Machine Learning for Football Match-Point Prediction. Algorithms Performance on Small Datasets

Marin FOTACHE[1], Irina COJOCARIU[1,2], Armand BERTEA[1,2]
[1]Alexandru Ioan Cuza University of Iasi, Romania
[2]Brandweb Design SRL, Iasi, Romania
fotache@uaic.ro, cojocariu.irina96@yahoo.com, armandbertea@gmail.com

Match-point prediction is an operationally relevant task in football analytics, supporting tactical preparation, squad management, and performance monitoring. Based on a dataset provided by InStat on the results of a struggling Romanian football team, this study proposes a reproducible and auditable decision-support workflow framed as a supervised binary classification problem that estimates the probability of securing at least one point by match end (90 minutes plus stoppage time) using match data. Predictors are engineered to represent three interpretable constructs: average team age, tactical deployment, and key-player exposure. The workflow was implemented in R using two intertwined frameworks for data processing and exploration (tidyverse) and Machine Learning (tidymodels). Five classification algorithms were benchmarked. Results provide some insights on the classification performance when applied to small datasets. Also, the average team age and key-player minutes emerge as the most important predictors in explaining the variability of point attainment for the reference team.

Keywords: *Football Analytics, Classification, Match outcome, Applied Machine Learning algorithms, Tidymodels, Feature engineering, Decision support*
DOI: 10.24818/issn14531305/30.2.2026.01

1 Introduction

Football clubs increasingly adopt data-driven capabilities to support decisions in match preparation, squad selection, and ongoing performance monitoring. In practice, analysts and coaching staff often require compact, match-level outputs that can complement scouting and post-match review with probabilistic assessments aligned to short-term competitive objectives. A particularly actionable objective in league contexts is the likelihood of securing at least one point (win or draw), as it directly supports points accumulation strategies and managerial targets related to competitive stability.

This paper develops a data analysis and Machine Learning (ML) workflow for predicting whether the reference football team will get at least one point in a given match during a regular season (in the first or the second Romania league). This is a typical classification problem where the outcome is binary. The empirical design followed two applied constraints frequently emphasized in business informatics and operational analytics:

- predictors should remain interpretable for coaching and performance staff; and
- predictors should be reliably derivable at match level without requiring high-frequency tracking data or granular event-stream infrastructures.

Accordingly, the feature set is engineered to capture three decision-relevant constructs: squad structure, tactical deployment, and key-player exposure. After the initial Exploratory Data Analysis and the examination of bivariate associations between the outcome (getting at lead one point in a match) and the predictors, the study conducts a controlled comparison between five classification algorithms under a unified ML framework (*tidymodels*). To support transparent interpretation without resorting to instance-level explanation methods, the analysis includes global diagnostic reporting and frames interpretation in line with acknowledged cautions concerning correlated predictors and missing values [1].

As the available dataset was limited due to

copyright constraints, another interest was to compare classification performance of the five algorithms when trained on small datasets.

2 Related Work

2.1 Classical approaches to match outcome modeling

Early football forecasting research commonly represented match outcomes through score-driven statistical models in which goals are treated as realizations of stochastic counting processes. Canonical Poisson formulations and later refinements (including correlation-aware and dynamic variants) have been used for both explanation and prediction under uncertainty, including applications focused on inefficiencies in betting markets [2], [3].

Alongside goal-based modeling, rating and ranking baselines offer parsimonious proxies for latent team strength. These approaches remain widely used because they are transparent, relatively easy to maintain, and compatible with operational decision cycles. In association football, systems have shown that simple rating dynamics can capture non-trivial predictive signal even when feature engineering is limited [4].

While these methods provide interpretable reference points, they typically rely on explicit assumptions about goal-generation mechanisms and/or the time evolution of latent strength. Such assumptions are not always aligned with feature-engineering frameworks that integrate heterogeneous match-level indicators, prioritize standardization and reproducibility, and are designed for incremental extension as new data assets become available.

2.2 Machine learning for football result prediction

ML methods have been extensively explored for football result prediction, especially in league settings where match-level datasets can be assembled and enriched with contextual predictors. Empirical work suggests ML models can match or outperform classical baselines when engineered features capture performance-relevant information and evaluation protocols avoid optimistic bias [5]. Survey studies further highlight that predictive accuracy depends critically on representation choices, feature construction, and validation rigor [6].

Beyond forecasting, applied research increasingly uses ML as a diagnostic tool to identify performance variables associated with winning or avoiding defeat in specific competitions. For example, supervised learning has been used to isolate key performance metrics that differentiate wins from non-wins in league contexts [7]. This aligns with a decision-support view in which predictive models are expected not only to classify outcomes, but also to generate signals that practitioners can interpret and act upon.

In operational environments, the main constraint is often governance rather than algorithm availability: teams must be able to engineer predictors reliably, control preprocessing inside resampling to reduce leakage risk, and maintain auditable workflows that support comparability across model families and seasons. These requirements closely match the economic informatics emphasis on reproducibility, workflow governance, and lifecycle management of analytics artifacts.

2.3 Performance indicators, context, and engineered predictors

Performance analysis research provides the conceptual basis for operationalizing match-level predictors as interpretable indicators. Seminal contributions emphasize two requirements: indicators should be conceptually relevant to the performance construct of interest and practically measurable within routine analytic processes, enabling the transformation of raw observations into decision-relevant features [8]. Subsequent work shows that technical performance is systematically moderated by contextual conditions—particularly match location, opposition quality, and match status—implying that the same indicator may reflect different underlying behavior across situations [9].

Complementary research identifies match

statistics that discriminate successful from unsuccessful teams, supporting compact match-level aggregates for classification and decision support when richer event-level representations are unavailable [10]. Related work has also proposed specific match actions (e.g., entries into the penalty area) as sensitive proxies of competitive superiority, which can be operationalized as actionable predictors [11].

Recent football analytics increasingly adopts probabilistic performance metrics such as expected goals (xG) to reduce outcome randomness by evaluating chance quality rather than relying only on realized goals. Evidence suggests xG-type measures can enable more stable performance evaluation and support tactical decision processes; this trend is reinforced by explainable xG formulations designed to improve interpretability and communication to analysts and coaching staff [12], [13], [14]. In parallel, tracking data has enabled fine-grained tactical descriptors (e.g., spatial compactness, inter-line distances, and network-based passing structure), facilitating both predictive and explanatory analyses [15], [16]. However, tracking infrastructures remain unevenly available, motivating research designs that remain effective using performance-intelligence datasets and engineered match-level predictors.

Within such constraints, formation choice can serve as a tractable tactical signature. Evidence indicates that common formations or arrangements (e.g., 4-2-3-1, 4-3-3, 3-5-2) impose distinct role requirements and influence technical and physical patterns, including passing involvement and attacking structure [17], [18]. Finally, squad composition and player attributes introduce an additional decision-relevant layer. Age structure and competitive experience have been linked to both individual contribution and aggregate team performance, and systematic reviews highlight the breadth of player attributes used in modern data-driven frameworks [19], [20]. Together, these findings motivate parsimonious constructs—such as team age structure and key-player minutes—as feasible match-level proxies for experience, stability, and availability.

