Analysis of 2022 – 2023 Football Players for Position Detection and Classification

Samet Kenar
Middle East Technical University
Ankara, Turkey

Abstract— **This paper's main goal is to analyze 2022 – 2023 Football Players Performance Metrics with various machine learning algorithms according to their position. A comprehensive feature engineering approach is utilized and formed several new variables under 10 categories. Five different models (Multinomial Logistic Regression, SVM, ANN, Random Forest and XGBoost) are compared and feature selection is applied with RFECV. The highest accuracy and generlization success is obtained with the SVM. This analysis provides valuable insights in terms of quantifying on-field roles.**

Keywords—Football Performance Metrics, Feature Engineering, PCA, Artificial Neural Network, Support Vector Machine, Random Forest, XGBoost, R, Python

I. INTRODUCTION

The modern football uses data-driven analysis methods to evaluate player performance more objectively. Understanding and correctly classifying the roles of players on the field plays an important role in the decision making processes of technical teams. This study aimed to predict the main positions of players using comprehensive individual performance data. After eliminating missing data, various machine learning models are trained to evaluate the classification performance for each position.

II. METHODOLOGY

A. *Dataset*

The dataset was obtained from Kaggle. The dataset's goal is to determine the performances of football players according to their various metrics. There are 2689 football players and a total of 124 variables in the dataset; the first variable is the "rank". There is no missing value (NA) in the data set. 5 variables are categoricel, the other variables are numerical. Categorical variables are: player, country, position, team, league (comp), and year of birth (born). "Rank" and "player" variables were removed from the analysis as they are like an identification number. In addition, the position variable was reduced to four groups as goalkeeper (GK), defense (DF), midfielder (MD) and forward (FW) for better understanding. Due to the high dimensionality of the data, a comprehensive feature engineering process was used to derive a total of 37 new variables from 118 performance metrics exlcuding age to better evaluate the performances of players. The remaining 118 numerical variables were classified into 10 different groups according to their definitions before starting the research. After the grouping process, Principal Component Analysis (PCA) was applied to each group and new variables were created.

The groups and newly created variables are given below. All of these features numeric and represent a score.

1) Playing Time and Participation into Game
   a. Playing Contribution (PC)
2) Shots and Goals
   a. Shot Efficiency Index (SEI)
   b. Penalty Impact Score (PIS)
   c. Shot Type Preference (STP)
   d. Free Kick Profile (FKP)
3) Basic Passess and Distance Base Passess
   a. Total Passing Volume (TPV)
   b. ShortvsLongPassProfile (SLPP)
   c. PassingAccuracy (PA)
   d. PassProfileVariation (PPV)
4) Luck and Creativity
   a. OffensivePlaymakerScore (OPS)
   b. ChainImpact (CI)
   c. OpenPlaayvsPieceCreativity (OPPC)
   d. DribbleandFoulCreation (DFC)
   e. IndividualCreativity (IC)
   f. CounterAttackInitiationScore (CAIS)
5) Dribbling & Take On
   a. DribbleActivityIndex (DAI)
   b. DribbleSuccessScore (DSS)
6) Ball Carrying
   a. BallProgressionIndex (BPI)
   b. RiskCarryingProfile (RCP)
   c. CarryEfficiencyScore (CES)
7) Contact and Touch with Ball
   a. OverallTouchActivity (OTA)
   b. TouchLocationIndex (TLI)
   c. FinalThirdPresence (FTP)
8) Defense Activities
   a. DefenseActivityIndex (DEI)
   b. DefensiveZoneRecoveryProfile (DZRP)
   c. DefensiveBlockingProfile (DBP)
   d. OnetooneDefensiveVulnerability (DV)
   e. PressingDefenderIndex (PDI)
   f. DefensiveReliabilityIndex (DRI)
9) Discipline and Faul Activities
   a. DisciplinaryAgressionIndex (DAGI)
   b. FouleandOffsideTendency (FOT)
   c. PenaltyLiabilityIndex (PLI)
   d. OffsidevsPenaltyOutcome (OPO)
   e. OwnGoalPropensity (OGP)
   f. FairPlayProfile (FPP)
10) Air Ball Activities
   a. AerialDuelActivity (ADA)
   b. AerialSuccessIndex (ASI)

B. *Descriptive Statistics*

Descriptive statistics are crucial tools in data analysis. They allow us to understand the data overall. Statistical measures such as mean, median, mod and quantiles describe the central tendency. Summaries of descriptive statistics for numerical variables are presented in Table 1.

