University of Salerno
Department of Computer Science

UFC-Predictor

A machine learning-based approach to predicting UFC fight winners

https://github.com/lucasalvaa/UFC-predictor

Student:

Salvatore Luca - 0512119210

Course: Machine Learning

Profs. Giuseppe Polese, Loredana Caruccio

Academic Year 2025-26

Abstract

The **UFC-predictor** is a machine learning module that aims to make predictions about UFC fights. The module uses information such as physical characteristics and historical results of the two contenders in a match. After several steps of data preparation and feature engineering were performed on a dataset containing information on historical UFC matches from 1994 to October 2025, several models have been trained and combined into an ensemble. The accuracy of the ensemble is 62.77%: although this value seems quite far from the ideal threshold of 70%, it is still an excellent result, as the priority was to ensure that the model was bias-free. Finally, the ensemble was deployed using a REST API and a simple graphical web user interface.

1 Introduction

What is UFC?

The **UFC** (Ultimate Fighting Championship) is an organization that promotes professional mixed martial arts (MMA) contests. The UFC was founded in 1993 and purchased by Zuffa Inc. in January 2001 for $2 million, and amateur boxer and entrepreneur **Dana White** was named UFC president that same year [1]. Thanks to his entrepreneurial spirit and some rules that made the sport less violent, White turned the mixed martial arts league into a global phenomenon. Today, the UFC is worth more than $15 billion, with 675 fighters under contract.

Document Structure

The project development follows the guidelines of the **CRISP-DM** reference model [2]. This model breaks down the life cycle of a data mining project in six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment, as shown in Figure 1. The document is structured to reflect the six phases of the model.

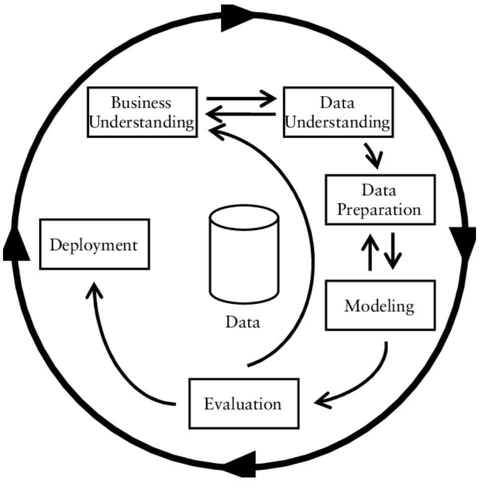


Figure 1: CRISP-DM reference model

2 Business Understanding

Terminology

Mixed martial arts (MMA) is a hybrid combat sport incorporating techniques from boxing, wrestling, judo, jujitsu, karate, muay thai, and other disciplines. MMA dates back to ancient Greece's Olympic Games in 648 BCE with *pankration*, a brutal combat sport that combines wrestling, boxing and street fighting. Although initially decried by critics as a brutal blood sport without rules, MMA gradually shed this negative image and became one of the world's fastest-growing spectator sports in the early 21st century [3].

Weight classes. The UFC currently recognizes a total of eight weight classes in men's MMA and four in women's. The upper weight limits of these classes are:

- Strawweight, 115 pounds (52 kg), W
- Flyweight: 125 pounds (57 kg), W and M
- Bantamweight: 135 pounds (61 kg), W and M
- Featherweight: 145 pounds (66 kg), W and M
- Lightweight: 155 pounds (70 kg), M
- Welterweight: 170 pounds (77 kg), M
- Middleweight: 185 pounds (84 kg), M
- Light Heavyweight: 205 pounds (93 kg), M
- Heavyweight: 265 pounds (120 kg), M

A **UFC card** is an event that consists in multiple bouts divided between a preliminary card and a main card. As the name suggests, the main card is the portion of the event that typically features the most important fights of the night. Closing the main card is the Main Event. There are two types of UFC cards: numbered events (e.g., UFC 325) and UFC Fight Nights. Although the former typically feature the highest-ranked fighters and multiple title bouts, UFC Fight Nights still feature elite athletes but often serve as a platform for rising prospects.

A **bout** (or fight, or match) is a singular contest between two fighters within a specific weight class. Standard bouts consist of 3 rounds, each lasting 5 minutes, with a 1-minute rest period between them. Main Events and Title Bouts are scheduled for 5 rounds of 5 minutes.

The **method of victory** (or result) is how a bout ends.

