

Vysoká škola ekonomická v Praze

Fakulta informatiky a statistiky



Využití statistických metod pro predikci výsledků v oblasti bojových sportů

BAKALÁŘSKÁ PRÁCE

Studijní program: Aplikovaná informatika

Autor: Adam Zacpal

Vedoucí bakalářské práce: Ing. Jan Fojtík, Ph.D.

Praha, květen 2024

Poděkování

Tímto bych chtěl poděkovat Ing. Janu Fojtíkovi, Ph.D. za odborné konzultace a vedení mé bakalářské práce. Také bych chtěl poděkovat rodině za významnou podporu ve studiu.

Abstrakt

Bakalářská práce se zabývá využitím statistických metod pro predikci výsledků ve smíšených bojových uměních (MMA), zejména v zápasech organizace UFC. Cílem práce bylo přispět k lepšímu porozumění faktorů ovlivňujících výsledky zápasů, implementovat prediktivní modely a poskytnout tak užitečné poznatky o tomto sportu. Práce je strukturována do pěti částí, které zahrnují pochopení problematiky sportu, získání, zpracování a čištění vstupních dat, analýzu klíčových proměnných ovlivňujících výsledky a způsoby ukončení zápasů a implementaci prediktivních modelů. V analýze byly zkoumány, mimo jiné, fyzické charakteristiky zápasníků, jejich předešlé zápasy, externí podmínky zápasu či historický vývoj způsobu ukončení zápasu. Pro predikci výsledků a způsobů ukončení zápasů byly vytvořeny vybrané modely strojového učení a Elo model. Práce představuje komplexní přístup k analýze a predikci sportovních výsledků pomocí moderních statistických technik a knihoven programovacího jazyka python. Hlavní přínos práce vidí autor zejména v přesnosti predikce jednoho z modelů pro způsob ukončení zápasu.

Klíčová slova

Strojové učení, analýza dat, Elo, prediktivní modely, web scraping, smíšená bojová umění, MMA, UFC

Abstract

The bachelor thesis deals with the use of statistical methods for the prediction of results in mixed martial arts (MMA), especially in UFC fights. The aim of the thesis was to contribute to a better understanding of the factors influencing match outcomes, to implement predictive models and thus provide useful insights into the sport. The thesis is structured in five parts, which include understanding the sport, obtaining, processing and cleaning the input data, analyzing the key variables affecting the outcomes and the way fights are finished, and implementing predictive models. The analysis examined, among other things, the physical characteristics of the fighters, their previous matches, external match conditions, and the historical development of match finishing methods. Selected machine learning models and an Elo model were created to predict match outcomes and finishing methods. This paper presents a comprehensive approach to the analysis and prediction of sporting outcomes using modern statistical techniques and python programming language libraries. The main contribution of the work is seen by the author in particular in the accuracy of the prediction of one of the models for the match finishing method.

Keywords

Machine learning, data analysis, Elo, predictive models, web scraping, mixed martial arts, MMA, UFC

Obsah

Úvod	10
1 Smíšená bojová umění a predikce výsledků	11
1.1 Historie MMA	11
1.2 Pravidla MMA	12
1.3 Predikce výsledků v MMA	13
1.4 Statistické prediktivní modely	14
1.4.1 Strojové učení	14
1.4.2 Elo model	15
1.4.3 Úspěšnost modelů	16
1.5 Nástroje pro práci s daty	16
1.5.1 Python	16
1.5.2 Další nástroje	17
2 Vstupní data	18
2.1 Web scraping z tapology.com	18
2.2 Web scraping z sherdog.com	20
2.3 Spojení datasetů	22
2.3.1 Úprava souboru „records.csv“	23
2.3.2 Připojení dat o počtu a typu ukončení	24
2.4 Čištění datasetu	25
2.4.1 Odstranění nepotřebných sloupců	26
2.4.2 Doplnění chybějících hodnot	26
2.5 Finální podoba datasetu	27
3 Analýza klíčových proměnných	29
3.1 Vliv fyzických charakteristik na výsledek zápasu	29
3.1.1 Rozdíl ve výšce	29
3.1.2 Rozdíl v rozpětí paží	30
3.1.3 Rozdíl ve věku	31
3.1.4 Srovnání vlivu fyzických charakteristik	32
3.2 Vliv jiných charakteristik na výsledek zápasu	33
3.2.1 Národnost	33
3.2.2 Titulové zápasy	35
3.2.3 Nesplnění stanovené váhy	36
3.3 Vliv předešlých výsledků	37

3.4 Správnost vypsaných kurzů	38
3.5 Vliv externích podmínek zápasu na způsob ukončení zápasu	40
3.5.1 Rozhodčí	40
3.5.2 Aréna	42
3.5.3 Počet kol	44
3.5.4 Váhové kategorie	44
3.6 Historické změny v rozložení způsobu ukončení zápasů	45
3.7 Vliv předešlých ukončení zápasníků na způsob ukončení zápasu	47
3.8 Rozdíly v ukončení tříkolových a pětikolových zápasů	48
4 Prediktivní modely	50
4.1 Predikce výsledku zápasu	50
4.1.1 Logistická regrese	50
4.1.2 Další modely strojového učení	52
4.1.3 Elo model	55
4.2 Predikce způsobu ukončení zápasu	57
4.2.1 Logistická regrese	59
4.2.2 SVM	60
4.2.3 Rozhodovací strom	61
4.2.4 k-NN	62
4.2.5 Náhodný les	63
5 Porovnání výsledků	64
5.1 Porovnání predikce výsledku zápasu	64
5.2 Porovnání predikce způsobu ukončení zápasu	65
Závěr	66
Použitá literatura	67
Přílohy	I
Příloha A: tapology_events.ipynb	I
Příloha B: sherdog_bouts.ipynb	I
Příloha C: creating_dataset.ipynb	I
Příloha D: ufc_analysis.ipynb	I
Příloha E: ufc_prediction.ipynb	I

Seznam obrázků

Obr. 2.1 Seznam turnajů UFC (www.tapology.com)	18
Obr. 2.2 Profil zápasníka (www.sherdog.com).....	22
Obr. 3.1 Histogram rozdílů ve výšce zápasníků (vlastní zpracování).....	30
Obr. 3.2 Histogram rozdílů v rozpětí zápasníků (vlastní zpracování)	31
Obr. 3.3 Histogram rozdílů věku zápasníků (vlastní zpracování).....	31
Obr. 3.4 Podíl výšky, rozpětí a věku na vítězství v zápase (vlastní zpracování)	32
Obr. 3.5 Podíl výher podle národnosti (vlastní zpracování)	35
Obr. 3.6 Poměr výher a proher šampionů (vlastní zpracování)	35
Obr. 3.7 Poměr výher a proher při nesplnění limitu (vlastní zpracování)	37
Obr. 3.8 Heatmapa porovnávající zastoupení výher a proher (vlastní zpracování)	38
Obr. 3.9 Poměr vítězů podle kurzů (vlastní zpracování).....	39
Obr. 3.10 Heatmapa porovnávající kurzové kategorie vítěze a poraženého (vlastní zpracování)	39
Obr. 3.11 Odchylka rozhodčích od průměru ve způsobu ukončení (vlastní zpracování)	42
Obr. 3.12 Porovnání způsobů ukončení pro různé arény (vlastní zpracování)	43
Obr. 3.13 Způsoby ukončení pro tříkolové a pětikolové zápasy (vlastní zpracování)	44
Obr. 3.14 Podíl způsobů ukončení pro váhové kategorie (vlastní zpracování)	45
Obr. 3.15 Vývoj způsobu ukončení (vlastní zpracování)	46
Obr. 4.1 Matice záměn výsledku zápasů pro model logistické regrese (vlastní zpracování)	51
Obr. 4.2 Matice záměn výsledku zápasu pro model podpůrných vektorů (vlastní zpracování)	52
Obr. 4.3 Matice záměn výsledku zápasu pro rozhodovací strom (vlastní zpracování)	53
Obr. 4.4 Matice záměn výsledku zápasu pro model k-NN (vlastní zpracování)	54
Obr. 4.5 Matice záměn výsledku zápasu pro model náhodný les (vlastní zpracování).....	55
Obr. 4.6 Matice záměn výsledku zápasu pro Elo model (vlastní zpracování).....	56
Obr. 4.7 Vývoj skóre nejlepších pěti zápasníků (vlastní zpracování).....	57
Obr. 4.8 Matice záměn způsobu ukončení zápasu pro model logistické regrese (vlastní zpracování)	59
Obr. 4.9 Matice záměn způsobu ukončení zápasu pro model podpůrných vektorů (vlastní zpracování)	60
Obr. 4.10 Matice záměn způsobu ukončení zápasu pro rozhodovací strom (vlastní zpracování)	61
Obr. 4.11 Matice záměn způsobu ukončení zápasu pro k-NN model (vlastní zpracování) .	62
Obr. 4.12 Matice záměn způsobu ukončení zápasu pro náhodný les (vlastní zpracování) .	63

Seznam tabulek

Tab. 2.1 Četnost metod ukončení v datasetu (vlastní zpracování)	23
Tab. 2.2 Popis sloupců tabulky "ufc_bouts" (vlastní zpracování)	27
Tab. 3.1 Zastoupení skupin podle výšky, rozpětí a věku (vlastní zpracování)	33
Tab. 3.2 Výsledky zápasů podle národnosti (vlastní zpracování)	34
Tab. 3.3 Zastoupení četností podle nesplnění limitu (vlastní zpracování)	36
Tab. 3.4 Způsoby ukončení zápasu (vlastní zpracování)	40
Tab. 3.5 Podíl způsobů ukončení pro každého rozhodčího (vlastní zpracování)	41
Tab. 3.6 Výšeč z korelační tabulky pro metody ukončení zápasu (vlastní zpracování)	47
Tab. 3.7 Podíl počtu ukončení pro tříkolové zápasy (vlastní zpracování)	48
Tab. 3.8 Podíl počtu ukončení pro pětikolové zápasy (vlastní zpracování)	49
Tab. 5.1 Srovnání modelů pro predikci výsledků (vlastní zpracování)	64
Tab. 5.2 Srovnání modelů pro predikci způsobu ukončení (vlastní zpracování)	65

Seznam výpisů programového kódu

Výpis 2.1 Procházení seznamu "page_list" a uložení turnajů do DF (vlastní zpracování)...	19
Výpis 2.2 Získávání informací z profilu zápasníka (Obr. 2.2) (vlastní zpracování)	21
Výpis 2.3 Funkce, která získá "record" a informace zápasníka a uloží ho do seznamu (vlastní zpracování)	24

Úvod

Bakalářská práce se zabývá využitím statistických metod pro predikci sportovních výsledků ve smíšených bojových uměních, konkrétně v zápasech nejprestižnější organizace tohoto sportu UFC. Smíšená bojová umění jsou unikátní sportovní disciplínou spojující různé bojové sporty do jednoho.

Práce představuje komplexní přístup k analýze a predikci sportovních výsledků pomocí moderních statistických technik a knihoven programovacího jazyka python. Cílem práce je přispět k lepšímu porozumění faktorů ovlivňujících výsledky zápasů ve smíšených bojových uměních, implementovat prediktivní modely a poskytnout tak užitečné poznatky o této disciplíně. Zajímavost práce spočívá i v tom, že je sport poměrně mladý a není zdaleka tak prozkoumaný jako některé starší a sledovanější sporty.

Práce je rozdělena do pěti částí, z nichž první se zaměřuje na pochopení problematiky sportu a způsobů predikce výsledků. V této části práce si také představíme různé statistické modely a nástroje pro práci s daty. Další část se bude věnovat vstupním datům, jejich získáním pomocí metod web scrapingu, zpracováním, čištěním a představením vytvořeného datasetu. Poté bude provedena analýza klíčových proměnných ovlivňujících výsledky a způsoby ukončení zápasů, včetně fyzických charakteristik zápasníků, jejich předešlých výsledků a dalších externích faktorů. Na základě této analýzy budou implementovány modely strojového učení a Elo model pro predikci výsledků a způsobů ukončení zápasů. V závěrečné části provedeme vyhodnocení a porovnání výsledků modelů.

Autor se od roku 2022 žije profesionálním sázením na zápasy smíšených bojových umění a v budoucnu by se chtěl věnovat právě práci s daty a tvorbě prediktivních modelů, téma je tak ideální kombinací těchto dvou činností. Práce by měla poskytnout hlubší vhled do zápasů smíšených bojových umění a v praxi by mohla pomáhat například při tvorbě kurzových nabídek sázkových kanceláří nebo identifikaci výhodných příležitostí.

1 Smíšená bojová umění a predikce výsledků

Smíšená bojová umění neboli MMA (z anglického Mixed Martial Arts), je plno kontaktní bojový sport kombinující techniky boxu, kickboxu, brazilského jiu-jitsu, zápasu, juda, thajského boxu a jiných disciplín (Britannica, 2024) zahrnující údery, kopy, strhy a další techniky boje na zemi. Zápasy se konají v kleci ze sloupků a pletiva, aby se zabránilo případnému vypadnutí a zranění zápasníků.

1.1 Historie MMA

Smíšená bojová umění mají svůj původ už v roce 648 př. n. l., kdy se ve starověkých olympijských hrách objevil pankration (Britannica, 2024). Tohle bojové umění kombinovalo zápas, box a pouliční boj. Zakázáno bylo pouze dloubání očí a kousání. Zápas končil, když jeden z bojovníků uznal porážku nebo upadl do bezvědomí. V některých případech zápasníci během zápasu zemřeli, nejčastěji v důsledku udušení (Wikipedia, 2011). Pankration byl jednou z nejoblíbenějších disciplín starověkých olympijských her. V roce 393 n. l. císař Theodosius I. zakázal konání olympijských her, což vedlo k poklesu popularity pankrationu.

Na začátku 20. století se toto bojové umění objevilo znovu v Brazílii pod názvem „vale tudo“¹. O jeho popularizaci se velkou mírou postarali bratři Carlos a Hélio Gracie (Britannica, 2024). Ti se učili judo od Japonského mistra Mitsuyo Maedy a sami rozvíjeli tento styl se zaměřením na boj na zemi. Později si založili vlastní školu, a aby dokázali, jak je jejich styl efektivní, vyzývali kohokoliv na zápasy bez pravidel, kde bylo vše povoleno. Jejich veřejné výzvy se staly populární a zápasy brzy vyprodávaly fotbalové stadiony (Britannica, 2024).

Do povědomí širší veřejnosti v Severní Americe se MMA dostalo v roce 1993, kdy se rodina Gracieových rozhodla rozšířit svou značku „brazilské jiu-jitsu“ a podílela se na organizaci turnaje „UFC 1“, kde se mělo rozhodnout, který styl zápasení je nejefektivnější (Britannica, 2024). Turnaje se zúčastnilo celkem osm zápasníků, každý reprezentující jeden z bojových stylů a to box, savate, sumo, kickbox, karate, taekwondo, shootfighting² a brazilské jiu jitsu, které reprezentoval Hélioův syn Royce Gracie. Ačkoliv byl Royce výrazně menší než všichni ostatní, zápasy poměrně jednoduše vyhrál a ihned získal přízeň fanoušků.

¹ „Vale tudo“ v překladu „vše dovoleno“.

² Shootfighting je termín, který předcházela MMA. Pochází z Japonských soutěží obdobného typu a dnes se již nepoužívá.

Organizace „Ultimate Fighting Championship“, která tento turnaj pořádala rychle získávala na popularitě. Turnaje sledovalo stále více lidí, s čímž přišlo i velké množství negativních reakcí zejména k brutalitě sportu a absenci dostatečných pravidel. V této době zápasy připomínaly pouliční rvačky a sport neměl dobrou pověst. Nové vedení „UFC“ v roce 2001 vytvořilo pravidla, která měla učinit sport bezpečnější. Byly přidány váhové kategorie, kola, časové limity a rozšířil se seznam zakázaných technik (Britannica, 2024).

V současnosti je MMA jedním z nejpobulárnějších sportů po celém světě a turnaje nejznámější organizace „UFC“ si kupují miliony lidí.

1.2 Pravidla MMA

Regulační orgány ve Spojených státech v roce 2009 přijaly soubor pravidel známý jako „Unified rules of MMA“, který brzy přijaly organizace po celém světě a sjednotily tak pravidla sportu i mimo USA (Britannica, 2024). Tento standard vychází z pravidel vytvořených organizací „UFC“ v roce 2001.

Podle těchto pravidel bojují zápasníci ve speciálních rukavicích, s odhalenými prsty, bez bot a chráničem na zuby. Utkání je ve většině případů vypsáno na tři kola po pěti minutách s minutovými pauzami mezi koly. U titulových, či hlavních zápasů je to pět kol po pěti minutách. Každé kolo hodnotí 3 rozhodčí na základě pevně daných kritérií samostatně, na konci zápasu sečte každý zvlášť své body a po porovnání bodů s dalšími dvěma rozhodčími se určí vítěz. Mezi kritéria, která se hodnotí, patří v tomto pořadí od nejdůležitějšího po nejméně důležité: efektivní údery a práce na zemi, efektivní agresivita, kontrola zápasu (ASSOCIATION OF BOXING COMMISSIONS AND COMBATIVE SPORTS, 2022). Zápas končí remízou jen výjimečně za velmi specifických podmínek.

Souboj může skončit také předčasně, a to v případě, kdy jeden ze zápasníků není schopen dále pokračovat v boji. Takový stav může nastat v případě „knockoutu“³, „submise“⁴ či například zranění.

Navzdory trvajícím přesvědčení odpůrců sportu o jeho brutalitě, stále více výzkumů ukazuje, že MMA je například bezpečnější než box. Výzkum „*Combative Sports Injuries: An Edmonton Retrospective*“ říká, že ačkoliv je výskyt zranění u zápasníků MMA větší, často se jedná pouze o pohmožděniny či jiné drobné zranění, zatímco u boxerů je vyšší pravděpodobnost vážných zranění, jako je otřes mozku, ztráta vědomí nebo poranění očí (Kapman, Reid, Phillips, Qin, Gross, 2016).

³ Knockout, zkráceně K.O., je situace, kdy zápasník po inkasovaném úderu či kopu buď plně ztratí vědomí nebo rozhodčí usoudí, že už se není schopen dále efektivně bránit. V takovém případě se jedná o takzvaný „technický knockout“.

⁴ Submission je způsob ukončení při, kterém jeden ze zápasníků nějakým chvatem donutí protivníka vzdát se. V případě, kdy se zápasník nechce vzdát a například při škrcení ztratí vědomí nebo utrží zranění, se kterým nelze pokračovat v zápase, rozhodčí ukončí zápas takzvanou „technickou submisí“.

1.3 Predikce výsledků v MMA

V MMA, podobně jako v jakémkoliv jiném sportu je schopnost předpovědět výsledek zápasu velmi hodnotná. Fanoušci, sázkaři, trenéři i analytici by si jistě přáli křišťálovou kouli, která by jim řekla, jak bude zápas probíhat a jak dopadne. Umět správně předpovědět zápas však není vůbec snadné a je to složitý proces, který vyžaduje zohlednění mnoha faktorů a na který se dá nahlížet z mnoha úhlů pohledů.

Metody predikce výsledků:

1. Analýza statistik: Studium předchozích výsledků zápasníků, počtu vítězství a porážek, způsobů vítězství a porážek, počtu úderů, takedownů⁵ a podobně.
2. Fyzické a technické hodnocení: Hodnocení fyzické kondice, síly, rychlosti a technických schopností zápasníka.
3. Hodnocení taktické připravenosti: Sledování videí, podrobná analýza stylu zápasení v postoji i na zemi a schopnosti držet se stanovené taktiky boje.
4. Zohlednění tréninkových a psychologických faktorů: Hodnocení aktuální formy, motivace, tréninkových metod a dalších faktorů, jako jsou například zranění či rivalita zápasníků.
5. Využití predikcí expertů: Sledování prognóz a připomínek profesionálních analytiků, komentátorů a novinářů.
6. Využití statistických modelů: Statistické modely, jako je například logistická regrese, mohou být použity k analýze historických dat a identifikaci proměnných, které jsou spojeny s určitým vyústěním zápasu.

Vzhledem k povaze sportu o úspěchu často rozhodují ty nejmenší detaily, které mohou uniknout i zkušenému analytikovi a ani kombinací všech těchto metod nelze zaručit správnost predikce. Důležité je dlouhodobé a soustavné udržení pozitivní výkonnosti.

V MMA můžeme předpovídat například vítěze zápasu, způsob ukončení nebo kolo ukončení a různé kombinace těchto možností.

Příklady různých úrovní predikce:

1. Zápasník „A“ vyhraje.
2. Zápasník „B“ vyhraje na body.
3. Zápasník „A“ vyhraje na K.O. ve třetím kole.

⁵ Takedown je v terminologii MMA označení pro strh/poraz bojovníka k zemi a následnou kontrolu.

1.4 Statistické prediktivní modely

Prediktivní modelování je statistický proces používaný k předvídání budoucích událostí nebo výsledků analýzou vzorců v daném souboru vstupních dat. Jedná se o klíčovou součást prediktivní analýzy, což je typ datové analytiky, který využívá aktuální a historická data k předpovídání aktivit, chování a trendů (Lawton, Carew, Burns, 2022).

1.4.1 Strojové učení

V rámci strojového učení existují dva základní přístupy: učení s učitelem a učení bez učitele.

Učení s učitelem je přístup, který je definován použitím označených datasetů, to znamená, že pro vstupní data je známý správný výstup. Datasety jsou určeny k trénování algoritmů, které poté přesně klasifikují data nebo správně předpovídají výsledky (Delua, 2021). Učení s učitelem lze dále rozdělit na dva typy problémů:

1. Klasifikační modely: Využívají algoritmus k přesnému zařazení testovacích dat do určitých kategorií, například k oddělení housek od rohlíků nebo k oddělení spamu od doručené pošty (Delua, 2021). Mezi běžné typy klasifikačních algoritmů patří metoda podpůrných vektorů nebo rozhodovací stromy.
2. Regrese: Využívá algoritmy k pochopení vazem mezi závislými a nezávislými proměnnými. Regresní modely jsou užitečné pro předpovídání číselných hodnot na základě různých datových bodů, například pro prognózy příjmů z prodeje pro danou firmu (Delua, 2021). Mezi oblíbené regresní algoritmy patří lineární regrese, logistická regrese a polynomiální regrese.

Učení bez učitele využívá algoritmy strojového učení k analýze a shlukování souborů neoznačených dat. Tyto algoritmy objevují skryté vzory v datech bez nutnosti lidského zásahu. Modely učení bez učitele se používají pro tři hlavní úlohy (Delua, 2021):

1. Shlukování (Clustering): Shlukování má za úkol rozdělit neoznačená data do různých skupin, tak aby podobné datové body spadaly do stejného shluku, a odlišné datové body do jiného (Kaushik, 2019). Technika je užitečná například pro kompresi obrázků či segmentaci trhu.
2. Asociace: Využívá různá pravidla k nalezení vztahů mezi proměnnými v dané množině dat. Tyto metody se často používají pro analýzu tržního koše a algoritmů, které například doporučují, který film by se vám mohl líbit na základě pozorovaných pravidel.
3. Redukce dimenzionality: Je to technika učení, která se používá v případě, že počet proměnných v daném souboru dat je příliš vysoký. Snižuje počet datových vstupů na zvládnutelnou velikost a zároveň zachovává integritu dat (Delua, 2021).

1.4.2 Elo model

Elo rating je statistický model používaný k hodnocení relativní síly hráčů v hrách s nulovým součtem („zero-sum games“). To jsou hry, při kterých zisk jednoho hráče znamená ekvivalentní ztrátu druhého hráče. Elo rating byl vyvinut maďarským profesorem Arpadem Elo pro hodnocení hráčů šachu, ale později byl adaptován i pro jiné hry a sporty. Hlavním cílem ratingu je poskytnout objektivní měřítko pro porovnání síly hráčů a předpovídání výsledků zápasů.

