# UFC fight prediction outcomes

Kristina Seibel
Majd Alshamandy

University of Leipzig

**Abstract**. This study investigates the feasibility of predicting Ultimate Fighting Championship (UFC) fight outcomes using only the physical attributes of the fighters. A dataset comprising over 6,500 historical matches was used to evaluate the performance of various supervised classification algorithms in forecasting the winner (Red vs. Blue corner). The baseline model, Gaussian Naive Bayes, demonstrated significant limitations due to high inter-feature correlations and class imbalance. To address these issues, logistic regression (standard and weighted), linear and quadratic discriminant analysis (LDA and QDA with modifications), and random forests were applied. The results underscore the necessity of robust modeling techniques and enriched feature sets when dealing with imbalanced and interdependent data in sports analytics.

## 1 Introduction

Combat sports such as mixed martial arts (MMA) present a compelling challenge for predictive modeling due to their inherently stochastic nature and multifactorial determinants. This project explores whether basic physical characteristics; height, reach, and weight, can serve as meaningful predictors for UFC match outcomes (Berthet (2023)).

We frame the task as a binary classification problem: predicting whether the fighter in the Red or Blue corner emerges victorious. By applying multiple supervised learning models under controlled assumptions, we assess the extent to which quantifiable physical metrics alone can inform outcome forecasting in highly dynamic athletic contests.

### 1.1 Motivation

While psychological factors, technical skills, and strategic variability greatly influence UFC matches, many of these attributes are not readily available or measurable prior to a bout. Physical stats, however, are consistently reported and accessible. This study investigates whether these basic measurements, used in isolation, can yield predictive insights (Berthet (2023)).

Additionally, the project aims to evaluate how simple models like Naive Bayes perform under idealized assumptions such as feature independence, and whether enhancements through discriminative or ensemble techniques lead to substantial gains.

### 1.2 Dataset Description

The dataset was sourced from publicly available UFC records and hosted on Kaggle (MaksBasher (2024)). It includes over 6,500 entries spanning several decades, with features categorized into:

- Fighter metadata: names, weight class, title bout flag
- Physical statistics: height, reach, weight
- Fight history: wins, losses, win streaks
- Outcome: binary indicator of the winning corner (Red = 1, Blue = 0)

## 2 Data Preprocessing

To build a robust foundation for analysis, the dataset was first cleaned and organized:

- A missing value heatmap was used to assess feature completeness across fight entries (see Appendix A).
- Attributes with excessive missingness, like rank-related stats, were excluded.
- Entries lacking critical physical metrics (height, reach, weight) or missing target labels were removed entirely.
- Categorical variables were numerically encoded where needed.
- The target label (winning corner) was binarized: 1 for Red win, 0 for Blue win.
- A stratified 70/30 train-test split with a fixed random seed ensured class balance and reproducibility.

### 2.1 Feature Distributions and Correlation Analysis

To better understand the relationships and spread of attributes, several visualizations were employed:

- Histograms and scatter matrices highlighted outliers and the general distribution of physical features.
- A correlation heatmap (see Appendix B) revealed strong inter-feature relationships:
  - RedHeightCms vs. RedReachCms: 0.90
  - BlueHeightCms vs. BlueReachCms: 0.88
  - RedWeightLbs vs. BlueWeightLbs: 0.97

These strong correlations suggest tight alignment in matchups and significant multicollinearity, posing challenges for models like Naive Bayes that assume feature independence. This supports choosing classifiers that can manage correlated data effectively.

- The scatter matrix (Appendix C) showed:
  - strong linear trends, especially between height and reach
  - tight clustering in weight due to standardized classes
  - a few notable outliers in reach and weight, possibly reflecting atypical fighter profiles

Overall, distributions appeared balanced, though some asymmetry between corners was present, valuable insight for future feature selection and model design.

### 2.2 Feature Engineering

Based on the Feature Distributions and Correlation Analysis and to capture relative physical advantage, the following variables were engineered:

- height_diff: difference in height between Red and Blue corner
- reach_diff: difference in arm reach
- weight_diff: weight difference prior to bout

### 2.3 Models Applied

We tested a range of classification algorithms:

- **Naive Bayes**: baseline with independence assumption
- **Gaussian Discriminant Analysis (GDA)**: Implemented in two variants:
  - LDA (Linear Discriminant Analysis): standard and with equal priors, threshold adjusted
  - QDA (Quadratic Discriminant Analysis): standard and with equal priors, threshold adjusted
- **Logistic Regression**: standard and weighted variants
- **Random Forest Classifier**

Each model was evaluated using standard metrics: accuracy, precision, recall, and F1-score. Confusion matrices were analyzed to assess class-specific performance.

