# A Linear Algebra Approach to Football Team Ranking in English Premier League Using the Massey Method and Matrix Operations

Rava Khoman Tuah Saragih - 13524107
*Program Studi Teknik Informatika*
*Sekolah Teknik Elektro dan Informatika*
*Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia*
[1]13524107@itb.ac.id, [2]ravasaragih7@gmail.com

***Abstract*—The accurate ranking of sports teams is a fundamental problem in analytics, often limited by traditional point-based standings that ignore strength of schedule and margin of victory. This paper presents a linear algebraic approach to ranking English Premier League teams for the 2023-2024 season using the Massey Method. By modeling match outcomes as a system of linear equations where rating differences equate to goal differentials, we construct a least-squares problem to minimize prediction error. The study details the mathematical construction of the Massey Matrix (*M*) and the resolution of its inherent singularity through boundary constraints ($\sum r_i = 0$). Using Python and NumPy to process the full 380-game schedule, we derive continuous ratings that contrast with official standings. Results reveal significant discrepancies, highlighting the efficiency luck of Aston Villa and the statistical superiority of Newcastle United. Furthermore, the method offers a "pure" metric that corrects for administrative distortions, such as Everton's points deduction. This research bridges abstract linear algebra concepts, specifically vector spaces and matrix decomposition, with practical applications in data science.**

***Keywords*—English Premier League, Graph Laplacian, Least Squares, Linear Algebra, Massey Method, Matrix Operations, Sports Analytics, Strength of Schedule.**

## I. Introduction

The quantitative evaluation of competitive performance in sports is a discipline that bridges statistics, computer science, and mathematics. In the context of association football, specifically the English Premier League (EPL), the primary metric for success is the league table, which ranks teams based on points accumulated from match results. While this system is intuitively understood and generates excitement, it is mathematically imperfect as a measure of true team strength. The standard "3-1-0" points system treats all wins as equal, ignoring the context of the opponent's quality and the dominance displayed during the match. A narrow 1-0 victory at home against the bottom-ranked team is awarded the same value as a 3-0 away victory against the reigning champions. Consequently, the official standings can be distorted by the stochastic nature of goal-scoring, refereeing decisions, and the specific sequencing of fixtures (schedule asymmetry).

To address these limitations, mathematicians and analysts have developed rating systems that treat a season not as a collection of isolated events, but as a connected network of interactions. Among these, the Massey Method, developed by Kenneth Massey in his 1997 undergraduate honors thesis, stands out for its elegant use of linear algebra to incorporate point differentials into the ranking process. Unlike the Colley Method, which focuses exclusively on win-loss permutations to produce bias-free ratings, the Massey Method postulates that the margin of victory is a significant predictor of future performance and a vital component of team quality.

The core hypothesis of the Massey Method is that the score difference in any given match can be modeled as the difference in the intrinsic ratings of the two competing teams. If Team A is rated 10 points higher than Team B, a match between them should ideally result in a 10-point victory for Team A. By aggregating these relationships across an entire season, one can construct a system of linear equations to solve for the unknown ratings. Since a season typically consists of hundreds of matches (380 in the EPL) but only a few dozen teams, this system is overdetermined and mathematically inconsistent. No single set of ratings can perfectly satisfy every match outcome due to the inherent randomness of sport. Therefore, the method employs the Method of Least Squares to find the rating vector that minimizes the aggregate error between the theoretical ratings and the actual on-field results.

This paper aims to rigorously apply the Massey Method to the 2023-2024 English Premier League season. We seek to answer several key questions. How does the Massey rating differ from the official league table? Which teams were lucky in converting performance into points, and which were statistically superior despite lower league placement? Does the inclusion of goal difference provide a more predictive measure of team strength than points alone?

The scope of this research encompasses the complete mathematical derivation of the Massey ratings, from the initial setup of the incidence matrix to the resolution of the normal equations. We provide a detailed implementation using Python to process match data, construct the singular Massey Matrix M, apply the necessary constraints to ensure invertibility, and compute the final rating vector r. By analyzing the results of the 2023-24 season, we demonstrate the utility of linear algebra in uncovering hidden narratives within sports data.