Table 1. Descriptive Summary of Numerical Data

	Age	PC	SEI	PIS	STP	FKP
Min	15.00	-2.67	-1.96	-6.75	-15.62	-7.88
1st Qu.	23.00	-1.90	-1.44	-0.58	-0.62	-0.51
Med	26.00	-0.13	-0.43	0.02	-0.10	-0.05
Mean	26.06	-0.01	0.00	0.00	0.00	0.00
3rd Qu.	29.00	1.66	1.05	0.40	0.71	0.40
Max.	41.00	4.07	11.85	19.50	5.30	22.57
NA's	269	269	269	269	269	269

	TPV	SLPP	PA	PPV	OPS	CI
Min	-8.41	-9.68	-7.75	-9.28	-33.82	-12.68
1st Qu.	-1.83	-0.51	-0.46	-0.5	-0.61	-0.33
Med	0.02	0.19	0.14	-0.03	0.32	-0.27
Mean	0.00	0.00	0.003	0.00	0.002	-0.005
3rd Qu.	1.66	0.77	0.67	0.46	1.09	0.11
Max.	26.65	10.06	4.18	3.63	1.48	32.55
NA's	269	269	269	269	269	269

	OPPC	DFC	IC	CAIS	DAI	DSS
Min	-14.93	-15.59	-15.09	-22.93	-1.67	-5.63
1st Qu.	-0.10	-0.20	-0.17	-0.11	-0.88	-0.48
Med	0.04	0.13	-0.13	0.03	-0.25	-0.14
Mean	0.01	0.01	0.02	0.00	0.03	0.008
3rd Qu.	0.27	0.22	0.04	0.09	0.62	0.45
Max.	19.44	26.46	27.00	10.40	19.47	9.26
NA's	269	269	269	269	269	269

	BPI	RCP	CES	OTA	TLI	FTP
Min	-3.39	-7.19	-10.71	-4.47	-4.47	-4.47
1st Qu.	-1.19	-0.81	-0.36	-1.24	-1.22	-1.17
Med	-0.27	-0.13	-0.03	-0.04	-0.03	0.00
Mean	0.01	-0.01	0.00	0.00	0.00	0.01
3rd Qu.	0.85	0.72	0.33	1.03	1.04	1.01
Max.	24.45	9.92	8.35	18.23	18.23	18.23
NA's	269	269	269	269	269	269

	DEI	DZRP	DBP	DV	PDI	DRI
Min	-3.43	-6.05	-14.39	-18.63	-14.99	-13.47
1st Qu.	-1.51	-0.71	-0.47	-0.25	-0.41	0.00
Med	-0.04	0.05	0.20	0.03	0.07	0.17
Mean	-0.02	0.00	0.00	0.00	0.00	0.00
3rd Qu.	1.10	0.54	0.64	0.36	0.50	0.39
Max.	26.30	16.91	11.70	14.72	8.76	3.76
NA's	269	269	269	269	269	269

	DAGI	FOT	PLI	OPO	OGP	FPP
Min	-0.96	-10.85	-22.93	-3.80	-6.54	-12.16
1st Qu.	-0.44	-0.36	-0.11	-0.31	-0.27	-0.17
Med	-0.17	0.19	0.13	-0.16	-0.08	0.06
Mean	0.00	0.00	-0.01	0.00	0.00	0.00
3rd Qu.	0.13	0.61	0.36	0.00	0.15	0.34
Max.	42.50	6.40	7.71	20.96	15.59	8.37
NA's	269	269	269	269	269	269

	ADA	ASI
Min	-15.21	-2.17
1st Qu.	-0.51	-0.6
Med	0.07	-0.04
Mean	-0.01	0.00
3rd Qu.	0.67	0.63
Max.	1.74	11.92
NA's	269	269

Table 1 represents descriptive statistics for various numerical variables created to measure the performance of football players. Playing Contribution has a mean close to 0 and a range between -2.67 and 4.07. Although the means are generally close to 0 under the Shots and Goals category (SEI, PIS, STP, FKP), the distributions are quite wide and the extreme values are high; there are significant performance differences between players. For Basic Passess and Distance Bases Passess (TPV, SLPP, PA, PPV), the values appear balanced with means and medians close to 0, but the maximum and minimum extremes show large differences. The metrics under the Luck and Creativity (OPS, CI, OPPC, DFC, IC, CAIS) also have balanced means around zero but there are extreme values; this shows that the creativity of the players is quite diverse. Dribbling & Take on and Ball Carrying categories (DAI, DSS, BPI, RCP, CES) have similar means close to zero, but there are extreme values that indicate that these abilities are much higher in some players. For Contact and Touch with Ball category (OTA, TLI, FTP) have stable means around zero, but the presence of significant outliers indicates that there are differences in the players' touch efficiency. In the Defense Activities (DEI, DZRP, DBP, DV, PDI, DRI), the distributions are wide with means close to zero and it is seen that some players exhibit serious differences in terms of their defensive performance. For Discipline and Foul Activities (DAGI, FOT, PLI, OPO, OGP, FPP) attract attention with their means close to zero and wide distributions. It can be understood that some players exhibit risky behaviors in terms of discipline. Finally, in the Aerial Ball Activities (ADA, ASI) category, while the variables are expressed with means close to zero, it can be seen that there are signifcant performance differences between tha players in terms of aerial ball challenges. There are 269 missing values in all variables.

