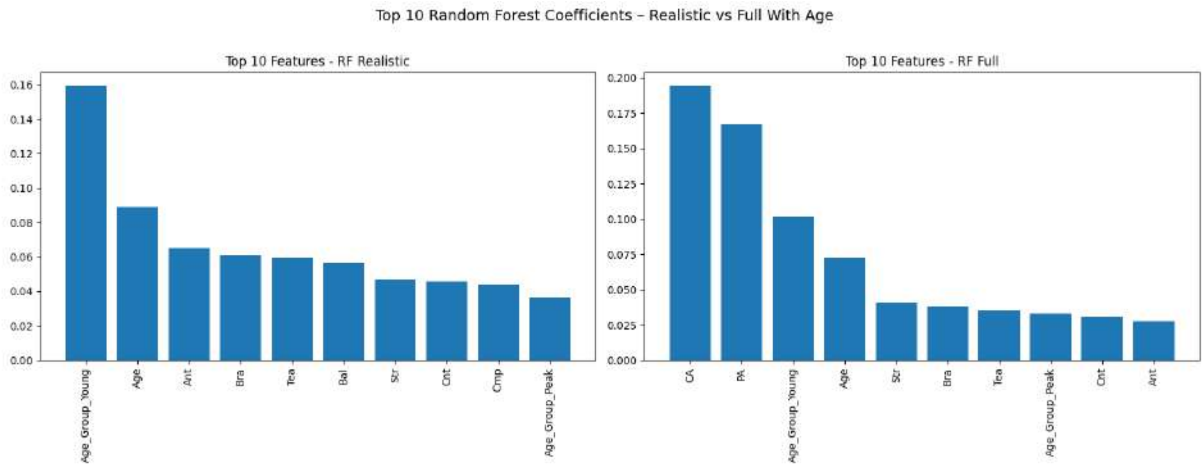


Figure 71: Top 10 Random Forest Feature Importances (Realistic Mode, With Age)

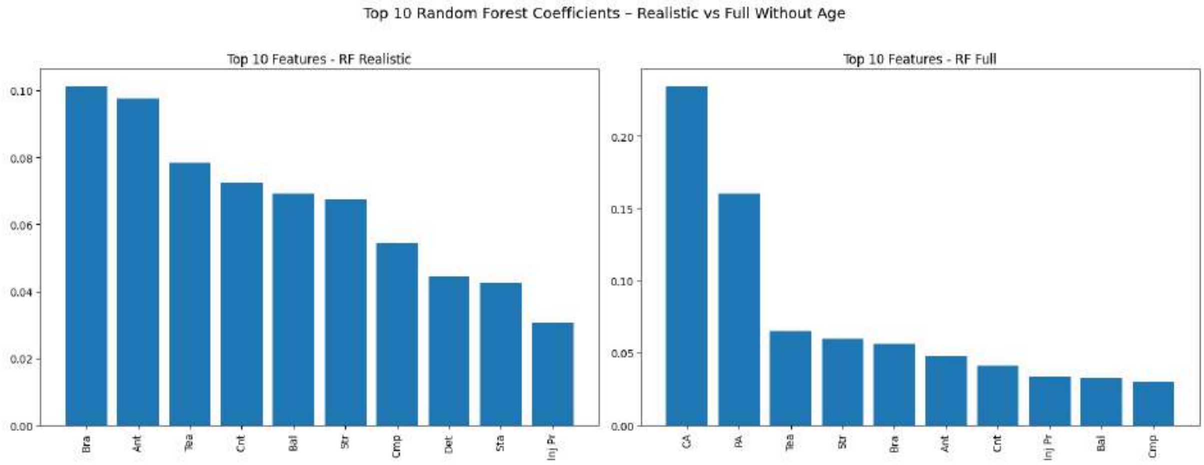


Figure 72: Top 10 Random Forest Feature Importances (Realistic Mode, Without Age)

Decision Tree (DT) — Feature Importance (Impurity-Based)