Princip modelu spočívá v relativním hodnocení hráče na základě jeho vlastního ratingu a ratingu jeho soupeře. Každý hráč začíná s určitým počátečním hodnocením, který se po každém zápase aktualizuje na základě jeho očekávaného a skutečného výsledku. Když hráč s horším ratingem porazí ratingově lepšího soupeře sebere mu hodně bodů, když hráč s lepším ratingem porazí horšího soupeře sebere mu málo bodů a když zápas skončí remízou, hráč s nižším ratingem získá malou část bodů od toho s vyšším (Mittal, 2020). Hodnocení hráčů se mění postupně v čase, výrazná změna v hodnocení vyžaduje dlouhodobě stabilní výkony.

Pravděpodobnost výhry neboli očekávané skóre hráče Alice se vypočítá pomocí vzorce:

$$EA = \frac{1}{1 + 10^{(RB-RA)/400}}$$

RB značí rating hráče Bob a RA rating hráče Alice.

Vzorec pro úpravu počátečních skóre, kde K je K-faktor úpravy skóre a SA počet bodů za výsledek zápasu (1 za výhru, 0 za prohru, 0,5 za remízu):

$$R'A = RA + K(SA - EA)$$

Řekněme tedy, že se utká Alice s ratingem 1600 a Bob s ratingem 1400.

Očekávané skóre Alice = 0,759746926:

$$EA = \frac{1}{1 + 10^{(1400-1600)/400}}$$

Očekávané skóre Boba = 0,240253074:

$$EA = \frac{1}{1 + 10^{(1600-1400)/400}}$$

Vidíme, že součet očekávaného skóre Alice a Boba je roven 1.

Řekněme dále, že K = 32 a vypočítejme nové ratingy obou hráčů.

$$R'A = 1600 + 32(1 - 0,759746926)$$

$$R'B = 1400 + 32(0 - 0,240253074)$$

Nový rating hráče Alice je nyní 1608 a Boba 1392.

V současnosti se Elo model využívá kromě šachu také ve sportu, deskových hrách, videohrách, ale také v různých seznamovacích aplikacích, které na základě Elo modelu hodnotí profily uživatelů a podle nich pak profily nabízejí méně či více.

1.4.3 Úspěšnost modelů

Se vzestupem popularity strojového učení vznikla řada modelů, které se snažily predikovat výsledky sportovních zápasů. Studie „*Sports prediction and betting models in the machine learning age: The case of tennis*“ (Wilkins, 2021), která byla provedena pro predikci výsledků v tenise říká, že průměrnou správnost modelů nebylo možné zvednout na více než asi 70 % a bez ohledu na použitý model je většina relevantních informací již obsažena v sázkových kurzech (Wilkins, 2021). Přidání dalších údajů o zápasech a hráčích nevede k výraznému zlepšení. Výnosy z aplikace predikcí na sázkovém trhu jsou velmi proměnlivé a z dlouhodobého hlediska většinou záporné. Jako nejslibnější se ukazuje predikce kombinují více přístupů (Wilkins, 2021).

1.5 Nástroje pro práci s daty

Práce s daty je nedílnou součástí statistické analýzy a má několik částí, které zahrnují sběr dat, porozumění datům, přípravu dat, prediktivní modelování, vyhodnocení výsledků a nasazení výsledků do praxe. Každá z těchto částí má svá specifika a existuje pro ně široká škála nástrojů a technik. Tato kapitola se zaměřuje na představení a popis některých klíčových nástrojů, které jsou k dispozici pro práci s daty. Dva nejpopulárnější jazyky v této oblasti jsou Python a R (Datacamp, 2022).

1.5.1 Python

Python je univerzální programovací jazyk, který se stal nedílnou součástí datové analýzy a datové vědy. Jeho jednoduchost, výkonnost, ale hlavně velké množství knihoven, za kterými stojí rozsáhlá komunita uživatelů, je příčinou jeho stále rostoucí popularity (Datacamp, 2022). Python nabízí téměř neomezené možnosti, které zprostředkovávají právě zmíněné knihovny v oblasti získávání, čištění a vizualizace dat nebo strojového učení.

1. Pandas: Open-source knihovna pro manipulaci s daty, která poskytuje nástroj „DataFrame“, což je snadno použitelná a výkonná tabulková struktura, která umožňuje snadné ukládání, načítání a indexování dat. Pandas poskytuje také jednorozměrnou datovou strukturu „Series“ a další funkce pro operace s nimi.
2. NumPy: NumPy je knihovna pro vědecké výpočty v Pythonu. Poskytuje vysokoúčinné datové struktury, jako jsou vícedimenzionální pole „ndarray“, a širokou škálu matematických funkcí pro práci s daty.
3. Matplotlib: Je to knihovna pro vizualizaci dat v Pythonu. Poskytuje širokou škálu funkcí pro tvorbu statických, interaktivních a animovaných grafů a vizualizací dat.
4. Seaborn: Seaborn je knihovna pro vizualizaci dat založená na Matplotlib, která usnadňuje tvorbu esteticky příjemných a informativních grafů. Často se používá pro vizualizaci výsledků analýzy dat a pro zkoumání vztahů mezi proměnnými.

5. Scikit-learn: Scikit-learn je knihovna pro strojové učení v Pythonu. Obsahuje širokou škálu algoritmů pro klasifikaci, regresi, shlukování, redukci dimenzionality a další techniky strojového učení, stejně jako nástroje pro evaluaci a předzpracování dat.
6. Requests: Requests je knihovna pro práci s HTTP požadavky v Pythonu. Je často používána pro komunikaci s webovými API a stahování dat z internetu.
7. BeautifulSoup: Populární knihovna, která slouží k extrakci dat z HTML a XML dokumentů. Je často používána pro web scraping, což je proces získávání strukturovaných dat z webových stránek.

1.5.2 Další nástroje

R je další programovací jazyk, který se často používá v statistice a analýze dat. Podobně jako Python obsahuje mnoho balíčků pro manipulaci s daty, vizualizaci a statistické analýzy, jako je například dplyr, ggplot2 a tidyr. Stejně jako Python i R dokáže pokrýt všechny oblasti datové vědy. Oproti Pythonu je však v R náročnější řešení složitějších úloh (Datacamp, 2022).

Excel je běžně používaný nástroj pro jednoduchou analýzu a zpracování dat pomocí vestavěných funkcí. Není však zdaleka tak robustní a pro velké datasety může být pomalý a neefektivní. Python oproti Excelu umožňuje reprodukci a sdílení kódu pomocí skriptů a notebooků nebo pokročilejší metody strojového učení.

Jupyter Notebook je interaktivní vývojové prostředí, které umožňuje kombinovat kód, text a vizualizace do jednoho dokumentu. Často se používá pro explorativní analýzu dat a podporuje různé programovací jazyky včetně Pythonu a R.

2 Vstupní data

Vstupní data pro bakalářskou práci byla získána pomocí technik web scrapingu veřejně dostupných dat z nejpopulárnějších webových stránek pro zaznamenávání výsledků MMA zápasů tapology.com a sherdog.com, které byly následně spojeny k vytvoření jednoho datasetu. Obě stránky poskytují informace o zápasnících, jejich zápasech a organizacích. Tento projekt byl omezen na data z nejprestižnější MMA organizace UFC, a to kvůli konzistenci úrovně zápasů, která se napříč organizacemi velmi liší. K vytěžení těchto webů byly použity knihovny Pythonu requests a BeautifulSoup, pro práci a ukládání dat knihovna pandas. Použité vývojové prostředí bylo Jupyter Notebook.

2.1 Web scraping z tapology.com

Tato kapitola popisuje proces extrakce dat v příloze A – „tapology_events.ipynb“.


Tento notebook má za cíl získat data o všech zápasech v UFC s dostupnými podrobnostmi o zápasnících, kteří se jich zúčastnili. Stránka tapology.com byla vybrána z důvodu přehledného členění turnajů podle organizací a bylo tak jednoduché získat potřebná data.

Nejprve se pošle request na url, která odkazuje na první stránku seznamu turnajů organizace UFC. Navracený obsah se přečte pomocí html parseru a uloží do BeautifulSoup objektu. Pomocí vyhledávací metody a cyklu se z BeautifulSoup objektu vždy vytěží odkaz na další stránku se seznamem turnajů (Obr. 2.1), uloží se do seznamu stránek a přejde na tento odkaz. Tohle se opakuje, dokud nedojdeme na poslední stránku. Výsledkem je seznam „page_list“ plný odkazů, které ukazují na stránky se seznamem turnajů.

MIXED MARTIAL ARTS PROMOTION

Ultimate Fighting Championship

[More MMA Promotions](#)










Name: Ultimate Fighting Championship

Headquarters: Las Vegas, Nevada, United States

Acronyms: UFC

Ownership: Endeavor

Promotion Links:



UFC EVENTS

« First « Prev 1 2 3 4 5 6 ... Next » Last »

UFC FIGHT NIGHT: ANKALAEV VS. WALKER • Saturday, January 13, 2024

UFC Apex • Las Vegas, NV • US West Region

★ Ankalaev vs. Walker II • 205 lbs • Show Bouts

UFC 296: EDWARDS VS. COVINGTON • Saturday, December 16, 2023

T-Mobile Arena • Las Vegas, NV • US West Region

★ Welterweight Title Fight • Edwards vs. Covington • 170 lbs • Show Bouts

UFC FIGHT NIGHT: SONG VS. GUTIERREZ • Saturday, December 09, 2023

UFC Apex • Las Vegas, NV • US West Region

★ Song vs. Gutierrez • 135 lbs • Show Bouts

UFC FIGHT NIGHT: DARIUSH VS. TSARUKYAN • Saturday, December 02, 2023

Moody Center • Austin, TX • US Southwest Region

Obr. 2.1 Seznam turnajů UFC (www.tapology.com)

Výpis 2.1 Procházení seznamu "page_list" a uložení turnajů do DF (vlastní zpracování)

```
events = pd.DataFrame(columns=["Event Link", "Event Name", "Date", "Venue", "Location"])
# Iterates through page_list
for page in tqdm(page_list):
    data = req.get(page, headers = headers).text
    soup = BeautifulSoup(data,"html.parser")
    # Finds all listings on a page
    listings = soup.find_all(class_="fcListing")
    # Iterates listings and saves each one in DataFrame
    for listing in listings:
        link = "https://www.tapology.com" + listing.find("a")["href"]
        name = listing.find("a").text
        date = listing.find(class_="datetime").text
        venue = ""
        if listing.find(class_="venue") != None:
            venue = listing.find("span", class_="venue").text
        location = ""
        if listing.find(class_="venue-location") != None:
            location = listing.find(class_="venue-location").text
        region = ""
        if listing.find(class_="region") != None:
            region = listing.find(class_="region").text
        events = events._append({"Event Link":link, "Event Name":name, "Date":date,
"Venue":venue, "Location":location}, ignore_index=True)
events = events.replace(r'\n', '', regex=True)
```

V druhém kroku se „page_list“ prochází cyklem, který podobným procesem, jako v předchozím odstavci, najde všechny turnaje na stránce (Obr. 2.1) a uloží je do DataFramu „events“ (Výpis 2.1).

DataFrame „events“ obsahuje sloupce: „Event Link“, „Event Name“, „Date“, „Venue“, „Location“.

Ve třetím kroku se konečně získávají cílová data o zápasech procházením odkazů DataFramu „events“. Každá stránka má informace o turnaji a zápasech na nich. Algoritmus nejdříve zjistí, zda se turnaj již uskutečnil, vyhledá obecné informace o turnaji a poté iteruje po jeho zápasech a hledá informace o zúčastněných zápasnících, výsledku zápasu, způsobu ukončení zápasu, váhové kategorii a dalších dostupných informacích. Dále najde tabulku s detaily zápasu, kde vyhledá informace například o vypsání kurzech, výsledku vážení či věku zápasníků. Na konci iterace se všechny získané informace uloží do DataFramu „bouts“, kde jeden řádek znamená jeden zápas.

DataFrame „bouts“ má 7577 řádků a obsahuje sloupce: "Bout Link", "Fighter A", "Link A", "Fighter B", "Link B", "Nickname A", "Nickname B", "Record A", "Record B", "Odds A", "Odds B", "Title A", "Title B", "Weight A", "Weight B", "Age A", "Age B", "Height A", "Height B", "Reach A", "Reach B", "Result", "Time", "Weightclass", "Rounds", "Event Link", "Venue", "Date", "Location", "Billing", "Event Name".

Pro další vyhledávání bylo nezbytné zjistit, zda v DataFramu nemáme zápasníky se stejným jménem. Víme, že zápasníci sice můžou mít stejná jména, ale odkaz na jejich profil musí být vždy unikátní. Pomocí sloupců „Fighter A“, „Fighter B“, „Link A“ a „Link B“ a metody „groupby“ jsme tedy prozkoumali, zda není jedno jméno propojeno s více odkazy a bylo zjištěno, že ve dvou případech se tak stalo. V jednom případě se jednalo o bratry se stejným prvním jménem, proto bylo přidáno jejich prostřední jméno pro jednoznačnou identifikaci. V druhém případě byla pro jednoznačnou identifikaci přidána přezdívka zápasníků. Upravený DataFrame s 7577 zápasy a 2479 unikátními zápasníky byl uložen do souboru „bouts_new.csv“ (příloha).

2.2 Web scraping z sherdog.com

Tato kapitola popisuje proces extrakce dat v příloze B – „sherdog_records.ipynb“.

Tato část má za cíl získat data primárně o předchozích zápasech všech zápasníků z datasetu „bouts_new.csv“, ale také doplňující data o samotných zápasnících, která budou později připojena k prvnímu datasetu. Stránka sherdog.com byla vybrána z důvodů konzistentního formátu zaznamenávání metod ukončení zápasu, jako tomu není u tapology.com.

Nejdříve byl vytvořen seznam unikátních jmen zápasníků z datasetu získaného v předchozí kapitole. Pro každé jméno se spustil algoritmus, který pomocí google vyhledávače našel odkaz na Sherdog profil zápasníka (Obr. 2.2). Pomocí html parseru byla stránka z odkazu opět přečtena a uložena do BeautifulSoup objektu. Z profilu byl opět pomocí vyhledávacích metod objektu vytěženy obecné informace o zápasnících a poté požadovaná tabulka se záznamy o jednotlivých zápasech zápasníka (Výpis 2.2). Tyto informace byly uloženy do DataFramu „records“, kde jeden řádek znamená opět jeden zápas.

DataFrame „records“ má 54860 záznamů a obsahuje sloupce: "Fighter", "URL", "Nationality", "Nickname", "Birthday", "Height ft", "Height cm", "Opponent", "Result", "Method", "Round", "Time", "Event", "Date", "Referee" byl uložen do souboru „records.csv“.

Testováním algoritmu bylo zjištěno, že pro některá jména vyhledávání profilu vykazovalo chybné chování, proto byla jména ještě před spuštěním algoritmu změněna. Jednalo se o jména „Ian Machado Garry“, „Sako Chivichitan“ a „Razak Al-Hassan“. Po dokončení byla jména nahrazena původními hodnotami.

Výpis 2.2 Získávání informací z profilu zápasníka (Obr. 2.2) (vlastní zpracování)

```
def get_record(fighter):
    fighter_data, url = get_website(fighter)
    fighter_soup = BeautifulSoup(fighter_data, "html.parser")
    # Finds fighter information
    nationality = ""
    if fighter_soup.find("strong", itemprop="nationality") != None:
        nationality = fighter_soup.find("strong", itemprop="nationality").text
    nickname = ""
    if fighter_soup.find("span", class_="nickname") != None:
        nickname = fighter_soup.find("span", class_="nickname").find("em").text
    birthday = ""
    if fighter_soup.find("span", itemprop="birthDate") != None:
        birthday = fighter_soup.find("span", itemprop="birthDate").text
    height_ft = ""
    height_cm = ""
    if fighter_soup.find("b", itemprop="height") != None:
        height_ft = fighter_soup.find("b", itemprop="height").text
        if fighter_soup.find("b", itemprop="height").nextSibling.nextSibling.nextSibling
        != None:
            height_cm = fighter_soup.find("b",
itemprop="height").nextSibling.nextSibling.nextSibling
    # Finds table with fighters record
    table = fighter_soup.find("table", class_="new_table fighter")
    fighter_record = pd.DataFrame(columns=["Fighter", "URL", "Nationality", "Nickname",
"Birthdate", "Height ft", "Height cm", "Opponent", "Result", "Method", "Round", "Time",
"Event", "Date", "Referee"])
    # Fills the DataFrame with fighter record
    for row in table.find_all("tr", class_=""):
        col = row.find_all("td")
        if (col != []):
            result = col[0].text
            opponent = col[1].text
            event = col[2].a.text
            date = col[2].find("span", class_="sub_line").text
            method = col[3].b.text
            referee = col[3].span.text
            rnd = col[4].text
            time = col[5].text
            fighter_record = fighter_record._append({"Fighter":fighter, "URL":url,
"Nationality":nationality, "Nickname":nickname, "Birthdate":birthday, "Height
ft":height_ft, "Height cm":height_cm, "Opponent":opponent, "Result":result,
"Method":method, "Round":rnd, "Time":time, "Event":event, "Date":date, "Referee":referee},
ignore_index=True)
    return fighter_record
```

DAVID DVORAK

"The Undertaker"

CZECH REPUBLIC

HRADEC KRÁLOVÉ

AGE

31 / JUN 5, 1992

HEIGHT

5'5" / 165.1 CM

WEIGHT

125 LBS / 56.7 KG

ASSOCIATION

ALL SPORTS ACADEMY

CLASS

FLYWEIGHT

WINS

20

LOSSES

6

KO / TKO

8 / 1

40%

SUBMISSIONS

8 / 0

40%

DECISIONS

4 / 5

20%

KO / TKO

1 / 1

17%

SUBMISSIONS

0 / 0

0%

DECISIONS

5 / 5

83%

FIGHT HISTORY - PRO

RESULT	FIGHTER	EVENT	METHOD/REFEREE	R	TIME
LOSS	Steve Erceg	UFC 289 - Nunes vs. Aldana Jun / 10 / 2023	Decision (Unanimous) Mitchell Cadlick VIEW PLAY-BY-PLAY	3	5:00
LOSS	Manel Kape	UFC Fight Night 216 - Cannonier vs. Strickland Dec / 17 / 2022	Decision (Unanimous) Keith Peterson VIEW PLAY-BY-PLAY	3	5:00
LOSS	Matheus Nicolau	UFC on ESPN 33 - Blaydes vs. Daukaus Mar / 26 / 2022	Decision (Unanimous) Chad Trukovich VIEW PLAY-BY-PLAY	3	5:00
WIN	Juancamilo Ronderos	UFC Fight Night 188 - Font vs. Garbrandt May / 22 / 2021	Submission (Rear-Naked Choke) Keith Peterson VIEW PLAY-BY-PLAY	1	2:18
WIN	Jordan Espinosa	UFC Fight Night 178 - Covington vs. Woodley Sep / 19 / 2020	Decision (Unanimous) Jason Herzog VIEW PLAY-BY-PLAY	3	5:00

Obr. 2.2 Profil zápasníka (www.sherdog.com)

2.3 Spojení datasetů

Tato kapitola popisuje propojení datasetů v příloze C – „creating_dataset.ipynb“.

V této části propojíme dva datasety získané v předchozích kapitolách do jednoho, s kterým budeme dále pracovat. Cíl, kterého chceme dosáhnout, je vypočítat pro každý záznam v souboru „bouts_new.csv“ (tedy pro každý zápas v UFC), počet vítězství a proher obou zápasníků na body, K.O. a submisi pro oba zápasníky a připojit k nim doplňující informace. Tyto hodnoty se budou počítat vždy pouze do posledního zápasu, který předcházel zápasu, pro který hodnoty počítáme. Abychom mohli tyto hodnoty vypočítat a datasety propojit, musíme nejdříve upravit některé sloupce v tabulce ze souboru „records.csv“.

2.3.1 Úprava souboru „records.csv“

Sloupec „Method“ reprezentuje způsob ukončení zápasu. V MMA existují pouze čtyři hlavní způsoby ukončení: na body, K.O., submitse a „no contest“⁶. Po prozkoumání sloupce bylo zjištěno, že obsahuje až 733 různých hodnot, a to z důvodu, že každý záznam, kromě metody ukončení, obsahuje také popis v závorce, který už může mít mnoho podob. Tento popis byl ze sloupce „Method“ odstraněn a uložen do nového sloupce „Description“. Sloupec „Method“ nyní z důvodu nekonzistence obsahoval už jen 20 různých hodnot, které byly dále sloučeny do požadovaných čtyř.

1. „KO“, „TKO“ a „K.O.“ bylo sloučeno pod „KO/TKO“.
2. „Submission“, „Technical Submission“, „Technical Submission“ a „Submission (Rear-Naked Choke)“ bylo sloučeno do „Submission“.
3. „Draw“ and „Technical Draw“ není metoda ukončení, ale výsledek zápasu a každá remíza končí na body, proto byly tyto hodnoty sloučeny do „Decision“.
4. "No Contest", "NC", "No Decision – Overturned by CSAC", "No Contest – Collard Failed Drug Test", "No Decision" a "ND" znamenají, že zápas nemá výsledek, tudíž je můžeme sloučit do „No Contest“.
5. „Technical Decision“ můžeme sloučit do „Decision“, jelikož je to pouze jiný druh toho samého.
6. Všechny diskvalifikace se v MMA počítají jako „KO/TKO“, proto můžeme záznamy "Disqualification", "DQ" a „DG“ přidat k této hodnotě.

Po provedení těchto operací zůstaly ve sloupci „Method“ pouze požadované čtyři různé hodnoty (Tab. 2.1). DataFrame byl uložen do souboru „records_clean.csv“.

Tab. 2.1 Četnost metod ukončení v datasetu (vlastní zpracování)

Metoda ukončení	Počet
KO/TKO	20223
Decision	17871
Submission	16213
No Contest	457

⁶ „No contest“ znamená, že zápas je bez výsledku. Nikam se nepočítá a je na něj nahlíženo, jako by se nestal.

2.3.2 Připojení dat o počtu a typu ukončení

Nyní použijeme DataFrame ze souboru „records_clean.csv“, který jsme upravili v minulé kapitole a přidáme ho k datasetu o zápasech v UFC „bouts_new.csv“.

Tato operace byla provedena na základě jmen zápasníků a dat zápasů. Jména zápasníků byla v obou datasetech totožná, data zápasů však měla v obou datasetech jiný formát, a musela tak být pomocí jednoduché funkce a knihovny „datetime“ převedena.