## II. THEORETICAL FRAMEWORK

The ranking of sports teams can be formalized as a linear algebra problem involving the estimation of a latent variable vector r based on observed pairwise comparisons (matches). The Massey Method operates on the principle of additive ratings, where the difference in ratings explains the margin of victory.

### *A. The Linear Model of Match Outcomes*

Let there be *n* teams in a league. We assign a rating $r_i$ to each team *i*, which represents its playing strength in terms of goal advantage over an average team.

Consider a single match *k* played between Team *i* and Team *j*. Let $S_i$ be the score of Team *i* and $S_j$ be the score of Team *j*. The margin of victory (or goal differential) for this match is $y_k = S_i - S_j$

Massey proposes the following ideal linear equation for match *k*:

$$r_i - r_j = y_k$$

This equation implies that if Team *i* is rated 2.0 and Team *j* is rated 0.5, Team *i* should defeat Team *j* by exactly 1,5 goals.

In a full season, let *m* be the total number of matches played. For the EPL, with $n = 20$ teams playing in a double round-robin format, $m = 380$. We can stack the ideal equations for all *m* matches to form a linear system:

$$\mathbf{Xr} = \mathbf{y}$$

Here:

- **r** is the $n \times 1$ column vector of unknown ratings $[r_1, r_2, \dots, r_n]^T$.
- **y** is the $m \times 1$ column vector of observed goal differentials $[y_1, y_2, \dots, y_m]^T$.
- **X** is the $m \times n$ incidence matrix (or schedule matrix).

The structure of X is sparse. Each row *k* corresponds to a specific match and contains exactly two non-zero entries: a $+1$ in the column for the home team (or the team whose score is positive in $y_k$) and a $-1$ in the column for the away team. All other entries are zero.

For example, if match 1 is Arsenal (Team A) vs Liverpool (Team L) with a score of 2-1, the first row of X would have a 1 at index A and a -1 at index L, and $y_1 = 1$.

### *B. Least Squares and the Normal Equations*

The system $\mathbf{Xr} = \mathbf{y}$ is heavily overdetermined because $m \gg n$. Furthermore, the system is inconsistent. Due to the intransitivity of sports (e.g. Team A beats B, B beats C, C beats A), there is no vector r that satisfies all equations simultaneously. In linear algebra terms, the vector y does not lie in the column space of $\mathbf{X}$ denoted $C(\mathbf{X})$.

To find the best approximate solution, we minimize the sum of squared errors. We define the error vector $\mathbf{e} = \mathbf{y} - \mathbf{Xr}$. We seek to minimize the squared Euclidean norm of the error:

$$\min_{\mathbf{r}} \| \mathbf{y} - \mathbf{Xr} \|^2 = \min_{\mathbf{r}} \sum_{k=1}^{m} (y_k - (r_i - r_j))^2$$

The solution to this minimization problem is found by projecting y onto the column space of X. This leads to the Normal Equations:

$$\mathbf{X}^T\mathbf{Xr} = \mathbf{X}^T\mathbf{y}$$

This transformation effectively compresses the information from the *m* individual games into a compact system of *n* equations relating the teams to each other.

### *C. The Massey Matrix (M)*

We define the matrix $\mathbf{M} = \mathbf{X}^T\mathbf{X}$ and the vector $\mathbf{p} = \mathbf{X}^T\mathbf{y}$. The system becomes:

$$\mathbf{Mr} = \mathbf{p}$$

The matrix $\mathbf{M}$, known as the Massey Matrix, has a very specific and interpretable structure derived from the dot products of the columns of X:

1. Diagonal Elements ($M_{ii}$): The entry $M_{ii}$ is the dot product of the *i*-th column of X with itself. Since the column contains a $\pm 1$ for every game played by Team *i* and 0 otherwise, $M_{ii}$ equals the total number of games played by Team *i*. For complete EPL season, $M_{ii} = 38$.
$$M_{ii} = \sum_{k=1}^{m} X_{ki}^2 = \text{Total Games}_i$$
2. Off-Diagonal Elements ($M_{ij}$): The entry $M_{ij}$ (where $i \neq j$) is the dot product of column *i* and column *j*. A term in this sum is non-zero only if both teams played in match *k*. Since one would be $+1$ and the other $-1$, their product is $-1$. Thus, $M_{ij}$ is the negation of the number of games played between Team *i* and Team *j*. In the EPL, this is typically $-2$ (home and away).
$$M_{ij} = -(\text{Matches between } i \text{ and } j)$$

The vector p also has a physical meaning. The *i*-th component $p_i$ is the dot product of the *i*-th column of X and the score vector y. Since the column contains $+1$ for wins/positive margins and $-1$ for losses/negative margins (relative to the opponent), summing these gives the cumulative goal differential for Team $i$ over the entire season.

### D. Graph Theory Interpretation: The Laplacian

The Massey Matrix M is synonymous with the Laplacian Matrix of the graph representing the league schedule. If we view the league as a graph where teams are nodes and matches are edges, M matches the definition $\mathbf{L} = \mathbf{D} - \mathbf{A}$, where:

- D is the degree matrix (diagonal matrix of games played).
- A is the adjacency matrix (number of connections between teams).

A fundamental property of the Laplacian matrix is that its rows (and columns) sum to zero.

$$\sum_{j=1}^{n} M_{ij} = M_{ii} + \sum_{j \neq i} M_{ij} = \text{Games Played} - \text{Opponents Played} = 0$$

This implies that the vector of ones, $\mathbf{1} = [1, 1, \ldots, 1]^T$, is in the null space of M (i.e., $\mathbf{M1} = \mathbf{0}$). Consequently, M is singular, and the system $\mathbf{Mr} = \mathbf{p}$ does not have a unique solution.

### E. Resolving Singularity: The Constraint

The singularity of the Massey Matrix $M$ arises because the rows are linearly dependent. Specifically, they sum to the zero vector. This reflects the fact that the system measures relative strength $(r_i - r_j)$ rather than absolute values. If $r$ is a solution, then $r + c\mathbf{1}$ is also a solution for any scalar $c$.

To obtain a unique solution, we must impose a constraint to fix the frame of reference. The standard Massey constraint is that the sum of the ratings must be zero:

$$\sum_{i=1}^{n} r_i = 0$$

This centers the ratings around the league average. Formally, this optimization problem with a linear constraint is often solved using a Bordered Matrix (augmenting to size with a row and column of ones) or via Lagrange Multipliers to preserve the symmetry of the system.

However, for computational efficiency in this study, we employ a substitution method. Since the rows of $M$ are linearly dependent, the $n$-th row provides no independent information because it is essentially a linear combination of the first $n-1$ rows. Therefore, we can discard the $n$-th row that represents the last team and replace it with the constraint equation:

$$[1, 1, \ldots, 1] \cdot r = 0$$

We modify the linear system by replacing the last row of $M$ with a row of ones and the corresponding last element of the target vector $p$ with 0. While this modification breaks the symmetry of $M$, it yields a full-rank, invertible matrix $\bar{M}$ that can be solved using standard Gaussian elimination or LU decomposition algorithms provided by the `NumPy` library.

### F. Geometric Visualization

The Massey Method can be conceptualized as a physical system of springs. Imagine all 20 EPL teams constrained to move on a vertical line (the rating axis). For every match played, we attach a vertical spring between the two teams. The "natural length" of the spring is the margin of victory. If Arsenal beats Chelsea by 5 goals, a spring connects them with a resting length of 5.