- **Knockout** (KO): A situation in which a fighter, after being knocked down by their opponent on the ring mat, is unable to get back to their feet within the time limit set for the fight to continue.
- **Technical Knockout** (TKO): A match is stopped by the referee for various reasons, e.g., a fighter is still conscious but unable to defend themselves from their opponent's strikes.
- **Submission**: Used primarily in ground fighting, these can be divided into *restraints* (e.g., chokes, chokes, and compressions) and *manipulations*. One of the two fighters forces their opponent to submit due to perceived pain or fear of injury.
- **Decision**: The result of a match determined by the judges if the fight does not end by knockout or submission. The decision can be unanimous or split.
- **No-Contest**: Occurs when an accidental injury interrupts the bout before enough time has elapsed to make a decision.
- **Disqualification** (DQ): Awarded for a serious intentional foul that ends a bout, multiple fouls, or flagrant disregard for rules/referee commands.

Reach is the distance from one tip of the middle finger to the other with the arms fully extended.

Stance is the physical orientation of the fighter. The most common stances are *Orthodox* (i.e. left hand/foot forward), *Southpaw* (i.e. right hand/foot forward) and *Switch* (i.e. the fighter is capable of changing stances fluently).

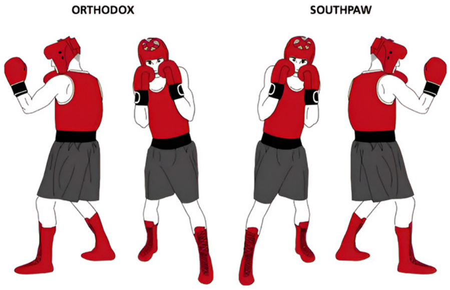


Figure 2: Orthodox vs Southpaw stances

Research Background

Several attempts to build machine learning models to predict the winner of a UFC match have already been documented in the literature.
In 2013, Ho explored the concept of "MMA Math" by training and comparing several machine learning models. Using a dataset containing more than 2,000 fights, Ho found that while simple transitivity (the "A beat B, B beat C, therefore A beats C" logic) is often flawed in combat sports, his best-performing model achieved a peak accuracy of **66.2%** using an SVM approach [4].
In 2019, Hitkul et al. trained seven different models and achieved a best accuracy of **62.8%** [5].
In 2023, Holmes et al. used Markov chains to simulate fights using information about the fighters' skills, rather than predicting a binary outcome. They achieved an accuracy of **61.77%** [6].
In 2024, Yin trained five models and combined their predictions through ensemble learning. It is curious to observe that in this research, although accuracy of the five models varies from 59.13% to 63.99%, the ensemble decision achieved an accuracy of **65.52%** [7].
In 2024, Yan et al. trained three different models and achieved an average accuracy of about **66%** [8].
Based on observations from these scientific articles, and other similar works found around the web [9, 10], it is deductible that accuracy rarely exceeds the 65% threshold. The reason behind this phenomenon is that mixed martial arts is a sport with a high coefficient of unpredictability. In combat sports that involve striking, such as MMA, there is the so-called "Puncher's chance": it means that a puncher always has a chance to win the bout by landing even one punch, even if he is the underdog of the match. Other factors of unpredictability are the possibility of injury during the fight and the psychological attitude of the fighters.

Business success criteria

The purpose of this experiment is to build a bias-free machine learning model capable of exceeding the 60% accuracy threshold. It is therefore essential that the training set does not contain data leakage. In particular, the training set must be free of look-ahead bias, e.g., if a fighter appears in a bout, his records must be calculated based on previous matches. Furthermore, characteristics related to the bout must not be taken into account in the training set: in fact, if you want to predict the outcome of a fight before it takes place, it is not possible to know how many punches a fighter will attempt to throw at his opponent or whether the fight ended early. In conclusion, the goal is not so much to maximize accuracy as it is to apply all possible techniques to ensure that the model does not contain bias.

3 Data Understanding

Dataset Selection

The dataset selected for this experiment is **UFC_full_data_silver**, accessible on the page [UFC Data: Stats & Rankings & Betting Odds] on kaggle.com [11]. According to its owner, the dataset contains "*raw fight facts straight from UFCStats: fighter bios, event details, per-round striking and grappling stats, control time, knockdowns, submissions, takedowns, and more*".