Table 2. Descriptive Statistics of Categorical Data

Levels	Nation	Squad	Comp	Born	Pos
1	370	34	490	242	964
2	365	34	550	220	878
3	230	34	565	212	683
4	206	33	540	207	164
5	184	33	544	200	
6	105	33		192	
Other	1229	2488		1416	

According to above table, for nation, the most players come from Spain (ESP, 370 players). Spain is followed by France (FRA, 365 players), Germany (GER, 230 players), Italy (ITA, 206 players) and England (ENG, 184 players). There are 1229 players from other countries. The teams with the most players are Hellas Verona, Schalke 04 and Cádiz (34 players each). These are followed by Ajaccio and Nottingham Foresst (33 players each). The distribution of players between the teams is generally balanced and no single team stands out clearly. For League (Comp), the most players come from France's Ligue 1 (565 players). This is followed by Spain's La Liga (550 players), Italy's Serie A (544 players), England's Premier League (540 players) and Germany's Bundesliga (490 players). The distribution of players between the leagues is generally quite balanced, but France's Ligue 1 stands out as the most densely populated league. The most players were born in 1997 (242 players). This is followed by players born in 1996 (220 players), 2000, (212) players, 1998 (207 players) and 1999 (200 players). The majority of players belong to the young age group. The majority of players are Defenders (DF, 964 players). Midfielders (MF) are second (878 players), forwards (FW) are third (683 players) and goalkeepers (GK) are the least represented position (164 players) naturally. It can be seen that the general distributions of players in terms of countries, teams and leagues are balanced but clearly concentrated in some categories. In the position distribution, the weight of defenders is quite evident.

C. *Exploratory Data Analysis*

In this study, 6 different research questions were adressed to examine the structure of the data set and the relationship between performance metrics.

1) Does the "Playing Contribution" score differ significantly by player position?

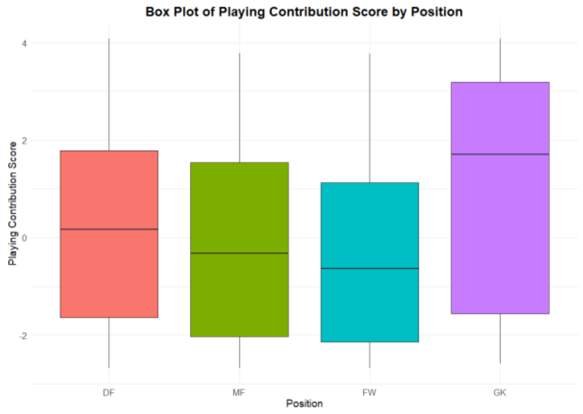

Fig 1.

When the boxplot is examined, there seems differences in the "Playing Contribution" scores between player positions. In particular, goalkeepers (GK) seem to have higher average contribution scores compared to other positions (DF, MF and FW). Forwards (FW), midfielders (MF) and defenders (DF) have lower contribution scores at closer levels. Although the distributions are generally wide, goalkeepers have higher median and quartile values than other positions.

Table 3. Kruskal-Wallis Rank Sum Table

Kruskal-Wallis Rank Sum Test		
Data: Playing Contribution by Position		
Chi-Sqr = 74.305	df = 3	p-value = 5.107e-16

According to the Kruskal-Wallis test results, Playing Contribution scores between the player positions differ significantly. This clearly shows that there are statistically significant differences in terms of "Playing Contribution" between different positions.

Table 4. Post-Hoc Wilcoxon Table

Pairwise Comparisons sign Wilcoxon rank sum test with continuity correction			
Data: Playing Contribution by Position			
	DF	MF	FW
MF	5.1e-4	-	-
FW	1.2e-7	6.76e-2	-
GK	1.4e-6	1.73e-10	1.70e-12
P value adjustment method: holm			

The post-hoc Wilcoxon test results were used to determine which positions these differences occurred between by p-value threshold 0.05. There is a significant difference between DF and MF. Also, the difference between DF and FW seems very strong. Interestingly, there is no significant difference between MF and FW since the p-value (0.0676) is greater than 0.05.

2) Is there a difference in the average Shot Efficiency Index and Passing Accuracy between players in different leagues?

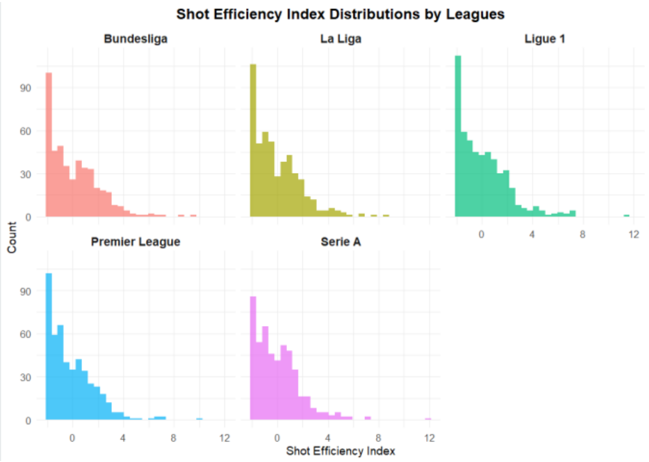

Fig 2.

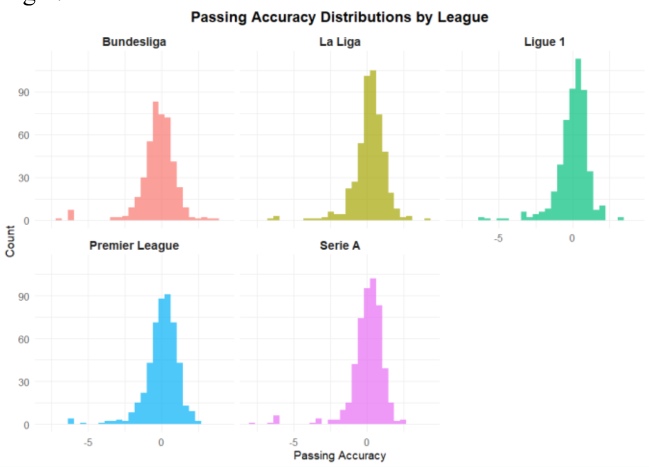

Fig 3.

