Obraz zawierający logo

Opis wygenerowany automatycznie

Studium Magisterskie

Łukasz Makaruk (115902)

**Wykorzystanie narzędzi uczenia maszynowego oraz optycznego rozpoznawania znaków do predykcji meczy e-sportowych**

Praca magisterska

Na kierunku: Big Data

Praca wykonana pod kierunkiem

Dr. Sebastiana Zająca

Warszawa 2023

# Streszczenie

Praca podejmuje problematykę sportów elektronicznych i specyfikacji rynku e-sportowego. Pokazuje wykorzystanie narzędzi uczenia maszynowego oraz optycznego rozpoznawania znaków do predykcji meczy e-sportowych na podstawie turnieju z gry League of Legends. Omówione zostają także tezy prezentowane przez literaturę naukową oraz profesjonalnych zawodników tego rodzaju sportów.

# Słowa kluczowe

e-sport, sporty elektroniczne, uczenie maszynowe, modele predykcyjne, optyczne rozpoznawanie znaków

**Dziedzina pracy (kody wg programu Erasmus)**

# Klasyfikacja tematyczna

# Tytuł pracy w języku angielskim

Use of machine learning and optical character recognition tools to predict e-sports matches

**SPIS TREŚCI**

[Streszczenie 2](#_Toc139835387)

[Słowa kluczowe 2](#_Toc139835388)

[Klasyfikacja tematyczna 2](#_Toc139835389)

[Tytuł pracy w języku angielskim 2](#_Toc139835390)

[WSTĘP 3](#_Toc139835391)

[II. Teoretyczna analiza omawianego zagadnienia 4](#_Toc139835392)

[2.1 Historyczny zarys e-sportu 4](#_Toc139835393)

[2.2 Czy e-sport może być rozpatrywany w kategorii sportu? 7](#_Toc139835394)

[2.3 Dlaczego warto zajmować się analizą branży e-sportowej? 10](#_Toc139835395)

[2.4 Przegląd literatury naukowej w temacie predykcji meczy e-sportowych 13](#_Toc139835396)

[2.5 Jakie jest zdanie profesjonalnych zawodników na temat tej analizy 17](#_Toc139835397)

[III. Omówienie danych wykorzystanych w analizie 24](#_Toc139835398)

[3.1 Sposób pozyskania bazy danych 24](#_Toc139835399)

[3.2 Wykorzystane zmienne 25](#_Toc139835400)

[3.3 Problemy ze zmiennymi 31](#_Toc139835401)

[IV. Analiza danych i modelowanie 40](#_Toc139835402)

[4.1 Przygotowanie końcowej bazy danych 40](#_Toc139835403)

[4.2. Analiza danych 44](#_Toc139835404)

[4.3. Modelowanie 57](#_Toc139835405)

[4.3.1 Modele zawierające zmienne dotyczących konkretnych postaci 58](#_Toc139835406)

[4.3.2 Modele pozbawione zmiennych dotyczących konkretnych postaci 69](#_Toc139835407)

[4.3.3 Modele w podziale na czas meczu 78](#_Toc139835408)

[V. Omówienie wyników najlepszych modeli 81](#_Toc139835409)

[5.1 Modele z podziałem na etap spotkania 81](#_Toc139835410)

[5.2 Model ogólny dla całego spotkania 84](#_Toc139835411)

[VI. Podsumowanie 87](#_Toc139835412)