Preliminary Exploration

The raw dataset has 8231 rows and 361 columns. By printing out a list of the names of all the columns to get an idea of how the dataset is structured, we can realize that there are four main categories of features: bout-, fighters-, match- and odds-related features. The features were then grouped into five clusters:

- 17 bout-related features
- 34 fighter_1-related features
- 34 fighter_2-related features
- 270 rounds-related features
- 6 odds-related features

Among these clusters, we are not interested in rounds- and odds-related features. The first group can cause a type of data leakage called **look-ahead bias**: it occurs when a study or simulation is based on data that were not yet available or known during the time period studied. In other words, the day before the bout it is impossible to know, for example, how many punches a fighter will attempt to throw at his opponent. As for the odds, it is best to exclude them, as they do not reflect reality but are just probabilistic estimates. At most, they could be used as an additional metric for comparison with the model results.
Table 1 shows a list of the bout- and fighters-related features. Obviously, fighter_1 and fighter_2 have the same features, even though the table shows them only once. It will be possible to know about some of these features, i.e. *result*, *result_details*, *finish_round finish_time*, *knockdowns*, *total_strikes_att*, *total_strikes_succ*, *sig_strikes_att*, *sig_strikes_succ*,

Table 1: Summary of Bout-Related and Fighters-Related Features

Category	Feature Names
Bout-Related	event_name, referee, winner, num_rounds, title_fight, weight_class, result, gender, result_details, finish_round, finish_time, fight_url, event_date, event_city, event_state, event_country, event_url
Fighters-Related	name, url, fighter_f_name, fighter_l_name, fighter_nickname, fighter_height_cm, fighter_weight_lbs, fighter_reach_cm, fighter_stance, fighter_dob, fighter_w, fighter_l, fighter_d, fighter_nc_dq, fighter_SlpM, fighter_Str_Acc, fighter_SApM, fighter_Str_Def, fighter_TD_Avg, fighter_TD_Acc, fighter_TD_Def, fighter_Sub_Avg, fighter_url, knockdowns, total_strikes_att, total_strikes_succ sig_strikes_att, sig_strikes_succ, takedown_att, takedown_succ, submission_att, reversals, ctrl_time_sec, ranking

Table 2: Summary of missing values across features

Feature	Count	%	Feature	Count	%	Feature	Count	%
event_state	559	6.79	f_1_fighter_nc_dq	6654	80.84	f_2_fighter_nc_dq	6781	82.38
referee	25	0.30	f_1_nickname	2324	28.23	f_2_nickname	2383	28.95
weight_class	10	0.12	f_1_reach_cm	419	5.09	f_2_reach_cm	903	10.97
			f_1_dob	78	0.95	f_2_dob	191	2.32
			f_1_stance	31	0.38	f_2_stance	84	1.02
			f_1_l_name	30	0.36	f_2_height_cm	35	0.43
			f_1_height_cm	13	0.16	f_2_weight_lbs	19	0.23
			f_1_weight_lbs	3	0.04	f_2_l_name	19	0.23

takedown_att, *takedown_succ*, *submission_att*, *reversals*, *ctrl_time_sec*, only at the end of the bout, which is why they should be excluded. The feature *winner* was not taken into consideration in this list because it is supposed to be the target variable. Likewise, the feature *result* was not excluded because it could be useful to calculate the amount of wins and losses by KO and submission for each fighter.

Missing Values

Table 2 shows the number of missing values for the features selected in the previous step. Zero-filling fighters' nc_dq columns, i.e. the count of bouts ended by no contest or disqualification, and ignoring fighter's nicknames, the amount of records with null values is 1590. By ignoring the feature *event_state* too, this number drops to 1147.

Data Quality

- **num_rounds**: contrary to what one might expect knowing the rules, there are 31 zero-round bouts, 167 one-round bouts, and 11 two-round bouts in the dataset. At first glance we might think that these 219 bouts were sampled incorrectly. However, in the early years, the UFC only allowed one-round bouts with no time limits. In 1999, a rulebook change was introduced that distinguished between two-, three-, and five-round bouts. Shortly thereafter, two-round matches were abolished. The dataset contains 7290 three-round bouts and 732 five-round bouts. We may consider removing bouts with fewer than three rounds from the dataset as they may be outliers compared to the majority.
- **gender**: the dataset contains 864 women's bouts and 7367 men's bouts. It is notable that the feature contains a larger number of male matches. Keeping such a small fraction of female bouts in the dataset could introduce noise into the data. For example, historically, the number of knockout wins in women's fights is lower than that in men's fights. Furthermore, women's physical characteristics are generally different from those of men: just think of the fact that there are no female fighters above 145 pounds.
- **result**: as mentioned above, this feature will be important for calculating the number of historical wins and losses by knockout and submission of each fighter in each fight. However, the feature's values are not well organized, i.e. in some rows it takes the value *Decision*, in others *Decision - Split* or *Decision - Unanimous*. It would be appropriate to unify the variants under *Decision*.
- **Fighters' records**: it is essential that fighters' records are up to date as of the date of the fight and not looking into the future. To verify that they have been sampled correctly, let's select all the bouts involving a fighter called Ilia Topuria and see what his record turns out to be. Ilia Topuria is a Spanish-Georgian fighter who is currently undefeated and holds the lightweight belt, which he won on June 28, 2025 at UFC 317. His actual record is 17-0-0. After verifying that the range of *event_date* in the