Výpis 2.3 Funkce, která získá "record" a informace zápasníka a uloží ho do seznamu (vlastní zpracování)

```
def calculate_records(fighter_name, bout_date):
    df = records.loc[records["Fighter"] == fighter_name]
    # Finds fighter information
    nationality, nickname, birthday, referee = "", "", "", ""
    nationality = df.iloc[0]["Nationality"]
    nickname = df.iloc[0]["Nickname"]
    birthday = df.iloc[0]["Birthday"]
    if not df.loc[df["Date"] == bout_date, "Referee"].empty:
        referee = df.loc[df["Date"] == bout_date, "Referee"].iloc[0]
    # Filters DataFrame for the date of a given fight
    df = df.loc[df["Date"] < bout_date]
    # Calculates fighters record
    counts = df.groupby(["Result", "Method"])["Fighter"].count()
    ko_win, ko_loss, sub_win, sub_loss, dec_win, dec_loss, draws, ufc_win, ufc_loss,
    ufc_draws = 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
    for index, value in counts.items():
        if index[0] + index[1] == "winKO/TKO":
            ko_win = value
        elif index[0] + index[1] == "lossKO/TKO":
            ko_loss = value
        elif index[0] + index[1] == "winSubmission":
            sub_win = value
        elif index[0] + index[1] == "lossSubmission":
            sub_loss = value
        elif index[0] + index[1] == "winDecision":
            dec_win = value
        elif index[0] + index[1] == "lossDecision":
            dec_loss = value
        elif index[0] == "draw":
            draws = value

    # Calculates fighters ufc record
    ufc_events = df[df['Event'].str.startswith('UFC')]
    ufc_counts = ufc_events.groupby("Result")["Fighter"].count()
    for index, value in ufc_counts.items():
```



```

        if index == "win":
            ufc_win = value
        elif index == "loss":
            ufc_loss = value
        elif index == "draw":
            ufc_draws = value
    info_list=[]
    info_list = [ko_win, ko_loss, sub_win, sub_loss, dec_win, dec_loss, draws, ufc_win,
ufc_loss, ufc_draws, nationality, birthday, nickname, referee]
    return info_list

```

Poté byla na každý záznam v tabulce „bouts_new“ zavolána funkce „calculate_records“ (Výpis 2.3), která pomocí jména zápasníka a data zápasu vypočítá informace o jeho předchozích zápasech a vrátí je spolu s doplňujícími informacemi o něm. Informace vrácené z funkce „calculate_records“ se následně nalepí k záznamu v tabulce „bouts_new“.

Výsledný dataset má tyto sloupce: „Bout Link“, „Fighter A“, „Link A“, „Fighter B“, „Link B“, „Nickname A“, „Nickname B“, „Record A“, „Record B“, „Odds A“, „Odds B“, „Title A“, „Title B“, „Weight A“, „Weight B“, „Age A“, „Age B“, „Height A“, „Height B“, „Reach A“, „Reach B“, „Result“, „Time“, „Weightclass“, „Rounds“, „Event Link“, „Venue“, „Date“, „Location“, „Billing“, „Event Name“, „KO win A“, „KO loss A“, „Sub win A“, „Sub loss A“, „Dec win A“, „Dec loss A“, „Draws A“, „UFC win A“, „UFC loss A“, „UFC draws A“, „Nationality A“, „Birthday A“, „Nick A“, „Referee“, „KO win B“, „KO loss B“, „Sub win B“, „Sub loss B“, „Dec win B“, „Dec loss B“, „Draws B“, „UFC win B“, „UFC loss B“, „UFC draws B“, „Nationality B“, „Birthday B“, „Nick B“, „Referee B“ a byl uložen do souboru „bouts_detailed.csv“.

2.4 Čištění datasetu

Konečně jsme získali dataset se všemi potřebnými informacemi k další analýze a modelování. Nyní musíme tato data vyčistit, upravit a doplnit chybějící hodnoty. Dataset „bouts_detailed“ nyní obsahuje padesát tři sloupců z nichž je velká část zdvojená pro zápasníka „A“ a zápasníka „B“.

Všechny sloupce byly podrobeny základní analýze, kde bylo zjištěno, že sloupce „Nickname A“ a „Nickname B“ se nepodařilo naplnit daty. Sloupce můžeme odstranit a použít místo nich „Nick A“ a „Nick B“, o které jsme data doplnili v minulém kroku.

Podobně jako v kapitole 2.3.1 musíme v tomto datasetu upravit sloupec „Result“, který kromě metody ukončení zápasu obsahuje také její popis. Vznikl tak nový sloupec „Description“ a sloupec „Result“ nyní obsahuje pouze metody ukončení, proto byl jeho název změněn na „Method“. Sloupec „Result“ naplníme skutečným výsledkem zápasu. Protože jsou data strukturovaná tak, že zápasník „A“ je vždy, mimo případů remízy či zápasu bez výsledku, vítěz, sloupec bude obsahovat pouze hodnoty „win“, „NC“ a „draw“. Hodnotu „Draw“ tak nyní můžeme v sloupci „Method“ změnit na „Decision“, jelikož ji už máme ve sloupci „Result“.

Dále bylo ze sloupce „Time“ vytaženo kolo ukončení zápasu a uloženo do sloupce „Finish“. V případě, že zápas skončil na body, byla do sloupce přidána hodnota 0, v opačných případech hodnoty 1 až 5 pro ukončení v prvním až pátém kole.

Pro každý zápas máme dva záznamy o ringovém rozhodčím, každý zápas však má pouze jednoho, proto byly doplněny chybějící hodnoty v sloupci „Referee“ hodnotami ve sloupci „Referee B“ a sloupec „Referee B“ byl smazán.

V dalším kroku byly upraveny sloupce „Height A“, „Height B“, „Reach A“ a „Reach B“, které obsahují hodnoty v palcích a centimetrech. Ponechány byly pouze číselné hodnoty v centimetrech pro snadnější manipulaci.

Sloupce „Weight A“ a „Weight B“, obsahují hmotnost v librách i v kilogramech, ponechány byly pouze číselné hodnoty v librách. Tohle rozhodnutí bylo učiněno z důvodu, že sloupec „Weightclass“ obsahuje také hodnoty v librách a při přepočítávání by mohlo dojít k nepřesnostem.

Ze sloupců „Age A“ a „Age B“ ponecháme pouze roky. Měsíce, týdny a dny jsou u věku zápasníků irelevantní.

Sloupce „Odds A“ a „Odds B“ obsahují americké kurzy zápasníků a jejich kategorii. Kategorii extrahujeme a uložíme do sloupců „Odds cat A“ a „Odds cat B“. V původních sloupcích zůstanou pouze číselné hodnoty.

2.4.1 Odstranění nepotřebných sloupců

Pro další analýzu a modelování už nebudeme potřebovat tyto sloupce, proto byly odstraněny:

1. „Bout Link“,
2. „Time“ – důležitou informaci jsme umístili do sloupce „Finish“,
3. „Event Link“ – ponecháme „Event Name“, který má podobný význam,
4. „Link A“ a „Link B“,
5. „Record A“, „Record B“ – detailní „record“ máme rozložen v jiných sloupcích,
6. „Birthday A“ a „Birthday B“ – ve sloupci „Age A“ a „Age B“ máme věk zápasníků, se kterým se bude lépe pracovat.

Dále bylo rozhodnuto smazat všechny záznamy zápasů, které se odehrály před adopcí „NJSACB“ pravidel, vytvořených v roce 2001 (1.1), protože obsahovali velké množství nekompatibilních dat se zbytkem datasetu. Smazáno bylo 283 řádků.

2.4.2 Doplnění chybějících hodnot

V sloupci „Rounds“ byly nalezeny dva záznamy, které nebyli v souladu s pravidly MMA od zmíněné změny. Tyto záznamy byly manuálně prověřeny a došlo se k závěru, že jsou pouze špatně zapsané a byly změněny na odpovídající hodnotu. 69 chybějících hodnot bylo nahrazeno hodnotou „3 x 5“ pro běžné zápasy a hodnotou „5 x 5“ pro titulové a hlavní zápasy.

Chybějící hodnoty ve sloupci „Odds A“ a „Odds B“ byly nahrazeny hodnotou „+100“, která v amerických kurzech znamená vyrovnané šance. Pro sloupce „Odds cat A“ a „Odds cat B“ to byla hodnota „Even“.

Ve sloupci „Weight A“ a „Weight B“ nahradíme chybějící hodnoty, hodnotou sloupce „Weightclass“. To znamená, že zápasník navážil přesně domluvenou hmotnost.

„Age A“ a „Age B“, tedy věk zápasníků, byl nahrazen průměrem datasetu „30“.

Výšku a rozpětí zápasníků „Height A“, „Height B“, „Reach A“ a „Reach B“ nahradíme průměrem v jejich váhové kategorii. U jednoho zápasu však stále chyběla hodnota „Reach B“, protože pro tuto váhovou kategorii nebyl v datasetu žádný záznam. Informace se nepodařilo dohledat, proto byla hodnota v záznamu nahrazena průměrem nejbližší nižší váhové kategorie.

Po doplnění hodnot byly změněny datové typy sloupců:

1. Na datový typ „int“: „Rounds“, „Odds A“, „Odds B“, „Age A“, „Age B“ a „Finish“.
2. Na datový typ „float“: „Weight A“, „Weight B“, „Height A“.

Dataset byl uložen do souboru „ufc_bouts.csv“.

2.5 Finální podoba datasetu

Dataset „ufc_bouts“ obsahuje 7294 záznamů a má 52 sloupců popsanych v tabulce (Tab. 2.2), z nichž je 19 kategoriálních a 33 numerických.

Kategoriální proměnné jsou: „Fighter A“, „Fighter B“, „Result“, „Venue“, „Date“, „Location“, „Billing“, „Event Name“, „Referee“, „Description“, „Method“, „Title A“, „Nationality A“, „Nick A“, „Odds cat A“, „Title B“, „Nationality B“, „Nick B“, „Odds cat B“.

Numerické proměnné jsou: „Weightclass“, „Rounds“, „Finish“, „Odds A“, „Weight A“, „Age A“, „Height A“, „Reach A“, „KO win A“, „KO loss A“, „Sub win A“, „Sub loss A“, „Dec win A“, „Dec loss A“, „Draws A“, „UFC win A“, „UFC loss A“, „UFC draws A“, „Odds B“, „Weight B“, „Age B“, „Height B“, „Reach B“, „KO win B“, „KO loss B“, „Sub win B“, „Sub loss B“, „Dec win B“, „Dec loss B“, „Draws B“, „UFC win B“, „UFC loss B“, „UFC draws B“.

Tab. 2.2 Popis sloupců tabulky "ufc_bouts" (vlastní zpracování)

Název sloupce	Popis	Název sloupce	Popis
Fighter A	Jméno zápasníka A	Location	Lokace zápasu
Fighter B	Jméno zápasníka B	Billing	Označení zápasu v rámci turnaje
Result	Výsledek zápasu	Event Name	Jméno turnaje

Weightclass	Váhová kategorie, ve které byl zápas vypsán	Referee	Rozhodčí v kleci
Rounds	Počet vypsání kol zápasu	Description	Popis ukončení zápasu
Venue	Aréna, ve které se zápasilo	Method	Metoda ukončení zápasu
Date	Datum zápasu	Finish	Kolo ukončení zápasu
Odds A	Kurz na vítězství zápasníka A	Odds B	Kurz na vítězství zápasníka B
Title A	Titulový status zápasníka A	Title B	Titulový status zápasníka B
Weight A	Výsledek vážení záp. A	Weight B	Výsledek vážení záp. B
Age A	Věk v době zápasu záp. A	Age B	Věk v době zápasu záp. B
Height A	Výška záp. A	Height B	Výška záp. B
Reach A	Rozpětí paží záp. A	Reach B	Rozpětí paží záp. B
KO win A	Počet výher KO/TKO záp. A	KO win B	Počet výher KO/TKO záp. B
KO loss A	Počet proher KO/TKO záp. A	KO loss B	Počet proher KO/TKO záp. B
Sub win A	Počet výher submisí záp. A	Sub win B	Počet výher submisí záp. B
Sub loss A	Počet proher submisí záp. A	Sub loss B	Počet proher submisí záp. B
Dec win A	Počet výher na body záp. A	Dec win B	Počet výher na body záp. B
Dec loss A	Počet proher na body záp. A	Dec loss B	Počet proher na body záp. B
Draws A	Počet remíz záp. A	Draws B	Počet remíz záp. B
Nationality A	Národnost záp. A	Nationality B	Národnost záp. B
Nick A	Přezdívka záp. A	Nick B	Přezdívka záp. B
Odds cat A	Kurzová kategorie na vítězství záp. A	Odds cat B	Kurzová kategorie na vítězství záp. B
UFC win A	Počet výher v UFC záp. A	UFC win B	Počet výher v UFC záp. B
UFC loss A	Počet proher v UFC záp. A	UFC loss B	Počet proher v UFC záp. B
UFC draws A	Počet remíz v UFC záp. A	UFC draws B	Počet remíz v UFC záp. B

3 Analýza klíčových proměnných

Tato část práce se zaměřuje na studium faktorů ovlivňujících výsledky zápasů v získaném datasetu. Zkoumá vliv fyzických charakteristik, jako je výška, rozpětí a věk zápasníků, stejně jako další faktory, jako je nesplnění váhy nebo předešlé výsledky. Dále analyzuje správnost kurzů a vliv externích podmínek na způsob ukončení zápasu. Tato analýza má za cíl poskytnout hlubší vhled do faktorů ovlivňujících výsledky zápasů UFC a přispět k lepšímu porozumění jejich dynamiky.

Analýza byla provedena pomocí Python knihoven pandas, numpy, matplotlib a seaborn v příloze D - „ufc_analysis.ipynb“.

Předměty analýzy:

1. Vliv fyzických charakteristik (výška, rozpětí paží, věk) zápasníků A a B na výsledek zápasu.
2. Vliv jiných charakteristik (nesplnění váhy, titul, národnost) zápasníků A a B na výsledek zápasu.
3. Vliv předešlých výsledků zápasníka A a B na výsledek zápasu.
4. Správnost vypsání kurzů.
5. Vliv externích podmínek zápasu (rozhodčí, váhová kategorie, počet kol, aréna) na způsob ukončení zápasu?
6. Historické změny v rozložení způsobu ukončení zápasů.
7. Vliv předešlých ukončení zápasníků na způsob ukončení zápasu.
8. Rozdíly v ukončení tříkolových a pětikolových zápasů.

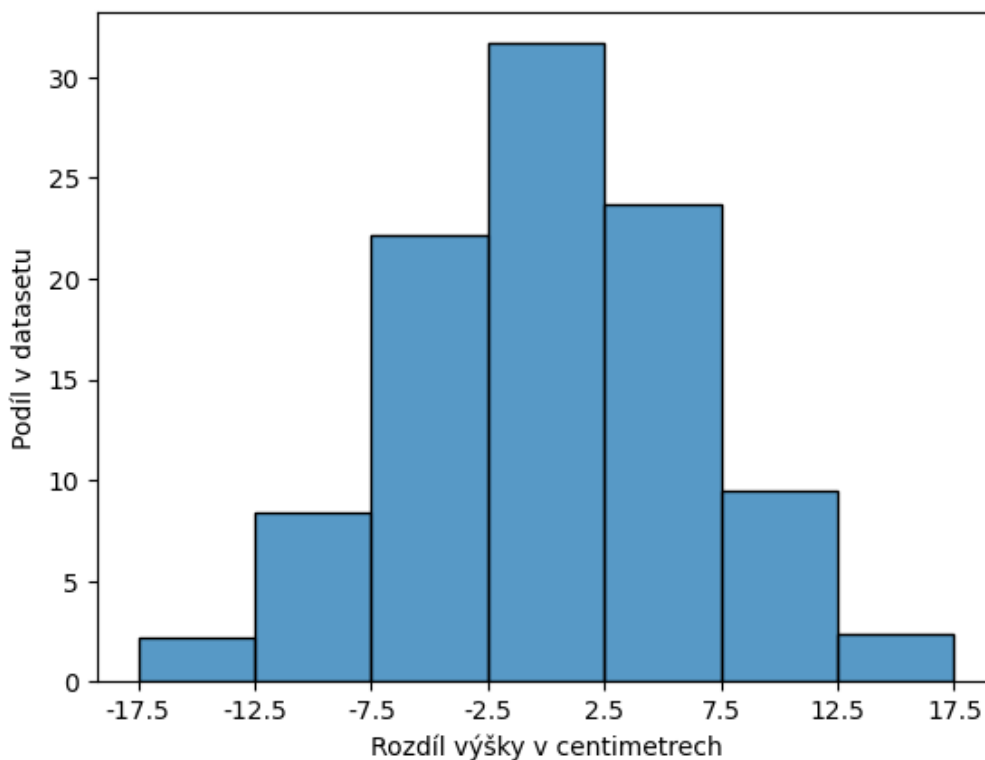
3.1 Vliv fyzických charakteristik na výsledek zápasu

V první části naší analýzy se zaměříme na fyzické charakteristiky zápasníků. Budeme zkoumat výšku, rozpětí a věk zápasníků a sledovat, jak tyto faktory mohou ovlivnit jejich výkonnost a šance na úspěch. Tuto analýzu omezíme pouze na zápasy, které skončili vítězstvím jednoho z bojovníků. Zápasy, které skončily remízou nebo bez výsledku budeme nyní ignorovat.

3.1.1 Rozdíl ve výšce

K prozkoumání, zda má výška vliv na výsledek zápasu byl nejdříve vytvořen sloupec „Height Diff“, který obsahuje rozdíl výšky vítěze („Height A“) a poraženého („Height B“). Na tomto rozdílu byl poté s použitím intervalů po pěti centimetrech od -17,5 cm do 17,5 cm a parametru, který převádí hodnoty do procent, vytvořen histogram (Obr. 3.1).

Z histogramu (Obr. 3.1), který se velmi blíží normálnímu rozdělení je patrné, že nejvyšší podíl zápasů spadá do střední kategorie -2,5 cm až 2,5 cm, tedy kategorie, která obsahuje zápasníky s minimálním výškovým rozdílem. Dále si můžeme všimnout jemného zešíkmení zleva značící, že vyšší zápasníci vítězí častěji, a to konkrétně v 51,8 % případů.



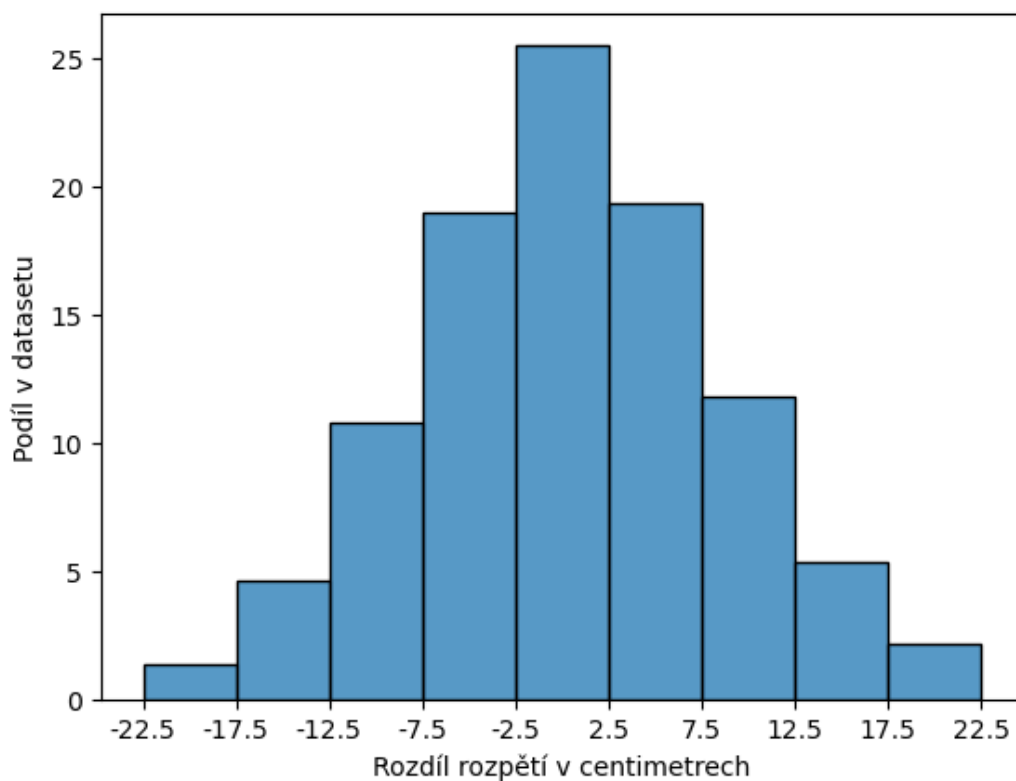
Obr. 3.1 Histogram rozdílu ve výšce zápasníků (vlastní zpracování)

3.1.2 Rozdíl v rozpětí paží

Vliv rozpětí na výsledek zápasu byl zkoumán stejným způsobem jako výška. Byl vytvořen sloupec „Reach Diff“, který obsahuje rozdíl rozpětí vítěze („Reach A“) a poraženého („Reach B“). Na tomto rozdílu byl poté vytvořen histogram (

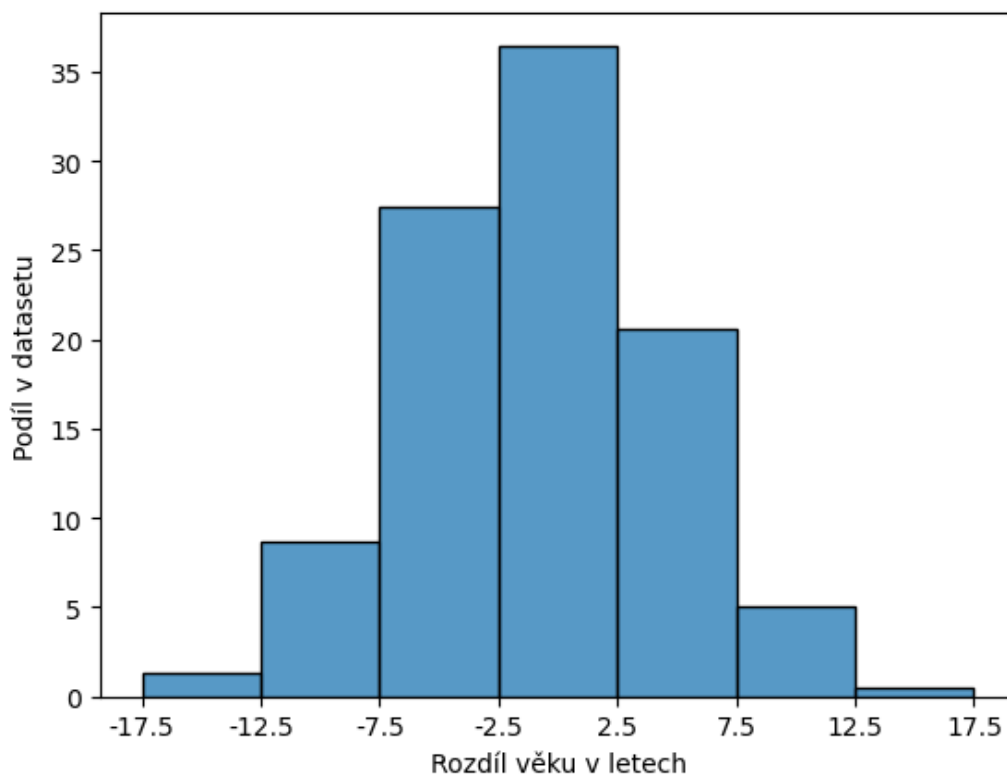
Obr. 3.2), nyní s použitím intervalů po pěti centimetrech, od -22,5 cm do 22,5 cm a parametru, který převádí hodnoty do procent.

Histogram (Obr. 3.2 Histogram rozdílu v rozpětí zápasníků (vlastní zpracování)) vypadá podobně, jako ten s rozdílem ve výšce, s tím rozdílem, že zde je zešíkmení zleva na první pohled vidět až od rozdílu 7,5 centimetrů. Bylo spočítáno, že zápasníci s větším rozpětím vítězí v 52,7 % případů.