3 Data and Feature Engineering

3.1 Data sources

The analysis in this paper used licensed InStat performance-intelligence data for Romanian professional football competitions over the 2016–2022 period. The unit of analysis was the match, and the reference team was Politehnica Iași which generally fights (unsuccessfully) for escaping the relegation to lower leagues. Raw data were available at multiple granularities, including match metadata, player participation records, and formation timelines.

To ensure consistency with the decision-support objective and the limited sample size, all information was transformed into a unified match-level analytical dataset. Player-related data were processed by computing minutes played for each participant and extracting minutes played by the top five ranked players, where ranking is based on an aggregated performance index provided by InStat. These variables operationalize key-player exposure as a proxy for availability and on-pitch influence. Tactical deployment was derived from formation logs and encoded as duration-weighted measures representing the total minutes spent in each predefined system of play during the match. Squad structure was summarized by average team age at match time, computed as a minutes-weighted mean of participating players, serving as a compact indicator of experience and lineup stability.

The target variable *get_points* was defined as a binary indicator of whether the reference team secured at least one point at match end (win or draw versus loss). The final dataset used for modeling contained 82 match observations.

3.2 Feature blocks. Variables

Predictors were engineered to operationalize three decision-relevant constructs, consistent with the performance-indicator perspective in football analytics [8] and evidence on the contextual determinants of technical

performance [9], [21]. Specifically, the feature set captures:
(A) squad structure, proxied by average team age at match time;
(B) tactical deployment, encoded through duration-weighted formation usage; and
(C) key-player exposure, measured via minutes played by top-ranked players according to an aggregated performance index.

The final data frame contained 82 match observations, after merging engineered predictors (average age, time-weighted formation usage, and key-player minutes) into a unified match-level dataset.
The complete list of variables used in the subsequent analysis and ML classification models is shown in **Table 1.** Every record refers to a match played by the reference team.

Table 1. Variables

Variable	Type	Description
get_points	Factor (yes/no)	The main variable in the classification models. Its value is *yes* if the reference team (i.e. Politehnica Iași) got at least one point in the current match (win or draw), and *no* if it was defeated.
arr__4_2_3_1	Numeric	Total number of minutes during which the team played in a 4-2-3-1 formation in the match (duration-weighted tactical deployment).
arr__4_4_2_classic	Numeric	Total number of minutes during which the team played in a classic 4-4-2 formation in the match.
arr__4_4_2_diamond	Numeric	Total number of minutes during which the team played in a 4-4-2 diamond formation in the match.
arr__4_1_4_1	Numeric	Total number of minutes during which the team played in a 4-1-4-1 formation in the match.
arr__3_4_3	Numeric	Total number of minutes during which the team played in a 3-4-3 formation in the match.
arr__4_3_3_down	Numeric	Total number of minutes during which the team played in a 4-3-3 (defensive) formation in the match.
arr__3_5_2	Numeric	Total number of minutes during which the team played in a 3-5-2 formation in the match.
home_away	Character	Whereas the match was played at home or on the opponent's stadium
opponent_match	Character	The name of the opponent team
average_player_age	Numeric	Average age of all players who participated in the match, weighted by minutes played (proxy for squad experience)
minutes_top_1	Numeric	Minutes played in the match by the highest-ranked player according to the aggregated performance index.
minutes_top_2	Numeric	Minutes played in the match by the second-ranked player according to the aggregated performance index.
minutes_top_3	Numeric	Minutes played in the match by the third-ranked player according to the aggregated performance index.
minutes_top_4	Numeric	Minutes played in the match by the fourth-ranked player according to the aggregated performance index.
minutes_top_5	Numeric	Minutes played in the match by the fifth-ranked player according to the aggregated performance index.

Note: The prefix *arr_* denotes tactical arrangements (formations). Values correspond to duration-weighted minutes derived from formation timelines.

Player-exposure variables were restricted to the top five ranked players, as exploratory analysis indicated that playing time is concentrated within a small core of

contributors. This design reduced dimensionality while retaining the most informative signal for point-attainment classification.

3.3 Dataset construction principles

A match-level data frame was assembled to ensure consistent variable definitions, explicit handling of missing values, and readiness for categorical encoding within the workflow. Given the scope and limited sample size, the methodological emphasis was placed on process rigor (standardized preprocessing and controlled tuning) and transparent reporting, rather than extensive stratification, threshold optimization, or strong claims of generalizability.

4 Methodology and Tools

The first objective of this paper was to examine bivariate relationships between the outcome variable (*get_points*, i.e. whether the reference team will get at least one point in the current match) and other variables in **Table 1.** Graphical examination of the relationships was accompanied by a series of classical statistical tests.

The second objective was to build models that predict whereas, given the predictors in Table 1 (all variables except *get_points*), the reference team will get at least one point in the current match. A series of ML models were built and tuned using five popular algorithms. The problem was particularly challenging since the available dataset was small, at least for ML modelling requirements. The best models (selected, for each algorithm by tuning the hyperparameters) performance was compared for each of the six algorithms on the test set.

The third objective was to assess the predictor's importance in the outcome variability, and to compare the estimations provided by best performing algorithms.

4.1 Data preparation and exploration

All analyses were implemented in R [22]. Data processing and preparation of the initial set provided by InStat relied mainly on the *tidyverse* ecosystem in R [23]. Packages *gtsummary* [24] and *ggplot2* [25] were deployed for the Exploratory Data Analysis [26], [27].

Inferential statistics tests were performed with the package *ggstatsplot* [28] which acts as a unified framework for the most popular parametric and non-parametric tests, since it provides all three main information on each statistical test, i.e., the p-value, the effect size and the confidence interval of the effect size [29].

4.2 Machine Learning workflow

All ML models were built and refined using the *tidymodels* ecosystem, [30], [31], which supports an integrated, auditable specification of preprocessing, resampling, tuning, and final fitting. To ensure strict comparability across algorithms and to mitigate data leakage, preprocessing was defined once as a shared recipe and executed within each resampling split and in the final model fit. To avoid data leakage [32], the initial dataset was split randomly into the training subset (75% of the observations) and the test subset (25%). Similar distribution of the outcome between train-test datasets was ensured with stratification. The test subset was not used during the process of training and tuning the models, but only in the final assessment of the ML models. Subsequently, the training dataset was split into five folds (using stratification) for cross-validation. In *tidymodels*, the train-test subsets and the cross-validation folds are built with package *rsample* [33].

The unified preprocessing pipeline included: (i) k-NN imputation for missing values; (ii) dummy encoding for nominal predictors, and (iii) removal of zero-variance predictors.

All preprocessing tasks were performed with package *recipes* [34]. Because diagnostics such as feature importance can be influenced by missingness mechanisms and collinearity, interpretation is framed as predictive contribution within the observed feature space rather than causal evidence [1].