### 2.4 Addressing Class Imbalance

Initial runs indicated bias favoring the Red corner due to dataset imbalance. To mitigate this:

- Equal Priors: enforced uniform class probabilities in LDA/QDA
- Threshold Adjustment: modified decision thresholds to favor minority class
- Class Weights: used in Logistic Regression for loss penalization
- Recall-based Evaluation: prioritized per-class recall over aggregate accuracy

## 3 Results and Comparative Analysis

**Baseline Model**

- Baseline Accuracy (majority class): 0.586
- Naive Bayes: accuracy 0.554; recall (Red): 0.820; recall (Blue): 0.153

**Discriminant Model, Logistic Regression, Random Forest**
The detailed Results of all matrics are in Appendix D.

### 3.1 Comparative Analysis

The comparative analysis involved evaluating the various used models against the baseline accuracy using only physical difference features. We visualized the results through three key figures: the Accuracy Comparison vs Baseline (Appendix D) showed how each model's overall correctness stacked up against simply predicting the majority class. The Recall Comparison by Class (Appendix D) highlighted how well each model identified instances of 'Red' wins (Class 1) versus 'Blue' wins (Class 0). Finally, the Accuracy vs F1-Score for the class 1 scatter plot (Appendix D) provided a combined view of overall accuracy and the balance between precision and recall for the 'Red' win class.

#### *3.1.1 Interpretation*

Standard LDA and Logistic Regression offered the highest accuracy and F1-scores for predicting Red victories but failed to capture Blue wins effectively. Adjustments for class balance led to better recall for the minority class but compromised overall performance. Random Forest achieved better recall balance while maintaining near-baseline accuracy. Most models performed only slightly better than or at the level of the baseline, suggesting that physical attributes alone are not strong predictors of fight outcomes and that additional features are needed for a more effective model.

# 4 Conclusion

## 4.1 Key Findings

- Models relying solely on physical attributes can marginally outperform baseline accuracy but struggle with class imbalance and feature dependence.
- Linear models with high accuracy are biased toward the majority class; mitigating strategies such as threshold tuning and weighting are essential but not sufficient.
- Random Forests show promise in handling interdependent features and offer greater fairness across classes.

## 4.2 Limitations

- Restricted feature space omits critical predictors such as fight history, skill-level indicators, or behavioral metrics.
- Observed biases reflect inherent imbalances in real-world UFC outcomes and dataset distributions.
- Interpretability in complex models remains a challenge for domain application.

## 4.3 Future Work

- Integration of dynamic features (recent performance, win/loss streaks, fighting style).
- Exploration of deep learning architectures for improved pattern recognition.
- Development of interpretable hybrid models with real-time input capability for scouting and betting applications.

Predicting UFC outcomes from physical attributes is possible to a limited extent, but model choice is crucial. Simpler models like Naive Bayes perform poorly due to feature correlations, while Random Forests outperform others by handling non-linear relationships and interdependencies effectively. Overall, physical attributes offer only a weak signal, so adding tactical, historical, or psychological features is essential for more accurate predictions.

## References

Vincent Berthet. Fighttracker: Real-time predictive analytics for mixed martial arts bouts. https://arxiv.org/pdf/2312.11067, 2023. [Online; retrieved on 27. Juli 2025].

MaksBasher. Ufc complete dataset (all events 1996-2024). https://www.kaggle.com/datasets/maksbasher/ufc-complete-dataset-all-events-1996-2024, 2024. [Online; retrieved on 27. Juli 2025].