Obr. 3.2 Histogram rozdílu v rozpětí zápasníků (vlastní zpracování)

3.1.3 Rozdíl ve věku



Obr. 3.3 Histogram rozdílu věku zápasníků (vlastní zpracování)

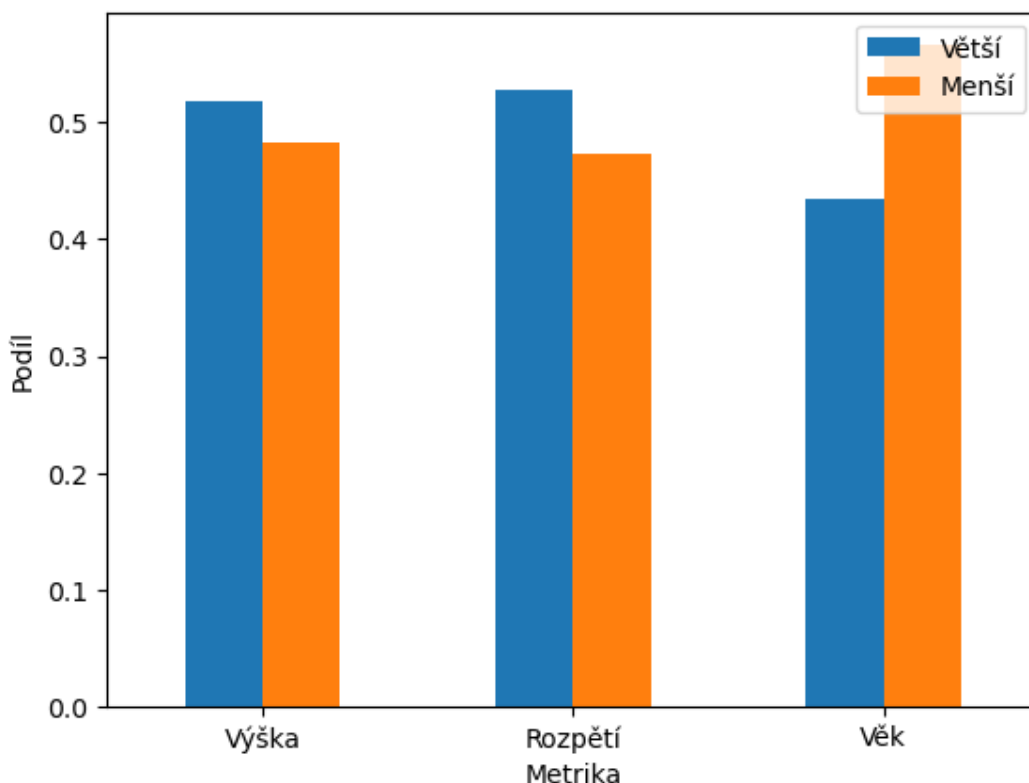
Pro vliv rozdílu ve věku na výsledek zápasu byl vytvořen sloupec „Age Diff“, který obsahuje rozdíl ve věku vítěze („Age A“) a poraženého („Age B“). Na tomto sloupci byl opět vytvořen histogram (

Obr. 3.3) s intervaly po pěti letech od -17,5 do 17,5 a obdobným parametrem.

Histogram je oproti předchozím dvěma výrazněji zešíkmený zprava, což značí výraznější vliv věku na výsledek zápasu. Rozdíl vidíme již na prvním intervalu od středu, kdy podíl vítězných zápasníků mladších o 3 až 7 let je 27,4 %, oproti 20,6 % zápasníků starších o 3 až 7 let. Pro rozdíl 7 až 12 let je to 8,7 % resp. 5,1 %. Celkový podíl vítězných mladších zápasníků proti vítězným starším zápasníkům je 56,6 %.

3.1.4 Srovnání vlivu fyzických charakteristik

V této podkapitole bylo provedeno porovnání vlivu fyzických charakteristik na výhru zápasníků. Graf (**Chyba! Nenalezen zdroj odkazů.**) porovnává vliv výšky, rozpětí a věku na podíl vítězství v zápase. Je z něj patrné, že věk má největší vliv na výhru či prohru zápasníka. Zatímco u výšky a rozpětí je větší hodnota výhodou, u věku je to naopak.



Obr. 3.4 Podíl výšky, rozpětí a věku na vítězství v zápase (vlastní zpracování)

Dále byla vytvořena tabulka (Tab. 3.1 Zastoupení skupin podle výšky, rozpětí a věku (vlastní zpracování)), která seskupuje výhry zápasníků na základě těchto tří charakteristik do skupin a porovnává počet výskytů skupin v datasetu. Pro každou z metrik byl vytvořen sloupec,

který porovnává, jestli je u vítězného zápasníka větší, menší nebo stejná. Pro přehlednost tabulka obsahuje pouze top 5 nejčetnějších skupin.

Největší počet výher měla skupina složená z vyšších, delších a mladších zápasníků, což odpovídá zjištěným skutečnostem. Druhá nejpočetnější skupina, byla skupina nižších, kratších a mladších zápasníků. Můžeme vidět, že i v dalších dvou skupinách, platí, že zápasník s větší výškou má i větší rozpětí a naopak. V těchto případech však mají zápasníci vyšší věk. Až v posledním řádku můžeme vidět neshodu výšky s rozpětím a výrazný pokles v četnosti této skupiny.

Tab. 3.1 Zastoupení skupin podle výšky, rozpětí a věku (vlastní zpracování)

Porovnání výšky	Porovnání rozpětí	Porovnání věku	Počet
Větší	Větší	Menší	1341
Menší	Menší	Menší	938
Menší	Menší	Větší	933
Větší	Větší	Větší	809
Menší	Větší	Menší	373

3.2 Vliv jiných charakteristik na výsledek zápasu

V další části analýzy se budeme věnovat doplňujícím informacím o zápasnících. Pokusíme se zjistit jaký vliv má národnost, nesplnění stanovené váhy či titul vliv na výsledek zápasu. Stejně jako v předešlé části budeme remízy a zápasy bez výsledku ignorovat.

3.2.1 Národnost

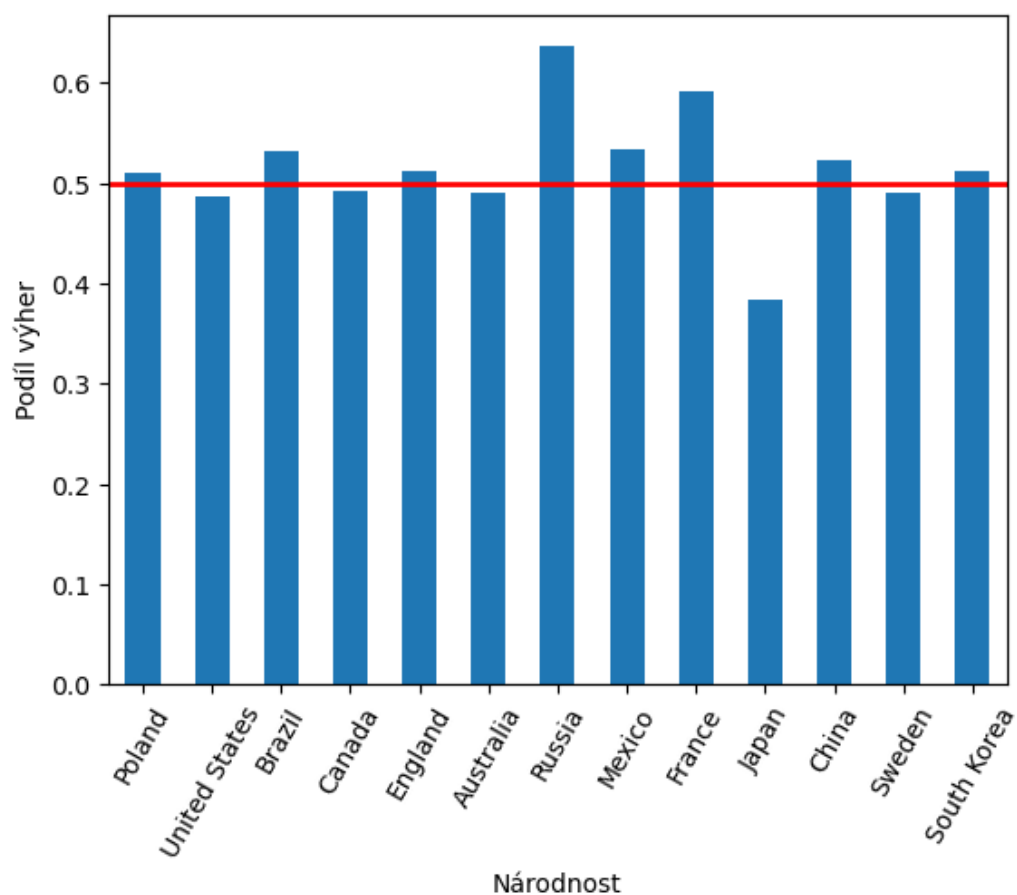
Existují národy, které jsou UFC úspěšnější než jiné? Na tuto otázku se pokusíme odpovědět především prostřednictvím sloupců „Nationality A“ a „Nationality B“.

Nejdříve byl z těchto sloupců získán seznam všech unikátních národností a spočítán počet výher a proher ke každé z nich. Dále byl vypočítán celkový počet zápasů a podíl vítězství. Z těchto informací byla vytvořena tabulka (Tab. 3.2 Výsledky zápasů podle národnosti (vlastní zpracování)) o 13 řádcích, obsahující pouze národnosti, jejichž počet zápasů je alespoň 100. Musíme podotknout, že je velký rozdíl mezi počtem zápasů jednotlivých národů, přičemž „United States“ mají 7743 a „Brazil“ 2127. Ostatních 11 zemí se pohybuje mezi 582 a 127.

Tab. 3.2 Výsledky zápasů podle národnosti (vlastní zpracování)

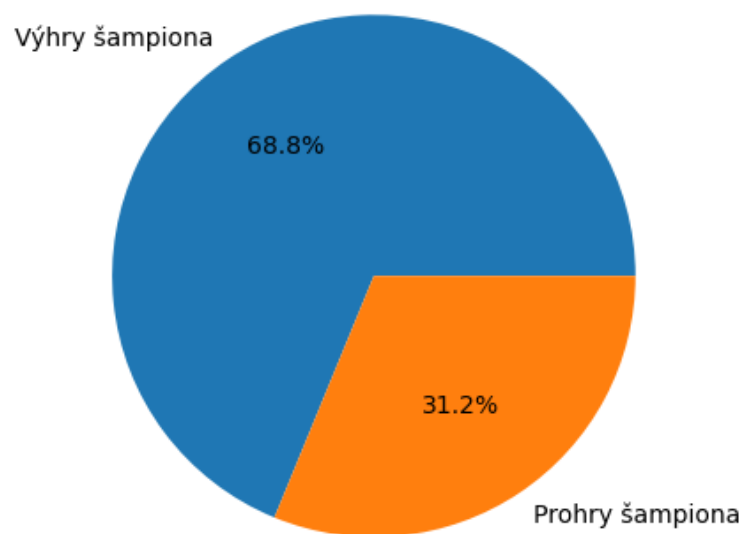
Národnost	Výhry	Prohry	Celkem	Podíl výher
United States	3770	3973	7743	0,487
Brazil	1133	994	2127	0,533
Canada	286	296	582	0,491
England	236	224	460	0,513
Russia	240	137	377	0,637
Australia	139	144	283	0,491
Japan	95	153	248	0,383
Mexico	105	92	197	0,533
Poland	95	91	186	0,511
Sweden	71	74	145	0,490
France	80	55	135	0,593
China	69	63	132	0,523
South Korea	65	62	127	0,512

Z této tabulky (Tab. 3.2) byl poté vytvořen graf (Obr. 3.5), který názorně ukazuje úspěšnost jednotlivých národností. Z grafu je patrné, že většina zemí má úspěšnost kolem 50 %. Je zde však několik zemí, které se od ostatních liší. Úspěšnost Ruska je přibližně 64 %, Francie těsně pod 60 % a Japonska asi 38 %. Vzhledem k vysokému počtu zápasníků z Brazílie je také jejich poměrně vysoká úspěšnost zajímavá. Zápasníci z USA mají naproti tomu druhou nejhorší úspěšnost.



Obr. 3.5 Podíl výher podle národnosti (vlastní zpracování)

3.2.2 Titulové zápasy



Obr. 3.6 Poměr výher a proher šampionů (vlastní zpracování)

V této části nás zajímalo, v kolika zápasech šampion obhájil titul a v kolika naopak o titul přišel. K tomu byly použity sloupce „Title A“ a „Title B“ a hlavně jedna hodnota „Champion“, ze které byl vypočítán počet výskytů ve sloupci „Title A“ (počet výher) a počet výskytů ve sloupci „Title B“ (počet proher). Další záznamy, kdy například titul držel prozatímní šampion, nebo byl titul volný, nebyly použity. Z dat bylo vypočítáno, že šampion vyhrál v 183 zápasech a prohrál pouze v 83 zápasech. Z hodnot byl následně vytvořen koláčový graf (

Obr. 3.6) znázorňující podíl výher šampionů k prohrám šampionů. Úspěšnost šampionů v titulových zápasech je 68,8 %.

3.2.3 Nesplnění stanovené váhy

V MMA komunitě existuje domněnka, že nesplnění stanovené váhy, pomáhá zápasníkům vyhrávat, protože mohou být v zápase těžší nebo méně vyčerpaní. MMA organizace si tuto výhodu uvědomují a v případě, že bojovník překročí stanovenou váhu, strhávají mu část výplaty. Druhý pohled ukazuje na to, že zápasníci, kterým se nepodařilo splnit váhový limit mohou být negativně ovlivněni různými faktory, které k tomuto nesplnění vedli či samotným procesem hubnutí.

Abychom mohli určit, zda bojovník nesplnil váhu, musíme rozlišovat mezi titulovými a netitulovými zápasy a zápasy v „catchweight“⁷. U titulových zápasů a zápasů v „catchweight“ není tolerance žádná, u běžných zápasů je tolerance 1 libra. Na základě těchto předpokladů a sloupců „Weightclass“, „Title A“ a „Title B“ byla vytvořena proměnná „Tolerance“ a naplněna korespondujícími hodnotami. V pracovním datasetu nyní máme 6704 záznamů s tolerancí „1“ a 453 záznamů s tolerancí „0“.

Poté byly vytvořeny 2 sloupce pro zápasníka „A“ a zápasníka „B“, které obsahují hodnotu „True“ pokud zápasník překročil váhový limit včetně tolerance a „False“ pokud váhu i s tolerancí splnil. Z těchto dat byla vytvořena tabulka (Tab. 2.1), která seskupuje hodnoty sloupců pomocí metody „groupby“ a vrací jejich počet.

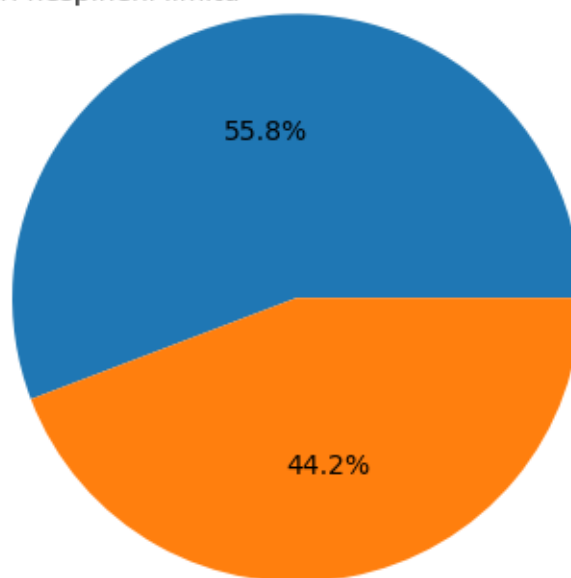
Tab. 3.3 Zastoupení četností podle nesplnění limitu (vlastní zpracování)

Nesplnění limitu „A“	Nesplnění limitu „B“	Počet
False	False	6883
False	True	149
True	False	118

⁷ „Catchweight“ je domluvený váhový limit, který nespádá do žádné ze standartních váhových kategorií.

V tabulce můžeme vidět, že v 6883 případech splnili váhový limit oba zápasníci. Zápasníci, kteří nesplnili váhu prohráli v 149 případech a vyhráli v 118 případech. V 7 případech nesplnili váhu oba zápasníci. Pomocí počtu výher a proher byl vytvořen graf (Obr. 3.7 Poměr výher a proher při nesplnění limitu (vlastní zpracování)), který ukazuje jejich poměr. Pokud zápasník nesplní váhu, je tedy větší pravděpodobnost (55,8 %), že zápas prohraje. Oba pohledy na věc tak nejspíš mají své opodstatnění.

Prohra při nesplnění limitu



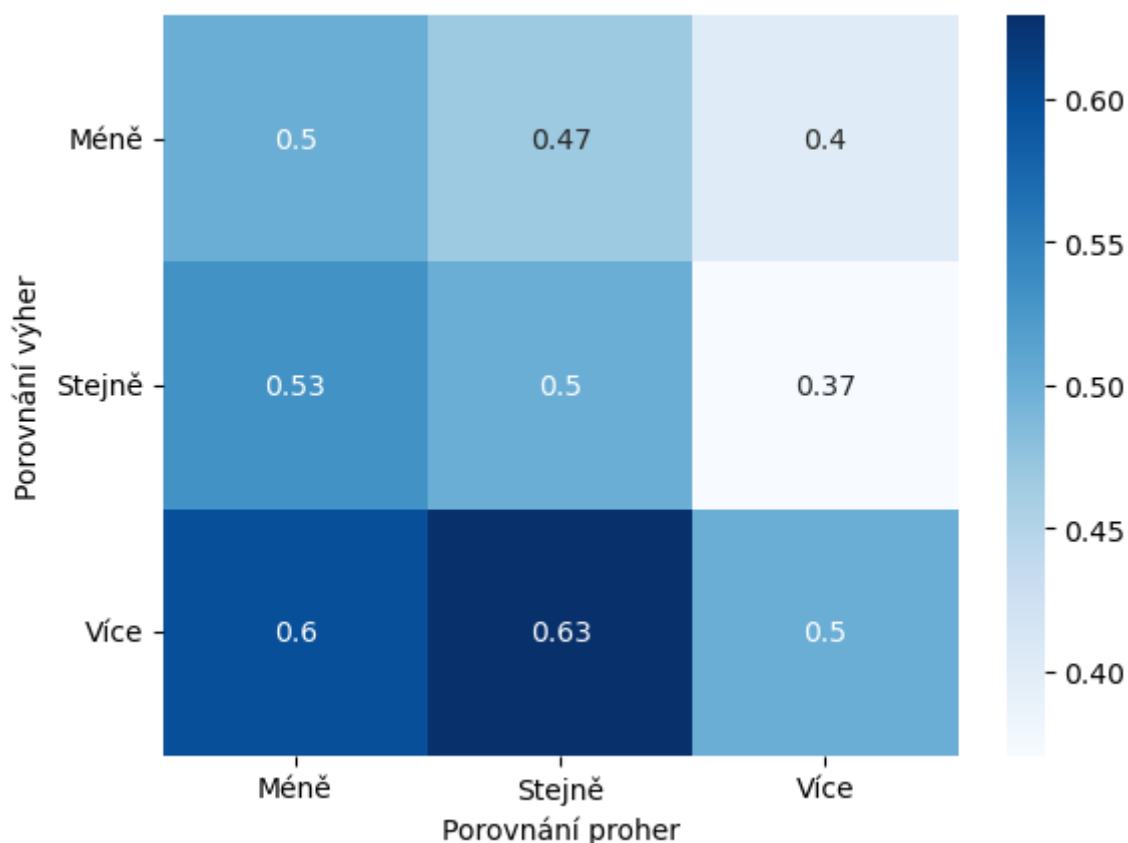
Výhra při nesplnění limitu

Obr. 3.7 Poměr výher a proher při nesplnění limitu (vlastní zpracování)

3.3 Vliv předešlých výsledků

V této části se podíváme na to, jaký vliv mají předešlé výsledky zápasníků „A“ a „B“ v UFC na výsledek zápasů. Stejně jako v předešlých částech budeme remízy a zápasy bez výsledku ignorovat.

Jako první byly vytvořeny sloupce „Win Comparison“ a „Loss Comparison“, které obsahují hodnoty „Více“, „Méně“ a „Stejně“. „Win Comparison“ porovnává, jestli má zápasník „A“ více, méně či stejně výher než zápasník „B“ a „Loss Comparison“ porovnává, jestli má zápasník „A“ více, méně či stejně proher než zápasník „B“. Z těchto sloupců byla poté vytvořena tabulka četností, která byla transformována do kontingenční tabulky. Pro každou hodnotu z kontingenční tabulky byla vypočítána její relativní hodnota a z tabulky vytvořena heatmapa (Obr. 3.8).



Obr. 3.8 Heatmapa porovnávající zastoupení výher a proher (vlastní zpracování)

Z heatmapy (Obr. 3.8) je vidět, že:

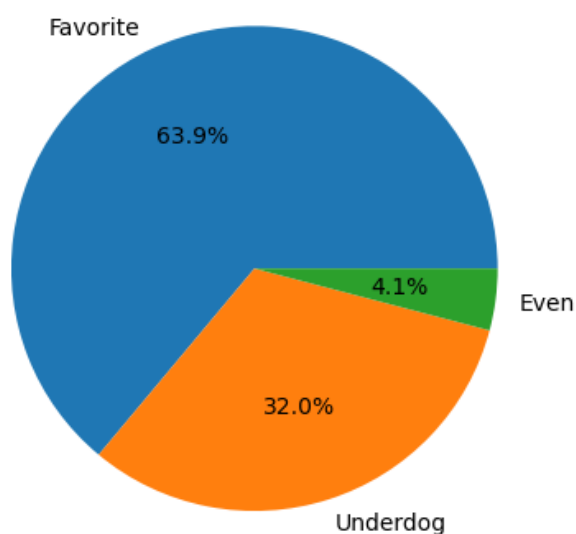
1. Bojovník, který měl stejný počet výher a méně proher než jeho soupeř, vyhrál v 53 % zápasů.
2. Bojovník, který měl větší počet výher a stejný počet proher jako jeho soupeř, vyhrál v 63 % zápasů.
3. Bojovník, který měl více výher a méně proher než jeho soupeř, vyhrál v 60 % zápasů.

Vidíme, že počet výher a počet proher mají vliv na výsledek zápasu. Očekávali bychom však, že výsledky bodů 2. a 3. budou opačné, protože menší počet proher by měl být při porovnání lepší než stejný počet proher. Fakt, že to tak není by mohl být způsoben tím, že některé prohry (do určitého bodu) mohou znamenat více zkušeností.

3.4 Správnost vypsaných kurzů

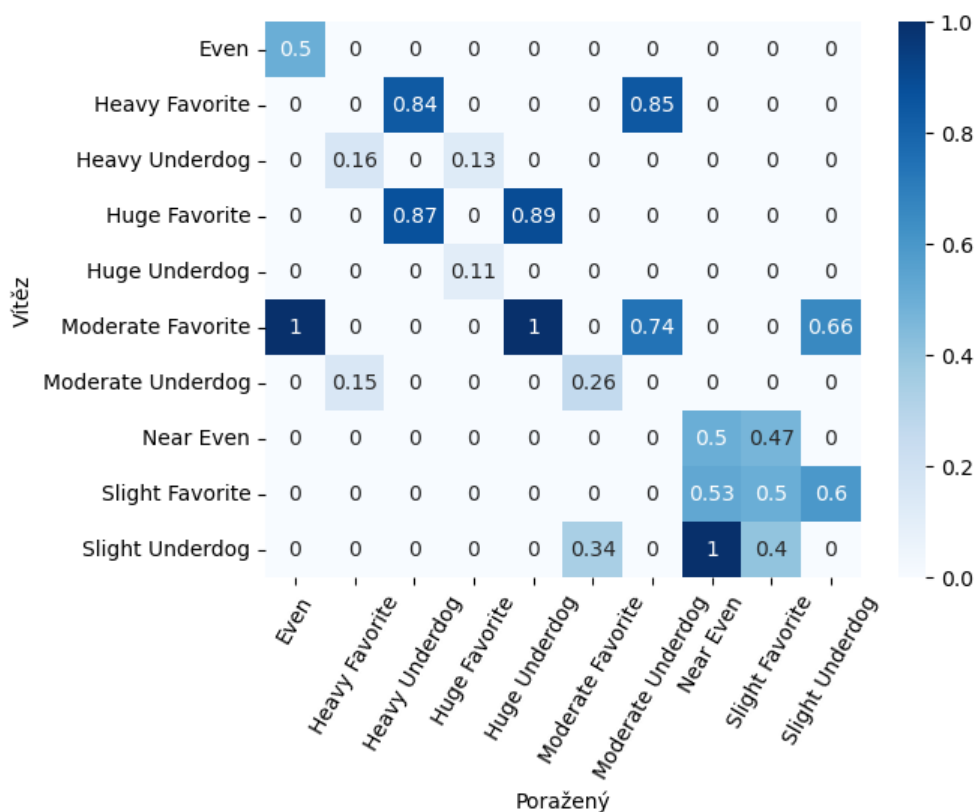
Ve čtvrté části analýzy se podíváme, jak predikovali výsledky zápasů kurzy. Opět použijeme pouze zápasy, které skončili vítězstvím jednoho ze zápasníků.