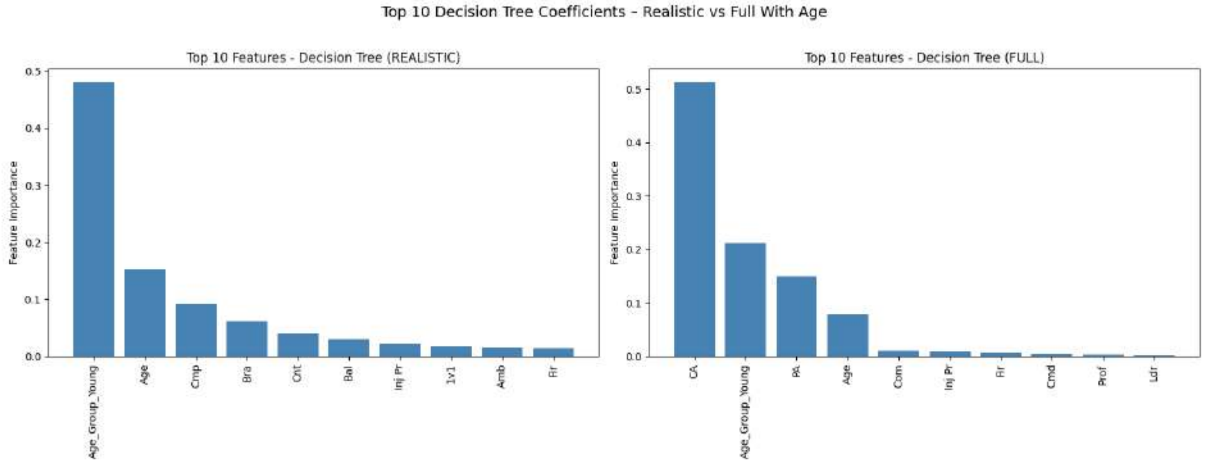


Figure 73: Top 10 Decision Tree Feature Importances (Realistic Mode, With Age)

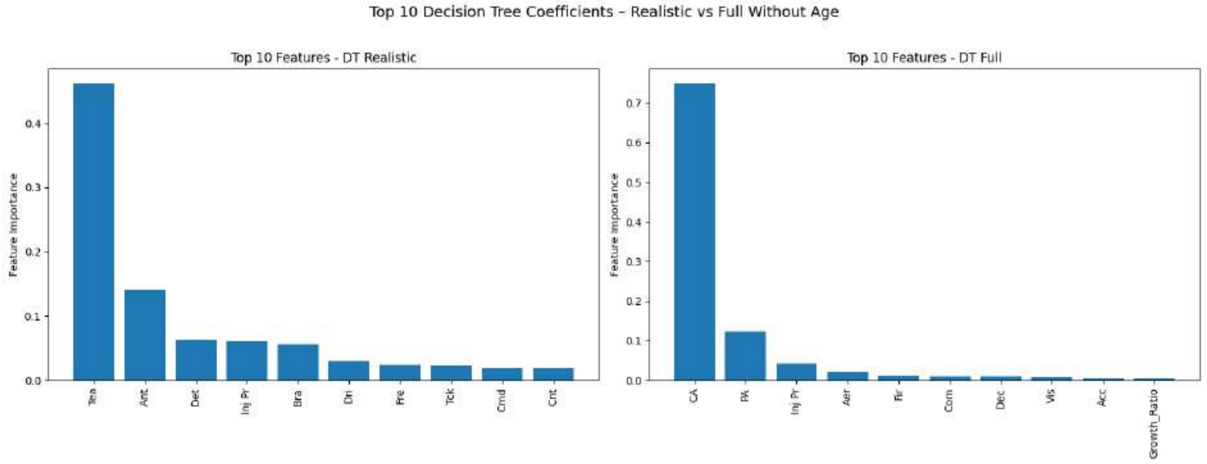


Figure 74: Top 10 Decision Tree Feature Importances (Realistic Mode, Without Age)

Support Vector Classifier (SVC) — Feature Importance (Permutation-Based)