For Shot Efficiency Index (Fig 2.), similar right-skewed distributions are observed across all leagues, but the values ofr Ligue 1 and Serie A players apper to be slightly more right-skewed than the other leagues. This may indicate that some players shoot with higher efficiency there. However, the overall pattern is quite similar.

For Passing Accuracy (Fig 3.), Ligue 1 players have a visually clear average pass accuracy. Bundesliga and Premier League players are concentrated at lower values. This suggest that there may be differences between leagues.

Table 5. Multivariate Normality Table

Multivariate Normality				
Test	Test Statistic	P-Value	Method	MVN
Henze-Zirkler	8.718	<0.001	asymptotic	Not Normal

Since test statistics is 8.718 and p < 0.001, the data are not multivariate normally distributed.

Table 6. Univariate Normality Table

Univariate Normality				
Test	Variable	Statistic	P-Value	Normality
Anderson-Darling	Shot Efficiency Index	13.797	<0.001	Not Normal
Anderson-Darling	Passing Accuracy	9.736	<0.001	Not Normal

Both Shot Efficiency Index and Passing Accuracy are not normally distributed. Therefore, it is appropriate to use non-parametric tests instead of parametric tests.

Table 7. Kruskal-Wallis Rank Sum Table

Kruskal-Wallis Rank Sum Test		
Data: Shot Efficiency Index by Comp		
Chi-Sqr = 3.0898	df = 4	p-value = 0.5429

There is no significant difference in Shot Efficiency Index according to leagues since p-value = 0.5429 > 0.05.

Table 8. Kruskal-Wallis Rank Sum Table

Kruskal-Wallis Rank Sum Test		
Data: Passing Accuracy by Comp		
Chi-Sqr = 54.859	df = 4	p-value = 3.478e-11

There is a significant difference in terms of passing accuracy according to leagues since p-value = 3.748e-11 < 0.001.

Table 9. Post-Hoc Wilcoxon Table

Pairwise Comparisons sign Wilcoxon rank sum test with continuity correction				
Data: Passing Accuracy by Comp				
	Bundesliga	La Liga	Ligue 1	Premier League
La Liga	4.6e-12	-	-	-
Ligue 1	5.61e-6	2.4e-1	-	-
Premier League	1.99e-5	2.7e-1	1.00	-
Serie A	3.31e-6	2.7e-1	1.00	1.00
P value adjustment method: bonferroni				

According to Post-Hoc Wilcoxon rank sum test, La Liga is significantly different from all other leagues. Also, there is no significant difference between Ligue 1, Premier League and Serie A. The data does not show a significant difference between leagues in terms of shooting efficiency. This suggest that players in different leagues have similar shooting efficiency levels. In constrast, there are statistically significant differences in terms of passing accuracy. In particular, La Liga players have significantly higher pass accuracy rates than others. Ligue 1, Premier League and Serie A exhibit similar passing performances. These results suggest that the style of play or tactical structures of some leagues may have an impact on passing success.

3) Do individual skill indicex – such as Dribble Success Score and Carry Efficiency Score – decrease as age increases?

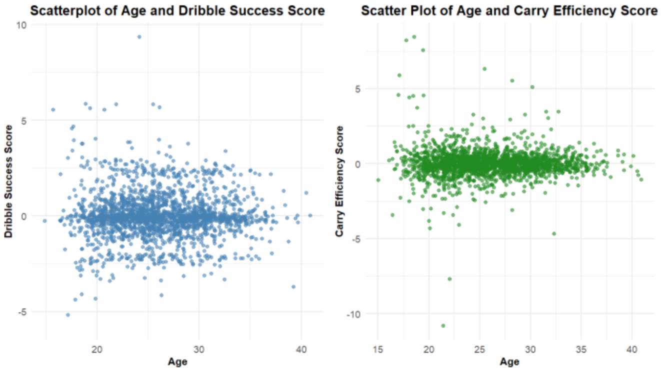

Fig 4.

The scatter plot does not show a clear trend of Dribbling Success scores decreasing with age. The values are quite spread out and there is no clear pattern between ages. Similarly, Ball Carrying Efficiency scores do not appear to decrease with age. In the graph, the values of young and ol players are distributed similarly.

Table 10. Shapiro Wilk Normality Table

Shapiro-Wilk Normality Test	
Data: Dribble Success Score	
W = 0.92781	p-value = 2.2e-16

Table 11. Shapiro Wilk Normality Table

Shapiro-Wilk Normality Test	
Data: Carry Efficiency Score	
W = 0.78836	p-value = 2.2e-16

For both variables, p-value < 0.001, which means the data are not normally distributed. Therefore, nonparametric correlation analysis (Spearman) was utilized.