dataset is from 1994-03-11 to 2025-10-04, a list of Topuria's record for each match has been printed. It turned out that his record is 16-0-0 in each bout and that, unfortunately, there are only nine out of seventeen bouts involving Topuria in the dataset. Out of curiosity, the golden dataset on the same kaggle page was also checked, but it only adds new features and not new bouts. Furthermore, even in this second version of the dataset, Topuria's record is 16-0-0 in every match.

- **Career-related statistics**: repeating the experiment also for the feature *fighter_SlpM* (i.e. Significant Strikes Landed per Minute in career), we discover that it also presents the same value for all the bouts. This is clearly a sampling error that can cause look-ahead bias in the model. We assume that the error is repeated in other fighters' career-related features, i.e. *fighter_Str_Acc*, *fighter_SApM*, *fighter_Str_Def*, *fighter_TD_Avg*, *fighter_TD_Acc*, *fighter_TD_Def*, *fig_Sub_Avg.*

4 Data Preparation

Data Selection

The selected **bout-related features** are: *winner*, *result*, *weight_class*, *gender*, *event_date*, *num_rounds.* The *winner* feature was selected to build the target variable. The *result* feature was selected to calculate the number of historical wins and losses by knockout and submission of each fighter in each fight. The *gender* feature was selected to drop women's bouts from the dataset. The *event_date* feature was selected for calculating historical statistics of each fighter up to the date of bout. The *num_rounds* feature was selected to drop the bouts with less than three rounds from the dataset.

The selected **fighters-related features** for both fighters are: *name*, *height_cm*, *weight_lbs*, *reach_cm*, *stance* and *dob* (i.e., *date of birth*). The *name* features were selected to compare it with the *winner* value and to calculate historical statistics of each fighter. The features related to physical attributes were selected to physically compare the fighters with each other. The *date_of_birth* features were selected to calculate the age of both fighters. For convenience, the prefixes f_1 and f_2 have been replaced by f1 and f2 in all features.

Several categorical features, e.g., URLs, referee and location-related features, have been excluded. Since most of the events take place in the USA and since we have no information about the fighters' nationalities in the dataset, information about event locations is considered useless. Features related to fighters' career records and statistics have also been excluded because, as seen in the previous section, they contain data leakage.

Data Cleaning

First, all UFC Women's Bouts have been removed from the dataset. The *gender* feature has also been removed.

The next step was to recalculate the fighters' records. For this purpose, the matches have been first sorted by ascending event date. Then, all fights in the dataset have been updated using a support dictionary. Whenever a new fighter was encountered in the dataset, it was entered into the dictionary with the entry 0-0-0. For each row, the fighters' records have been iteratively updated according to the dictionary and, immediately after, the dictionary was updated too according to the bout's result.

Once that was done, it was time to clean up the *result* feature. The results *Decision - Split* and *Decision - Unanimous* have been replaced with *Decision*, while the result *TKO - Doctor's Stoppage* has been replaced with *KO/TKO*. The only bout with the result *Could Not Continue* has been dropped. After this operation, the *result* feature has only four values: *Decision* (3262), *KO/TKO* (2496), *Submission* (1378) and *DQ* (21). This last group of bouts (DQ) has been removed from the dataset as it could contain potential outliers.

The next step was to recalculate the wins and losses by KO and Submission for each fighter, bout by bout. The approach used is similar to that used to calculate fighter records. Once this was done, it has been possible to remove bouts with fewer than three rounds from the dataset since, as mentioned in the previous section, they could be outliers. This operation was not previously arranged to avoid compromising the statistics of fighters who participated in bouts both before and after the new rounds' rules.