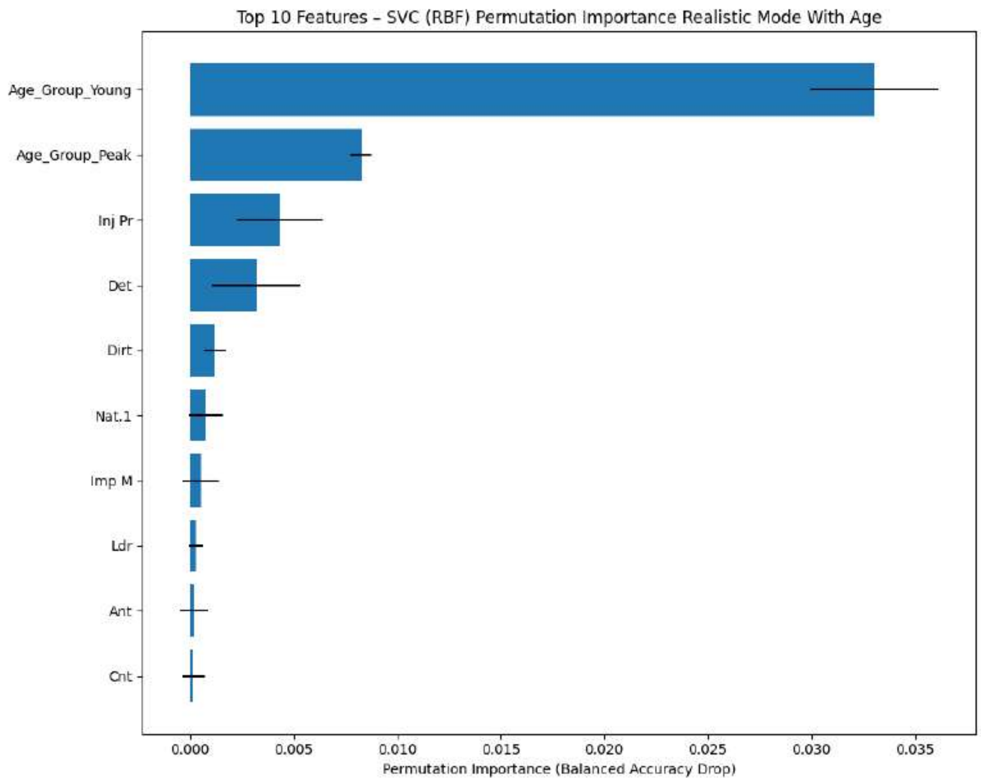


Figure 75: Top 10 SVC (RBF) Permutation Importances (Realistic Mode, With Age)

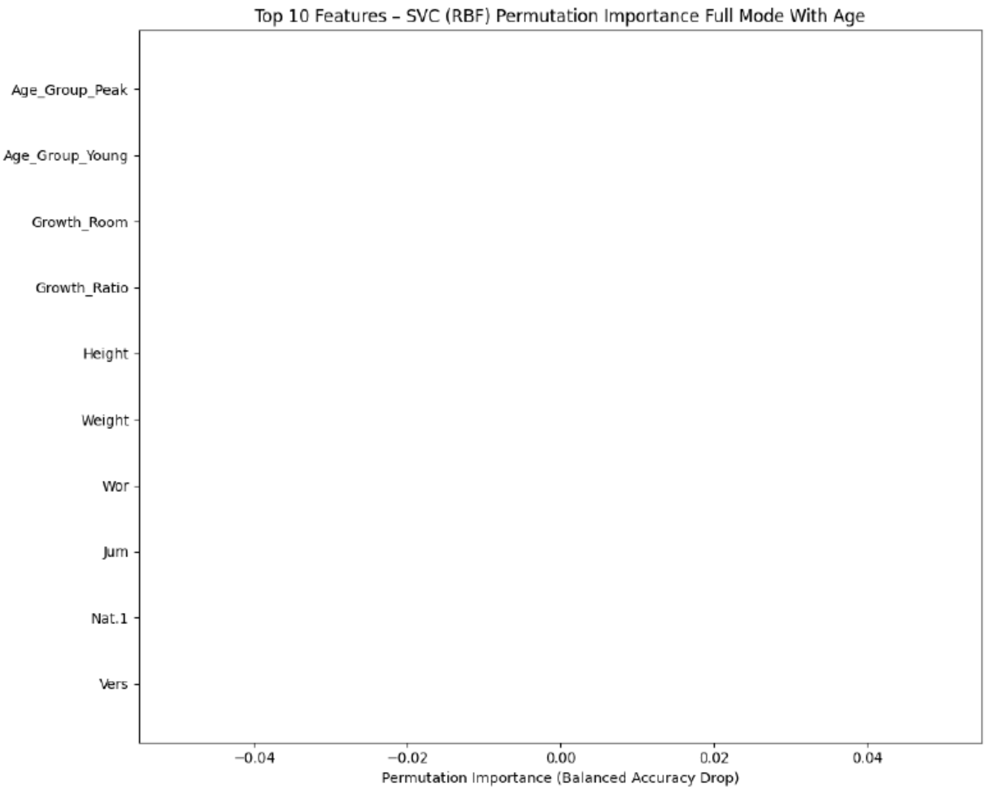


Figure76: Top 10 SVC (RBF) Permutation Importances (Full Mode, With Age)