The prediction task was addressed using five established algorithms for classification (see

[35] for a detailed description of each algorithm):

- Logistic regression [36] via the *glmnet* engine [37], [38], by tuning two hyperparameters, *penalty* (regularization) and *mixture* (proportion of Lasso penalty).
- Random Forest, which aggregates decision trees trained on bootstrap samples with randomized feature selection [39]. The number of trees was fixed at 700. Models were fitted with the ranger engine [40] and tuned by using two hyperparameters: *mtry* (the number of randomly selected predictors to continue with the node split) and *min_n* (minimal number of observations in anode to continue with the split).
- Gradient boosting, which sequentially combines weak learners into a stronger predictor [41]; XGBoost [42] was preferred from the gradient boosting approaches, using the xgboost engine [43], fixing the number of trees at 1000, and tuning six hyperparameters: *tree_depth* (maximal number of levels in the tree), *min_n* (the same as in random forest models), *loss_reduction* (minimum loss reduction), *sample_size* (proportion observations sampled), *mtry* (the same as in random forest models), and *learn_rate* (learning rate, i.e., the step size).
- Radial basis function support vector machines (SVMs) via kernlab engine [44], [45] with three hyperparameters: *cost*, *rbf_sigma* (Radial Basis Function sigma), and *margin* (Insensitivity Margin).
- Multilayer perceptron (MLP) via *nnet* engine [46] with three hyperparameters: *hidden_units* (number of hidden units), *penalty* (regularization), and *epochs*.

Hyperparameters were optimized via random grid-based search. Package *dials* [47] (part of the *tidymodels* framework) is instrumental in generating the appropriate candidate values for each hyperparameter. The size of grid search data frame varied on the number of tuned hyperparameters: 200 for logistic regression, random forest, support vector machine and multi-layer perceptron models, and 600 for the extreme gradient boosting models.

For all algorithms, cross-validation folds and grid search combinations of hyperparameters, the models were fitted with package *tune* [48]. Execution was parallelized with package future [49]. Function *tune_grid* in package *tune* computes a set of performance metrics for each model, e.g. accuracy and ROC AUC, using package *yardstick* [50]. For each algorithm, the best combination of hyperparameters was chosen based on the highest ROC AUC averaged along the cross-validation training folds.

Best selected model of each algorithm was then assessed on the new data, i.e. the test dataset. Predictors importance in the outcome variability was assessed with the *vip* package [51].

5 Results

5.1 Exploratory Data Analysis (EDA

5.1.1 Distribution of nominal variables

The distribution of the nominal variables used in the analysis: match outcome (*get_points*), match location (*home_away*), and opponent identity (*opponent_match*) is shown in **Fig. 1.**

The outcome variable is relatively balanced, with 52% of matches resulting in no points and 48% in at least one point. This near-balance is suitable for "regular" binary classification, as it reduces the risk of majority-class bias in model training.

The distribution of match location is also well balanced, with 51% of matches played away and 49% at home. This symmetry suggests that venue effects are not confounded with outcome frequency in the dataset and can be analyzed without strong structural imbalance.

Opponent frequencies are more heterogeneous, reflecting the multi-season nature of the dataset and the fact that some teams were encountered more often than others across competitions and seasons. No single opponent dominates the sample, which supports the use of opponent-related information as contextual input rather than as a primary explanatory driver.

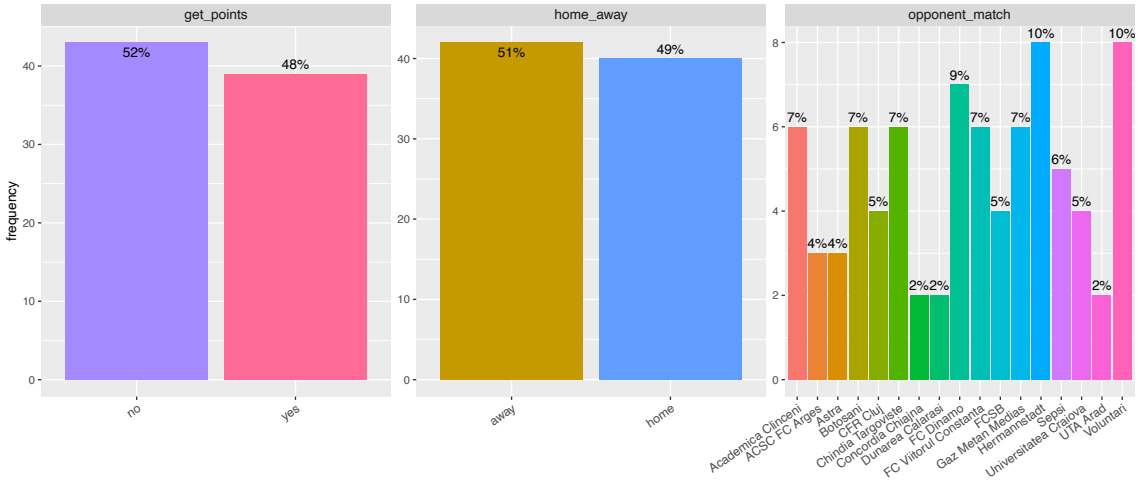

Fig.1. Distribution of nominal variables

Overall, the distributions indicate that the dataset does not suffer from severe imbalance in its categorical variables, supporting the validity of subsequent analyses.

5.1.2 Competitive profile and outcome patterns

Seasonal win/draw/loss distributions indicate a deterioration in competitive outcomes over the analyzed interval, with an increase in losses preceding the team's transition to the second league in the 2021–2022 season. This observation provides contextual grounding for the modeling task, as the probability of securing points is unlikely to be constant across seasons.

5.1.3 Tactical profile: formation arrangement over time

Formation selection (team arrangement) is relatively concentrated rather than dispersed across many systems. Across seasons, a small set of formations accounts for most tactical deployment (notably 4-2-3-1, 4-3-3, and 4-4-2), supporting the operationalization of formation behavior through match-level, time-weighted predictors.

Descriptive statistics (**Table 2.**) further indicate strong right-skewness and non-normal distributions for all formation variables, with medians equal to zero for most systems and large gaps between upper quartiles and maxima. This pattern reflects the fact that, in many matches, a given formation is not used at all, while in others it may dominate substantial portions of the match.

Table 2. Descriptive statistics on the variables related to the tactical deployment

Variable	Min	Q1	Median	Q3	Max	Mean	SD	Shapiro-Wilk Test
arr__4_2_3_1	0	0	567	2307	4267	1121.00	1241.000	W:0.807, p-value:0.000
arr__4_4_2_classic	0	0	0	1107	3139	633.00	1021.000	W:0.661, p-value:0.000
arr__4_4_2_diamond	0	0	0	0	1362	16.61	150.408	W:0.087, p-value:0.000
arr__4_1_4_1	0	0	0	0	844	10.96	93.327	W:0.096, p-value:0.000
arr__3_4_3	0	0	0	0	2757	63.32	338.190	W:0.188, p-value:0.000
arr__4_3_3_down	0	0	0	2221	3273	894.00	1187.000	W:0.729, p-value:0.000
arr__3_5_2	0	0	0	0	2045	40.30	263.790	W:0.141, p-value:0.000

The Shapiro–Wilk tests confirm significant deviations from normality for all tactical variables, justifying the use of non-parametric methods in exploratory comparisons and reinforcing the suitability of tree-based learners in subsequent modeling stages.

At an aggregate level, the tactical variables capture meaningful between-match variability in system deployment without introducing excessive dimensionality. Their inclusion allows the models to represent tactical continuity and change in a parsimonious and interpretable manner, consistent with the match-level decision-support focus of the study.

5.1.4 Player exposure and squad age structure

Because the model includes "key-player exposure" features, we inspected player usage distributions to confirm whether minutes are concentrated in a small core of players. The minutes-by-season profile indicates that certain players accumulate very high exposure (e.g., peak seasonal minutes around ~3500 for a top-minute player), while multi-season totals highlight a continuity nucleus (e.g., one player reaching ~6000 minutes across two seasons).