# A Heatmap for the missing value

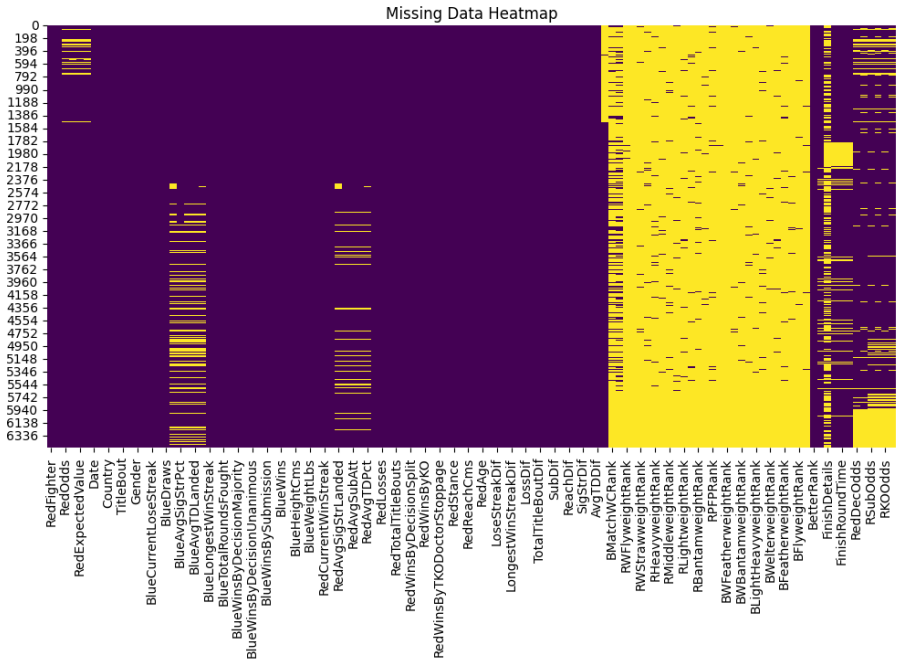


Fig. 1: Heatmap for the missing value

## B Correlation between physical attributes

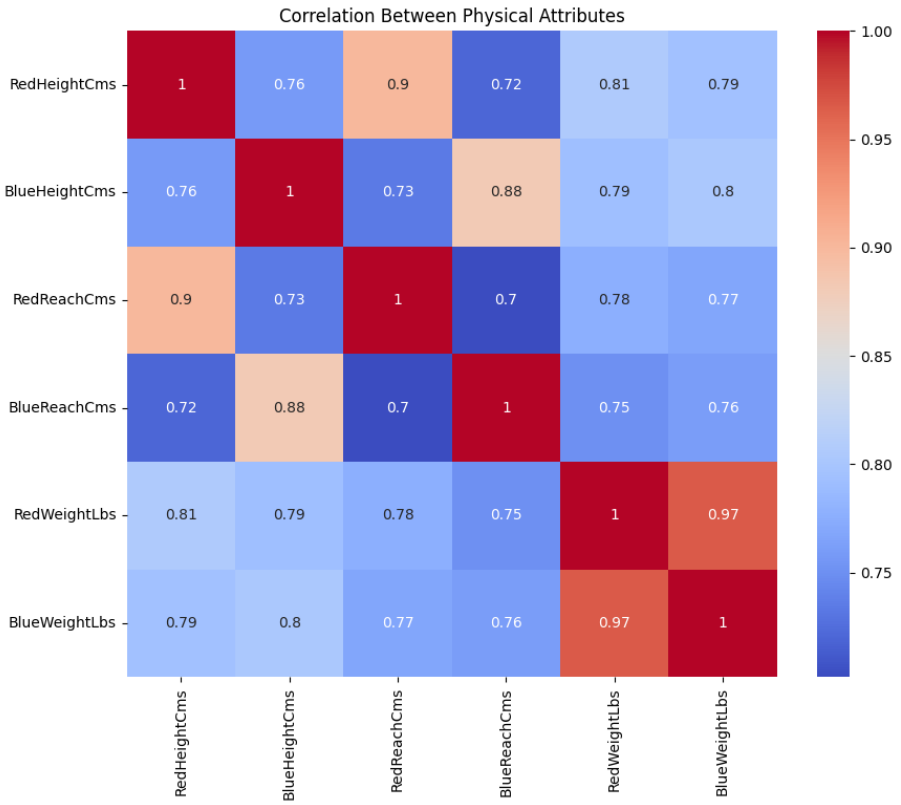


Fig. 2: Correlation Between Physical Attributes

## C Scatter Matrix of physical attributes

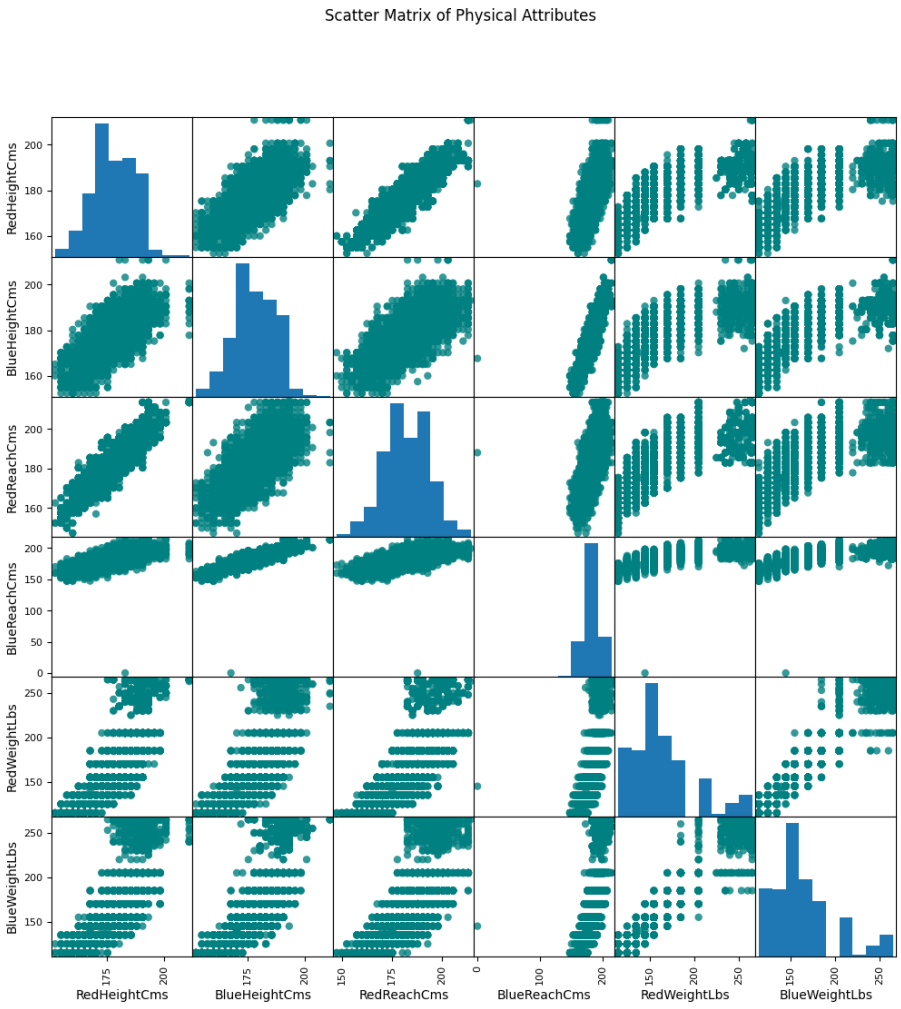


Fig. 3: Scatter Matrix of Physical Attributes

## D Results and Comparative Analysis

| Modell | Accuracy | Precision (cl.1) | Recall (cl.1) | Recall (cl.0) | F1-Score (cl.1) |
|---|---|---|---|---|---|
| LDA (Standard) | 0.593 | 0.591 | 0.987 | 0.036 | 0.739 |
| LogReg (Standard) | 0.593 | 0.591 | 0.987 | 0.036 | 0.739 |
| QDA (Standard) | 0.564 | 0.586 | 0.873 | 0.128 | 0.701 |
| Naive Bayes | 0.549 | 0.582 | 0.815 | 0.172 | 0.679 |
| Random Forest | 0.554 | 0.595 | 0.748 | 0.28 | 0.663 |
| LDA (Equal Priors) | 0.503 | 0.592 | 0.486 | 0.527 | 0.534 |
| LogReg (Weighted) | 0.502 | 0.594 | 0.47 | 0.547 | 0.525 |
| QDA (Equal Priors) | 0.473 | 0.601 | 0.296 | 0.722 | 0.397 |
| LDA (Threshold 0.6) | 0.469 | 0.61 | 0.255 | 0.77 | 0.36 |
| QDA (Threshold 0.6) | 0.465 | 0.616 | 0.229 | 0.798 | 0.334 |