# WSTĘP

Pojęcie sportów elektronicznych można z pewnością uznać za pojęcie niebędące szeroko znanym, jak również budzącym duże kontrowersje, nawet w XXI wieku. Pomimo postępującej globalizacji oraz cyfryzacji społeczeństwa, wiele osób ma problem z poważnym traktowaniem e-sportu. Zjawisko to niesie ze sobą zarówno bardzo dużo korzyści jak i zagrożeń, dlatego tak ważna jest rzetelna analiza tego tematu. Celem tej pracy jest zbadanie możliwości predykcji meczy e-sportowych z wykorzystaniem narzędzi uczenia maszynowego oraz optycznego rozpoznawania znaków. Podjęta analiza koncentruje się w obrębie jednej gry- League of Legends, wydanej przez firmę Riot Games. Dane meczowe dotyczą spotkań rozgrywanych w ramach jednego turnieju- League of Legends European Championship 2022 lato. Początkowo planowano wykorzystać większą ilość turniejów z różnych regionów, jednak ograniczenie w postaci nieujednoliconego interfejsu rozgrywek oraz wymaganej mocy obliczeniowej zadecydowało o porzuceniu tego pomysłu. W prezentowanym badaniu postarano się sprawdzić, czy początkujący badacz jest w stanie dokonać zadowalających predykcji zwycięzcy meczu oraz zweryfikować rozmaite tezy prezentowane w rozmowie z profesjonalnym zawodnikiem oraz w literaturze naukowej. Cały przebieg analizy- od momentu pozyskania bazy danych do momentu dokonania predykcji został dokładnie opisany, a każda zmiana planowanych kroków została dogłębnie wyjaśniona i uargumentowana. W podsumowaniu zebrano wnioski płynące z tego badania oraz przedstawiono możliwe jego ulepszenia oraz potencjalne przyszłościowe wykorzystania.

# II. Teoretyczna analiza omawianego zagadnienia

## **2.1 Historyczny zarys e-sportu**

Termin „sport elektroniczny” lub „e-sport” jest terminem stosunkowo nowym. Autor artykułu „On the Scientific Relevance of eSports” Michael G. Wagner[[1]](#footnote-1) jako czas pierwszego użycia tego terminu w wiarygodnym źródle wskazuje końcówkę lat dziewięćdziesiątych XX wieku, a dokładniej rok 1999. Wtedy to została wydana notka prasowa dotycząca powstania Online Gamers Association (OGA). Aby przybliżyć czytelnikom temat sportów elektronicznych, autor notki Matt Bettington podjął się próby porównania e-sportu do sportów tradycyjnych. Zbiegło się to również w czasie z nieudaną próbą organizacji pierwszego turnieju e-sportowego w Wielkiej Brytanii (UK Professional Computer Gaming Championship). Sprawiło to, że temat e-sportu na stałe zagościł w dyskusjach opinii publicznej. Na zachodzie (a dokładniej w Europie oraz Stanach Zjednoczonych) początek kompetetywnych rozgrywek sportów elektronicznych wiąże się z powstaniem gier z gatunku FPS (ang. *first person shooter*, strzelanek pierwszoosobowych). Głównymi reprezentantami tego gatunku w tamtych czasach był wydany w 1993 roku Doom oraz jego kontynuacja wydana trzy lata później- Quake. Rok po powstaniu Quake’a założona została liga CPL (*ang. Cyberathlete Professional League*) - jedna z najstarszych i zarazem najbardziej znaczących dla historii e-sportu. Kolejnym kamieniem milowym dla branży e-sportowej było wydanie na rynek kultowej już gry „Counter-Strike” (CS), które miało miejsce w 1999 roku. Popularny CS bardzo szybko wyprzedził w rozpoznawalności protoplastę gatunku i wraz ze swoją kontynuacją „Counter-Strike: Global Offensive”, która została wydana w roku 2012 pozostaje jedną z najpopularniejszych oraz najbardziej lubianych gier z gatunku strzelanek pierwszoosobowych. Według serwisu cybersport.pl[[2]](#footnote-2) turniej PGL Major Stockholm 2021 w szczytowym momencie śledziło ponad 2,7 miliona osób (2 748 850). Wartym zaznaczenia jest fakt, że do danych tych nie wliczamy widowni pochodzącej z chińskich platform streamingowych. Sporty elektroniczne rozwijały się nie tylko na zachodzie. Jako protoplastę e-sportu w rejonach wschodnich uznaje się Koreę, a stało się tak dzięki rozwojowi koreańskiej infrastruktury szerokopasmowej. Jednak gusta graczy w tamtych regionach świata różniły się znacząco od tych na zachodzie. We wschodniej kulturze największym zainteresowaniem cieszyły się gry z gatunku RTS (*ang. Real Time Strategy*) oraz MMORPG (*ang.* Multiplayer *Online Role Playing Games*). Strategie czasu rzeczywistego z początku zdominowały koreański rynek e-sportowy do tego stopnia, że gry, takie jak wydany w 1994 roku przez firmę Blizzard „Warcraft: Orcs & Humans”, bądź wydany cztery lata później „StarCraft” stały się nieodłącznym elementem wschodnioazjatyckiej popkultury. Obecny rynek e-sportowy jest zdecydowanie bardziej złożony i rozbudowany niż miało to miejsce w początkowych fazach jego rozwoju. Każdego roku powstają coraz to nowe gry, jak również i ich gatunki. Obecnie nie ma możliwości wskazać gatunków zdecydowanie dominujących żaden z rynków. Każdy typ gier ma swoich fanów na całym świecie, co pozwala również nawiązywać międzykontynentalne znajomości w oparciu o te same pasje czy upodobania. Zgodnie z informacjami zawartymi na oficjalnej stronie Uniwersytetu Maryville[[3]](#footnote-3) głównymi gatunkami gier wchodzących w branżę e-sportową są:

* Gry walki, w których dwóch graczy rywalizuje ze sobą za pomocą kombinacji różnego rodzaju ciosów, kopnięć oraz ataków specjalnych. Gry te są bardzo podobne do wszelkiego rodzaju sportów walki (oczywiście z pominięciem elementów fikcyjnych takich jak ataki specjalne). Przykładami gier z tego gatunku są serie gier takie jak Mortal Kombat, Tekken, Street Fighter czy Super Smash Bros.
* Gry wyścigowe, które pozwalają graczom na wcielenie się w rolę kierowców różnego rodzaju pojazdów oraz rywalizację z jednym bądź wieloma przeciwnikami. Gry z tego gatunku mogą być zarówno realistycznymi symulacjami prawdziwych wyścigów (np. seria gier F1), jak również produkcjami wykorzystującymi elementy fikcyjne (np. Mario Kart).
* Gry sportowe, które oparte są na rzeczywistych sportach takich jak piłka nożna czy koszykówka, w których gracze mają możliwość rywalizacji w ukochanych dyscyplinach sportowych bez wychodzenia z domu. Przykładem tego rodzaju gier mogą być serie gier takie jak FIFA czy NBA.
* Cyfrowe gry karciane, które wykorzystując format turowy pozwalają graczom na toczenie zaciętych pojedynków z wykorzystaniem kart. Znanymi grami z tego gatunku są produkcje takie jak Heartstone, czy stworzona przez polskie studio CD Projekt Red gra Gwint: Wiedźmińska gra karciana.
* Strategie czasu rzeczywistego (RTS), w których gracze rywalizują ze sobą jednocześnie z wykorzystaniem różnego rodzaju budynków, zasobów czy jednostek. Popularnymi grami z tego gatunku są: StarCraftII, Total War i Age of Empires.
* Strzelanki pierwszoosobowe (FPS), w których gracze posiadają pierwszoosobowy widok postaci dysponującej różnego rodzaju bronią (jak noże czy pistolety). Celem w tych grach jest najczęściej wyeliminowanie przeciwnej drużyny bądź przejęcie określonych celów na mapie. Przykładami gier z tego gatunku jest wspomniany wyżej Counter-Strike: Global Offensive, jak również seria gier Call of Duty, Halo czy Battlefield.
* Strzelanki trzecioosobowe (*ang. Third-Person Shooter,*TPS), które od FPS-ów różnią się tym, że kąt widzenia kamery pozwala graczowi dostrzec całą postać, a nie tylko jej broń. Grą, która dominuje ten gatunek jest Fortnite Battle Royale.
* Gry MOBA (*ang. Multiplayer Online Battle Arena*), które zestawiają przeciwko sobie dwie drużyny (najczęściej złożone z 3-5 graczy) na określonym obszarze. Każda z postaci wykorzystywanych w rozgrywce ma unikalny zestaw umiejętności oraz rolę. Popularnymi grami z tego gatunku są: League of Legends, Dota 2 czy SMITE.