Nejdříve byl vytvořen sloupec „Prediction“, který porovnával kurzy uložené ve sloupcích „Odds A“ a „Odds B“ a obsahoval hodnoty „Underdog“, „Favorite“ a „Even“. Tento sloupec byl poté promítnut do grafu (Obr. 3.9), ze kterého můžeme vidět, že kurzový favorit vyhrál v 63,9 % případů a outsider vyhrál v 32 % případů. Ve 4,1 % případů byly kurzy vyrovnané.



Obr. 3.9 Poměr vítězů podle kurzů (vlastní zpracování)

Ve zkoumání kurzů můžeme jít hlouběji a použít podrobnější kategorie kurzů ve sloupcích „Odds cat A“ a „Odds cat B“. Z těchto sloupců byla stejně jako v předešlé analýze vytvořena tabulka četností, která byla transformována do kontingenční tabulky. Pro každou hodnotu z kontingenční tabulky byla vypočítána její relativní hodnota a z tabulky vytvořena heatmapa (Obr. 3.10).



Obr. 3.10 Heatmapa porovnávající kurzové kategorie vítěze a poraženého (vlastní zpracování)

Z heatmapy je patrný jeden velmi zajímavý poznatek. „Heavy Favorite“ vyhrává méně procent zápasů proti „Heavy Underdog“ než proti „Moderate Underdog“, což by se nemělo stát, kdyby byly kurzy sestaveny správně.

V mapě mají tři pole hodnotu 1. Ve všech třech případech je to však způsobeno pouze jedním záznamem, proto to není zajímavé. Všechny ostatní hodnoty dávají smysl a jsou predikovány korektně.

3.5 Vliv externích podmínek zápasu na způsob ukončení zápasu

Externí podmínky zápasu, jako je rozhodčí, váhová kategorie, počet kol a místo konání, budou předmětem naší pozornosti v další části analýzy. Budeme zkoumat, jak tyto faktory mohou ovlivnit způsob ukončení zápasu.

Nejprve se podíváme na způsoby ukončení v celém datasetu, abychom měli s čím porovnávat. Z tabulky (Tab. 3.4 Způsoby ukončení zápasu (vlastní zpracování)) můžeme vidět, že nejvíce zápasů končí v 47 % případů na body („Decision“), 33 % zápasů končí knockoutem („KO/TKO“), 20 % submisí („Submission“) a 1 % bez výsledku („No Contest“).

Tab. 3.4 Způsoby ukončení zápasu (vlastní zpracování)

Způsob ukončení	Podíl
Decision	0,469
KO/TKO	0,325
Submission	0,195
No Contest	0,012

Dále se pokusíme zjistit, jaký na to mají vliv jiné faktory.

3.5.1 Rozhodčí

Rozhodčí hraje určitou roli v každém zápase. Může však mít vliv na způsob ukončení? K této analýze bude použit primárně sloupec „Referee“.

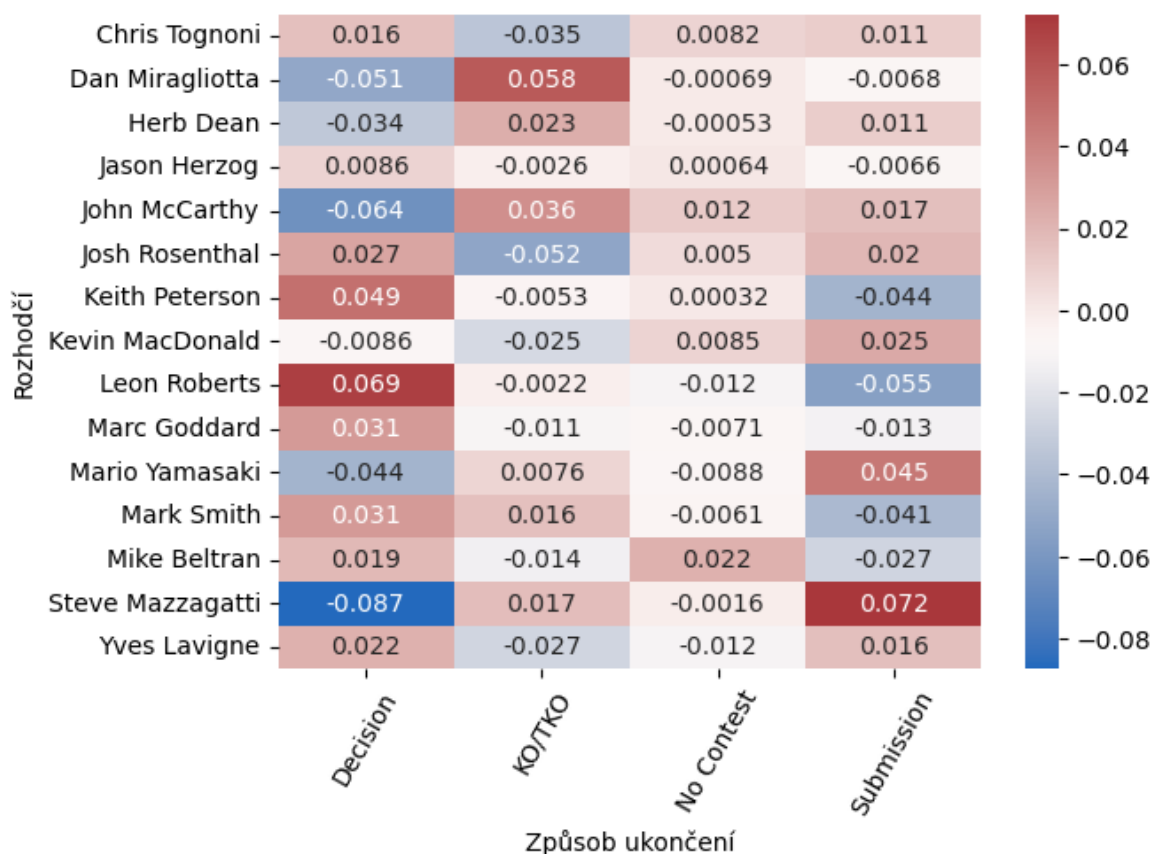
Nejdříve byli vyfiltrováni pouze rozhodčí, kteří figurovali alespoň u 100 zápasů. Takových rozhodčích bylo v datasetu 15 a rozhodovali 5195 z 7294 zápasů. Byla použita funkce „groupby“, která pro každou kombinaci rozhodčího a způsobu ukončení spočítala počet četností. Z těchto dat byla vytvořena kontingenční tabulka (Tab. 3.5), která obsahuje podíl způsobů ukončení pro každého rozhodčího.

Tab. 3.5 Podíl způsobů ukončení pro každého rozhodčího (vlastní zpracování)

Rozhodčí	Decision	KO/TKO	No Contest	Submission
Chris Tognoni	0,485	0,290	0,020	0,206
Dan Miragliotta	0,418	0,383	0,011	0,188
Herb Dean	0,435	0,348	0,011	0,206
Jason Herzog	0,477	0,322	0,012	0,188
John McCarthy	0,405	0,360	0,023	0,212
Josh Rosenthal	0,496	0,273	0,017	0,215
Keith Peterson	0,518	0,320	0,012	0,151
Kevin MacDonald	0,460	0,300	0,020	0,220
Leon Roberts	0,538	0,323	0,000	0,140
Marc Goddard	0,500	0,314	0,004	0,182
Mario Yamasaki	0,424	0,332	0,003	0,241
Mark Smith	0,500	0,341	0,005	0,154
Mike Beltran	0,487	0,311	0,034	0,168
Steve Mazzagatti	0,381	0,342	0,010	0,267
Yves Lavigne	0,491	0,298	0,000	0,211

Pro každou hodnotu v tabulce, Tab. 3.5 Podíl způsobů ukončení pro každého rozhodčího (vlastní zpracování), byla odečtena korespondující hodnota z tabulky, Tab. 3.4 Způsoby ukončení zápasu (vlastní zpracování), k určení odchylky rozhodčích od průměru. Nově vzniklá tabulka pak byla vizualizována pomocí heatmapy (Obr. 3.11 Odchylka rozhodčích od průměru ve způsobu ukončení (vlastní zpracování)).

Je nutno podotknout, že z principu věci, rozhodčí hraje roli hlavně u „KO/TKO“ ukončení, protože je často na něm, aby určil, zda je bojovník schopen se bránit nebo ne. V případě ukončení na „Submission“ je zastavení zápasu ve většině případů na samotných bojovnících.



Obr. 3.11 Odchylka rozhodčích od průměru ve způsobu ukončení (vlastní zpracování)

V heatmapě (Obr. 3.11) můžeme vidět, že:

1. „Dan Miragliotta“ má asi 5,8 % kladnou odchylku v počtu „KO/TKO“.
2. „John McCarthy“ má asi 3,6 % kladnou odchylku v počtu „KO/TKO“.
3. „Jack Rosenthal“ má asi 5,2 % zápornou odchylku v počtu „KO/TKO“.
4. „Chris Tognoni“ má asi 3,5 % zápornou odchylku v počtu „KO/TKO“.
5. „Mike Beltran“ má 2,2 % kladnou odchylku v „No Contest“ ukončeních, což je v absolutních číslech téměř trojnásobek, než jaká je norma pro celý soubor dat.
6. „Jason Herzog“ vykazuje nejmenší celkovou odchylku, což by mohlo znamenat, že nejlépe určuje, zda by měl být zápas zastaven.

Kladná odchylka v počtu „KO/TKO“ by mohla znamenat, že rozhodčí zastavuje zápasy dříve, než by měl. Naproti tomu záporná odchylka v počtu „KO/TKO“ by mohla znamenat opačný problém.

3.5.2 Aréna

Zde nás zajímá především to, jak jedno konkrétní místo ovlivňuje způsob, jakým zápas skončí. Po vypuknutí Covidu-19 začalo UFC pořádat některé turnaje v malé tréninkové aréně zvané „UFC Apex“. Samotná klec, ve které se zápasí, je v této aréně menší a zápas sleduje jen asi stovka fanoušků (většina z nich jsou přátelé a rodiny bojovníků).

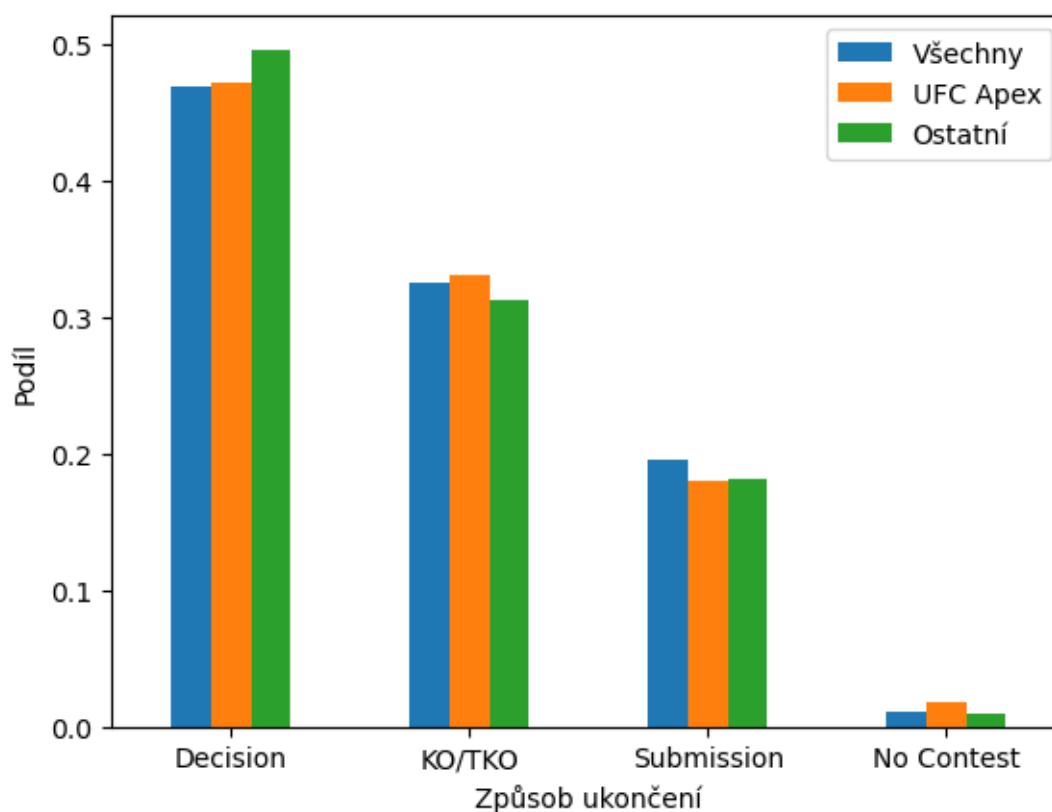
V našem datasetu je tato aréna reprezentována dvěma názvy „UFC Apex“ a „UFC APEX“, které budeme od tohoto momentu nazývat pouze prvním výrazem. V prvním kroku byla data rozdělena právě na „UFC Apex“ a všechny ostatní. Abychom byli korektní a eliminovali změnu v poměru ukončení v čase, vybíráme pouze zápasy, které se odehrály až po prvním turnaji v „UFC Apex“. První zápas v „UFC Apex“ se odehrál v roce 30. května 2020, takže na základě tohoto data filtrujeme všechny ostatní zápasy. Z dat je patrné, že po zavedení „UFC Apex“ se zde odehrálo více zápasů – 1080, než zápasů v jiných arénách dohromady - 901.

Pomocí těchto dat byl vytvořen seskupený sloupcový graf (Obr. 3.12), který porovnává způsoby ukončení pro všechny turnaje, turnaje v „UFC Apex“ a turnaje mimo „UFC Apex“.

Vidíme, že:

1. V „UFC Apex“ končí přibližně o 2,5 % méně zápasů bodovým rozhodnutím než v ostatních arénách.
2. V „UFC Apex“ je téměř dvojnásobný počet zápasů „No Contest“ než v jiných arénách i v celém souboru dat.

Tyto skutečnosti by mohly být způsobeny menším prostorem pro zápasy, který nutí bojovníky k většímu nasazení.

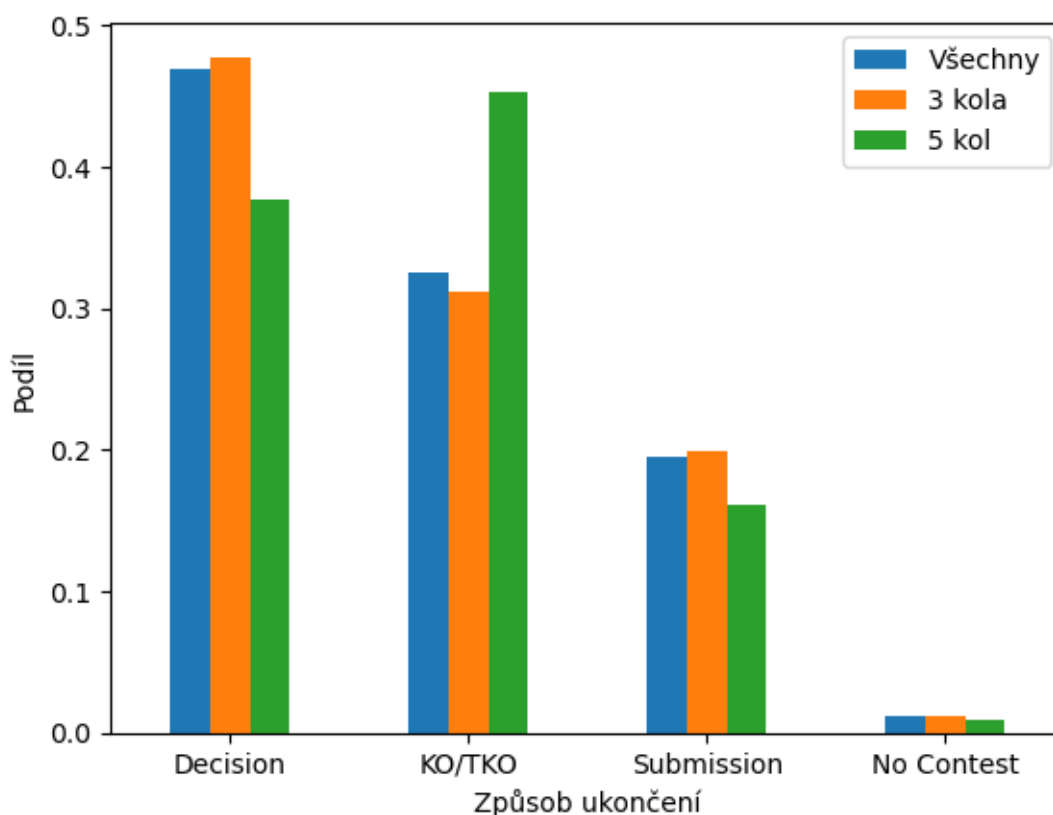


Obr. 3.12 Porovnání způsobů ukončení pro různé arény (vlastní zpracování)

3.5.3 Počet kol

V této části se zaměříme na to, jak se liší způsoby ukončení pro tříkolové a pětikolové zápasy. Pro oba typy zápasů byly spočítány podíly způsobů ukončení, které jsme poté vložili do grafu (Obr. 3.14). Mezi pětikolovými a tříkolovými zápasy je velký rozdíl v počtu zápasů ukončených „KO/TKO“ a bodovým rozhodnutím. Znatelný rozdíl můžeme pozorovat i v ukončeních na „Submission“.

1. Pětikolové zápasy končí v 38 % případů „Decision“, v 45 % případů na „KO/TKO“ a v 16 % případů na „Submission“.
2. Tříkolové zápasy končí v 48 % případů „Decision“, v 31 % případů na „KO/TKO“ a v 20 % případů na „Submission“.



Obr. 3.13 Způsoby ukončení pro tříkolové a pětikolové zápasy (vlastní zpracování)

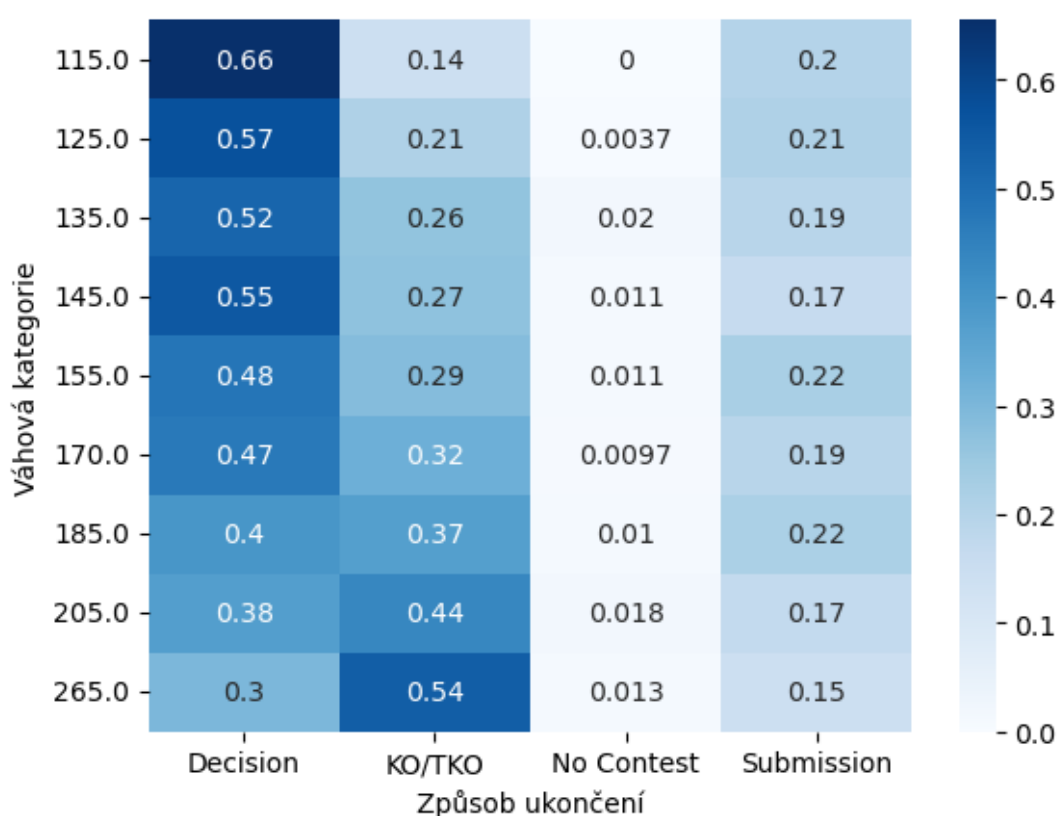
3.5.4 Váhové kategorie

V datasetu je 9 hlavních váhových kategorií („115“, „125“, „135“, „145“, „155“, „170“, „185“, „205“, „265“) a několik dalších „Catchweight“ kategorií, které pro tuto část analýzy odfiltrujeme.

V prvním kroku byla použita funkce „groupby“, která pro každou kombinaci váhové kategorie a způsobu ukončení spočítala počet četností. Z těchto dat byla vytvořena kontingenční tabulka a heatmapa (Obr. 3.14), která ukazuje podíl způsobů ukončení pro každou kategorii.

V heatmapě (Obr. 3.14 Podíl způsobů ukončení pro váhové kategorie (vlastní zpracování)) můžeme vidět:

1. Klesající trend v počtu ukončení na „Decision“ s rostoucí váhovou kategorií.
2. Se zvyšující se váhovou kategorií roste podíl „KO/TKO“.
3. U zápasů končících na „No Contest“ a „Submission“ nepozorujeme žádný významný vzorec.
4. Podíl zápasů končících na „KO/TKO“ převyšuje podíl zápasů končících na „Submission“ ve váhové kategorii „135“ a podíl zápasů končících na „Decision“ ve váhové kategorii „205“.