The last feature affected by data cleaning was *weight_class.* In addition to the eight weight classes currently present in the UFC men's rulebook, there were the following ones: *Catch Weight* (72), *Open Weight* (2), *Nieznana* (1), *UFC Middleweight Title* (1), *UFC Bantamweight Title* (1) and *UFC Light Heavyweight Title* (1). The three title bouts have been mapped into their respective weight classes. *Nieznana* is a Polish term that means "unknown"; *Open* and *Catch Weight* are terms used to describe a weight limit that does not adhere to the traditional limits for weight classes. Therefore, bouts of these two weight classes and the unknown one have been removed from the dataset as they could represent possible outliers.

After all these data cleaning operations, the dataset has 7061 rows and 31 columns.

Null values imputation

At this point in the process, there are only eight features with missing values: *f2_fighter_reach_cm* (668), *f1_fighter_reach_cm* (228), *f2_fighter_dob*

(56), *f2_fighter_stance* (40), *f1_fighter_stance* (16), *f2_fighter_height_cm* (15), *f1_fighter_dob* (11), *f1_fighter_height_cm* (8). The 65 rows with missing values for the fighters' birth dates were removed from the dataset. The rationale behind this choice was that, without knowing the fighters' age difference, imputing the date could have caused noise in the data. By performing this removal, the null values of the other six features were decreased. To impute the null values of height and reach, the respective median values for the fighter's weight class were calculated by joining the columns of the two fighters' characteristics. As for the imputation of null values of the stances, the mode was calculated by joining the relevant columns in the same way. The mode turned out to be the *Orthodox* stance.

Creation of Derived Attributes

- **f1_win**: the target variable; it equals 1 if fighter 1 won the match, otherwise 0.
- **weight_class_id**: the *weight_class* feature's values have been mapped to a range of values from 0 to 7 from the lightest to the heaviest weight class.
- **date**: the *event_date* values have been converted in Unix Time, i.e. the count of seconds since January 1, 1970 [12].
- **delta_age**: the age difference of fighters. Fighters' ages were calculated as the difference between the event_date and their date of birth.
- **delta_height**: the height difference of fighters.
- **delta_weight**: the weight difference of fighters.
- **delta_reach**: the reach difference of fighters.
- **delta_win_rate**: difference between the win rates of the two fighters. A fighter's *win rate* is the ratio between their number of wins and the number of fights they have had before the bout.
- **delta_experience**: difference in the number of fights had by the two fighters before the bout.
- **delta_submission_threat**: difference between the submission threat of the two fighters. A fighter's *submission threat* is the ratio between his number of wins by submission and the total number of wins he has achieved before the bout.
- **delta_ko_power**: difference between the K.O. power of the two fighters. A fighter's *K.O. power* is the ratio between his number of victories by K.O. and the total number of victories he has achieved before the bout.
- **delta_chin_durability**: difference between the chin durability of the two fighters. A fighter's *chin durability* is the ratio between his number of defeats by KO and the total number of defeats he suffered before the match. A fighter's *chin* is the ability to tolerate physical trauma to the head.
- **same_stance**: equals 1 if the two fighters historically fight with the same stance, 0 otherwise.

Target variable distribution

The target variable *f1_win* takes the value *1* 4495 times and the value *0* 2502 times. In fact, in the UFC, corner assignments are not random: the red corner (fighter 1) is almost always reserved for the reigning champion in title matches, the highest-ranked fighter, or the favorite of the host city's audience. It is therefore necessary to decide whether or not to balance the feature. With balancing, the model would learn only on the basis of fighter statistics. This can be useful when the model is used for betting. If the model favors the fighter in the blue corner, whose odds should be lower, a good profit could be made. However, not balancing the variable could improve the accuracy of the model and would not constitute a bias, since, as already mentioned, the fighter in the red corner should have greater skill and therefore a greater probability of winning. In conclusion, it was decided to balance the dataset in order to avoid the bias whereby the model would learn that fighter 1 can win more often.

Final dataset

By removing all categorical features and features no longer considered useful, e.g. the height and weight of the two fighters, the final dataset contains the following 22 features:
f1_ko_w, *f1_sub_w*, *f1_ko_l*, *f1_sub_l*, *f2_ko_w*, *f2_sub_w*, *f2_ko_l*, *f2_sub_l*, *f1_win*, *weight_class_id*, *date*, delta_age, *delta_height*, *delta_weight*, *delta_reach*, *delta_win_rate*, *delta_experience*, *delta_sub_threat*, *delta_ko_power*, *delta_chin_durability*, *same_stance*.
Figure 3 shows the correlation matrix of the final dataset. To best visualize the image, the correlation percentages between the pairs have been hidden. As can be seen from the warmer color, the only pairs exceeding 50% of correlation are (delta_height, delta_reach) and (f1_ko_w, f1_ko_l).