Table 12. Spearman's Rank Correlation Table

Spearman's Rank Correlation Rho	
Data: Dribble Success Score and Age	
S = 3151945756	p-value = 0.1563
Alternative hypothesis: true rho is not equal to 0	
Sample estimates:	
Rho	
0.02734765	

Table 13. Spearman's Rank Correlation Table

Spearman's Rank Correlation Rho	
Data: Carry Efficiency Score and Age	
S = 3231893366	p-value = 0.8897
Alternative hypothesis: true rho is not equal to 0	
Sample estimates:	
Rho	
0.00267688	

For both Dribble Success Score and Carry Efficiency Score, there is no statistically significant relationship with age since the p-values are greater than 0.001. Both scatter plots visually showed that there was no significant relationship between age and individual skill scores. This impression was also statistically confirmed by Spearman correlation analysis. The rho values were very closed to zero and p-values were well above the significant limit. These results shows us these individual skills do not automatically decline with age; other factors such as tranining, playing style or position may have a greater impact on these skills.

4) How different are players offensively, as measured by Shot Efficiency Index, Offensive Playmaker Score and Ball Progression Index according to their position?

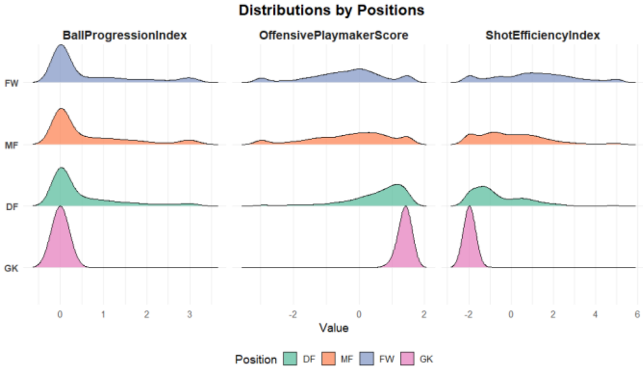

Fig 5.

According to the graph above, for Ball Progression Index, MF (midfielder) players seems the most effective group in terms of Ball Progression. FW players (forward) are behind MF. For Offensive Playmaker Score, the highest contribution in terms of OPS comes from MF players. FWs are in the middle, DF and GK are at the lowest. For Shot Efficiency Index, FW players are clearly ahead in terms of shooting efficiency. The other positions are much lower.

Table 14. Kruskal-Wallis Rank Sum Table

Kruskal-Wallis Rank Sum Test		
Data: Shot Efficiency Index by Position		
Chi-Sqr = 654.35	df = 3	p-value = 2.2e-16

Table 15. Kruskal-Wallis Rank Sum Table

Kruskal-Wallis Rank Sum Test		
Data: Ball Progression Index by Comp		
Chi-Sqr = 270.53	df = 4	p-value = 2.2e-16

Table 16. Kruskal Wallis Rank Sum Table

Kruskal-Wallis Rank Sum Test		
Data: Offensive Play Maker Score by Comp		
Chi-Sqr = 599.56	df = 4	p-value = 2.2e-16

In all tests (for all three variables), there are statistically significant differences between positions as the p-value is <0.001.

Table 17. Post-Hoc Wilcoxon Table

Pairwise Comparisons sing Wilcoxon rank sum test with continuity correction			
Data: Shot Efficiency Index by Position			
	DF	MF	FW
MF	2.0e-16	-	-
FW	2.0e-16	2.0e-16	-
GK	2.0e-16	2.0e-16	2.0e-16
P value adjustment method: bonferroni			

According to Post-Hoc Wilcoxon Test Results for Shot Efficiency Index, there is a significant difference between all position pairs. This test shows that especially FW (forward) players clearly differ from all other positions in terms of shooting efficiency.

Table 17. Post-Hoc Wilcoxon Table

Pairwise Comparisons sing Wilcoxon rank sum test with continuity correction			
Data: Ball Progression Index by Position			
	DF	MF	FW
MF	0.542	-	-
FW	0.966	0.064	-
GK	2.0e-16	2.0e-16	2.0e-16
P value adjustment method: bonferroni			

There is no significant difference between MF and FW players meaning that these two groups have similar performance in terms of ball progression. GK players significantly different from all groups.

Table 18. Post-Hoc Wilcoxon Table

Pairwise Comparisons sing Wilcoxon rank sum test with continuity correction			
Data: Offensive Play Maker Score by Position			
	DF	MF	FW
MF	2.0e-16	-	-
FW	2.0e-16	0.023	-
GK	2.0e-16	2.0e-16	2.0e-16
P value adjustment method: bonferroni			

MF players produced significantly higher scores than FW and DF players. There is a difference between MF and FW, indicating that MF players have more creative playmaking roles. Overall results shows that the offensive performances of the players differ significantly according to the position. FW (forward) players are the most successful group in shooting efficiency while MF (midfielder) players clearly stands out in offensive play making. In terms of ball carrying, midfielders and forwards are similarly successful.

5) How does the relationship between the Disciplinary Agression Index and the Fair Play Profile vary across player positions, and what role does touch location play?