Table 3. shows a highly heterogeneous distribution of minutes played by top-ranked players. For most exposure variables, the median is zero, indicating that key players are absent or marginally involved in many matches, while maximum values approach full-match duration when they are selected. The distributions are strongly right-skewed and deviate significantly from normality ($p < 0.001$), with large standard deviations relative to means, reflecting substantial between-match variability in player usage. These patterns indicate that key-player exposure is episodic and context-dependent rather than stable across matches.

Table 3. Descriptive statistics on the variables related to the key-player exposure

Variable	Min	Q1	Median	Q3	Max	Mean	SD	Shapiro-Wilk Test
minutes_top_1	0	0	0	94	100	38	44	W:0.695, p-value:0.000
minutes_top_2	0	0	0	93	97	38	44	W:0.689, p-value:0.000
minutes_top_3	0	93	94	95	100	80	32	W:0.541, p-value:0.000
minutes_top_4	0	0	0	95	98	44	47	W:0.652, p-value:0.000
minutes_top_5	0	0	23	94	98	43	45	W:0.705, p-value:0.000

Variable *average_player_age* exhibits low dispersion across matches, with values concentrated around the mean of 27.39 years and a standard deviation of only 0.71. Quartile values are tightly clustered (Q1 = 26.86, Median = 27.46, Q3 = 27.90), indicating that age structure is relatively stable across seasons and matches. This stability supports the interpretation of average player age as a structural characteristic of the squad rather than a volatile match-specific artifact.

Although the Shapiro–Wilk test indicates a mild deviation from normality ($p = 0.043$), the distribution remains approximately symmetric and continuous, especially when compared with the highly skewed exposure variables. As a result, average squad age can be treated as a compact and robust descriptor of experience and lineup composition, making it a suitable predictor for match-level modeling and subsequent interpretability analysis.

5.2. Bivariate relationships of the outcome variable

Bivariate associations between the outcome variable (*get_points*) and selected predictors were examined to provide descriptive context for the multivariate modeling stage. The analysis covers contextual variables (match location and opponent), tactical deployment, key-player exposure, and squad age structure. Non-parametric tests and graphical comparisons are used due to the non-normal distribution of most predictors.

Match outcome proportions are similar for home and away fixtures (**Fig. 2.**), with no clear separation between matches with and without point attainment. Consistent with this visual pattern, the chi-square test indicates no statistically significant association between match location and outcome ($\chi^2(1) = 0.76$, p = 0.38).

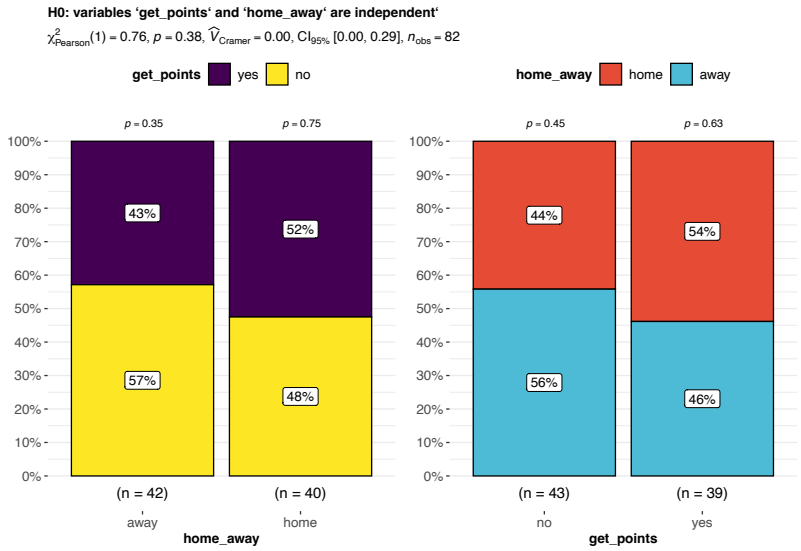

Fig.2. Independence test between the outcome and the match location (home_away)

Match outcome distributions (**Fig. 3.**) vary across opponents, reflecting the heterogeneous competitive context of the dataset. However, no statistically significant association is observed between opponent identity and point attainment ($\chi^2(16) = 15.85$, p = 0.46), suggesting that these differences are largely due to sampling variability.

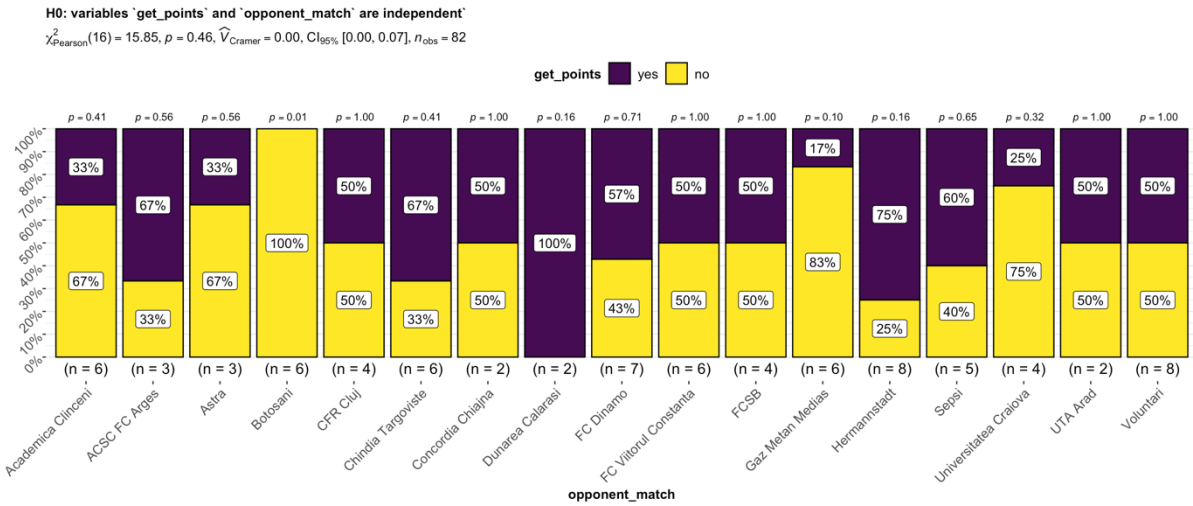

Fig. 3. Independence test between the outcome and the opponent team

In contrast, average squad age differs between outcome groups (**Fig. 4.**), with higher median values observed in matches where at least one point was secured. The Mann–Whitney U test confirms a statistically significant difference (W = 536.00, p = 0.005), with a moderate effect size, indicating a descriptive association between age structure and point attainment.