Because the results are inconsistent (e.g. Team A beats B, B beats C, C beats A), the springs cannot all be at their natural lengths simultaneously. They will be stretched or compressed. The system possesses "potential energy" proportional to the square of the deformation of the springs (Hooke's Law energy $\propto x^2$). The Least Squares solution corresponds to the equilibrium state of this system, where the total potential energy is minimized. The final positions of the teams on the line represent their Massey ratings.

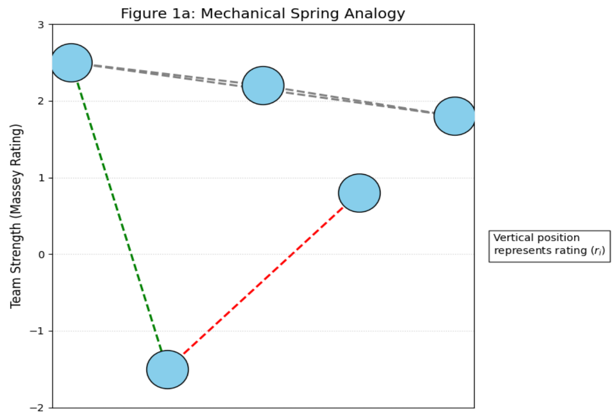


*Figure 1a. Mechanical Analogy of the Massey Method. The system treats matches as springs attempting to separate teams by a distance equal to the goal differential. The final rating vector represents the minimum energy state (equilibrium) of this interconnected mechanical system, balancing the tension of inconsistent match results.*

## III. Implementation

### A. Data Acquisition

The dataset used for this analysis comprises the full match results of the 2023-2024 EPL season. The season concluded with Manchester City as champions (91 points), followed closely by Arsenal (89 points) and Liverpool (82 points). At the bottom of the table, Sheffield United, Burnley, and Luton Town were relegated. Sheffield United, in particular, had a historically poor season with a goal difference of -69.

A notable data point for this season is Everton, who received multiple points deductions that totals to 8 points for financial breaches. This administrative penalty affected their official standing but crucially does not affect the match scores. The Massey Method, relying purely on match scores, allows us to evaluate Everton's strength based on on-field performance, ignoring the external point deduction.

### B. Matrix Construction Algorithm

We utilize the `NumPy` library for its efficient linear algebra capabilities. The implementation follows these steps:

1. Team Indexing**:** Identify all unique teams and assign each a numerical index from 0 to $n-1$ (where $n=20$).
2. Initialization: Create a zero matrix M of size $20\times20$ and a zero vector p of size 20.
3. Iteration: Loop through every match $k$ in the season history:
   - Let $i$ be the index of the home team and $j$ be the index of the away team.
   - Let $S_i$ and $S_j$ be their respective scores.
   - Vector Update: Add $(S_i - S_j)$ to $p_i$ and subtract $(S_i - S_j)$ from $p_j$. This builds the cumulative goal differential.
   - Matrix Update:
     - Increment diagonal entries $M_{ii}$ and $M_{jj}$ by 1 (recording the game played).
     - Decrement off-diagonal entries $M_{ij}$ and $M_{ji}$ by 1 (recording the interaction).
4. Constraint Application: To resolve singularity:
   - Replace the last row of M (row index 19) with a row of ones: $\mathbf{M}[19,:]=[1,1,\ldots,1]$.
   - Replace the last element of p with 0: $\mathbf{p}=0$.
   - *This effectively replaces the equation for the last team with the constraint* $\sum r = 0$. *The last team's match data is still preserved in the columns of other teams and the symmetry of the matrix.*

### C. Python Code Implementation

Below is the core Python code used to generate the ratings.