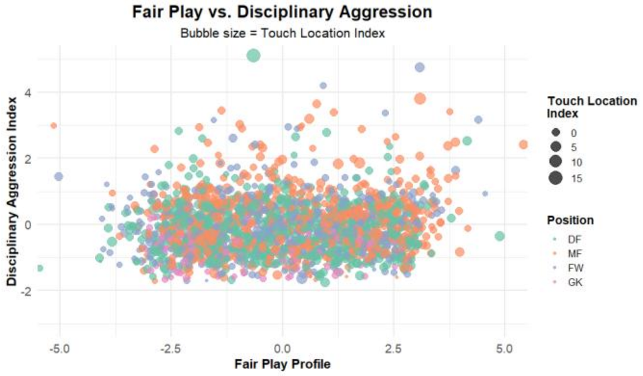

Fig 6.

According to figure , generally, disciplinary agression appears to decrease as Fair Play scores increases slightly. Also, MF and FW positions show more widespread distribution and players with larger Touch Location Indexes. Finally, GK's have lower density and generally lower agression-fairplay profiles.

Table 19. Robust Regression Results for Fair Play Profile

Term	Estimate	Std.Error	T value
(Intercept)	-0.0174	0.0147	-1.18
DisciplinaryAgressionIndex	-0.1222	0.0094	-13.05
Pos4MF	0.2643	0.0202	13.07
Pos4FW	0.0416	0.0252	1.65
Pos4GK	-0.4978	0.1122	-4.44
TouchLocationIndex	0.0168	0.0092	1.82
DisciplinaryAgressionIndex:pos4MF	0.4518	0.0153	29.60
DisciplinaryAgressionIndex:pos4FW	0.5095	0.0218	23.38
DisciplinaryAgressionIndex:pos4GK	-0.4088	0.1689	-2.42
DisciplinaryAgressionIndex:TouchLocationIndex	0.0311	0.0050	6.21
Pos4MF:TouchLocationIndex	0.0032	0.0123	0.26
Pos4FW:TouchLocationIndex	0.0538	0.0149	3.62
Pos4GK:TouchLocationIndex	-0.0116	0.0678	-0.17
DisciplinaryAgressionIndex:pos4MF:TouchLocationIndex	-0.0603	0.0080	-7.57
DisciplinaryAgressionIndex:pos4FW:TouchLocationIndex	-0.0681	0.0086	-7.94
DisciplinaryAgressionIndex:pos4GK:TouchLocationIndex	-0.0407	0.1063	-0.38

This robust regression analysis is applied to examine the relationship between Disciplinary Agression Index and Fair Play Profile by condisering the effects of player positions and Touch Location Index. According to the results, in general, as aggression increases fair play profile decreases ($\beta = -0.1222$, $p < 0.001$). However this relationship changes significantly according to position. While the effect of agression is weaker in midfielders and forwards (interaction $\beta > 0$, $p < 0.001$), this effect is more negative in goalkeepers. In addition, a high Touch Location Index reduces the negative effect of aggression on fair play; this effect is especially evident in MF and FW positions (triple interaction $p < 0.001$). The findings show that position and on

D. Missingness

Since there were no missing observations in the original data set, missing data are created. There are 2689 observations and 38 numerical variables in the data set, and 269 missing values are created randomly in each numerical variable. This structure creates approximately 10% missigness for each variable. MCAR test is applied to test whether the missing data are distributed in a completely random manner, and the obtained p-value statistically confirmed that the missing data are MCAR.

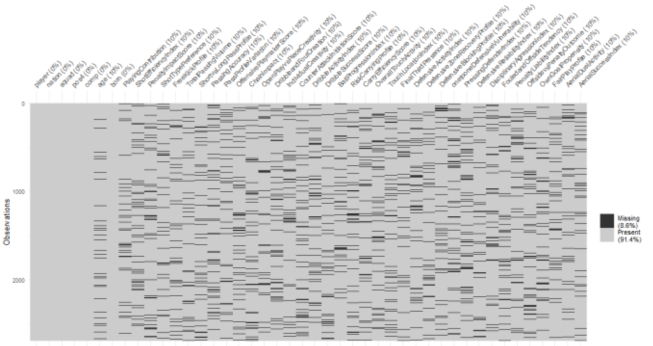

Fig 7. Missingness Mechanism Plot

In addition, the Missigness Mechanism Plot presented in the figure shows the distribution of missing observations in the data set according to variables. Black areas represent missing values. In the graph, missingnesses shows a random pattern rather than being concentrated in a specific variable. There is no systematic pattern and missing data are homogeneously distributed among the variables. These findings show that missing data can be used safely.

Table 20. The Output of MCAR Test

	statistic	df	p.value	missing.patterns
1	33880	92702	1	2330

After EDA, multiple imputation method, Predictive Mean Matching (PMM), is carried out through MICE package. PMM based on estimating the values of each variable according to similar observations. By doing so, missing data imputation has high statistical consistency and appropriateness.

E. *Modelling*

1. *Multinomial Regression*

For position classification, the Multinomial Logistic Regression model is applied. The hyperparameters of the model are optimized with the GridSearchCV method and the best reesults are obtained with the values C: 0.01, solver: lbfgs and max_iter: 500. First, in the model where all 38 numerical variables are included, the training set accuracy rate is 80.3%, the test accuracy is 77.3%, f1 scores for train and test sets are 80.3% and 77.4% respectively, and Cohen's Kappa values are 0.7176 for train and 0.6743 for test. The model is particularly successful in distinguishing all positions; while it also showed balanced and strong classification.