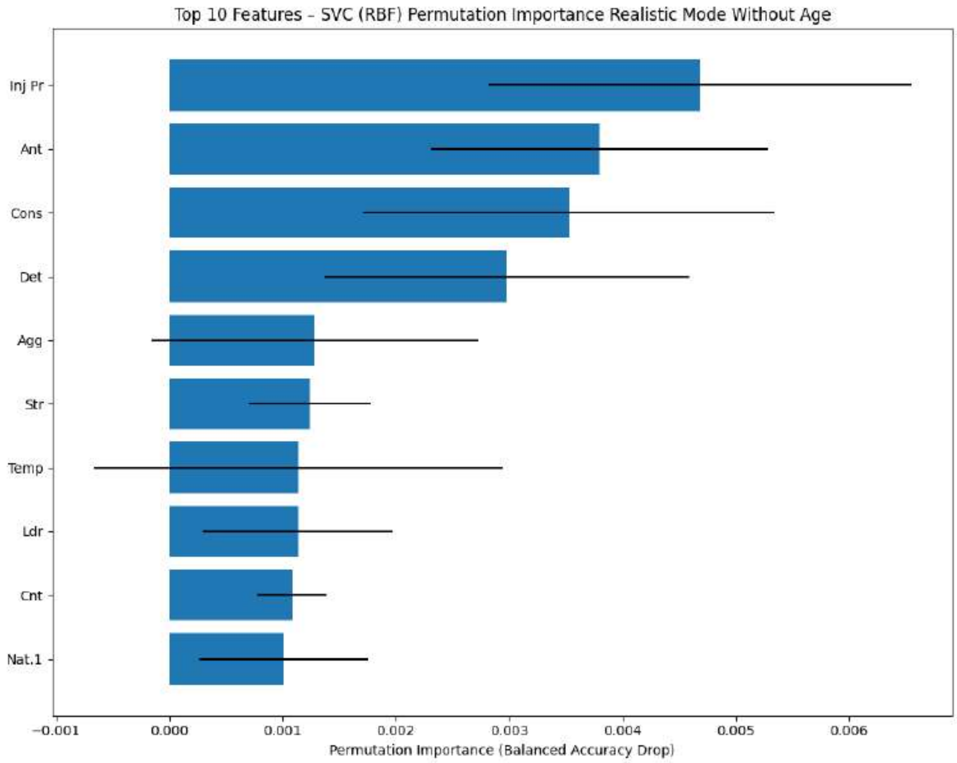


Figure 77: Top 10 SVC (RBF) Permutation Importances (Realistic Mode, Without Age)

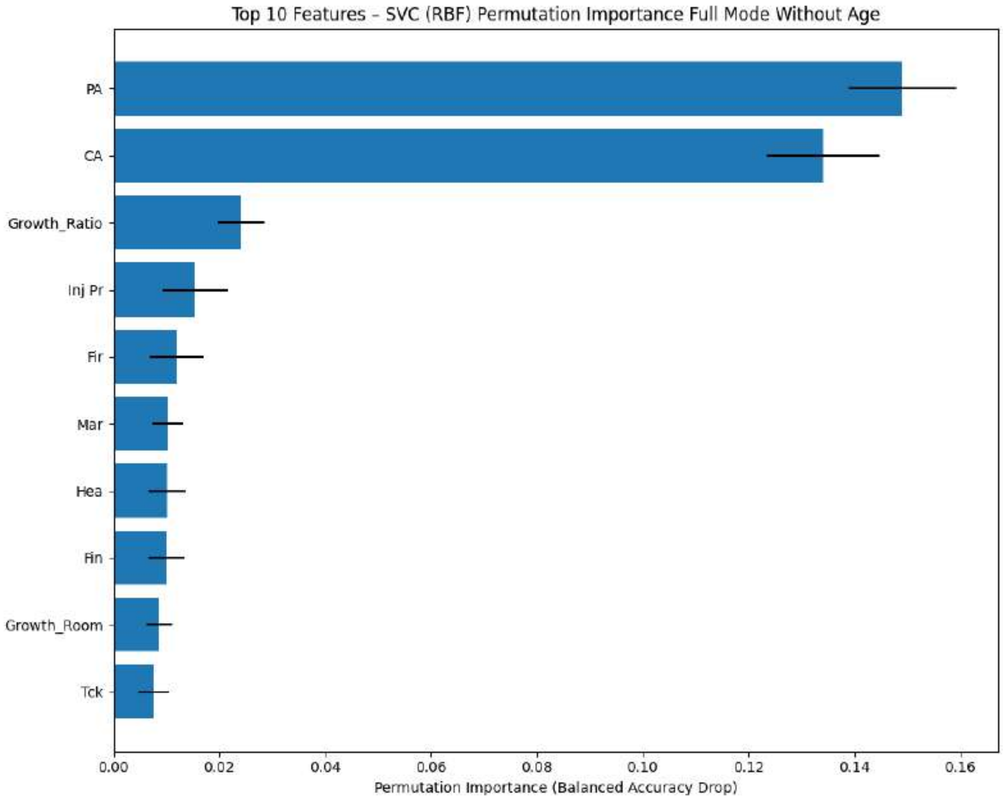


Figure 78: Top 10 SVC (RBF) Permutation Importances (Full Mode, Without Age)