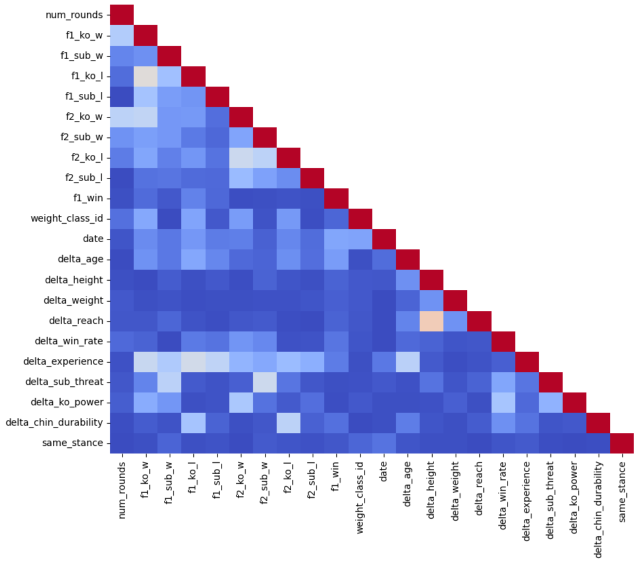


Figure 3: Correlation Matrix

5 Modeling

Modeling Technique

The modeling technique chosen is **ensemble learning** of five models: RandomForestClassifier, LGBMClassifier, XGBClassifier, LogisticRegression, and SVC. While the first three models are based on decision trees, LogisticRegression is a linear model that examines the relationship between predictor variables and a target variable [13], and SVC is a geometric model that classifies data by finding an optimal line or hyperplane that maximizes the distance between each class in an N-dimensional space [14]. By ensembling the probabilistic outputs of tree-based, linear, and distance-based models, the system would mitigate the risk of overfitting to the specific noise of a single algorithm's decision boundary. However, if one of the models achieves an accuracy greater than 1.5% compared to the ensemble decision, then that model will be preferred over the ensemble for deployment.

Cross Validation

Unlike standard cross-validation (K-Fold), which assumes the independence of observations, in the UFC context it is necessary to respect temporal causality. It was therefore decided to adopt **Time-Series Cross-Validation** [15]. In this scheme, the dataset is divided into k sequential folds (in our experiment, k equals 3) where the training set always precedes the test set in time. This prevents the look-ahead bias, ensuring that the model does not use ”future” information (such as the athletic or technical evolution of a fighter still unknown at a given time) to predict past events.

Time Decay

One of the main obstacles in sports prediction is Concept Drift, i.e., the change over time in the statistical relationships between variables (Widmer & Kubat, 1994 [16]). To mitigate this phenomenon, a Time Decay function based on the *date* feature was applied. A weight was assigned to each observation according to an exponential function. This technique forces the learning algorithm to minimize the error on recent matches more, giving them greater statistical relevance than historical matches, reflecting the current state of technical evolution in mixed martial arts in the UFC.

Hyperparameter Tuning

For each model in the ensemble, a stochastic search of hyperparameters was performed using Randomized Search. Unlike Grid Search, it allows an optimal subset of the parameter space to be explored with reduced computational cost, while maintaining virtually identical performance (Bergstra & Bengio, 2012 [17]). The hyperparameter tuning was guided by temporal inner cross-validation to ensure that the chosen parameters were robust and not subject to overfitting on individual folds.

6 Evaluation

Accuracy

The models were evaluated using the **accuracy** metric on the third fold of the *time series split*. As the table shows, while individual model performance hovered between 59.41% and 61.81%, the Soft Voting Ensemble provided a stabilized prediction of 61.98%. The Random Forest Classifier achieved the highest individual accuracy (61.81%).