To improve models performance, Recursive Feature Elimination with Cross-Validation applied to optimize the number of variables in the model. 36 variables are found to be sufficient for the best performance. Individual Creativity (IC) and Own Goal Propensity (OGP) are removed due to their low contribution to the model. The training accuracy of the model created using selected features is calculated as 80.2%, test accuracy as 77.1%, f1 scores as 80.2% (train) and 77.2% (test), and kappa values as 0.7156 (train) and 0.6716 (test). After RFECV, the model did not show a decrease in performance despite the reduced number of variables. This shows us the selected variables carry sufficient information in terms of position classification.

2. *Support Vector Machine*

Support Vector Machines (SVM) method is utilized as the second model in position classification. As a result of GridSearchCV applied for hyperparameter optimization, the best parameter combination is determined as C: 1, gamma: auto and kernel: rbf. In the model where all variables are used, the accuracy rate in the training test is 89.9% and the test accuracy is 80.3%. The f1 scores of the model are 89.9% (train) and 80.5% (test) respectively. Cohen's Kappa values are 0.8545 (train) and 0.7166 (test). The model is successful and a high and balanced classification performance is observed in all positions.

As a result of the RFECV method applied to reduce the complexity of the model and obtain a more efficient feature set, the optimum number of variables are determined as 35. During this process, variables Penalty Impact Score (PIS), Open Play vs Piece Creativity (OPPC), and Individual Creativity (IC), which have low exploratory power, removed from the model. The test accuracy of the model re-established with the selected features are calculated as 89.96% train accuracy, 81.04% test accuracy, train f1 score as 89.98%, test f1 score as 81.19% and Cohen's Kappa Values as 0.8559 and 0.7275 respectively for both train and test sets. These results show that higher performance is achieved with fewer variables and the generalization of the model is increased.

3. *Artificial Neural Networks*

The third model used for position classification is Artifical Neural Network (ANN). The model is implemented with a multilayer perceptron (MLPClassifier) structure and as a result of hyperparameter optimization, the best result is obtained with the parameters alpha: 0.01, hidden layer sizes: 100 and solver: adam. In the model where all variables are used, training accuracy is calculated as 100%, test accuracy as 76.2%, training f1 score as 100%, test f1 score as 76.15%, training Cohen's Kappa Value as 1, and test kappa value as 0.6588. While giving excellent results in the training set, the decrease in performance in the test set shows that the model is seriously overfitting.

To get rid of this overfitting problem and establish a more balanced structure, Recursive Feature Elimination with Cross-Validation (RFECV) method is applied and 33 variables are selected. In this selection process, Counter Attack Initation Score (CAIS), Overall Touch Activity (OTA), Touch Location Index (TLI), Final Third Presence (FTP) and Open Play vs Piece Creativity (OPPC) variables are excluded from the model. In the model retrained after RFECV, training accuracy is 83.73%, test accuracy is 79% ,traning f1 score is 83.76%, test f1 score is 79.16%, training kappa value is 0.7665 and test kappa value is 0.6992. These values show that recursive feature selection increases the generalization ability of the model and significantly reduces the overfitting.

4. Random Forests

Another method applied for position classification is the Random Forest (RF) algorithm. As a result of hyperparameter optimization, the best results are obtained with the parameters n-estimators: 500, max-depth: None, max_features:log2, and min-samples-leaf: 2. The training accuracy of the model trained with all variables are calculated as 99.81%, test accuracy as 76.95%, training F1 score as 99.81%, test F1 score as 77.22%, training Cohen's Kappa Value as 0.9973 and test Cohen's Kappa Value as 0.6687. These results show that the model provides very high success on the training data but has a tendency to overfitting by showing lower performance on the test data.

In order to improve this situtation and obtain a simpler model, the RFECV (Recursive Feature Elimination with Cross-Validation) method applied. The RFECV result revealed that the most suitable number of variables for the model is 38. This shows that all independent variables utilized at the beginning were fond to be significant for the mode and none of them are eliminated. The training accuracy of the model created after RFECV is calculated as 98.05%, test accuracy as 76.80%, training F1 score as 98.07%, test F1 score as 77.10%, training Kappa Value as 0.9617 and test Kappa Value as 0.6663. These results show that the model has largely maintained its performance.

5. XgBoost

Last model used for position classification is the XGBoost algorithm. As a result of hyperparameter optimization, the best performance of the model is obtained with the following parameters, colsample-bytree:0.6, learning-rate:0.1, max-depth:3, n-estimators:200, and subsample:0.8. The training accuracy of the model trained with all variables are calculated as 97.68%, test accuracy as 79.74%, training F1 score as 97.67%, test F1 score as 79.82%, training Cohen's Kappa Value as 0.9666 and test Cohen's Kappa Value as 0.7088. These results show that the model fits the training data quite well and also shows high performance in test data.

To obtain a simpler and more generalizable modele, the RFECV method applied and the optimum number of variables are determined as 31. The removed 7 variables are age, Open Play vs Piece Creativity, Touch Location Index, Penalty Impact Score, Overall Touch Activity, Offside vs Penalty Outcome and Counter Attack Initiation Score. In the XGBoost model trained with these selected features, training accuracy is 97.35%, test accuracy is 79.18%, training F1 score is 97.35%, test F1 Score is 79.22%, training Kappa Value is 0.9620 and test Kappa Value is 0.7007. As a result of the RFECV process, despite the reduction in the number of variables in the model, only a very small decrease in performance is observed, which revealed that the model maintained a strong generalization ability.