XGBoost

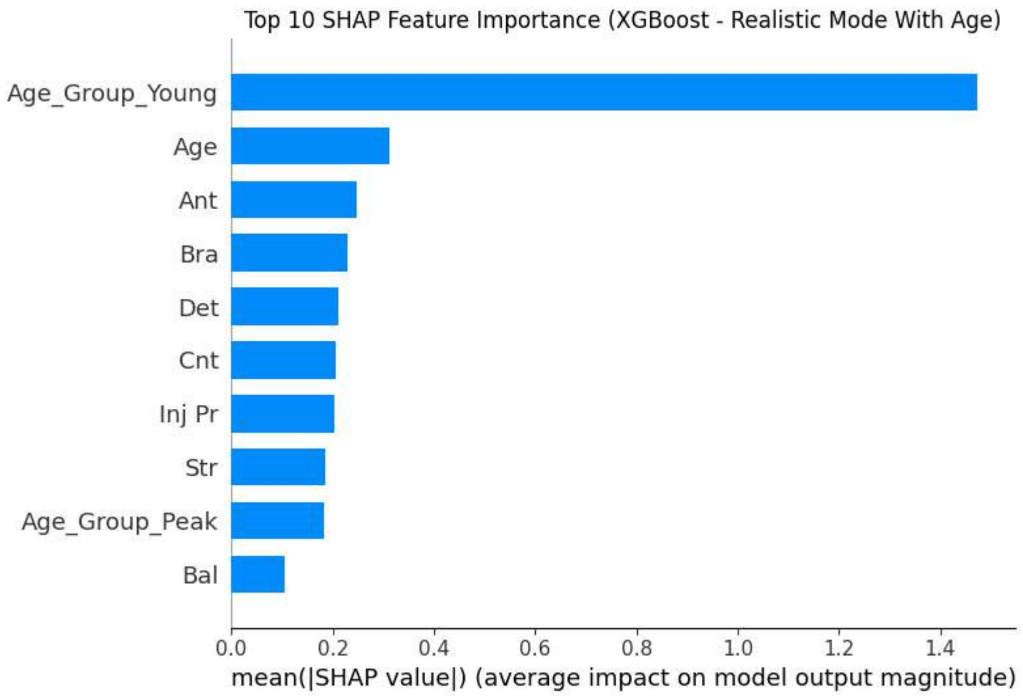


Figure 79: Top 10 XGBoost Feature Importances (Realistic Mode, With Age)

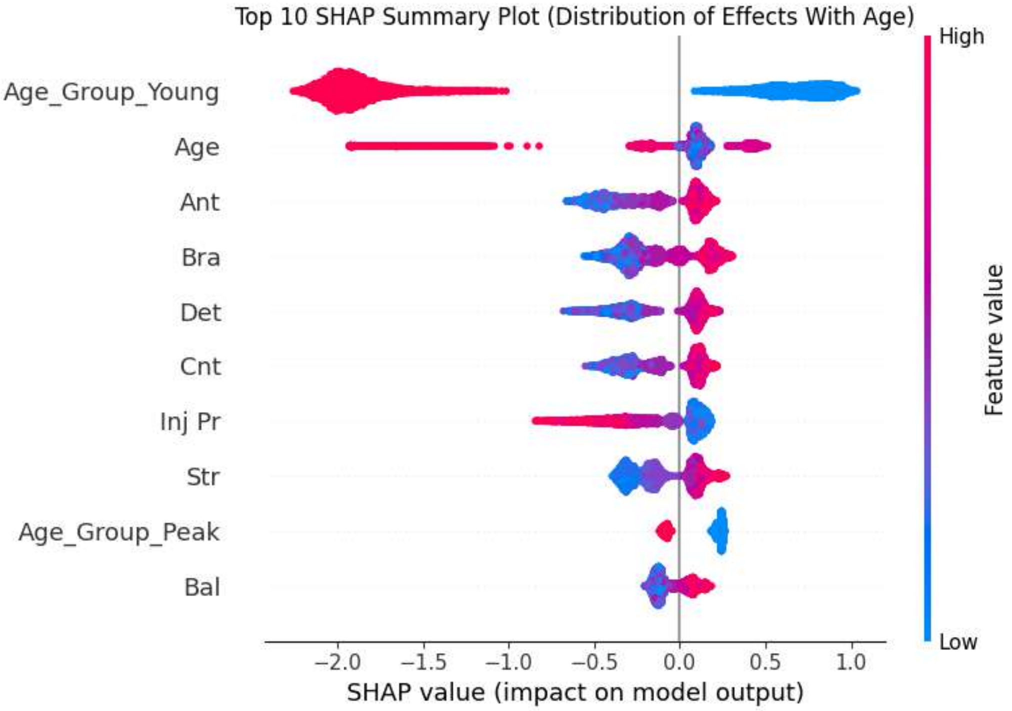


Figure 80: SHAP Summary Plot (Distribution of Effects, Realistic Mode, With Age)