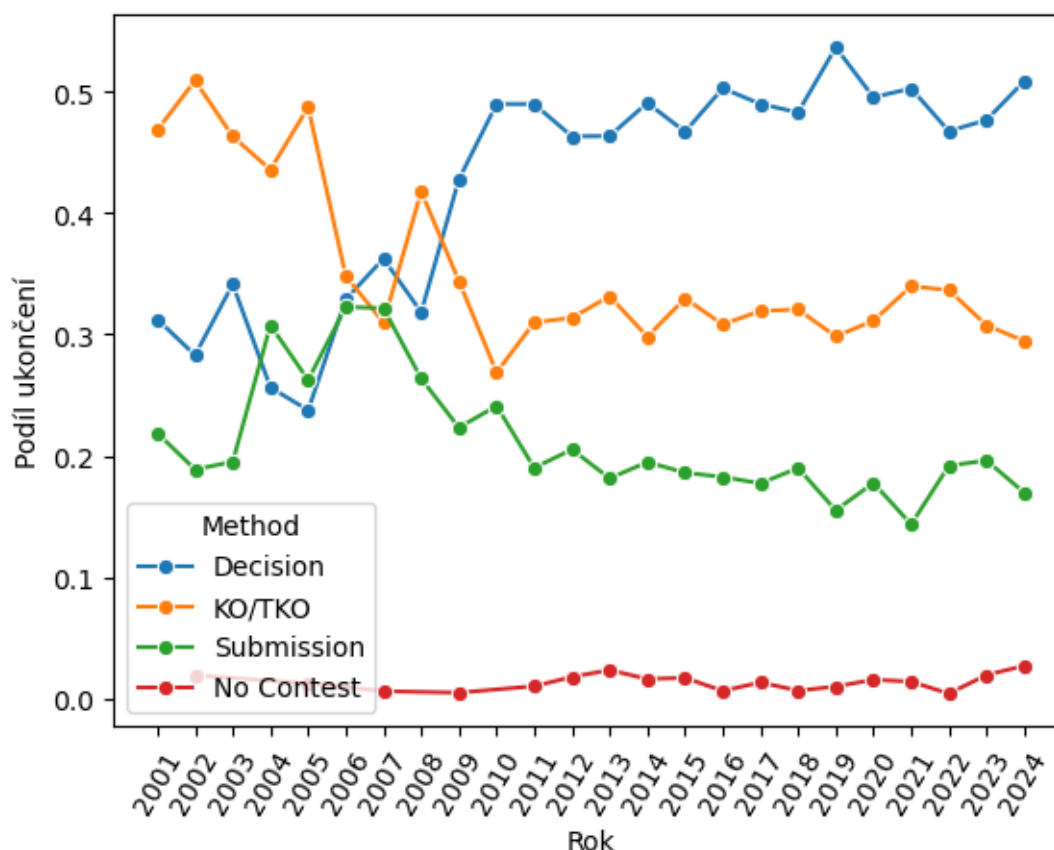


Obr. 3.14 Podíl způsobů ukončení pro váhové kategorie (vlastní zpracování)

3.6 Historické změny v rozložení způsobu ukončení zápasů

MMA je poměrně nový sport a stále se velmi rychle vyvíjí. I přes to, že jsme již data omezili na zápasy po představení společných pravidel, se sport dost změnil. V této části prozkoumáme změnu ve způsobech ukončení zápasů.

Ze sloupce „Date“ byl extrahován pouze rok a uložen do sloupce „Year“. Poté byl pomocí funkce „groupby“ nalezen počet ukončení pro každou kombinaci roku a způsobu ukončení. Pro každý rok byl poté vypočítán relativní podíl způsobů ukončení. Tato data se následně vložila do linkového grafu (Obr. 3.15), který vývoj způsobu ukončení přehledně ukáže.



Obr. 3.15 Vývoj způsobu ukončení (vlastní zpracování)

Z grafu (Obr. 3.15) vidíme, že rozložení metod se výrazně měnilo hlavně mezi lety 2003 a 2011.

1. V roce 2001 bylo nejrozšířenějším způsobem ukončení zápasu „KO/TKO“ s přibližně 48 %, následovalo bodové rozhodnutí s přibližně 30 % a submise s přibližně 22 %.
2. V roce 2004 došlo k velkému nárůstu počtu submisí, který dosáhl přibližně 30 %, a k poklesu bodových rozhodnutí na 25 %.
3. V roce 2006 mají všechny hlavní způsoby ukončení stejný podíl, přibližně 33 %.
4. Od roku 2007 do roku 2011 počet submisí klesal, a nakonec se ustálil mírně pod úrovní 20 %.
5. V roce 2008 došlo k velkému vzestupu ukončení na „KO/TKO“, které dosáhlo 40 %, ale v následujících letech klesalo a ustálilo se na úrovni kolem 30 %.
6. Vítězství bodovým rozhodnutím rostla od roku 2005 až do roku 2010 s mírným poklesem v roce 2008.

Od roku 2011 se metody stabilizovaly: „Decision“ kolem 50 %, „KO/TKO“ kolem 30 % a „Submission“ kolem 20 %.

3.7 Vliv předešlých ukončení zápasníků na způsob ukončení zápasu

V této části byla provedena korelační analýza mezi proměnnými uchovávanými záznamy o předešlých způsobech ukončení zápasníků a samotným způsobem ukončení zápasu. Analýza byla provedena pouze na datech, kde byl určen vítěz. Remízy a zápasy bez výsledku byly vynechány.

Nejdříve byl pomocí „sklearn“ knihovny převeden sloupec „Method“, tak aby pro každou hodnotu vznikl sloupec s hodnotou „True“ nebo „False“. Poté byly zvoleny všechny sloupce, které obsahují záznamy o předešlých ukončeních zápasníků („KO win A“, „KO loss A“, „Sub win A“, „Sub loss A“, „Dec win A“, „Dec loss A“, „Draws A“, „KO win B“, „KO loss B“, „Sub win B“, „Sub loss B“, „Dec win B“, „Dec loss B“, „Draws B“), sloupce nově vzniklé z „Method“ („Encoded_Decision“, „Encoded_Submission“, „Encoded_KO/TKO“) a na nich byla pomocí funkce „corr“ knihovny pandas vytvořena korelační tabulka (Tab. 3.6).

Tab. 3.6 Výše z korelační tabulky pro metody ukončení zápasu (vlastní zpracování)

Proměnná	Encoded_Decision	Encoded_KO/TKO	Encoded_Submission
KO win A	-0,059	0,188	-0,148
KO loss A	-0,048	0,039	0,013
Sub win A	-0,066	-0,118	0,222
Sub loss A	-0,018	0,032	-0,014
Dec win A	0,159	-0,081	-0,103
Dec loss A	0,034	-0,035	-0,002
Draws A	-0,002	-0,016	0,021
KO win B	-0,065	0,071	-0,003
KO loss B	-0,094	0,141	-0,049
Sub win B	0,017	0,017	-0,042
Sub loss B	-0,052	-0,020	0,089
Dec win B	0,105	-0,052	-0,069
Dec loss B	0,134	-0,068	-0,087
Draws B	0,021	-0,001	-0,024

Z korelační tabulky (Tab. 3.6) je patrné, že ačkoliv jsou korelační koeficienty slabé, existují některé proměnné, které korelují s metodami ukončení více než jiné.

1. Nejvyšší korelační koeficienty ve sloupci pro bodové rozhodnutí jsou „Dec win A“ (0,16), „Dec loss B“ (0,13) a „Dec win B“ (0,1). To znamená, že pro metodu „Decision“ existuje nejsilnější vztah k vítězstvím na „Decision“ obou zápasníků a prohrám na „Decision“ poraženého zápasníka.
2. Nejvyšší korelační koeficienty pro „KO/TKO“ jsou „KO win A“ (0,19) a „KO loss B“ (0,14). To znamená, že pro metodu „KO/TKO“ existuje nejsilnější vztah s „KO/TKO“ vítězstvími vítězného bojovníka a „KO/TKO“ prohrami poraženého bojovníka.
3. Nejvyšší korelační koeficienty ve sloupci pro submisi jsou „Sub win A“ (0,22), „KO win A“ (-0,15), „Dec win A“ (-0,1). To znamená, že pro metodu „Submission“ existuje nejsilnější vztah k výhrám na submisi, knockout a na body vítězného bojovníka.

Je zajímavé, že sloupce pro „KO/TKO“ a „Decision“ mají alespoň nějaký vztah (nad 0,1) s proměnnými vítězného i poraženého bojovníka, ale sloupec pro „Submission“ má vztah pouze s proměnnými vítězného bojovníka.

3.8 Rozdíly v ukončení tříkolových a pětikolových zápasů

V předešlé části analýzy jsme zjistili, že počet kol má vliv na způsob ukončení zápasu. Nyní nás bude zajímat, v jakém kole tyto zápasy končí, na základě počtu kol.

Nejdříve byla data rozdělena podle sloupce „Rounds“ na tříkolové a pětikolové zápasy. Z těchto dat a sloupce „Finish“, který obsahuje kolo ukončení, se poté vytvořily tabulky (Tab. 3.7,

Tab. 3.8) obsahující podíl počtu ukončení v daném kole.

Můžeme vidět, že jak v pětikolových, tak v tříkolových zápasech s každým dalším kolem, klesá pravděpodobnost, že zápas skončí předčasně. Podíl zápasů, které skončili v prvním, nebo druhém kole je u tříkolových zápasů větší než u pětikolových, ale podíl zápasů, které skončili ve třetím kole je větší u pětikolových zápasů.

Tab. 3.7 Podíl počtu ukončení pro tříkolové zápasy (vlastní zpracování)

Kolo ukončení	Podíl na ukončení
0 ⁸	0,489
1	0,269
2	0,164

⁸ „0“ v tabulce u kola ukončení znamená, že zápas nebyl předčasně ukončen.

3	0,077
---	-------

Tab. 3.8 Podíl počtu ukončení pro pětikolové zápasy (vlastní zpracování)

Kolo ukončení	Podíl na ukončení
0	0,386
1	0,236
2	0,161
3	0,108
4	0,064
5	0,045

4 Prediktivní modely

Tato část představuje důležitý krok výzkumu, neboť se zaměřuje na aplikaci teoretických poznatků a analýz na praktické modely, které nám pomohou s predikcí MMA zápasů.

Tato kapitola se bude věnovat aplikaci několika různých statistických modelů, mezi které patří zejména modely logistické regrese a Elo model, který byl původně vyvinut pro hodnocení hráčů v šachu. Tyto modely porovná s dalšími modely strojového učení, jako jsou například SVM, rozhodovací stromy, k-NN a náhodný les. Budeme zkoumat, jak modely zohledňují různé faktory ovlivňující výsledky zápasů, včetně fyzických charakteristik zápasníků, jejich historických výsledků, národnosti, a dalších relevantních proměnných. Na základě těchto modelů budeme predikovat jak výsledky zápasů, tak i způsob jejich ukončení. Modely budou implementovány pomocí knihoven „sklearn“ a „skelo“ v příloze E – „ufc_prediction.ipynb“.

4.1 Predikce výsledku zápasu

V první části se budeme zabývat predikcí výsledku zápasu. Pro tuto analýzu byly z datasetu odstraněny všechny remízy a zápasy bez výsledku. Jak již bylo zmíněno, zápasy MMA končí remízou jen velmi zřídka za velmi specifických podmínek, které mohou nastat až v průběhu zápasu, tudíž je nelze spolehlivě predikovat.

4.1.1 Logistická regrese

Úprava dat

K vytvoření strojově učených modelů, musíme nejdříve náš dataset upravit. Kvůli způsobu získávání dat je zápasník „A“ vždy vítěz a zápasník „B“ poražený. To bylo nutno před aplikací modelů upravit. Data byla zkopírována, zápasník „A“ a „B“ byli prohozeni a výsledek zápasu byl změněn z „win“ na „loss“. Tato obrácená kopie pak byla „nalepena“ na existující data.

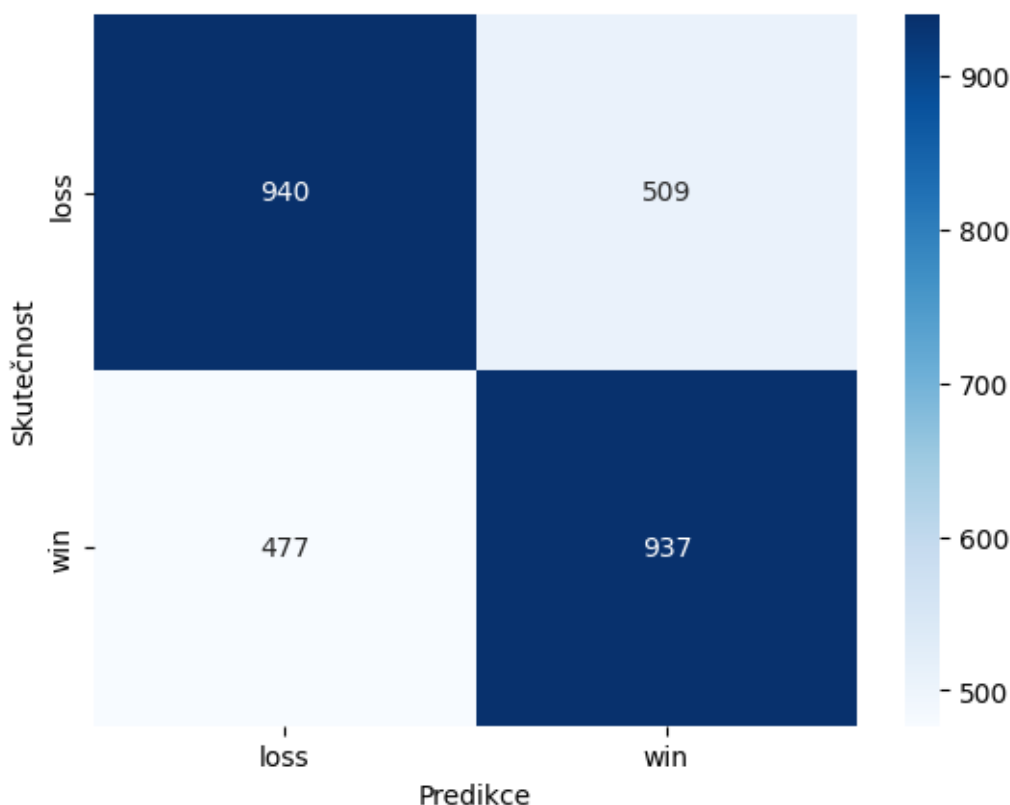
Poté byly vytvořeny sloupce „Weight Miss A“ a „Weight Miss B“ udávající, zda zápasník nedodržel stanovený váhový limit, stejně jako v kapitole „3.2.3 Nesplnění stanovené váhy“. Pomocí numerických proměnných pro zápasníka „A“: „Age A“, „Height A“, „Reach A“, „KO win A“, „KO loss A“, „Sub win A“, „Sub loss A“, „Dec win A“, „Dec loss A“, „Draws A“, „UFC win A“, „UFC loss A“, „UFC draws A“ a zápasníka „B“: „Age B“, „Height B“, „Reach B“, „KO win B“, „KO loss B“, „Sub win B“, „Sub loss B“, „Dec win B“, „Dec loss B“, „Draws B“, „UFC win B“, „UFC loss B“, „UFC draws B“, jsme získali jejich rozdíly. Všechny chybějící hodnoty ve sloupcích „Title A“ a „Title B“ byly nahrazeny hodnotou „No“.

Tvorba modelu

Pro samotnou tvorbu modelu byly vybrány sloupce „Title A“, „Title B“, „Odds A“, „Odds B“, „Odds cat A“, „Odds cat B“, „Nationality A“, „Nationality B“, „Weight Miss A“, „Weight Miss B“, „Age Diff“, „Height Diff“, „Reach Diff“, „KO win Diff“, „KO loss Diff“, „Sub win Diff“, „Sub loss Diff“, „Dec win Diff“, „Dec loss Diff“, „Draws Diff“, „UFC win Diff“, „UFC loss Diff“, „UFC draws Diff“ jako nezávislé proměnné. Sloupec „Result“ byl uložen s hodnotou 0 pro „loss“ a 1 pro „win“ do závislé proměnné. Z kategoriálních proměnných byly vytvořeny „umělé proměnné“ obsahující pouze hodnoty „True“ a „False“. Nezávislé proměnné byly standardizovány a data rozdělena na trénovací (80 %) a testovací (20 %) část. Model logistické regrese byl vytvořen pomocí funkce „GridSearchCV“, která se používá pro ladění modelu a výběru nejlepších parametrů.

Predikce

Jako nejlepší parametry byly vybrány „C“: 0,00001, „penalty“: „l2“ a „solver“: „liblinear“. Nízká hodnota „C“ značí silnou regularizaci. Tento model predikoval z 2863 hodnot 940 správně jako „loss“, 937 správně jako „win“, 477 špatně jako „loss“ a 509 špatně jako „win“ (Obr. 4.1). Celková správnost byla 65,6 %. Model měl o něco lepší přesnost pro predikci prohry (66,3 %) než výhry (64,8 %) a lepší úplnost pro výhru (66,3 %) než pro prohru (64,9 %).



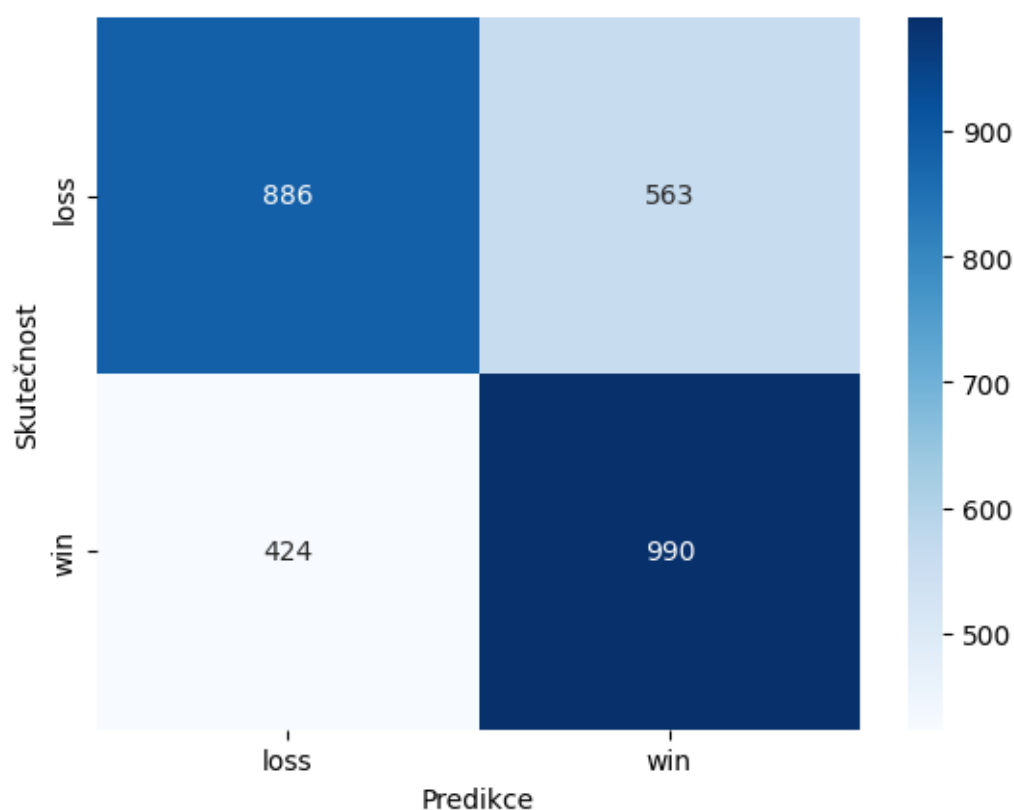
Obr. 4.1 Matice záměn výsledku zápasů pro model logistické regrese (vlastní zpracování)

4.1.2 Další modely strojového učení

Jako další modely strojového učení pro porovnání s logistickou regresí byly vybrány modely SVM, k-NN, rozhodovací strom a náhodný les. Tyto modely prošly stejnou úpravou dat a procesem tvorby modelu jako logistická regrese.

SVM

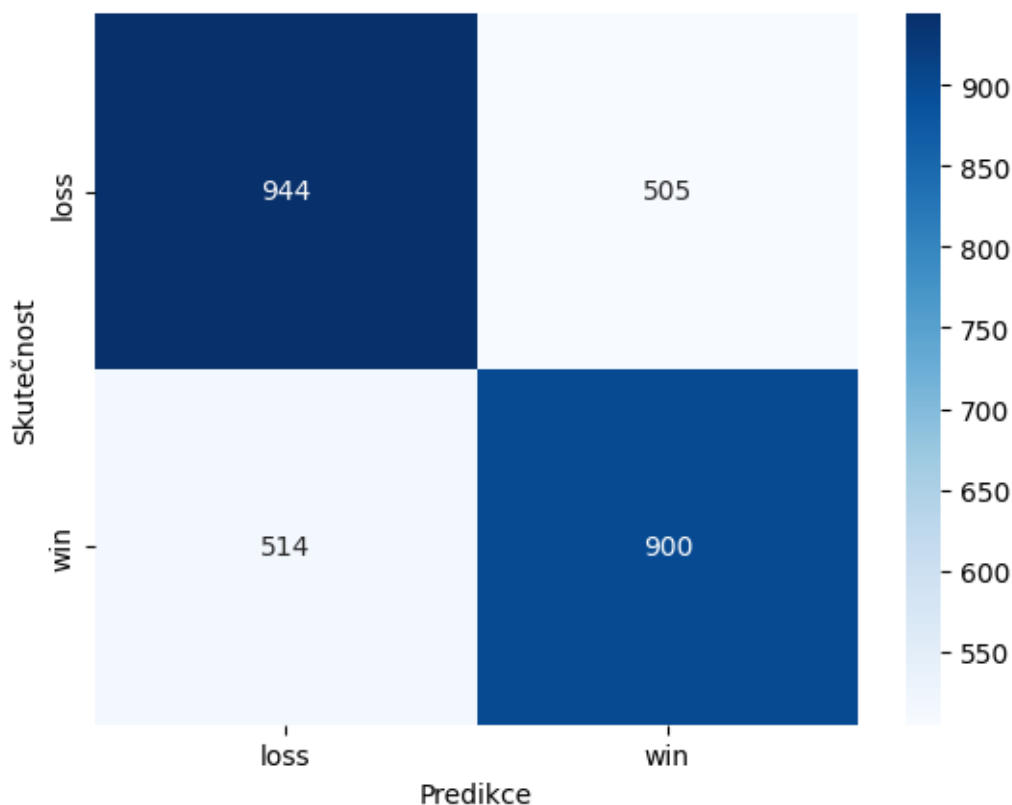
Jako nejlepší parametry pro model podpůrných vektorů byly zvoleny „C“: 0,001, a „kernel“: „linear“. Model SVM byl však pro takové množství dat velmi pomalý a možnosti výběru parametrů velmi omezené. Z 2863 hodnot model predikoval 886 správně jako „loss“, 990 správně jako „win“, 424 špatně jako „loss“ a 563 špatně jako „win“ (Obr. 4.2). Tento model má stejnou správnost jako model logistické regrese, ale častěji predikuje výhru. Přesnost pro „loss“ (67,8 %) je však vyšší než pro „win“ (63,7 %). Úplnost pro výhru je přesně 70 %, pro prohru pak 61,1 %



Obr. 4.2 Matice záměn výsledku zápasu pro model podpůrných vektorů (vlastní zpracování)

Rozhodovací strom

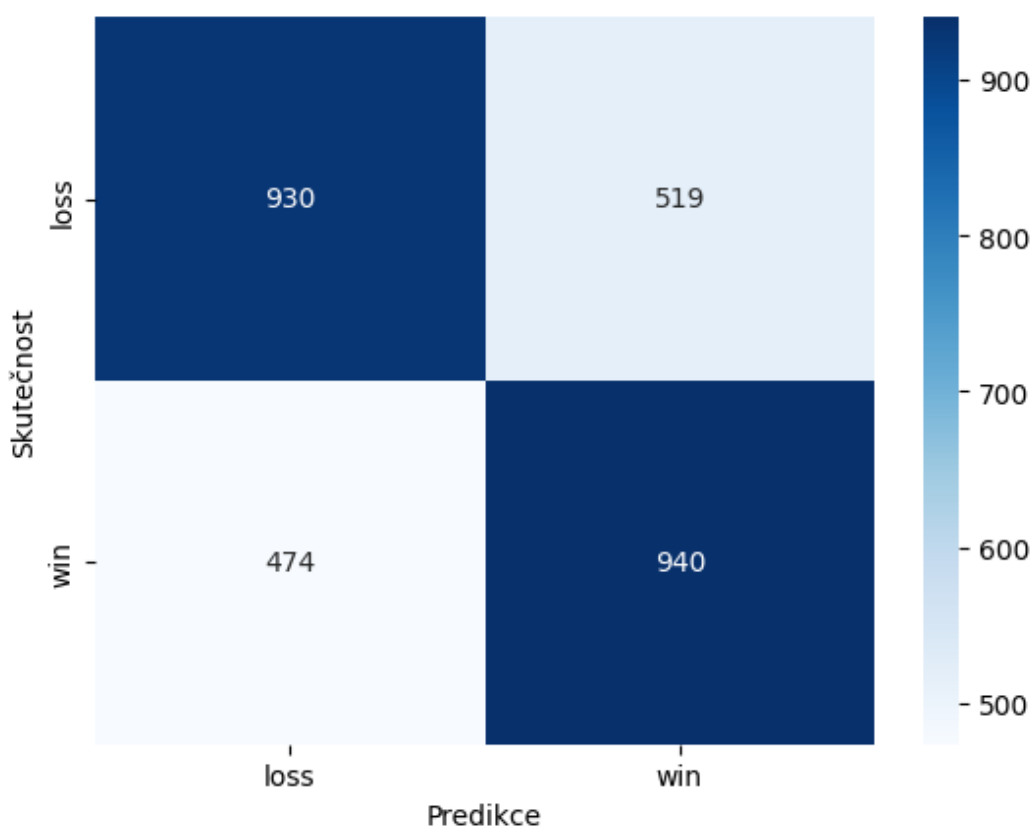
Pro rozhodovací strom byly jako nejlepší parametry zvoleny „class_weight“: „balanced“, „criterion“: „entropy“, „max_depth“: 24, „max_features“: „sqrt“, „min_samples_leaf“: 4, „min_samples_split“: 2 a „splitter“: „random“. I přes to, že je model optimalizován pro velké množství parametrů, má nižší celkovou správnost 64,4 %. Z 2863 predikovaných hodnot bylo 944 správně jako „loss“, 900 správně jako „win“, 514 špatně jako „loss“ a 505 špatně jako „win“ (Obr. 4.3). Přesnost pro prohru je 64,7 % a pro výhru 64,1 %. Úplnost pro prohru je 65,1 % a pro výhru 63,6 %.