W każdej z tych gier sposób rozgrywki oraz wymagane od gracza umiejętności znacząco się różnią. Grafika przedstawiona na rysunku 1 pokazuje wizualne różnice między tymi gatunkami.

Rys. 1. Prezentacja gatunków gier e-sportowych.

Obraz zawierający Gra komputerowa, zrzut ekranu, Strategiczna gra wideo, Oprogramowanie gier wideo

Opis wygenerowany automatycznie

*Źródło:* Opracowanie własne. Legenda: 1- Mortal Kombat X (gra walki), 2- Mario Kart (gra wyścigowa), 3- FIFA 21 (gra sportowa), 4- Heartstone (gra karciana), 5- StarCraft II (RTS), 6- Counter-Strike: Global Offensive (FPS), 7- Fortnite Battle Royale (TPS), 8- League of Legends (MOBA).

Jak można zauważyć na rysunku 1 każdy rodzaj gier oferuje graczowi zupełnie inne doznania. Warto pamiętać, że gatunki gier mogą się również przeplatać, a jedynym co ogranicza rozwój nowych formatów rozgrywki jest kreatywność developerów. Mnogość gatunków gier oraz stylów rozgrywki pozwala zakładać, że e-sport jest w stanie zaspokoić zamiłowanie do różnorodności statystycznego konsumenta. Obecnie organizowanych jest wiele turniejów e-sportowych, które oprócz wspólnej rywalizacji pozwalają również na przenikanie się wzajemnie kultury zachodniej oraz wschodniej, co może prowadzić do szeroko pojętej globalizacji. Warto jednak odnotować, że pomimo wspólnej integracji na różnego rodzaju wydarzeniach różnice rynku wschodniego i zachodniego są dalej widoczne. Oba te rynki wydają się być od siebie niezależne, co niekiedy zmusza wydawców do wprowadzania różnic w grach, w zależności od rynku docelowego.

## **2.2 Czy e-sport może być rozpatrywany w kategorii sportu?**

E-sport z uwagi na bycie zjawiskiem stosunkowo nowym, jest również tematem budzącym liczne kontrowersje. Wielu ludzi oburza się słysząc o sportach elektronicznych i wzbrania się przed jakimkolwiek zaliczaniem ich do kategorii sportu. Warto więc zastanowić się, czy e-sport może być rozpatrywany tak, jak sporty tradycyjne. Pierwszym aspektem, który należy poruszyć w celu odpowiedzenia na to pytanie jest sposób wpisywania się cybersportów w definicje sportu. Według „Małej encyklopedii sportu”[[4]](#footnote-4) sport określany jest jako „świadoma, dobrowolna działalność człowieka, podejmowana głównie dla zaspokojenia potrzeby zabawy, popisu, walki, a także wewnętrznego doskonalenia się w drodze systematycznego rozwoju cech fizycznych, umysłowych i wolicjonalnych”. Zgodnie z tą definicją sporty elektroniczne mogą być uznawane za sport, ponieważ dostarczają rozrywki i rywalizacji, a także pozwalają doskonalić się zarówno pod kątem umysłowym jak i fizycznym (pamięć mięśniowa, spostrzegawczość). Bardziej rozbudowaną definicję sportu przedstawiają M. Demel oraz A. Skład[[5]](#footnote-5). Według ich definicji sportem możemy nazwać wszystko, co spełnia następujące punkty:

* niezależność bezpośrednich motywów od podstawowych potrzeb życiowych
* dodatnie emocje towarzyszące postanowieniom i działaniom
* brak materialnego efektu, który zawsze jest wynikiem pracy produkcyjnej
* ruchowy charakter działania z akcentem specjalistycznym
* systematyczność w dążeniu do osiągnięcia maksymalnych wyników
* jaskrawy moment współzawodnictwa jako jeden z zasadniczych bodźców doskonalenia
* ścisłe stosowanie się do przepisów określających normy sprzętu, warunki i formy ruchowe, co umożliwia wymierność i porównywalność wyników.

W przypadku tej definicji e-sport również wydaje się spełniać normy sportu, jednak warto zaznaczyć, że dużo zależy od interpretacji powyższych punktów. Zdanie autorów popiera również Andrzej Stępnik, autor publikacji „E-sport z perspektywy teorii sportu”[[6]](#footnote-6). Stwierdza on, że e-sport spełnia wszystkie warunki przewidziane dla sportu, dlatego można rozpatrywać go w tej kategorii. Następnym ważnym aspektem, który może pomóc w odpowiedzi na pytanie postawione w tytule tego podrozdziału są możliwości oferowane przez e-sport dla branży sportowej. Autorzy publikacji „eSports venues: A new sport business opportunity”[[7]](#footnote-7) podkreślają duży potencjał tej nowej gałęzi rynku dla biznesu sportowego. Jest to spowodowane między innymi tym, że sporty elektroniczne (podobnie jak sporty tradycyjne) posiadają rozbudowaną infrastrukturę turniejową, pozwalającą na organizacje wydarzeń zarówno na szczeblu lokalnym, jak i międzynarodowym oraz oddane grono fanów, którzy gotowi są wspierać ten rodzaj sportu. Wartym odnotowania jest również fakt posiadanego potencjału rozwoju sportów elektronicznych. Jak podkreślają autorzy publikacji „The rise of E-Sports and potential for Post-COVID continued growth”[[8]](#footnote-8) sporty elektroniczne ze względu na swoją specyfikę posiadają dużo mniejsze bariery rozwojowe od sportów tradycyjnych, a różnego rodzaju sytuacje nadzwyczajne takie jak pandemie (np. pandemia COVID-19) nie wpływają na tą gałąź rynku tak znacząco, jak ma to miejsce w przypadku na przykład koszykówki czy piłki nożnej. Ostatnim aspektem, który może pomóc w odpowiedzi na pytanie z tytułu tego podrozdziału jest wartość dodana płynąca z emocji, których dostarczają sporty elektroniczne. Według analizy zawartej w publikacji „Video game play is positively correlated with well-being”[[9]](#footnote-9) granie w gry wideo wpływa pozytywnie na ocenę jakości życia oraz jest w obecnych czasach uważane za jeden z najbardziej popularnych sposobów spędzania wolnego czasu. Warto jednak zaznaczyć, że e-sport może również prowadzić do szeregu negatywnych skutków zdrowotnych, zarówno na płaszczyźnie fizycznej, jak i psychicznej. Według artykułu opublikowanego na oficjalnej stronie Światowej Organizacji Zdrowia (*ang. World Health Organization*, WHO) zbyt długie granie może prowadzić do uzależnienie od gier, które zostało zaklasyfikowane jako zaburzenie psychiczne. Zaburzenie to charakteryzuje się przedkładaniem rozgrywki ponad codziennymi obowiązkami oraz stratą kontroli nad ilością czasu spędzanego przed monitorem[[10]](#footnote-10). Podobnie jest w przypadku negatywnych skutków fizycznych. Nadmierna ilość czasu przeznaczana na granie może skutkować w rozmaitych chorobach takich jak: zespół cieśni nadgarstka, bóle i zapalenia rąk oraz ramion, otyłość czy stres psychiczny[[11]](#footnote-11). Co wskazuje, że w e-sporcie również mogę występować kontuzje podobne to tych obecnych w sporcie tradycyjnym. Rozwój e-sportu spowodował również powstanie nowych zagrożeń dla pasjonatów tego rodzaju sportu. Zarówno wydawcy gier jak i podmioty zewnętrzne dostrzegły możliwość spieniężenia zamiłowania fanów. Od wielu lat możemy zauważyć pojawienie się elementów hazardowych w grach przyciągających miłośników e-sportu. Do wielu produkcji została wprowadzona mechanika tak zwanych *Loot Boxów,* czyli skrzyń z losową zawartością, które gracze, skuszeni potencjalną wygraną, mogą nabyć (najczęściej) za realne pieniądze. Skala problemu była na tyle duża, że przyciągnęła również uwagę organów legislacyjnych. Podobnie sprawa ma się w przypadku zakładów bukmacherskich. Od wielu lat użytkownicy serwisów bukmacherskich takich jak STS czy LV Bet mają możliwość obstawiania meczy e-sportowych. Implementacja narzędzi hazardowych do gier e-sportowych może prowadzić do uzależnienia od hazardu, tym bardziej, że hazard ten jest często podawany w przyjaznej młodemu odbiorcy formie (tak jak w przypadku *loot boxów*), co sprawia, że nie jest on do końca świadomy zagrożenia. Podsumowując, pomimo licznych zagrożeń płynących z rozwoju branży e-sportowej, sporty elektroniczne dobrze wpisują się w definicję sportu, więc rozgrywki te należy klasyfikować jako sport, co więcej jako sport, posiadający dużo większe predyspozycje do rozwoju aniżeli sporty tradycyjne.