Model	**Fold 3 Accuracy**
Random Forest	**61.81%**
LightGBM	61.75%
XGBoost	59.41%
Logistic Regression	61.18%
SVC	59.81%
Ensemble	**61.98%**

Additional Metrics

Figure 4: Confusion Matrix

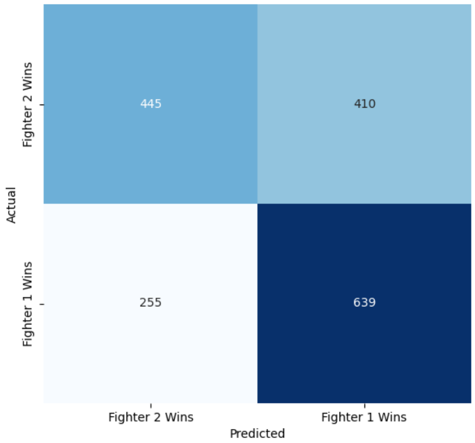


Figure 4 shows the confusion matrix of the ensemble. From the confusion matrix, the *Precision*, *Recall*, and *F1 Score* metrics were calculated.

	Precision	**Recall**	**F1-score**
Fighter 0	64%	52%	57%
Fighter 1	61%	71%	66%

There is a notable divergence in the Recall metrics between Fighter 0 (0.52) and Fighter 1 (0.71). This asymmetry suggests that the ensemble is significantly more effective in identifying winners in the Red Corner category. The high recall for Fighter 1 indicates that the features associated with these victories (e.g., specific reach or age advantages) are consistently recognized by the ensemble, capturing 72% of the actual occurrences. Conversely, the lower recall for Fighter 0, coupled with a higher precision (0.64), implies a ”conservative” predictive behavior. The model only assigns the victory to Blue Corner when the evidence is substantial, thereby avoiding false positives but failing to capture more unexpected wins.

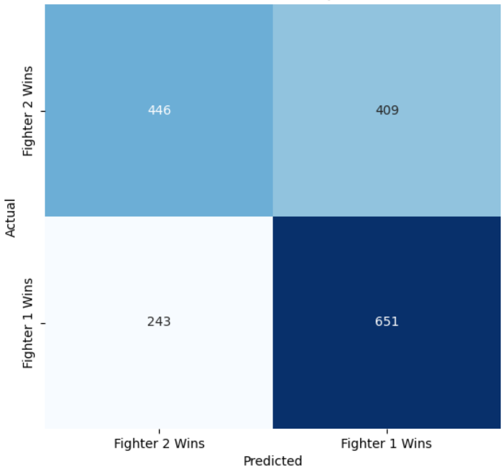
Lighten the ensemble

With a view to continuous learning and deployment, removing even one of the models from the ensemble could speed up training time and lighten the API with which the ensemble will be released.

The experiment has been repeated excluding the model with the lowest accuracy, i.e. XGBoost, from the ensemble. This not only lightened the ensemble, but improved its accuracy by 0.46.%, for a total of **62.44%**.

A further experiment was carried out by also excluding the SVC model from the ensemble. Again, there was an increase in accuracy, specifically 0.28%, for a total of **62.77%**. In both cases, there was no radical change in the other metrics; at most, some decreased or increased by 1%.

Figure 5: Conf. Matrix excluding XGBoost and SVC



7 Deployment

To make the ensemble usable, a simple web GUI and a REST API have been implemented. The REST API was created using the **FastAPI** Python framework [18] and exposes the `/predict` endpoint, which expects to receive the names of the two fighters and the number of rounds as input. After that, for both fighters, an HTTP request is sent to *ufcstats.com* to obtain the URL of the fighter's page and a second request to obtain the content of the page itself. The fighter's page contains information about the fighter and the history of their bouts. The HTML scripts are processed using Beautiful Soup, a Python library for pulling data out of HTML and XML files.

The graphical user interface was implemented using the **Streamlit** Python framework [19]. It allows the user to enter the names of the two fighters and call the API's `/predict` endpoint by clicking a button. It also contains a checkbox to indicate that the match to be predicted will consist of five rounds.

Finally, the two frontend and backend modules were uploaded to Docker Hub as **Docker** images. A `docker-compose.yml` file was added to the project to allow downloading the images, creating a Docker network, and creating containers with a single command, i.e., `docker compose up -d`.

Future Works

- As seen in the *Data Understanting* section, the dataset does not contain all bouts in UFC history. The training dataset could be improved by implementing a script to scrape data of all matches. In addition, this would also allow career statistics to be calculated, e.g., SLpM (Significant Strikes Landed per Minute) and SApM (Significant Strikes Absorbed per Minute).

- The dimensionality of the dataset could be reduced using a Recursive Feature Extraction (RFE) algorithm. The goal of RFE is to select features by recursively considering smaller and smaller sets of features. First, the estimator is trained on the initial set of features to obtain the importance of each feature. Then, the least important features are pruned from the current set of features. That procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached [20].