Obr. 4.3 Matice záměn výsledku zápasu pro rozhodovací strom (vlastní zpracování)

Model k-NN

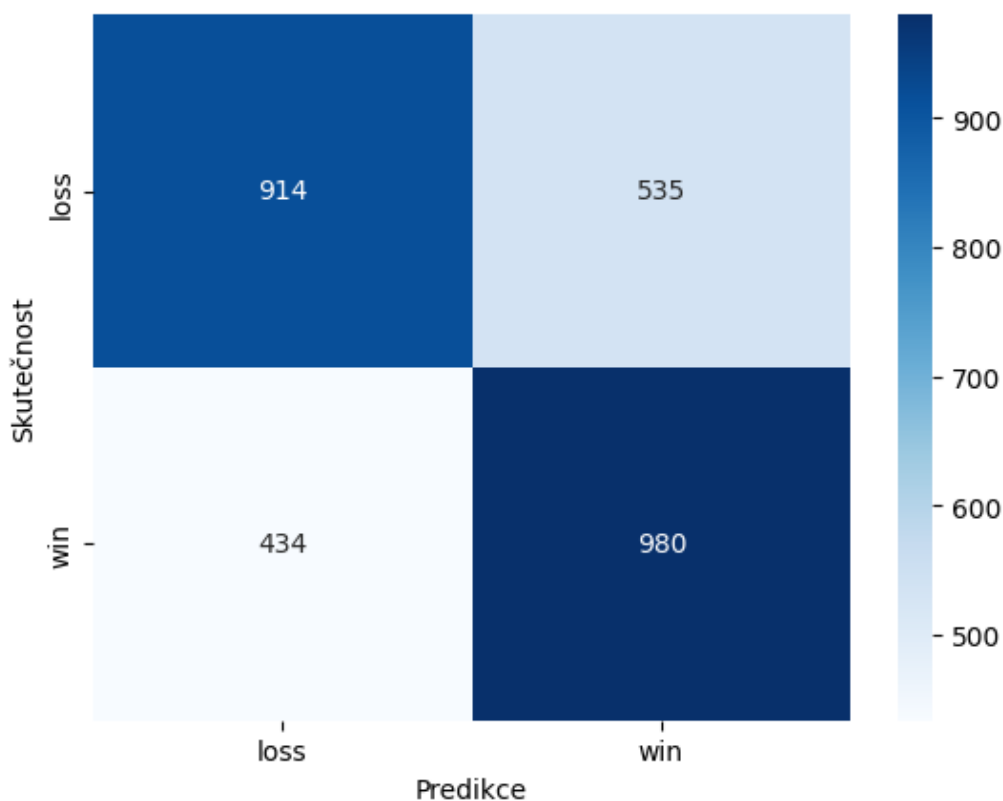
Pro model k-nejbližších sousedů byly vybrány tyto parametry: „n_neighbors“: 300, „algorithm“: „auto“, „p“: 2 a „metric“: „mankowski“. Tento model predikoval z 2863 hodnot 930 správně jako „loss“, 940 správně jako „win“, 474 špatně jako „loss“ a 519 špatně jako „win“ (Obr. 4.4). Celková správnost byla 65,3 %. Model měl o něco lepší přesnost pro predikci prohry (66,2 %) než výhry (64,4 %) a lepší úplnost pro výhru (66,5 %) než pro prohru (64,2 %).



Obr. 4.4 Matice záměn výsledku zápasu pro model k-NN (vlastní zpracování)

Náhodný les

Parametry vybrané pro tento model byly: „criterion“: „gini“, „max_depth“: 10, „min_samples_leaf“: 4, „min_samples_split“: 10 a „n_estimators“: 50. Náhodný les predikoval 914 hodnot správně jako „loss“, 980 správně jako „win“, 434 špatně jako „loss“ a 535 špatně jako „win“ (Obr. 4.5). Model dosáhl správnosti 66,2 %, přesnosti pro prohru 67,8 % a přesnosti pro výhru 64,7 %. Úplnost pro prohru byla 63,1 %, pro výhru pak 69,3 %.



Obr. 4.5 Matice záměn výsledku zápasu pro model náhodný les (vlastní zpracování)

4.1.3 Elo model

Elo model, dříve popsán v kapitole 1.4.2 je model, který na základě předchozích výsledků počítá skóre, v našem případě zápasníka MMA, a podle toho potom predikuje výsledek zápasu.

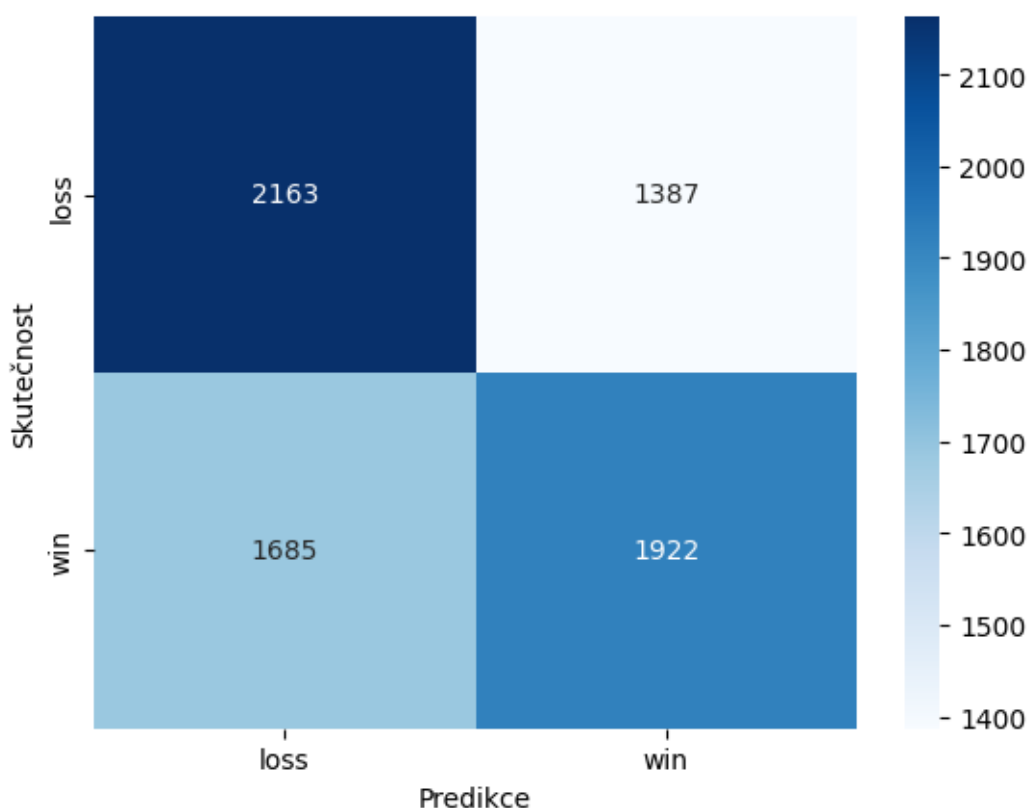
Tvorba modelu

Pro implementaci modelu byl využit balíček „skelo“. K vytvoření tohoto modelu potřebujeme pouze sloupce „Fighter A“, „Fighter B“ a „Date“. Pro měření linearity kalibrace klasifikátoru je lepší mít k dispozici údaje o výhře i prohře, proto změnilme pořadí u vybraných zápasníků a uložíme novou hodnotu výhry/prohry do proměnné „label“. Model umožňuje nastavení dvou hlavních parametrů, a to koeficientu K a počáteční hodnoty. Další možností, jak model upravit je například nastavení minimálního počtu zápasů jednotlivých zápasníků. Testováním bylo nalezeno nejlepší nastavení těchto parametrů. Počáteční

hodnota neměla vliv na výkon modelu a byla nastavena na 1500. Pomocí funkce „GridSearchCV“ byl nalezen nejlepší koeficient rozvoje $K = 150$. Hodnota koeficientu je poměrně vysoká, to je způsobeno poměrně nízkým počtem zápasů jednotlivých zápasníků MMA oproti jiným sportům či hrám. Ze stejného důvodu byl minimální počet zápasů nastaven na 0.

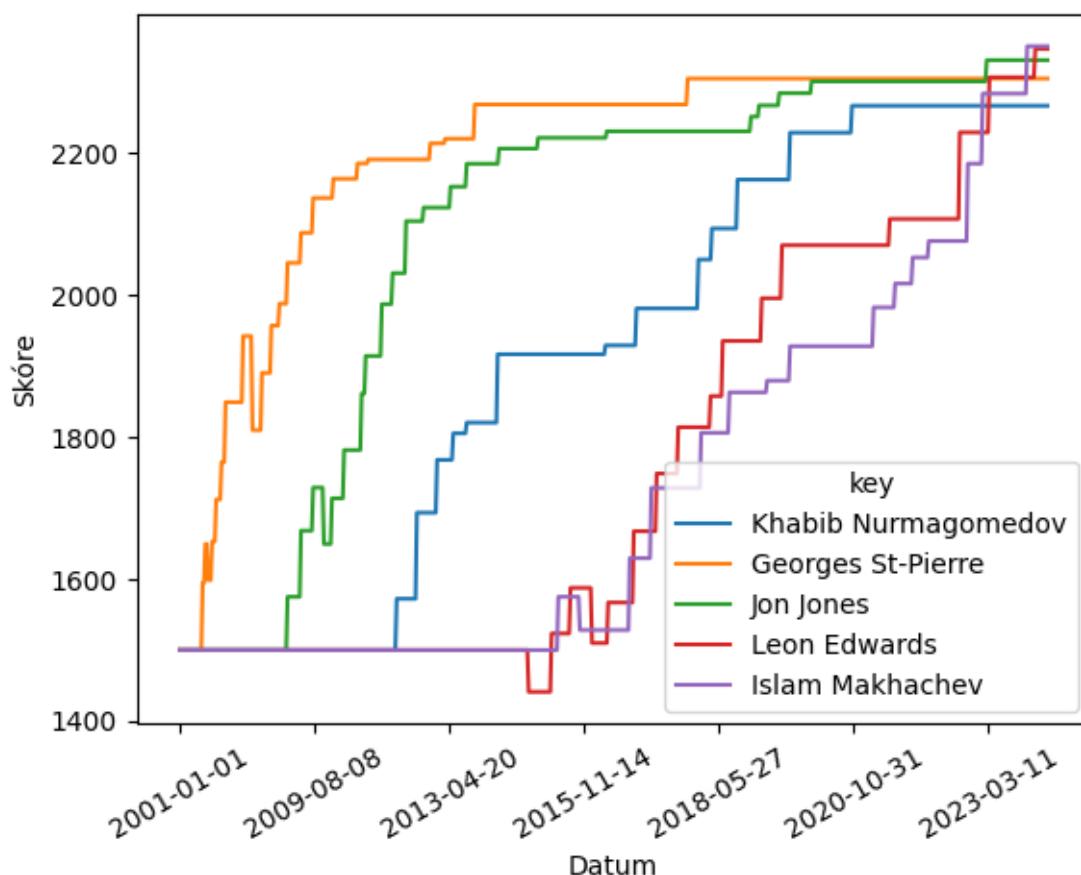
Predikce

Model s tímto nastavením predikuje 2163 hodnot správně jako „loss“, 1922 hodnot správně jako „win“, 1685 hodnot špatně jako „loss“ a 1387 hodnot špatně jako „win“ (Obr. 4.6). Celková úspěšnost modelu je 57,1 %, přesnost pro hodnotu „loss“ 56,2 % a pro hodnotu „win“ 58,1 %. Úplnost je pro prohru 60,9 % a pro výhru 53,3 %.



Obr. 4.6 Matice záměn výsledku zápasu pro Elo model (vlastní zpracování)

Na obrázku „Obr. 4.7 Vývoj skóre nejlepších pěti zápasníků (vlastní zpracování)“ můžeme vidět vývoj skóre zápasníků s nejvyšším dosaženým skóre. Tři ze zápasníků (Islam Makhachev, Leon Edwards a Jon Jones) jsou aktuálně podle žebříčku „fightmatrix.com“ v top 5 nejlepších na světě (fightmatrix.com, 2024). Zbývající dva (Georges St-Pierre a Khabib Nurmagomedov) již ukončili aktivní kariéru, ale podle stejného rankingu se řadí mezi nejlepší zápasníky všech dob. Tyto skutečnosti napovídají tomu, že byl Elo model sestaven správně.



Obr. 4.7 Vývoj skóre nejlepších pěti zápasníků (vlastní zpracování)

4.2 Predikce způsobu ukončení zápasu

V druhé části se budeme zabývat predikcí způsobu ukončení zápasu. Pro tuto analýzu byly z datasetu odstraněny zápasy bez výsledku. Remízy v tomto případě můžeme použít, protože spadají do běžného ukončení na body. Způsob ukončení budeme predikovat pouze pomocí modelů strojového učení, Elo model v tomto případě není vhodný.

Úprava dat

Data byla upravena podobně jako u predikce výsledků. Nejdříve se vytvořila obrácená kopie, která se připojila k existujícím datům, poté byly vytvořeny sloupce pro nesplnění váhového limitu a rozdíly numerických proměnných a doplněny chybějící hodnoty pro proměnné „Title A“ a „Title B“. Úprava dat se pro predikci způsobu ukončení liší v těchto detailech:

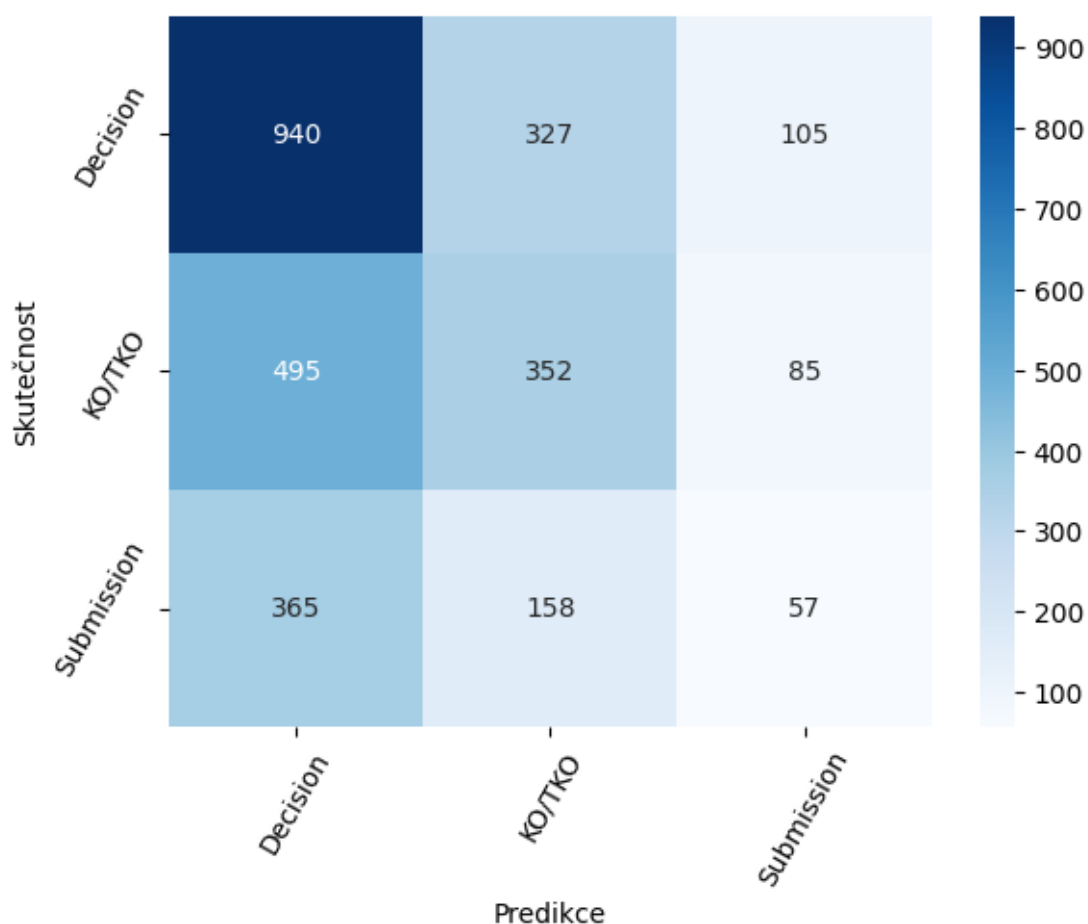
1. Hodnoty proměnné „Venue“ se pro arénu „UFC Apex“ změnili na „1“ a na „0“ pro všechny ostatní.
2. Z data byl extrahován rok.
3. Proměnná „Method“ byla transformována na numerické hodnoty.
4. Byly změněny datové typy tak, aby odpovídali novým hodnotám.

Tvorba modelu

Pro tvorbu modelu byly vybrány všechny nezávislé proměnné z modelů predikujících výsledky, tedy: „Title A“, „Title B“, „Odds A“, „Odds B“, „Odds cat A“, „Odds cat B“, „Nationality A“, „Nationality B“, „Weight Miss A“, „Weight Miss B“, „Age Diff“, „Height Diff“, „Reach Diff“, „KO win Diff“, „KO loss Diff“, „Sub win Diff“, „Sub loss Diff“, „Dec win Diff“, „Dec loss Diff“, „Draws Diff“, „UFC win Diff“, „UFC loss Diff“, „UFC draws Diff“. K nim však byly přidány ještě proměnné „Weightclass“, „Rounds“, „Venue“ a „Year“. Transformovaný sloupec „Method“ byl uložen s hodnotou „0“ pro „Decision“, „1“ pro „KO/TKO“ a „2“ pro „Submission“ do závislé proměnné. Z kategoriálních proměnných byly, stejně jako v předešlé kapitole, vytvořeny „umělé proměnné“ obsahující pouze hodnoty „True“ a „False“. Nezávislé proměnné byly standardizovány a data rozdělena na trénovací (80 %) a testovací (20 %) část. Modely byly vytvořeny pomocí funkce „GridSearchCV“ a všechny, vyjma k-NN, používají parametr „class_weight“, který byl nastaven tak, aby reflektoval četnost způsobů ukončení v datech viz. Tab. 3.4 Způsoby ukončení zápasu (vlastní zpracování).

4.2.1 Logistická regrese

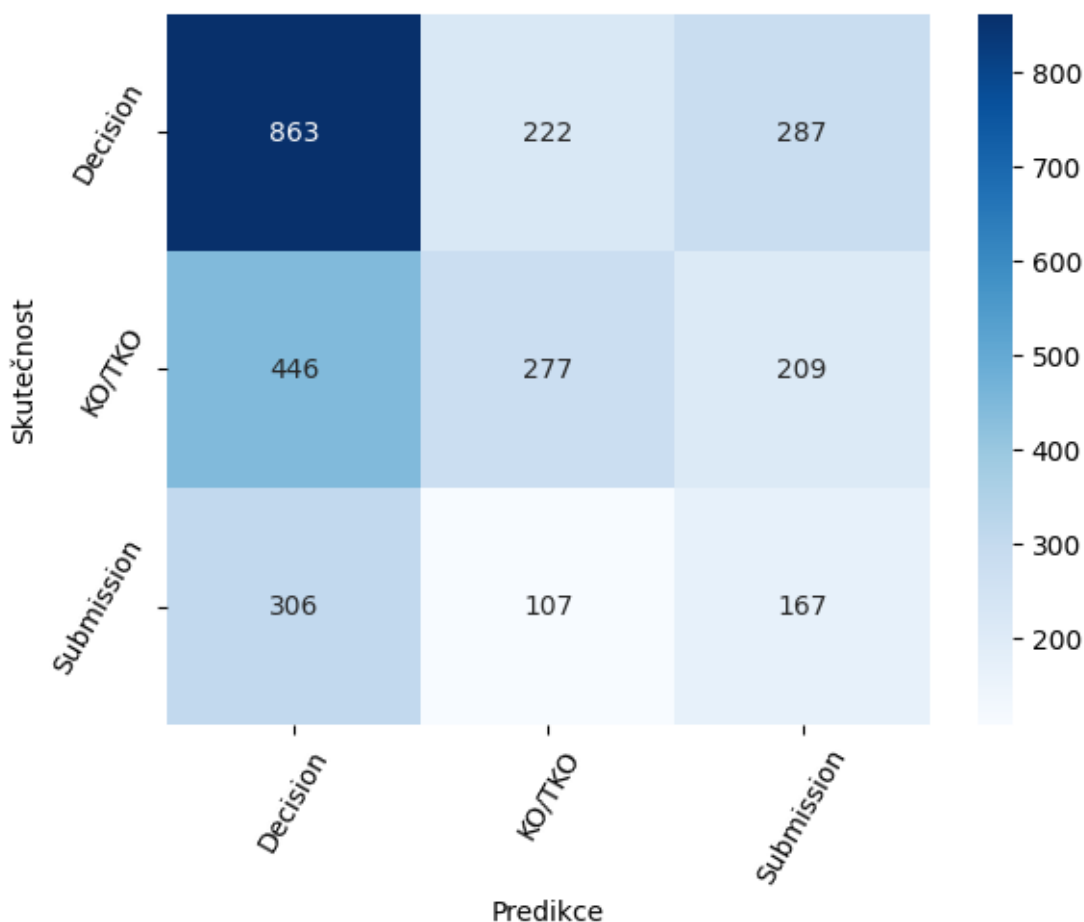
Pro model logistické regrese byly pomocí funkce „GridSearchCV“ vybrány tyto parametry: „C“: 0,00001, „penalty“: „l2“ a „solver“: „liblinear“. Model predikoval z 2884 hodnot 940 správně jako „Decision“, 352 správně jako „KO/TKO“, 57 správně jako „Submission“ 860 špatně jako „Decision“, 485 špatně jako „KO/TKO“ a 162 špatně jako „Submission“ (Obr. 4.8 Matice záměn způsobu ukončení zápasu pro model logistické regrese (vlastní zpracování)Obr. 4.8). Celková správnost modelu byla 46,8 %. Přesnost pro „Decision“ 52,2 %, „KO/TKO“ 42,1 % a „Submission“ 23,1 %. Úplnost byla pro bodové rozhodnutí 68,5 %, knockout 37,8 % a submisi 9,8 %.