## **2.3 Dlaczego warto zajmować się analizą branży e-sportowej?**

Jak zostało wspomniane w poprzednich rozdziałach tej pracy, e-sport jest zjawiskiem stosunkowo nowym, ale posiadającym wierną grupę fanów, śledzących poczynania swoich idoli. Warto więc zastanowić się nad zasadnością analizy tej gałęzi rynku. Pierwszym aspektem, na który należy zwrócić uwagę jest rosnąca świadomość społeczeństwa w tematyce e-sportu. Według raportu przygotowanego przez firmę Newzoo „Global Esports & Live Streaming Market Report 2022”[[12]](#footnote-12) sytuacja rynku sportów elektronicznych z roku na rok jest coraz lepsza. Tabela 1 prezentuje wysokopoziomowe dane dotyczące tego rynku za rok 2022.

Tabela 1. Dane rynku e-sportowego za rok 2022.

|  |  |
| --- | --- |
| Całkowita populacja | 7,911.5 miliona |
| Populacja posiadająca dostęp do Internetu | 5,115.9 miliona |
| Populacja świadoma e-sportu | 2,469.5 miliona |
| Widownia gier nadawanych na żywo | 921.2 miliona |
| Entuzjaści e-sportu | 261,2 miliona |
| Przychody z e-sportu | 1,384 miliona |
| Średni roczny przychód na każdego entuzjastę | 5.30 dolara |

Źródło: Opracowanie własne na podstawie: Newzoo 2022 Global Esports & Live Streaming Market Report

Jak wynika z danych zaprezentowanych w tabeli 1 ludzie świadomi e-sportu stanowią prawie 50% (48,3%) populacji posiadającej dostęp do Internetu, a 18% wszystkich ludzi posiadających dostęp do sieci globalnej stanowi widownia gier nadawanych na żywo. Wartym odnotowania jest fakt, że aż 5% ogółu ludzi posiadających dostęp do Internetu deklaruje się jako entuzjaści sportów elektronicznych, co biorąc pod uwagę średni roczny przychód z każdego entuzjasty (na poziomie 5,3 dolara) daje roczne przychody z e-sportu na poziomie 1,384 miliarda dolarów. Kolejnym ważnym aspektem jest dynamika rozwoju tej branży. Dzięki raportom przygotowanym przez firmę Newzoo można zwizualizować tempo rozwoju branży e-sportowej. Rysunek 2 prezentuje wzrost wartości rynku e-sportowego na przestrzeni lat oraz prognozowane przyszłe wzrosty.