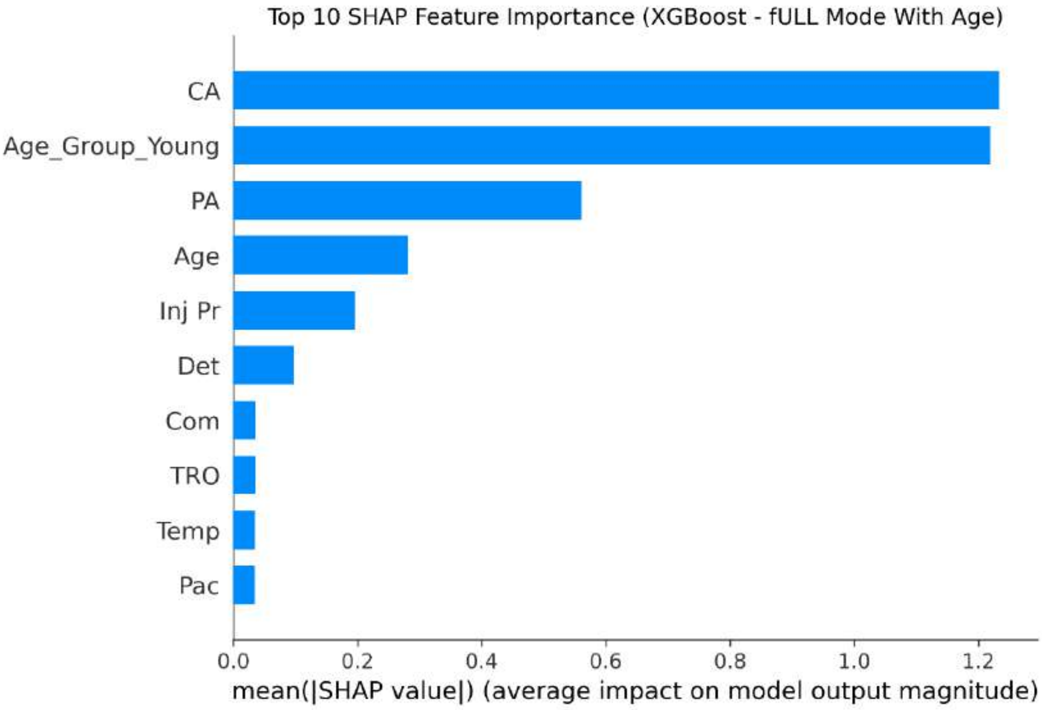


Figure 81: Top 10 XGBoost Feature Importances (Full Mode, With Age)

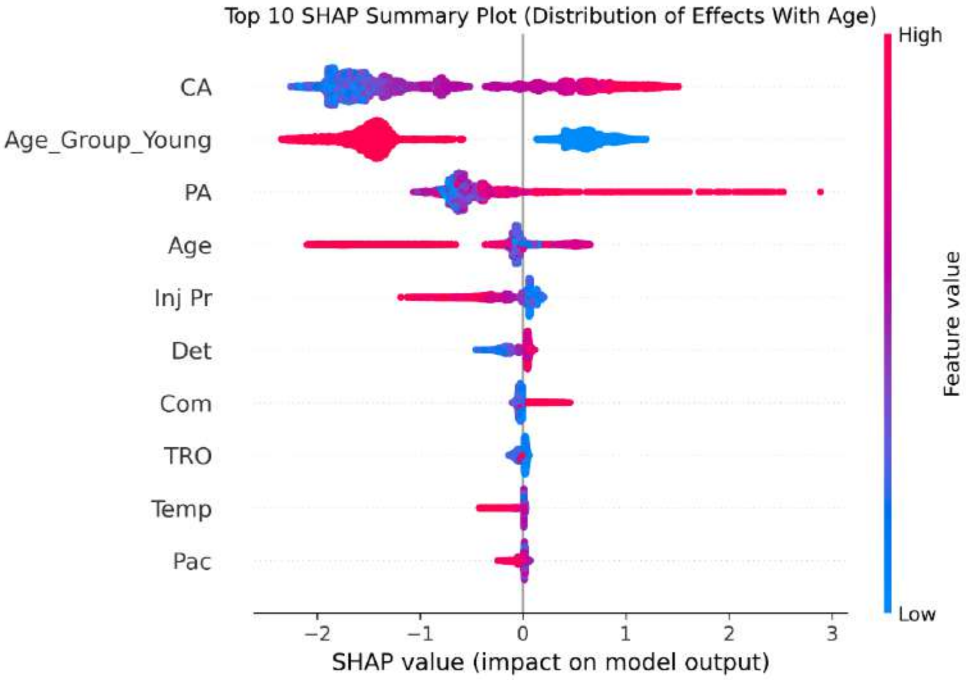


Figure 82: SHAP Summary Plot (Distribution of Effects, Full Mode, With Age)