Obr. 4.8 Matice záměn způsobu ukončení zápasu pro model logistické regrese (vlastní zpracování)

4.2.2 SVM

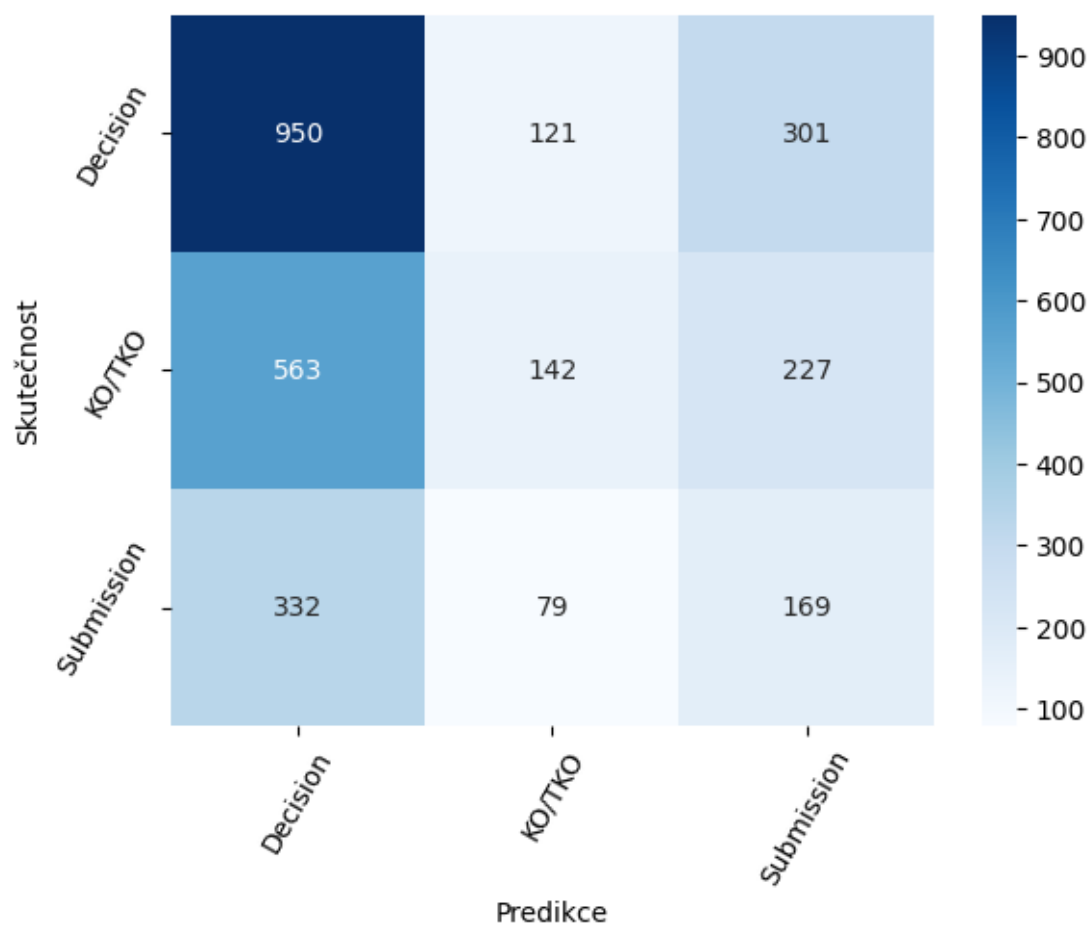
Pro SVM model byly zvoleny stejné parametry jako v případě predikce výsledku zápasu „C“: 0,001, a „kernel“: „linear“. Model bylo v domácích podmínkách, z důvodu vysoké výpočetní náročnosti, opět složité testovat pro větší množství parametrů. Model odhadl správně 863 hodnot jako „Decision“, 277 hodnot jako „KO/TKO“ a 167 hodnot jako „Submission“. Špatně predikoval 752 hodnot jako „Decision“, 329 hodnot jako „KO/TKO“ a 496 hodnot jako „Submission“ (Obr. 4.9). Model dosáhl správnosti 45,3 %. Přesnost byla pro „Decision“ 53,4 %, pro „KO/TKO“ 45,7 % a pro „Submission“ 25,2 %. Úplnost byla 62,3 % pro bodové rozhodnutí, 29,7 % pro knockout a 28,8 % pro submisi.



Obr. 4.9 Matice záměn způsobu ukončení zápasu pro model podpůrných vektorů (vlastní zpracování)

4.2.3 Rozhodovací strom

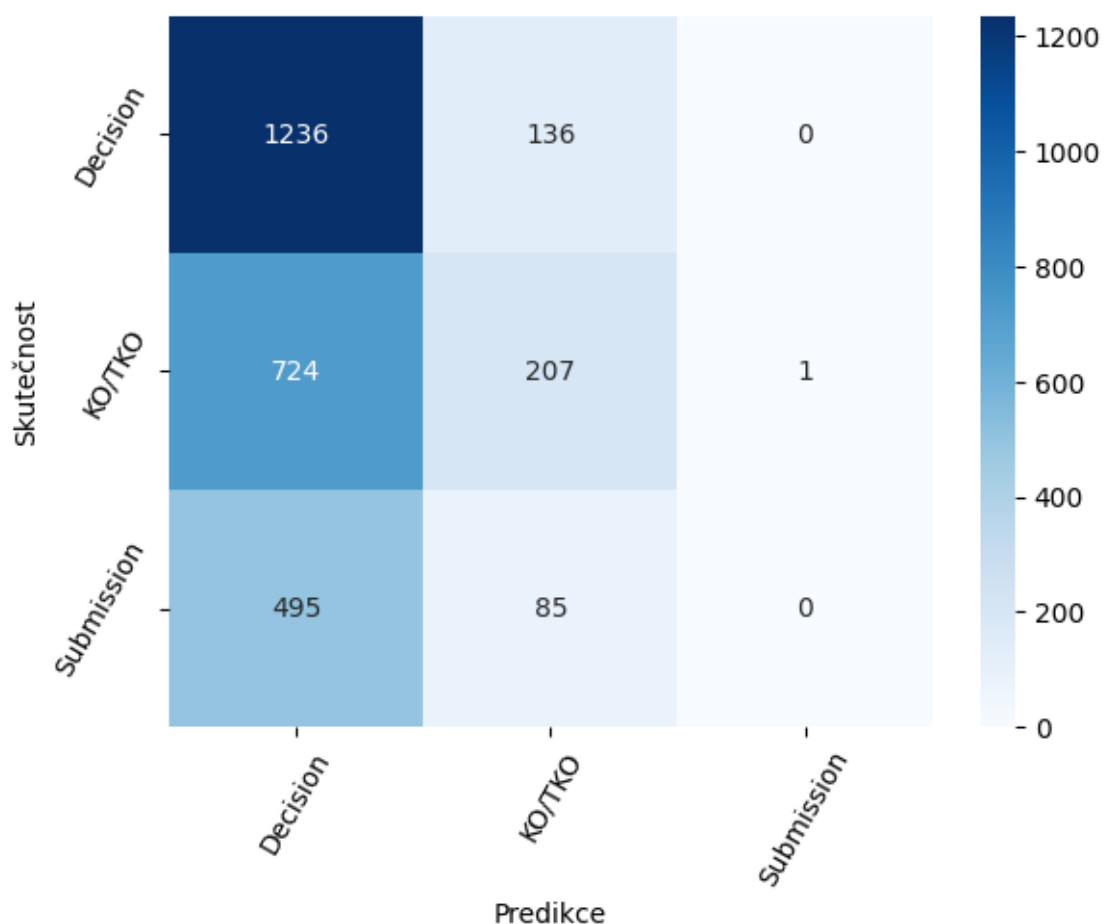
Parametry vybrané pro tento model, kromě „class_weight“, byly: „criterion“: „entropy“, „max_depth“: 12, „max_features“: „sqrt“, „min_samples_leaf“: 8, „min_samples_split“: 2 a „splitter“: „random“. Rozhodovací strom správně predikoval 950krát „Decision“, 142krát „KO/TKO“, 169krát „Submission“ a špatně 895krát „Decision“, 200krát „KO/TKO“ a 528krát „Submission“ (Obr. 4.10). Správnost modelu vyšla na 43,7 %, přesnost pro „Decision“ 41,6 %, pro „KO/TKO“ 15,2 % a pro „Submission“ 24,2 %. Úplnost rozhodovacího stromu pak byla 69,2 % pro „Decision“, 15,2 % pro „KO/TKO“ a 29,1 % pro „Submission“.



Obr. 4.10 Matice záměn způsobu ukončení zápasu pro rozhodovací strom (vlastní zpracování)

4.2.4 k-NN

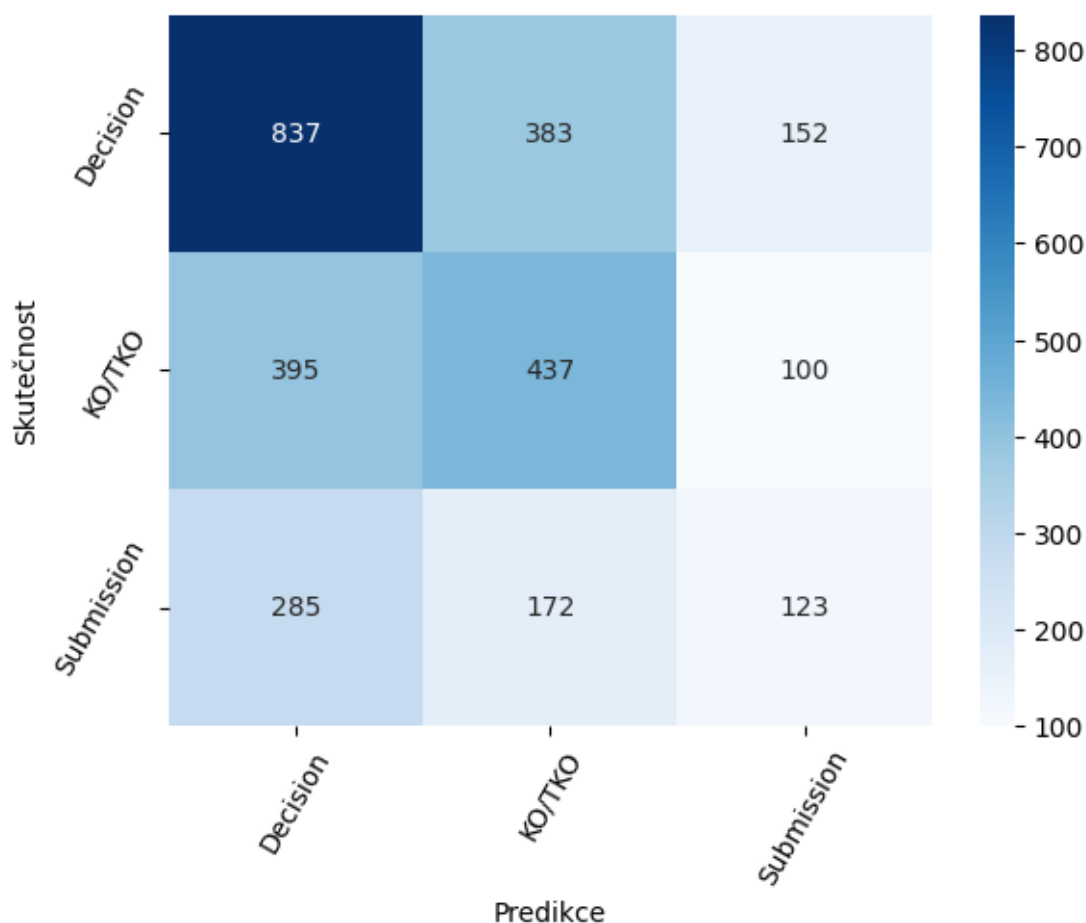
Pro model k-NN nelze použít parametr „class_weight“ jako v ostatních případech. Byly proto vybrány tyto parametry: „weights“: „distance“, „n_neighbors“: 100, „algorithm“: „auto“, „p“: 1 a „metric“: „mankowski“. Model predikoval 1236 hodnot správně pro „Decision“, 207 pro „KO/TKO“, žádnou pro „Submission“, 1219 špatně pro „Decision“, 221 pro „KO/TKO“ a jednu pro „Submission“ (Obr. 4.11). Celková správnost n-nejbližších sousedů byla přesně 50 %. Přesnost pro „Decision“ vyšla 50,3 % a 48,4 % pro „KO/TKO“. Úplnost byla pro „Decision“ 90 % a 22,2 % pro „KO/TKO“. Tento model predikoval pouze jednu submisi, a proto je přesnost i úplnost 0 %.



Obr. 4.11 Matice záměn způsobu ukončení zápasu pro k-NN model (vlastní zpracování)

4.2.5 Náhodný les

Jako nejlepší parametry pro model náhodného lesa byly zvoleny parametry „criterion“: „entropy“, „max_depth“: 15, „min_samples_leaf“: 1, „min_samples_split“: 2 a „n_estimators“: 100. Model predikoval z 2884 hodnot správně 837 jako „Decision“, 437 jako „KO/TKO“, 123 jako „Submission“ a špatně 680 jako „Decision“, 555 jako „KO/TKO“ a 275 jako „Submission“ (Obr. 4.12). Přesnost pro bodové rozhodnutí byla 55,2 %, pro knockout 44,1 % a pro submisi 32,8 %. Úplnost vyšla 61 % pro „Decision“, 46,7 % pro „KO/TKO“ a 21,2 % pro „Submission“. Celková správnost modelu pak byla 48,4 %.



Obr. 4.12 Matice záměn způsobu ukončení zápasu pro náhodný les (vlastní zpracování)

5 Porovnání výsledků

Tato kapitola se věnuje vyhodnocení a porovnání výsledků prediktivních modelů vytvořených v předchozí části práce. Modely se budou hodnotit na základě správnosti, přesnosti a úplnosti a jejich možnostech využití v reálném světě.

5.1 Porovnání predikce výsledku zápasu

Pro predikci výsledku zápasu bylo vytvořeno pět modelů strojového učení a jeden Elo model. Výsledky modelů strojového učení se vzájemně příliš nelišily. Největší správnost měl náhodný les, který predikoval 65,8 % hodnot správně, nejnižší správnost z nich měl pak rozhodovací strom, který predikoval 64,4 % hodnot správně. Elo model za modely strojového učení zaostával s pouze 57,1 % správnými predikcemi (Tab. 5.1).

V závislosti na použití modelu jsou dalšími důležitými prvky přesnost a úplnost. Pokud by měl být model použit například pro využití existujících sázkových příležitostí je nejdůležitější hodnota přesnost. Pokud by měl být model použit pro vytváření sázkových kurzů, důležitá je jak přesnost, tak úplnost i správnost. Největší přesnosti dosáhl pro prohru model podpurných vektorů, logistická regrese pak pro výhru. Největší úplnost pro prohru potom měl rozhodovací strom a pro výhru SVM (Tab. 5.1).

Je složité říct, který z modelů strojového učení je obecně nejlepší, jelikož žádný z nich není nejlepší ve všech hodnotách. I přes to, že Elo model správně ohodnotil nejlepší zápasníky, měl oproti ostatním modelům nejhorší výsledky, což bylo nejspíš způsobeno poměrně nízkým počtem zápasů jednotlivých zápasníků.

Tab. 5.1 Srovnání modelů pro predikci výsledků (vlastní zpracování)

Model	Přesnost pro prohru	Přesnost pro výhru	Úplnost pro prohru	Úplnost pro výhru	Správnost
Log. regrese	0,663	0,648	0,649	0,663	0,656
SVM	0,676	0,637	0,611	0,700	0,655
Rozhodovací strom	0,647	0,641	0,651	0,636	0,644
k-NN	0,662	0,644	0,642	0,665	0,653
Náhodný les	0,672	0,645	0,634	0,683	0,658
Elo rating	0,562	0,581	0,609	0,533	0,571

5.2 Porovnání predikce způsobu ukončení zápasu

Způsob ukončení zápasu byl predikován pouze pomocí modelů strojového učení. Tyto modely zahrnují logistickou regresi, metodu podpůrných vektorů, rozhodovací strom, metodu n-nejbližších sousedů a náhodný les.

Model k-NN jako jediný nemá parametr „class_weight“, a proto výrazně více predikoval ukončení na body oproti jiným způsobům. Vzhledem k velkému počtu ukončení na body v datech měl však nejlepší správnost 50 % (Tab. 5.2). Model měl dále ze všech nejvyšší přesnost pro ukončení na „KO/TKO“, a to 48,4 %. Model predikoval pouze 1 submisi, proto je pro ni přesnost i úplnost modelu 0 %.

Z modelů, které používali parametr „class_weight“, má největší správnost (48,9 %) náhodný les, který má také největší přesnost pro body a submise (35,5 %) a úplnost pro „KO/TKO“. Nejnižší celkové správnosti (43,7 %) a přesnosti pro body a „KO/TKO“ dosáhl rozhodovací strom (Tab. 5.2).

Tab. 5.2 Srovnání modelů pro predikci způsobu ukončení (vlastní zpracování)

Model	Přesnost body	Přesnost KO/TKO	Přesnost submise	Úplnost body	Úplnost KO/TKO	Úplnost submise	Správnost
Log. regrese	0,522	0,421	0,231	0,685	0,378	0,098	0,468
SVM	0,534	0,457	0,252	0,629	0,297	0,288	0,453
Rozhod. strom	0,515	0,415	0,242	0,692	0,152	0,291	0,437
k-NN	0,503	0,484	0,000	0,901	0,222	0,000	0,500
Náhodný les	0,544	0,440	0,355	0,638	0,448	0,200	0,489

Závěr

Cílem práce bylo přispět k lepšímu porozumění faktorů ovlivňujících výsledky zápasů ve smíšených bojových uměních, implementovat prediktivní modely a poskytnout tak užitečné poznatky o této disciplíně. K dosažení tohoto cíle byla provedena důkladná analýza klíčových proměnných a vytvořeny modely logistické regrese, SVM, rozhodovacího stromu, k-NN, náhodného lesa a Elo model.

V analýze bylo zjištěno například, že z fyzických charakteristik má velký podíl na výsledku zápasu především věk zápasníků, největší úspěšnost mají v UFC zápasníci z Ruska, titulové zápasy vyhrávají v přibližně 69 % případů úřadující šampioni a nesplnění váhového limitu je spíše nevýhodou. V dostupných kurzech byla odhalena chyba, kdy „Heavy Favorite“ vyhrával méně procent zápasů proti „Heavy Underdog“ než proti „Moderate Underdog“. Dále bylo zjištěno, že zápasy v aréně „UFC Apex“ končí méně často na body než v ostatních arénách, a že váhová kategorie má velký vliv na způsob ukončení zápasu, který se také měnil poměrně výrazně historicky.

Hodnocení modelů bylo provedeno na základě jejich správnosti, přesnosti a úplnosti. Každá z metrik hraje jinou roli v závislosti na použití modelu a vzhledem k tomu, že žádný model nebyl nejlepší ve všech hodnotách, nebylo možné určit nejlepší z nich. Strojově učené modely predikovali správný výsledek v přibližně 65 % případů, což přibližně odpovídalo správnosti vypsaných kurzů (viz. Správnost vypsaných kurzů). Elo model předpovídal výsledky zápasů nejhůře s 57 % správností.

Správnost modelů pro predikci způsobu ukončení nebyla příliš vysoká s ohledem na početnost způsobu ukončení „Decision“. I kdyby model predikoval všechny hodnoty jako „Decision“, dosáhl by velmi podobné správnosti. Mezi limity práce patří také chybějící rozdělení na muže a ženy. Ze zkušeností autora končí ženské zápasy častěji na body, tohle rozdělení by tak mohlo pomoci úspěšnosti modelů právě v predikci způsobu ukončení zápasu. Ani jeden ze zdrojů dat však tuto informaci neuváděl.

Autor vidí velkou hodnotu primárně v přesnosti predikce náhodného lesa ukončení na „Submission“. Model měl pro tuto hodnotu přesnost 35,5 %, což je opravdu skvělý výsledek při uvážení, že zápasy končí na „Submission“ jen v asi 20 % případů a sázkové kanceláře na tuto příležitost, ze zkušeností autora, vypisují nejvyšší kurzy.

Výsledky práce již v současnosti slouží autorovi v analýze a predikci zápasů a hodnocení výhodnosti sázkových příležitostí. Další výzkum by se mohl zaměřit na konkrétní porovnání pravděpodobnosti výsledku či způsobu ukončení s reálným kurzem sázkových kanceláří a identifikovat tak výhodné sázkové příležitosti ještě lépe, popřípadě samotné kurzy přímo vytvářet.

Použitá literatura

ASSOCIATION OF BOXING COMMISSIONS AND COMBATIVE SPORTS. (2022). *Official Unified rules of MMA* [Oficiální unifikované pravidla MMA]. ASSOCIATION OF BOXING COMMISSIONS AND COMBATIVE SPORTS. <https://www.abcboxing.com/wp-content/uploads/2022/08/unified-rules-mma-july-2022.pdf>

Britannica, T. Editors of Encyclopaedia (2024). *mixed martial arts* [smíšená bojová umění]. In Encyclopedia Britannica. <https://www.britannica.com/sports/mixed-martial-arts>

Příspěvatelé projektu Wikimedia. (2023). Pankrátion. *Wikipedie: Otevřená encyklopedie*. <https://cs.wikipedia.org/w/index.php?title=Pankr%C3%A1tion&oldid=22966709>

Karpman, S., Reid, P. A., Phillips, L., Qin, Z., & Gross, D. P. (2016). Combative Sports Injuries. *Clinical Journal of Sport Medicine*, 26(4), 332–334. <https://doi.org/10.1097/jsm.0000000000000235>

Lawton, G., Carew, J. M., Burns, E. (2022). *What is Predictive Modeling?* [Co je prediktivní modelování?]. EnterpriseAI. <https://www.techtarget.com/searchenterpriseai/definition/predictive-modeling>

Delua, J., (2021). *Supervised vs. Unsupervised Learning: What's the Difference?* [Učení s učitelem vs bez učitele: Jaký je rozdíl?]. IBM blog. <https://www.ibm.com/blog/supervised-vs-unsupervised-learning/>

Kaushik, S. (2024). *Clustering | Different methods, and applications* [Shlukování | Různé metody a aplikace]. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/>

Mittal, R. (2022). *What is an ELO Rating?* [Co je ELO rating?]. Medium. <https://medium.com/purple-theory/what-is-elo-rating-c4eb7a9061e0>

Luna, J. C. (2022). *Python vs R for Data Science: Which Should You Learn?* [Python vs R pro Data Science: Který by ses měl naučit?]. Datacamp. <https://www.datacamp.com/blog/python-vs-r-for-data-science-whats-the-difference>

Fight Matrix. (2024). *All-Time Absolute* [Absolutní ranking všech dob]. Fight Matrix. <https://www.fightmatrix.com/all-time-mma-rankings/all-time-absolute/>

Wilkens, S. (2021). *Sports prediction and betting models in the machine learning age: The case of tennis*. *Journal of Sports Analytics*, 7(2), 99–117. <https://doi.org/10.3233/jsa-200463>

Tapology. (2024). *Tapology*. <https://www.tapology.com/>

Sherdog. (2024). *Sherdog*. <https://www.sherdog.com/>

Přílohy

Příloha A: tapology_events.ipynb

Příloha obsahuje proces získání dat z tapology.com.

Příloha B: sherdog_bouts.ipynb

Příloha obsahuje proces získání dat z sherdog.com

Příloha C: creating_dataset.ipynb

Příloha obsahuje proces spojení a čištění dat.

Příloha D: ufc_analysis.ipynb

Příloha obsahuje analýzu získaných dat.

Příloha E: ufc_prediction.ipynb

Příloha obsahuje proces tvorby prediktivních modelů.