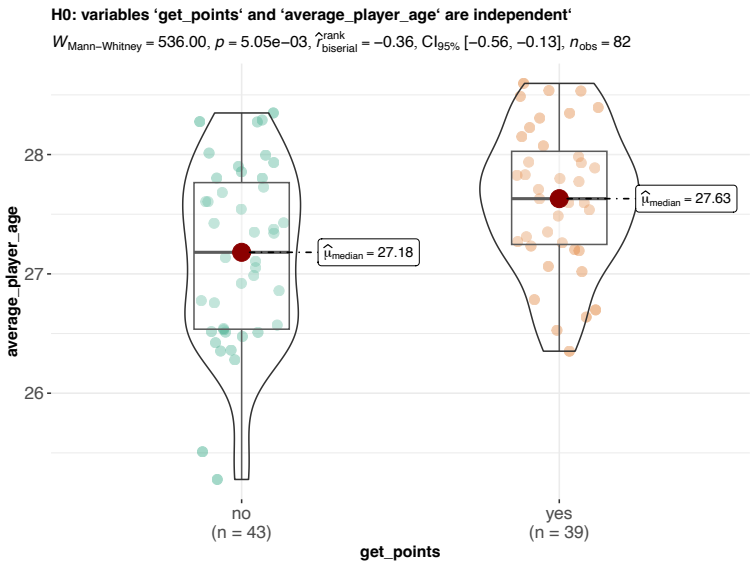

Fig. 4. Distribution of *average_player_age* by get_points

As **Table 4.** shows, bivariate tests for tactical deployment variables do not reveal statistically significant differences between matches with and without point attainment.

Table 4. Results of Mann–Whitney U tests for tactical deployment vs. match outcome

Outcome (get_points) vs. variable…	**Test statistic (W)**	**P-value**	**Effect Size**	**95% CI for Effect Size**
arr__4_2_3_1	997	0.12	0.19	[-0.06,0.42]
arr__4_4_2_classic	911	0.43	0.09	[-0.16,0.33]
arr__4_4_2_diamond	858	0.35	0.02	[-0.22,0.27]
arr__4_1_4_1	796	0.14	0.05	[-0.29,0.20]
arr__3_4_3	792	0.25	-0.06	[-0.30,0.19]
arr__4_3_3_down	705	0.18	-0.16	[-0.39,0.09]
arr__3_5_2	878	0.18	0.05	[-0.20,0.29]

Effect sizes are uniformly small and the confidence intervals include zero for all formations, indicating that formation usage alone does not differentiate outcomes at the marginal level.

Similarly, minutes played by top-ranked players do not show statistically significant bivariate differences between outcome groups (**Table 5.**).

Table 5. Results of Mann–Whitney U tests for key-player exposure vs. match outcome

Outcome vs. variable…	**Test statistic (W)**	**P-value**	**Effect Size**	**95% CI for Effect Size**
minutes_top_1	755	0.40	-0.10	[-0.34,0.15]
minutes_top_2	729	0.27	-0.13	[-0.37,0.12]
minutes_top_3	884	0.67	0.05	[-0.19,0.30]
minutes_top_4	864	0.81	0.03	[-0.22,0.27]
minutes_top_5	789	0.62	-0.06	[-0.30,0.19]

Effect sizes are small and unstable across players, suggesting that key-player exposure does not produce simple linear separation between wins/draws and losses when examined in isolation.

Taken together, the bivariate analysis indicates that most predictors carry limited marginal signal when considered independently. This finding motivates the use of multivariate machine learning models capable of capturing non-linear relationships and interactions between tactical, structural, and exposure-related variables.

5.3. Correlation among predictors

Whereas in the classical regression modeling larger correlation among predictors is a major concern, for many ML algorithms the model performance is not altered by the predictors' collinearity [52]. Nevertheless, examination of correlations among predictors is advisable and must be considered, especially when the most important predictors (as ranked by the algorithms per se, or by other techniques of interpretable ML) are correlated among them or with other predictors.

Since the Shapiro-Wilk test of normality rejected the normality hypothesis of the distribution of all variables (*average_player_age* was the only variable with a distribution closer to normality), the bivariate relationships among variables were assessed with the non-parametric Spearman test of correlation [29]. Results in **Fig. 5.** are presented as correlation coefficients.

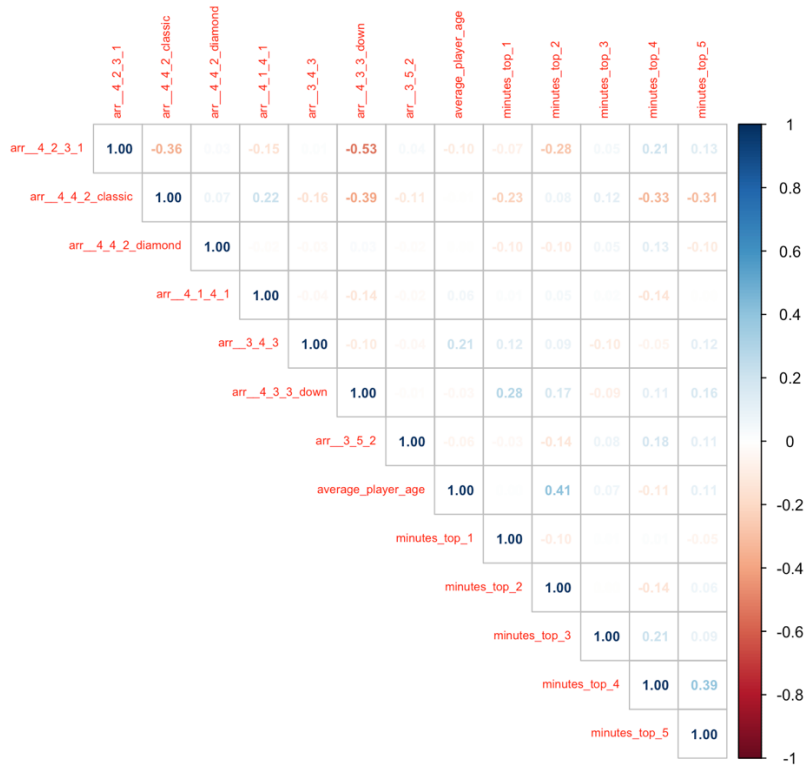

Fig.5. Correlation matrix of numeric predictors

The absolute values of the coefficients do not exceed 0.53, and most values are within the [-0.40, 0.40] range. Consequently, collinearity seems not be an issue in subsequent ML models.

5.3. Selecting the best (tuned) model for each ML algorithm

Models were trained and tuned under a unified preprocessing, resampling, and hyperparameter-optimization protocol to ensure comparability across algorithms. For each model family, the best configuration was

selected based on cross-validated ROC AUC on the training set. The resulting tuned hyperparameters are reported in **Table 6.**, illustrating the diversity of optimal configurations across algorithmic families.

Table 6. Best hyperparameter combinations (on columns) for each algorithm (train dataset)

Hyper-parameter	Extreme Gradient Boosting	Logistic Regression	Multi-Layer Perceptron	Random Forest	Support Vector Machines
cost					8.209957655
epochs			707		
hidden_units			7		
learn_rate	0.004430591				
loss_reduction	1.90523E-06				
margin					0.102854643
min_n	5			37	
mixture		0.254123671			
mtry	3			9	
penalty		0.363016288	5.34812E-07		
rbf_sigma					1.27436E-06
sample_size	0.753788572				
tree_depth	14				

To facilitate a direct visual comparison of discriminative performance, the ROC curves of the best-tuned model for each algorithm were evaluated on the test set and are shown in **Fig. 6.**