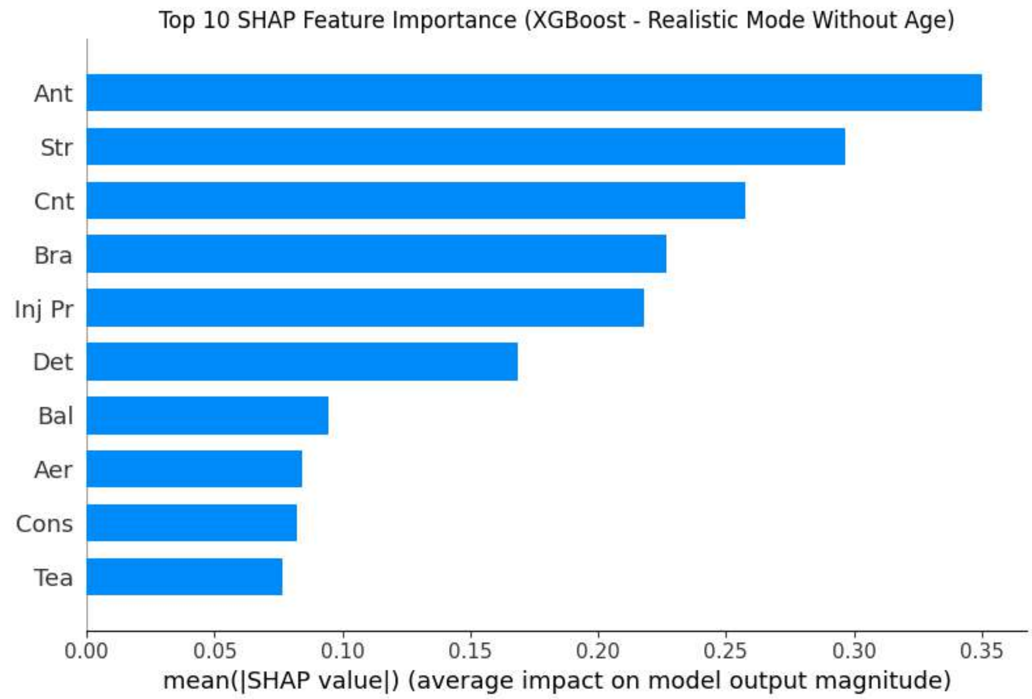


Figure 83: Top 10 XGBoost Feature Importances (Realistic Mode, Without Age)

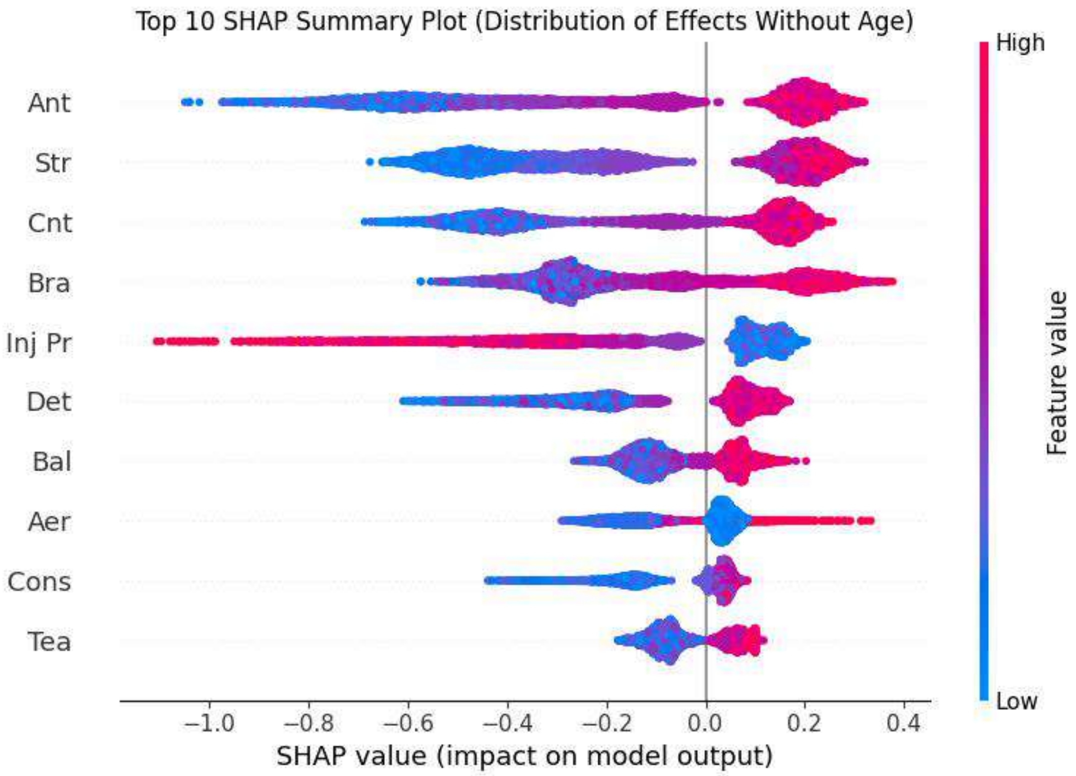


Figure 84: SHAP Summary Plot (Distribution of Effects, Realistic Mode, Without Age)