Rys. 2. Wartość rynku e-sportowego na przestrzeni lat 2017-2025.

Źródło: Opracowanie własne na podstawie: Newzoo 2022 Global Esports & Live Streaming Market Report i Newzoo 2018 Global Esports & Live Streaming Market Report

Jak wskazuje rysunek 2 rynek e-sportowy rozwija się w bardzo szybkim tempie i nawet w scenariuszu bazowym przewidywany jest dalszy wzrost. Podobną tendencję wzrostową wykazuje również liczebność publiki e-sportowej. Rysunek 3 przedstawia wzrost ilości widzów sportów elektronicznych.

Rys. 3. Liczebność widowni e-sportowej na przestrzeni lat 2017-2025.

Źródło: Opracowanie własne na podstawie: Newzoo 2022 Global Esports & Live Streaming Market Report i Newzoo 2018 Global Esports & Live Streaming Market Report

Jak wskazuje rysunek 3 ilość widzów e-sportu z roku na rok rośnie. Na przestrzeni lat 2017-2022 nastąpił wzrost aż o 58,8%. Warto też zwrócić uwagę na podział widowni e-sportowej pod względem płci i wieku.

Rys. 4. Podział widowni e-sportowej pod względem typu, płci i wieku.

Źródło: Opracowanie własne na podstawie: Newzoo 2022 Global Esports & Live Streaming Market Report

Rysunek 4 przedstawia podział widowni e-sportowej ze względu na typ (entuzjaści i widzowie okazjonalni) oraz płeć i wiek. Zgodnie z oczekiwaniami najliczniejszą grupę stanowią mężczyźni w wieku 21-35 lat, jednak wyraźnie widać, że sportami elektronicznymi interesują się ludzie w każdym wieku i niezależnie od płci. Z rozwojem branży e-sportowej wiąże się również rozwój innych branż z nią związanych- takich jak streamingi (czyli transmisje gier nadawane na żywo). Na platformach streamingowych takich jak YouTube czy Twitch istnieje możliwość prowadzenia transmisji ze swoich rozgrywek w czasie rzeczywistym. Wielu fanów e-sportu poza oficjalnymi meczami turniejowymi ogląda również transmisje swoich ulubionych graczy. Rysunek 5 przedstawia liczbę widowni transmisji gier na żywo w latach 2020-2025.

Rys. 5. Liczba widowni transmisji gier na żywo w latach 2020-2025.

Źródło: Opracowanie własne na podstawie: Newzoo 2022 Global Esports & Live Streaming Market Report

Jak wynika z rysunku 5 również w branży transmisji gier na żywo przewidywany jest znaczący wzrost. Jak wskazują przedstawione wyżej statystyki branża e-sportowa, pomimo bycia stosunkowo świeżą gałęzią rynku rozwija się bardzo dynamicznie, co tłumaczy zasadność jej analizy. Kolejną przesłanką podkreślającą sens analizy rynku e-sportowego jest również fakt obecności tej branży w codziennym życiu. Jak wspomniano w rozdziale 2.2 od wielu lat bukmacherzy oferują możliwość obstawiania wyników meczy e-sportowych. Zbudowanie modelu umożliwiającego predykcję meczy sportów elektronicznych może więc dać możliwość wygrywania większej ilości zakładów bukmacherskich oraz co za tym idzie, wzbogacenia się. Ostatnim aspektem, który wskazuje na potrzebę analizy branży e-sportowej jest fakt, że jest ona stosunkowo nowa, co sprawia, że ilość prac badawczych w tej tematyce jest bardzo ograniczona.