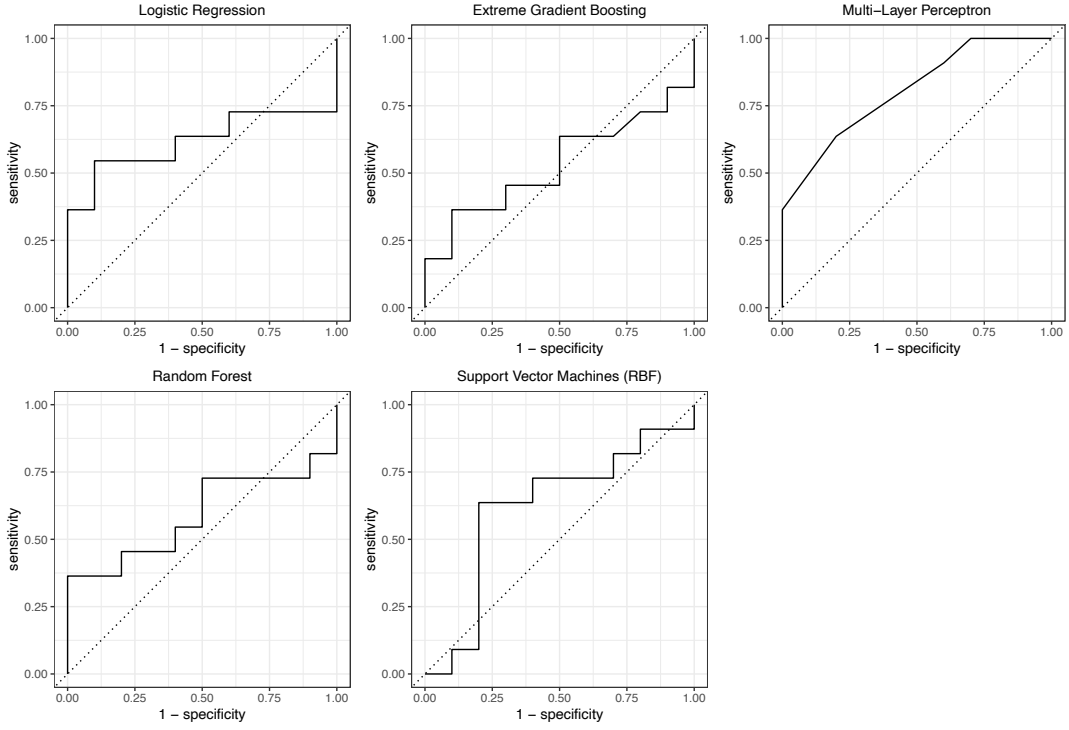

Fig.6. Model performance comparison for the selected (best) model of each algorithm

Visual inspection indicates clear differences in discriminative performance, with the Multi-Layer Perceptron and Support Vector Machine models exhibiting steeper curves and higher sensitivity–specificity trade-offs compared to tree-based and linear models.

Quantitative out-of-sample performance is summarized in **Table 7.** The Multi-Layer Perceptron achieves the highest ROC AUC (0.8045) and the lowest Brier score (0.2072), indicating superior discriminative capacity and calibration in this pilot setting. Support Vector Machines show comparable accuracy (0.7143) but lower ROC AUC, while Random Forest and Logistic Regression achieve moderate performance levels. Extreme Gradient Boosting performs weakest among the evaluated models, with ROC AUC close to random classification.

Table 7. Three performance metrics for the selected models

Algorithm	**Accuracy**	**ROC AUC**	**Brier_Class**
Logistic Regression	0.5714	0.6182	0.2373
Random Forest	0.5714	0.5909	0.2478
Extreme Gradient Boosting	0.5238	0.5318	0.2509
Support Vector Machines (RBF)	0.7143	0.6182	0.2432
Multi-Layer Perceptron	0.7143	0.8045	0.2072

Taken together, the results suggest that non-linear models with higher representational flexibility (MLP and SVM) generalize better than tree-based ensembles in this small-sample setting, while boosting appears more sensitive to noise and tuning constraints. These findings highlight the importance of controlled model comparison and justify the use of multiple algorithmic families when sample size is limited.

5.4 Global diagnostic analysis via feature importance

Global feature-importance diagnostics show that the predictive signal is concentrated in a very small subset of variables. For both the Multi-Layer Perceptron and Random Forest models, average squad age and minutes played by the third-ranked player emerge as the dominant predictors, indicating that demographic structure and sustained exposure of core players carry the strongest signal for point attainment.

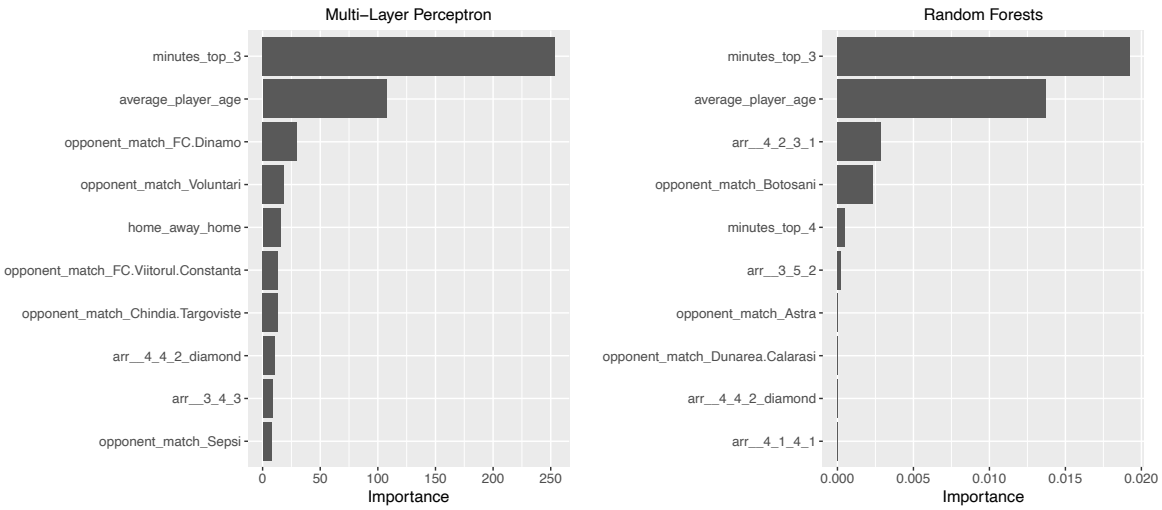

Fig.7. Feature importance for Multi-Layer Perceptron and Random Forest models

Importance values drop sharply, suggesting limited marginal contribution from most tactical and contextual predictors. Random Forest assigns moderate importance to time spent in the 4-2-3-1 formation and to selected opponent indicators, whereas the Multi-Layer Perceptron places relatively more weight on player-exposure variables and match-level structural features.

Consistent with established cautions for non-linear learners, feature-importance scores are interpreted as measures of predictive contribution rather than causal effects, particularly in the presence of correlated predictors and potential non-random missingness [1].

6 Discussion

In this pilot setting, match-point probability is most consistently captured through two related dimensions: squad age structure and key-player exposure. The prominence of average team age across models suggests that age structure functions as a compact proxy for experience, role stability, and match-to-match consistency—properties that fit the performance-indicator perspective emphasizing actionable and measurable constructs [8]. Key-player minutes provide a complementary signal, indicating that point attainment is associated not only with having high-impact players available, but also with their sustained involvement during the match.