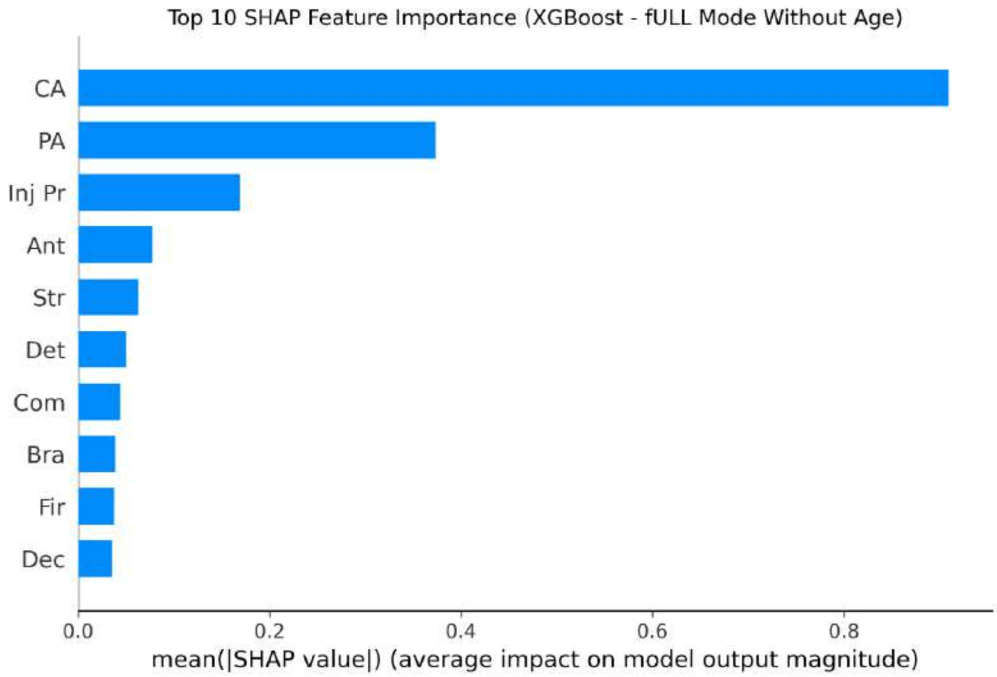


Figure 85: Top 10 XGBoost Feature Importances (Full Mode, Without Age)

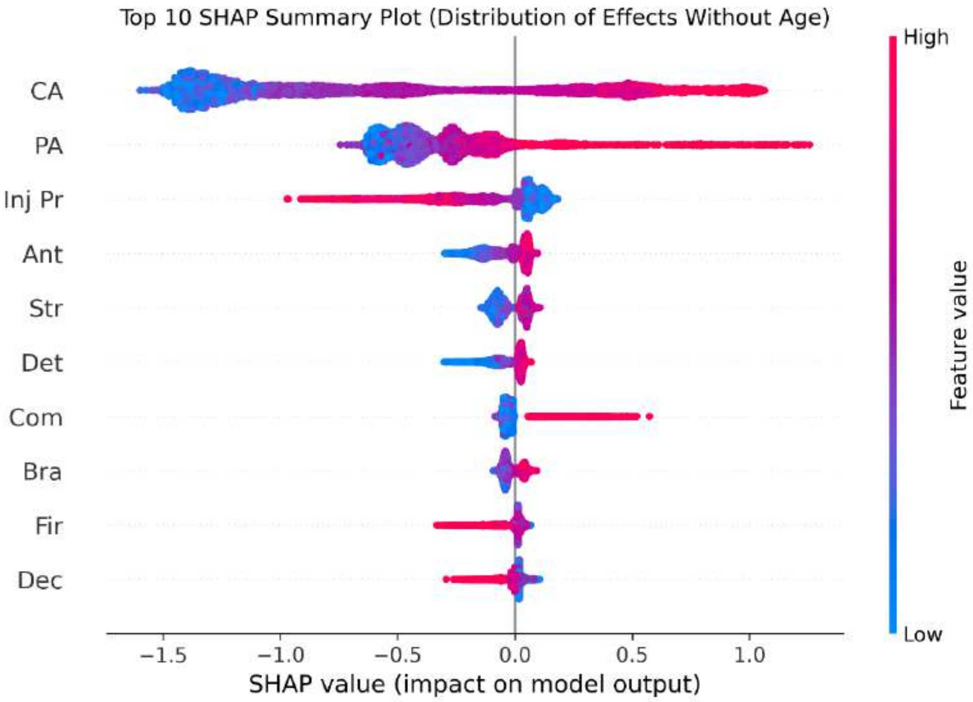


Figure 86: SHAP Summary Plot (Distribution of Effects, Full Mode, Without Age)