III. RESULTS

Table 20. Forward (FW) Position RFECV Results

	Precision	Recall	F1-Score	Support
Multinomial Logistic Regression	0.76	0.73	0.74	136
SVM	0.78	0.71	0.74	136
ANN	0.72	0.71	0.72	136
Random Forests	0.80	0.71	0.75	136
XGBoost	0.76	0.71	0.73	136

Table 21. Midfielder (MF) RFECV Results

	Precision	Recall	F1-Score	Support
Multinomial Logistic Regression	0.68	0.71	0.69	176
SVM	0.71	0.82	0.76	176
ANN	0.70	0.77	0.73	176
Random Forests	0.64	0.75	0.69	176
XGBoost	0.71	0.76	0.73	176

Table 22. Defender (DF) RFECV Results

	Precision	Recall	F1-Score	Support
Multinomial Logistic Regression	0.84	0.82	0.83	193
SVM	0.92	0.84	0.88	193
ANN	0.91	0.83	0.87	193
Random Forests	0.86	0.79	0.82	193
XGBoost	0.86	0.85	0.86	193

Table 23. GoalKeeper RFECV Results

	Precision	Recall	F1-Score	Support
Multinomial Logistic Regression	1.00	1.00	1.00	33
SVM	1.00	1.00	1.00	33
ANN	0.94	1.00	0.97	33
Random Forests	1.00	1.00	1.00	33
XGBoost	1.00	1.00	1.00	33

When the model performance metrics obtained after RFECV for all positions are evaluated, it is seen that the classification success of model varies according to the position. For the goalkeepere (GK) position, all models showed an exceptional performance by reaching 100% f1-score; this shows that goalkeeper features are more discriminatory compared to other positions. For the defense (DF) position, the most successful model is SVM with 0.92 precision and 0.88 f1-score, while ANN and XGBoost models also performed with similar high accuracy. In the midfield (MF) position, the SVM model gives the strongest results with 0.76 f1-score, followed by ANN and XGBoost. In this position, random forests showed limited success with a lower f1-score compared to other models. For the forward classification, the random forest model stood out with an f1-score of 0.75, while the f1-sccore values of the other models ranged from 0.72 to 0.74. In general, the SVM model sstand out especially in the midfield and defense classifications, while the XGBoost and ANN models attracted attention with their balancing performances. These results reveal that each position has its own distinctive features and that position-based strategies can play an important role in model selection.

Where the results of five different classification models generated using the variables selected with RFECV are evaluated in general, the Support Vector Machines (SVM) model stands out as the method that shows the most balanced and powerful performance in all metrics. This model provided high train accuracy (0.8996) and test accuracy (0.8114) and also kept the classification performance at the highest level with train f1-score (0.8998) and test f1-score (0.8119) values. Cohen's kappa value is also 0.7275 which is one of the most successful results in terms of classification. The ANN model shows high performance in terms of test accuracy (0.79) and test f1-score (0.7916); however, it falls slightly behind SVM with train accuracy (0.8373). The XGBoost model also achieved high learning success such as train (0.9735), it showed a performance close to ANN with accuracy (0.7918) and f1-score (0.7922) on the test set. However, the train-test set difference is relatively high, the risk is overfitting is considerable. The Logistic Regression model showed a more modest performance with (0.77) accuracy and (0.7721) f1-score on the test set. Finally the Random Forest model clearly shows overfitting with (0.99) training accuracy. The test accuracy is (0.76) and the f1-score (0.7722). In addition, the kappa value is one of the lowest with 0.6687.

When we look at the general picture, the SVM model stands out as the most consistent and successful algorithm in the position classification with its high test accuracy, strong f1-score and balanced kappa value.

IV. CONCLUSION

In this report, various machine learning models are applied using a football dataset consisting of a large number of performance metrics to classify the field position of football players. First, comprehensive feature engineering for 10 categories and 118 numerical variables applied with PCA to each category to make inferences in a more convenient way. Then, missing values are created and completed with the PMM (Predictive Mean Matching). Reduced feature set consisting of 45 variables optimized with the Recursive Feature Elimination with Cross-Validation (RFECV) andn the most efficient subsets of variable obtained for each model.

In the comparative analysis of the models, SVM is the most successful method. ANN and XGBoost models also attracted attention. In contrast, Random Forest model experienced a serious overfitting problem, while Logistic Regression provided relatively lower but balanced results. Consequently, advanced machine learning methods applied together with feature engineering provided football position classification.

V. DISCUSSION

The performance metrics used in this study have shown that they provide significant differences in the classification of football positions. Feature selection with RFECV has both increased the accuracy of the models and simplified the model by eliminating unnecessary variables. While the SVM model provided the most consistent and high performance in all metrics; ANN and XGboost gave similar results. In contrast, the Random Forest model overfitted the training data and showed lower success in the test set. In general, it has been observed that the selected features are at a level that will support tactical and technical distinction and that SVM is a strong method for such classification problems.