Except for the average player age predictor, all bivariate associations between the outcome and the predictors were qualified as statistically non-significant by the Chi-square and Mann-Whitney tests. Nevertheless, ML model comparison indicates that algorithms with higher representational flexibility generalize better even under the current small-sample constraints. The Multi-Layer Perceptron achieves the highest ROC AUC and the lowest Brier score, suggesting both superior discrimination and calibration, while Support Vector Machines deliver comparable accuracy but weaker ranking performance. Tree-based ensembles (Random Forest) show moderate performance, and boosted trees (XGBoost) perform close to random, which is consistent with the sensitivity of boosting to noise, tuning choices, and limited sample size under a fixed evaluation protocol [42]. These results reinforce the value of evaluating multiple model families under controlled preprocessing and resampling, rather than assuming a single "best" algorithm a priori.

Importantly, the reported signals should be interpreted as predictive rather than causal. For instance, higher key-player minutes may partly reflect selection and match-management decisions (e.g., stronger line-ups remaining on the pitch longer) and may also embed contextual influences not explicitly modeled, such as injuries, opponent quality, or match status [1]. Future extensions with larger samples and richer contextual covariates are therefore necessary to assess the stability and generalizability of these relationships.

7 Limitations and Future Work

This study is explicitly a pilot and should be interpreted within its empirical scope. The modeling dataset includes 82 matches, limiting statistical power and external validity. The predictor set is intentionally compact and match-level to support feasibility and interpretability, but it likely omits important contextual drivers known to influence outcomes, such as home/away effects, opponent strength, match status, injuries and availability constraints, fixture congestion, and within-match event dynamics.

Evaluation is reported using accuracy and ROC AUC without specifying an operational threshold or cost-sensitive objective. In practical deployment, threshold selection depends on decision context and risk preferences; additional analyses such as calibration assessment, threshold optimization, and utility-based evaluation would typically be required. Future work should:

- expand the dataset across seasons, teams, and competitions;
- incorporate controlled contextual covariates (e.g., location, opponent quality, match importance); and

- adopt temporal validation schemes that better approximate real-world forecasting (training on earlier fixtures and testing on later ones).

8 Conclusions

This paper proposes a reproducible decision-support workflow for match-point classification in football using match-level engineered predictors derived from InStat. A unified tidymodels framework integrates standardized preprocessing, resampling, and hyperparameter tuning, enabling controlled comparison across multiple algorithmic families. In the evaluated pilot setting, the Multi-Layer Perceptron and Support Vector Machine models achieve the strongest out-of-sample performance, with the MLP attaining the highest discriminative capacity (ROC AUC = 0.804) and best calibration (Brier score = 0.207), while tree-based ensembles show more modest results and boosted trees perform close to random.

Global feature-importance diagnostics consistently identify average squad age as the dominant predictive signal, with minutes played by core contributors emerging as secondary drivers of point attainment. Tactical deployment variables exhibit limited marginal influence when considered independently, reinforcing the importance of multivariate, non-linear modeling for capturing interactions between structural and exposure-related factors. Overall, the findings support the feasibility of compact and interpretable match-level predictors for exploratory decision support, while highlighting the need for larger samples and richer contextual information before operational deployment.

Acknowledgement

Data processing and analysis in this paper were supported by the Competitiveness Operational Program Romania, under project SMIS 124759 - RaaS-IS (Research as a Service Iasi).

References

[1] J. Strobl, A. Malley, and G. Tutz, "An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests," *Psychological Methods*, vol. 14, no. 4, pp. 323–348, 2009, doi: 10.1037/a0016973.

[2] M. Dixon and S. Coles, "Modelling association football scores and inefficiencies in the football betting market," *Applied Statistics*, vol. 46, no. 2, pp. 265–280, 1997.

[3] M. Crowder, M. Dixon, A. Ledford, and M. Robinson, "Dynamic modelling and prediction of English football league matches for betting," *The Statistician*, vol. 51, no. 2, pp. 157–168, 2002, doi: 10.1111/1467-9884.00258.

[4] J. Lasek, Z. Szlávik, and S. Bhulai, "The predictive power of ranking systems in association football," *International Journal of Applied Pattern Recognition*, vol. 1, no. 1, 2013, doi: 10.1504/IJAPR.2013.052339.

[5] R. Baboota and H. Kaur, "Predictive analysis and modelling football results using machine learning approach for English Premier League," *International Journal of Forecasting*, vol. 35, no. 2, pp. 741–755, 2019, doi: 10.1016/j.ijforecast.2019.01.004.

[6] K. Langaroudi and M. Yamaghani, "Sports result prediction based on machine learning and computational intelligence approaches: A survey," *Journal of Advances in Computer Engineering and Technology,* vol. 5, no. 1, pp. 27–36, 2019.

[7] H. Liu, W. G. Hopkins, and M. A. Gómez, "Machine learning models reveal key performance metrics of football players to win matches in Qatar Stars League," *International Journal of Sports Science & Coaching*, vol. 16, no. 2, pp. 227–236, 2021.

[8] M. D. Hughes and R. M. Bartlett, "The use of performance indicators in performance analysis," *Journal of Sports Sciences*, vol. 20, pp. 739–754, 2002.

[9] J. B. Taylor, S. D. Mellalieu, N. James, and D. A. Shearer, "The influence of match location, quality of opposition, and match status on technical performance in

professional association football," Journal of Sports Sciences, vol. 26, no. 9, pp. 885–895, 2008, doi: 10.1080/02640410701836887.

[10] J. Castellano, D. Casamichana, and C. Lago, "The use of match statistics that discriminate between successful and unsuccessful soccer teams," *Journal of Human Kinetics*, vol. 31, pp. 139–147, 2012.

[11] C. Ruiz-Ruiz, L. Fradua, Á. Fernández-García, and A. Zubillaga, "Analysis of entries into the penalty area as a performance indicator in soccer," *European Journal of Sport Science*, vol. 13, pp. 241–248, 2013.

[12] D. Link and M. de Lorenzo, "Examining shooting performance and decision making in soccer using expected goals," in *Proc. MIT Sloan Sports Analytics Conference*, Boston, MA, USA, 2016.

[13] M. Brechot and R. Flepp, "Dealing with randomness in match outcomes: Rethinking performance evaluation in European club football using expected goals," *Journal of Sports Economics*, vol. 21, pp. 335–362, 2020.

[14] M. Cavus and P. Biecek, "Explainable expected goal models for performance analysis in football analytics," arXiv:2206.07212, 2022.

[15] A. Bialkowski et al., "Large-scale analysis of soccer matches using spatiotemporal tracking data," in *Proc. IEEE International Conference on Data Mining*, Shenzhen, China, 2014, pp. 725–730.

[16] F. R. Goes, S. W. M. Brink, A. Elferink-Gemser, and C. Visscher, "The tactics of successful attacks in professional soccer: Large-scale spatiotemporal analysis of attacking patterns," *International Journal of Sports Science & Coaching*, vol. 14, no. 2, pp. 241–250, 2019.

[17] H. Broich et al., "Statistical analysis for the first Bundesliga in the current soccer season," *Progress in Applied Mathematics*, vol. 7, pp. 1–8, 2011.

[18] C. Yang et al., "Exploring the effect of match situations on the technical and physical performance in elite soccer players," *Journal of Sports Sciences*, 2018.

[19] R. Poli, L. Ravenel, and R. Besson, "Ten years of demographic analysis of the football players' labour market in Europe," *Football Observatory Monthly Report*, no. 39, 2018.

[20] E. Wakelam, V. Steuber, and J. Wakelam, "The collection, analysis and exploitation of footballer attributes: A systematic review," *Journal of Sports Analytics*, pp. 1–37, 2022.