Fig. 4: Metrics of all models

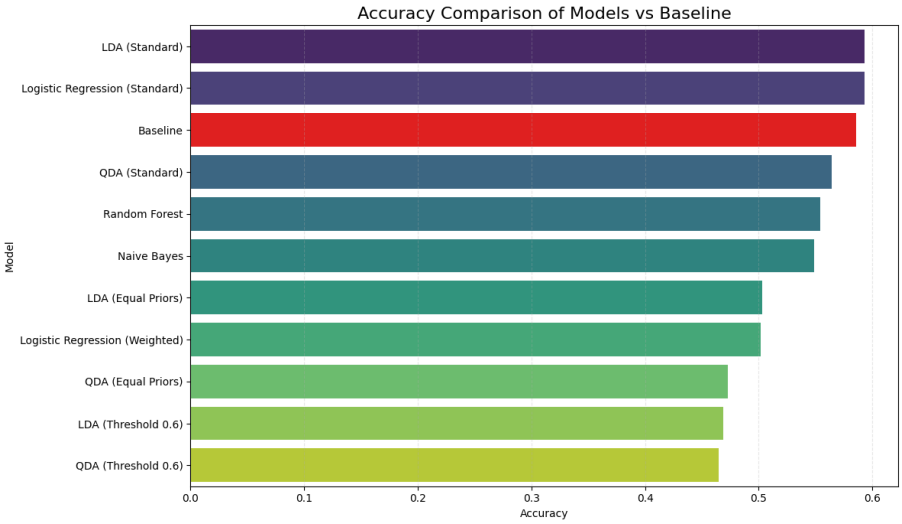


Fig. 5: Accuracy Comparison of Models vs Baseline

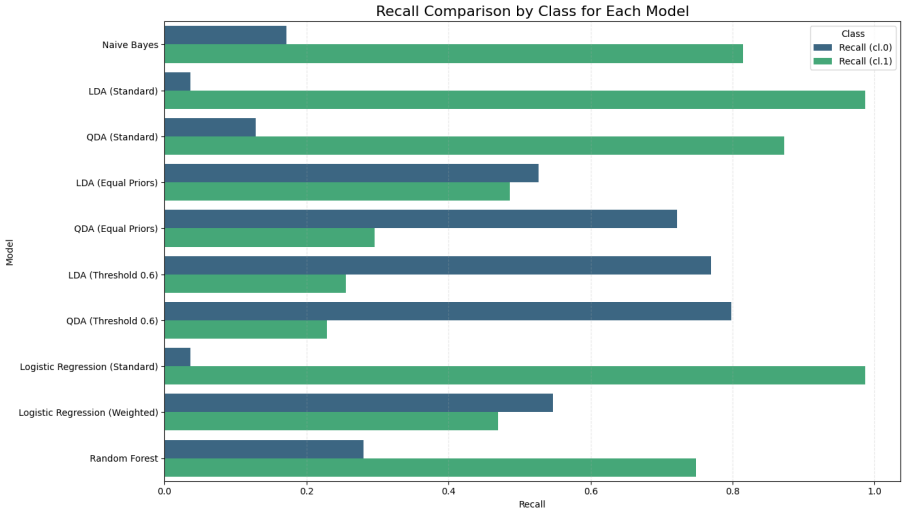


Fig. 6: Recall Comparison by Class for Each Model

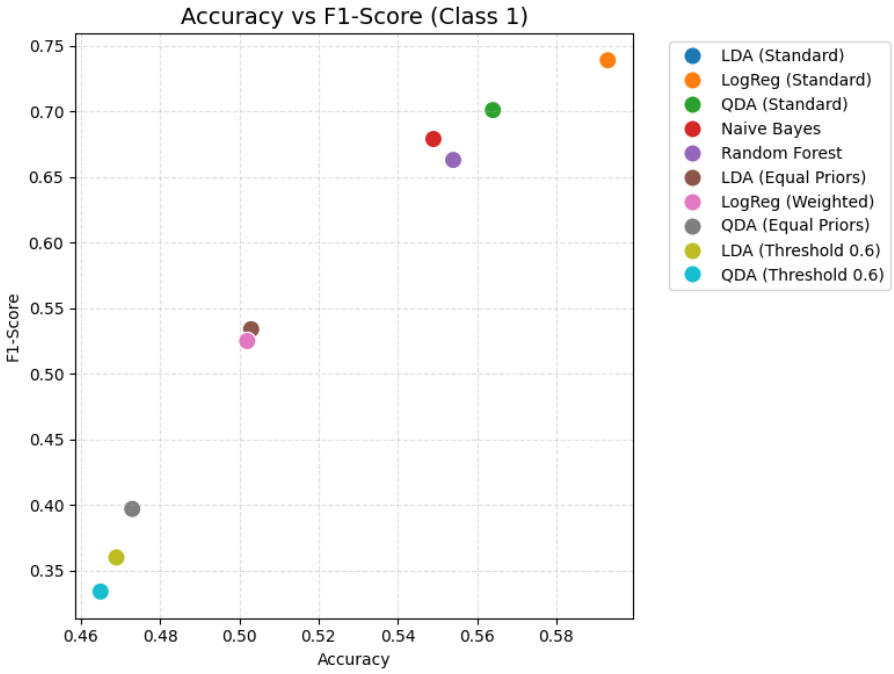


Fig. 7: Accuracy vs F1-Score (Class 1)