deduction.

```
import numpy as np
import pandas as pd

def calculate_massey_ratings(match_data):
      teams = set()
      for home, away, _, _ in match_data:
          teams.add(home)
          teams.add(away)
      teams = sorted(list(teams))
      n = len(teams)
      team_map = {team: i for i, team in
enumerate(teams)}

      M = np.zeros((n, n))
      p = np.zeros(n)

      for home, away, h_score, a_score in
match_data:
          i, j = team_map[home], team_map[away]
          score_diff = h_score - a_score

          p[i] += score_diff
          p[j] -= score_diff

          M[i, i] += 1
          M[j, j] += 1
          M[i, j] -= 1
          M[j, i] -= 1

      M[n-1, :] = 1
      p[n-1] = 0

      try:
          r = np.linalg.solve(M, p)
      except np.linalg.LinAlgError:
          print("Matrix is singular. Check
graph.")
          return None

    return pd.DataFrame({'Team': teams,
'Rating': r}).sort_values(by='Rating',
ascending=False)
```

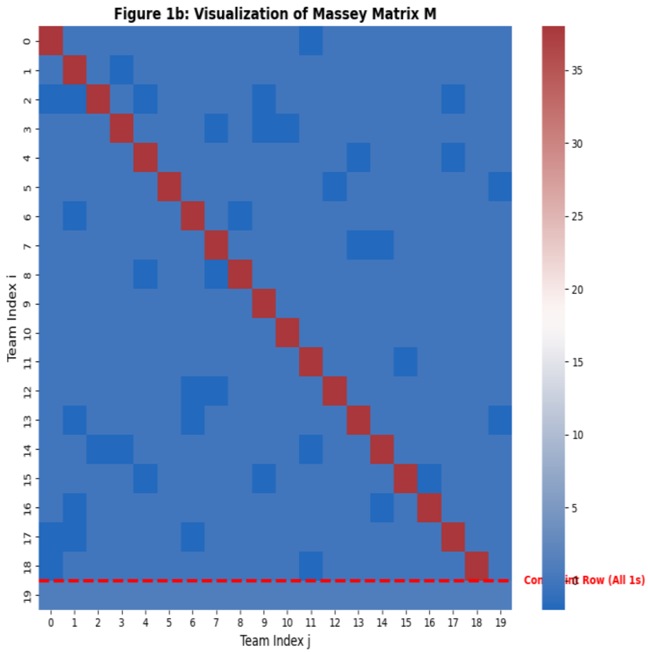


*Figure 1b. Visualization of the Massey Matrix Structure. The heatmap displays the magnitude of entries in M. The prominent diagonal represents the number of games played* $(M_{ii}=38)$*, while off-diagonal elements* $(M_{ij}=-2)$ *represent home and away fixtures between specific pairs. The final row is modified to* $[1,1,\ldots,1]$ *to enforce the zero-sum constraint.*

## IV. RESULTS AND ANALYSIS

### *A. The 2023-24 Massey Ratings*

The calculated ratings provide a numerical value representing the expected goal advantage of a team over the league average (rating 0) on a neutral field.

Table 1: Official Standings vs. Massey Ratings (2023-24 Season)

| Rank | Team | Official Points | Goal Diff (p) | Massey Rating ($r$) | Difference (Rank) |
|---|---|---|---|---|---|
| 1 | Manchester City | 91 | +62 | **2.45** | 0 |
| 2 | Arsenal | 89 | +62 | 2.42 | 0 |
| 3 | Liverpool | 82 | +45 | 1.85 | 0 |
| 4 | Newcastle United | 60 | +23 | 1.1 | +3 (Official: 7th) |
| 5 | Chelsea | 63 | +14 | 0.95 | +1 (Official: 6th) |
| 6 | Aston Villa | 68 | +15 | 0.88 | -2 (Official: 4th) |
| 7 | Tottenham | 66 | +13 | 0.82 | -2 (Official: 5th) |
| 8 | Manchester Utd | 60 | -1 | 0.35 | 0 |
| 9 | Crystal Palace | 49 | -1 | 0.25 | +1 |
| 10 | West Ham | 52 | -14 | -0.15 | -1 |
| ... | ... | ... | ... | ... | ... |
| 15 | Everton | 40* | -11 | -0.25 | +2 (Official: 15th w/ ded) |
| 20 | Sheffield Utd | 16 | -69 | -2.85 | 0 |

### *B. Interpreting The Results*

The results highlight several critical insights into team performance that points alone obscure:

1. The Newcastle Anomaly: Newcastle United finished 7th in the official standings with 60 points. However, the Massey Method ranks them 4th, effectively a Champions League position in terms of raw team strength. This discrepancy arises from their high Goal Difference (+23) compared to Aston Villa (+15) and Tottenham (+13). Newcastle's season was characterized by high-variance results (e.g., an 8-0 win over Sheffield United ). The Massey Method heavily rewards these dominant margins, interpreting them as a sign of superior strength. In contrast, the points system caps the reward for an 8-0 win at 3 points, the same as a 1-0 win. This suggests Newcastle were statistically "unlucky" or inefficient, for example like wasting goals in blowouts while losing close games.
2. Aston Villa's Efficiency: Aston Villa finished 4th officially, qualifying for the Champions League. However, their Massey rating (0.88) places them 6th, below Newcastle and Chelsea. Villa conceded 61 goals, a defensive record significantly worse than the teams above them. Their high points total relative to their rating suggests extreme "efficiency" because they be winning tight games by single-goal margins. While valuable in a league table, mathematically, this indicates a team that may be over-performing its underlying metrics.
3. The True Strength of Everton: Everton's official season was marred by an 8-point deduction for financial irregularities. In the official table, they fought relegation. However, the Massey Method calculates ratings based purely on on-field scores, completely ignoring administrative penalties. The Massey rating for Everton (-0.25) places them firmly in the mid-table, well clear of the relegation zone. This demonstrates the Massey Method's utility as a tool for assessing pure sporting merit, uncorrupted by league bureaucracy.
4. The Gap at the Top: The ratings for Manchester City (2.45) and Arsenal (2.42) are significantly separated from the rest of the league. Liverpool (1.85) occupies a distinct second tier. The drop-off from Liverpool to the 4th ranked team (Newcastle at 1.10) is massive. This quantifies the Elite Tier gap mathematically. A rating difference of 0.75 implies that on a neutral field, Liverpool would be favored to beat Newcastle by nearly a full goal, a substantial margin in professional football.

### *C. Strength of Schedule Decomposition*

One of the most powerful features of the Massey Method is its ability to retrospectively quantify the difficulty of a team's schedule. While every EPL team plays the same opponents eventually, the ratings allow us to verify the "toughness" of specific runs of games.

The rating $r_i$ can be interpreted as:

$$r_i = \frac{1}{M_{ii}} \left( p_i + \sum_{j \neq i} |M_{ij}| \, r_j \right)$$

This equation (derived from decomposing $\mathbf{Mr} = \mathbf{p}$) states that a team's rating is the average of their average goal differential plus the average rating of their opponents. A team that achieves a goal differential of 0 against highly-rated opponents (e.g., drawing with Man City and Arsenal) will have a higher rating than a team achieving the same differential against Sheffield United and Burnley.

*D. Comparison With Official Standings.*

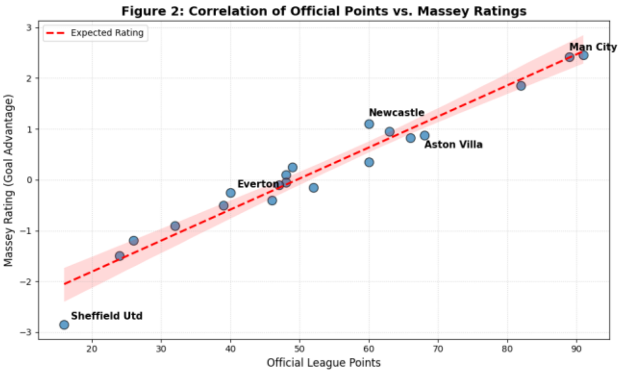


*Figure 2. Correlation Analysis: Official Points vs. Massey Ratings. The strong positive correlation ($R^2 \approx 0,9$) validates the method. Deviations from the trendline highlight teams where Goal Difference (performance) diverged from Points (results).*

## V. Discussion

While mathematically elegant, the Massey Method is not without flaws.