[21] C. Lago, "The influence of match location, quality of opposition, and match status on possession strategies in professional association football," *Journal of Sports Sciences*, vol. 27, no. 13, pp. 1463–1469, 2009, doi: 10.1080/02640410903131681.

[22] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, version 4.5.2, 2025. Available: https://www.R-project.org.

[23] H. Wickham, M. Averick, J. Bryan, W. Chang, L. D'Agostino McGowan, R. François, G. Grolemund, A. Hayes, H. Henry, and J. Hester, "Welcome to the tidyverse," *Journal of Open Source Software*, vol. 4, no. 43, Art. 1686, 2019, doi: 10.21105/joss.01686.

[24] D. D. Sjoberg, K. Whiting, M. Curry, J. A. Lavery, and J. Larmarange, "Reproducible summary tables with the gtsummary package," *The R Journal*, vol. 13, pp. 570–580, 2021, doi: 10.32614/RJ-2021-053.

[25] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*, Springer, New York, NY, USA, 2016.

[26] J. W. Tukey, "We need both exploratory and confirmatory," *The American Statistician*, vol. 34, no. 1, pp. 23–25, 1980, doi: 10.2307/2682991.

[27] J. T. Behrens, "Principles and procedures of exploratory data analysis," *Psychological Methods*, vol. 2, no. 2, pp. 131–160, 1997, doi: 10.1037/1082-989X.2.2.131.

[28] I. Patil, "Visualizations with statistical details: The ggstatsplot approach," *Journal of Open Source Software*, vol. 6, no. 61, Art. 3167, 2021, doi: 10.21105/joss.03167.

[29] L. Hatcher, *Advanced Statistics in Research: Reading, Understanding, and Writing Up Data Analysis Results*, Shadow Finch Media, 2013.

[30] M. Kuhn and H. Wickham, "Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles," 2020. Available: https://www.tidymodels.org.

[31] M. Kuhn and J.Silge,Tidy Modeling with R: A Framework for Modeling in the Tidyverse. O'Reilly, Sebastopol, California, USA, 2022.

[32] M. Kuhn and K. Johnson, *Applied Predictive Modeling*, Springer, New York, NY, USA, 2013, doi: 10.1007/978-1-4614-6849-3.

[33] H. Frick, F. Chow, M. Kuhn, M. Mahoney, J. Silge, and H. Wickham, "rsample: General resampling infrastructure," R package version 1.3.1, 2025. Available: https://CRAN.R-project.org/package=rsample.

[34] M. Kuhn, H. Wickham, and E. Hvitfeldt, "recipes: Preprocessing and feature engineering steps for modeling," R package version 1.3.1, 2025. Available: https://CRAN.R-project.org/package=recipes.

[35] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, New York, NY, USA, 2001.

[36] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning with Sparsity*, CRC Press, Boca Raton, FL, USA, 2015.

[37] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," Journal of Statistical Software, vol. 33, no. 1, pp. 1–22, 2010, doi: 10.18637/jss.v033.i01.

[38] K. K. Tay, B. Narasimhan, and T. Hastie, "Elastic net regularization paths for all generalized linear models," *Journal of Statistical Software*, vol. 106, no. 1, pp. 1–31, 2023, doi: 10.18637/jss.v106.i01.

[39] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[40] M. N. Wright and A. Ziegler, "ranger: A fast implementation of random forests for high-dimensional data in C++ and R," *Journal of Statistical Software*, vol. 77, no. 1, pp. 1–17, 2017, doi: 10.18637/jss.v077.i01.

[41] R. E. Schapire and Y. Freund, "A short introduction to boosting," *Journal of the Japanese Society for Artificial Intelligence*, vol. 14, no. 5, pp. 771–780, 1999.

[42] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[43] T. Chen, T. He, M. Benesty, V. Khotilovich, and Y. Tang, "xgboost: Extreme gradient boosting," R package version 3.1.2.1, 2025. Available: https://CRAN.R-project.org/package=xgboost.

[44] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis, "kernlab: An S4 package for kernel methods in R," *Journal of Statistical Software*, vol. 11, no. 9, pp. 1–20, 2004, doi: 10.18637/jss.v011.i09.

[45] A. Karatzoglou, A. Smola, and K. Hornik, "kernlab: Kernel-based machine learning lab," R package version 0.9-33, 2024. Available: https://CRAN.R-project.org/package=kernlab.

[46] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, 4th ed., Springer, New York, NY, USA, 2002.

[47] M. Kuhn and H. Frick, "dials: Tools for creating tuning parameter values," R package version 1.4.2, 2025. Available: https://CRAN.R-project.org/package=dials.

[48] M. Kuhn, "tune: Tidy tuning tools," R package version 2.0.1, 2025. Available:

https://CRAN.R-project.org/package=tune.
[49] H. Bengtsson, "A unifying framework for parallel and distributed processing in R using futures," *The R Journal*, vol. 13, no. 2, pp. 208–227, 2021, doi: 10.32614/RJ-2021-048.
[50] M. Kuhn, D. Vaughan, and E. Hvitfeldt, "yardstick: Tidy characterizations of model performance," R package version 1.3.2, 2025. Available: https://CRAN.R-project.org/package=yardstick.
[51] B. Greenwell, B. Boehmke, and B. Gray, "vip:Variable importance plots," R package, 2020. Available: https://CRAN.R-project.org/package=vip.
[52] A. Cutler, R. D. Cutler, and J. R. Stevens, "Random forests," in *Ensemble Machine Learning*, C. Zhang and Y. Q. Ma, Eds. New York, NY, USA: Springer, 2012, pp. 157–175, doi: 10.1007/978-1-4419-9326-7_5.

Marin FOTACHE graduated from the Faculty of Economics at Alexandru Ioan Cuza University of Iasi, Romania in 1989. He holds a PhD diploma in Business Information Systems (Business Informatics) from 2000. He had gone through all teaching positions since 1990 when he joined the staff of Al. I. Cuza University, from teaching assistant in 1990, to full professor in 2002. Currently he is professor within the Department of Accounting, Business Informatics and Statistics in the Faculty of Economics and Business Administration at Alexandru Ioan Cuza University. He is the (co)author of books and journal/conference articles in areas such as SQL, database design, NoSQL, Big Data, Data Engineering and Machine Learning.

Irina COJOCARIU is a graduate of the Master's program in Business Information Systems (2020) and currently a PhD candidate at the Doctoral School of Economics and Business Administration, Iași, with the research theme "Predictive models applicable in sport and sports event management". She is a Teaching Assistant at the Faculty of Economics and Business Administration and Project Manager at the IT company Brandweb Design SRL (Iași, Romania). Her work addresses interdisciplinary research at the intersection of machine learning, football analytics, and data science, with an emphasis on reproducible workflows and decision-support applications.

Armand BERTEA is co-founder and Chief Technology Officer (CTO) of Brandweb Design SRL (Iași, Romania), where he leads the development of software solutions and e-commerce projects. He is also a co-founder of Football Business Inside and of The Football Brain platform, initiatives focused on digitalization and sports analytics. He holds a PhD and has participated in international excellence programs in football management, including Business Excellence in Football Management. His professional activity integrates research, entrepreneurship, and software engineering, with an emphasis on innovative technological applications, reproducible data workflows, and decision-support systems in competitive environments.