- The GUI could be updated to extract upcoming UFC events from *ufcstats.com* and automatically predict the winner's name, relieving the user from having to manually enter the names of the fighters and the number of rounds.

References

[1] T. Kuthiala, “Ultimate Fighting Championship,” 2026. [Online]. Available: https://www.britannica.com/topic/Ultimate-Fighting-Championship

[2] R. Wirth and J. Hipp, “Crisp-dm: Towards a standard process model for data mining,” *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 2000. [Online]. Available: https://api.semanticscholar.org/CorpusID:1211505

[3] Britannica Editors, “Mixed martial arts (MMA),” 2026. [Online]. Available: https://www.britannica.com/sports/mixed-martial-arts

[4] C. Ho, “Does mma math work? a study on sports prediction applied to mixed martial arts,” 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:15357264

[5] Hitkul, K. Aggarwal, N. Yadav, and M. Dwivedy, “A comparative study of machine learning algorithms for prior prediction of ufc fights,” in *Harmony Search and Nature Inspired Optimization Algorithms*, N. Yadav, A. Yadav, J. C. Bansal, K. Deep, and J. H. Kim, Eds. Singapore: Springer Singapore, 2019, pp. 67–76.

[6] B. Holmes, I. G. McHale, and K. Żychaluk, “A markov chain model for forecasting results of mixed martial arts contests,” *International Journal of Forecasting*, vol. 39, no. 2, pp. 623–640, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169207022000073

[7] J. Yin, “Data-driven mma outcome prediction enhanced by fighter styles: A machine learning approach,” in *2024 4th International Conference on Machine Learning and Intelligent Systems Engineering (MLISE)*, 2024, pp. 346–351. [Online]. Available: https://ieeexplore.ieee.org/document/10674447

[8] S. Yan, L. Liu, and C. Ubaldo, “Artificial intelligence in ufc outcome prediction and fighter strategies optimaztion,” in *Proceedings of the 2024 9th International Conference on Intelligent Information Processing*. Association for Computing Machinery, 2024, p. 96–100. [Online]. Available: https://doi.org/10.1145/3696952.3696966

[9] R. Schols, “Can data science predict ufc fights? building a leak-free model with random forest.” [Online]. Available: https://medium.com/data-science-collective/can-data-science-predict-ufc-fights-building-a-leak-free-model-with-random-forest-4b6a1cf0945e

[10] Y. Tian, “Predict ufc fights with deep learning ii — data collection and implementation in pytorch.” [Online]. Available: https://medium.com/@yuan_tian/predict-ufc-fights-with-deep-learning-ii-data-collection-and-implementation-in-pytorch-ff7a95062554

[11] jerzyszocik on kaggle.com, “UFC Data: Stats & Rankings & Betting Odds.” [Online]. Available: https://www.kaggle.com/datasets/jerzyszocik/ufc-fight-forecast-complete-gold-modeling-dataset

[12] GNU.org, “Seconds since the Epoch.” [Online]. Available: https://web.archive.org/web/20251004005156/https://www.gnu.org/software/findutils/manual/html_node/find_html/Seconds-since-the-Epoch.html

[13] F. Lee, “What is logistic regression?” [Online]. Available: https://www.ibm.com/think/topics/logistic-regression

[14] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, no. 3, p. 273–297, Sep. 1995. [Online]. Available: https://doi.org/10.1023/A:1022627411411

[15] R. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice, 3rd edition*. OTexts: Melbourne, Australia, 2021. [Online]. Available: OTexts.com/fpp3

[16] G. Widmer and M. Kubat, “Learning in the presence of concept drift and hidden contexts,” *Machine Learning*, vol. 23, Nov. 1994. [Online]. Available: https://link.springer.com/article/10.1007/BF00116900

[17] J. Bergstra and Y. Bengio, “Random search for hyper-parameter optimization,” *J. Mach. Learn. Res.*, vol. 13, no. null, p. 281–305, Feb. 2012. [Online]. Available: https://dl.acm.org/doi/10.5555/2188385.2188395

[18] @tiangolo, “FastAPI documentation.” [Online]. Available: https://fastapi.tiangolo.com/

[19] Snowflake Inc., “Streamlit documentation.” [Online]. Available: https://docs.streamlit.io/

[20] scikit-learn developers, “RFE - scikit-learn documentation.” [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html