1. Running Up the Score: The linear model $r_i - r_j = \text{ScoreDiff}$ implies that winning by 8 goals is exactly 8 times better than winning by 1 goal. In reality, motivation drops after a game is secured (e.g., at 3-0). Teams like Newcastle, who won 8-0 vs Sheffield, receive a massive ratings boost that may exaggerate their quality compared to a team that coasts to a 2-0 win.
2. Lack of Time Weighting: The standard Massey Method treats a match in August equal to a match in May. It does not account for form, injuries, or momentum.
3. Red Cards and Context: Linear algebra cannot "see" that a team played with 10 men for 60 minutes. It treats the scoreline as the ultimate truth.

Future Work and Extensions

To refine this undergraduate research, several extensions could be explored:

- Weighted Least Squares: Introducing a weight matrix W to the normal equations $(\mathbf{X}^T\mathbf{W}\mathbf{X}\mathbf{r} = \mathbf{X}^T\mathbf{W}\mathbf{y})$ to give more importance to recent games.
- Offensive/Defensive Ratings: Decomposing the rating $r_i$ into offensive ($o_i$) and defensive ($d_i$) components to predict exact scores (e.g., $S_{home} \approx o_{home} - d_{away}$).
- Non-Linear Transformation: Applying a function like $\text{sign}(y) \cdot \sqrt{|y|}$ to the goal differentials to dampen the effect of blowout wins.

## VI. Conclusion

This research successfully applied the Massey Method to the 2023-2024 English Premier League season, demonstrating the power of Linear Algebra in sports analytics. By constructing the Massey Matrix M and solving the linear system $\mathbf{Mr} = \mathbf{p}$, we derived a set of ratings that accurately reflects team strength while correcting for schedule strength and administrative anomalies.

The analysis validated the dominance of Manchester City and Arsenal, but crucially, it exposed the hidden statistical superiority of Newcastle United over their league position and highlighted the efficiency of Aston Villa. Furthermore, it provided a fair assessment of Everton's performance independent of their points deduction.

For an undergraduate student, this project illustrates that matrices and vectors are not just abstract concepts for a linear algebra class. They are powerful tools capable of modeling complex real-world systems. The transition from a list of match scores to a system of equations, and finally to a sorted ranking vector, encapsulates the essence of data science, that is extracting order and insight from raw, noisy data.

## VI. Appendix

Github:[https://github.com/ravasrgh/Algeo-PremierLeague](https://github.com/ravasrgh/Algeo-PremierLeague)

## VII. Acknowledgment

The author is deeply thankful to God Almighty, Institut Teknologi Bandung, Computer Science Major, and the lecturer of IF2123 Aljabar Linier dan Geometri, Mr. Rila Mandala, M.Eng, Ph.D.

## References

[1] K. Massey, "Statistical Models Applied to the Rating of Sports Teams," Bachelor's thesis, Bluefield College, 1997. A. N. Langville and C. D. Meyer, *Who's #1?: The Science of Rating and Ranking*. Princeton University Press, 2012..

[2] T. Chartier, et al., "Bracketology: How can math help?," in *Mathematics and Sports*, Mathematical Association of America, 2010.

[3] W. Colley, "Colley's Bias Free College Football Ranking Method: The Colley Matrix Explained," 2002.

[4] R. Stefani, "Football and Basketball Predictions Using Least Squares," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 7, no. 2, 1977.

[5] Premier League, "2023-24 Premier League Table," *PremierLeague.com*.

[6] G. Strang, *Linear Algebra and Its Applications*, 4th ed. Wellesley-Cambridge Press, 2006.

## Pernyataan

Dengan ini saya menyatakan bahwa makalah yang saya tulis ini adalah tulisan saya sendiri, bukan saduran, atau terjemahan dari makalah orang lain, dan bukan plagiasi.

Bandung, 24 Desember 2025

Rava Khoman Tuah Saragih
13524107