## **2.4 Przegląd literatury naukowej w temacie predykcji meczy e-sportowych**

Zważając na fakt, że sporty elektroniczne są zjawiskiem nowym, to również i stopień ich zbadania jest stosunkowo niski. Znalezienie prac naukowych w temacie predykcji meczy e-sportowych jest więc zadaniem wymagającym. Jak zostało wspomniane we wstępie analiza ta dotyczy konkretnej gry (League of Legends), co jeszcze bardziej zawęża krąg prac badawczych, na których można się opierać. Kluczowym pytaniem, które należy zadać jest wykonalność zaproponowanej predykcji. W tym przypadku literatura naukowa stawia jasną tezę, że taka analiza jest możliwa. Jak stwierdzają autorzy publikacji „E-Sports Player Performance Metrics for Predicting the Outcome of League of Legends Matches Considering Player Roles”[[13]](#footnote-13) przewidywanie wyników meczy e-sportowych jest skomplikowane pod wieloma względami. Przede wszystkim wskazują na dużą trudność w pozyskaniu odpowiedniej bazy danych, ponieważ mecze na średnim poziomie (względem rangi graczy) różnią się w znacznym stopniu od tych rozgrywanych pomiędzy drużynami mistrzowskimi, a do predykcji najlepiej jest wykorzystać zarówno dane meczowe jak i przedmeczowe. Kolejnym aspektem, na który autorzy zwracają uwagę, jest fakt nieprzewidywalności wydarzeń sportowych- jedno, z pozoru nic nieznaczące zagranie może prowadzić do zmiany rezultatu meczu. Przykładem tego może być sytuacja, w której na początku dany gracz zdobywa jedno dodatkowe zabójstwo. Może to prowadzić do zdobycia przewagi we wczesnych etapach gry, która zgodnie z efektem kuli śnieżnej (sytuacja, kiedy jakieś zjawisko, proces najpierw ma stosunkowo małe natężenie, ale stopniowo jego natężenie rośnie aż do bardzo dużych rozmiarów[[14]](#footnote-14)) przełoży się na większą dominację w późniejszej fazie meczu, co z kolei może skutkować wygraniem całego pojedynku. Dodatkowo, autorzy artykułu stwierdzają, że pomimo panującego przekonania o tym, że każda rola w drużynie jest sobie równa, to zgodnie z ich analizą posiadają one różne znaczenie dla wygrania meczu. Autorzy publikacji „Continuous outcome prediction of league of legends competitive matches using recurrent neural networks”[[15]](#footnote-15) zdają się również podzielać pogląd dotyczący możliwości dokonania predykcji meczy sportów elektronicznych. Wykorzystane przez nich rekurencyjne sieci neuronowe uzyskały dokładność predykcji wahającą się od 63,91% do 83,54%. Czynnikiem, który w przypadku ich analizy spowodował wahania trafności był czas- wyniki różniły się w zależności od minuty meczu. Może to wynikać z faktu, że etap meczu, w którym dany heros osiąga maksimum swojego potencjału jest indywidualny dla każdej postaci. Autorzy sugerują, że ich analiza może być wykorzystana do określenia, kiedy dany zespół, grający konkretną kompozycją osiąga moment największej siły. Na podstawie tego, można założyć, że wybór konkretnych postaci wpływa w znacznym stopniu na szansę danej drużyny na zwycięstwo. Podobnego zdania zdają się być autorzy artykułu „Using Machine Learning to Predict Game Outcomes Based on Player-Champion Experience in League of Legends”[[16]](#footnote-16). Podkreślają oni, jak ważnym czynnikiem jest opanowanie danej postaci. W League of Legends występuje ponad 160 postaci, z których każda dysponuje zestawem różnych umiejętności, więc kluczowym dla wyniku jest używanie postaci, których mechaniki gracz ma opanowane. W przypadku rozgrywek na najwyższym szczeblu należy założyć, że gracze operują tylko dobrze przetrenowanymi bohaterami, więc ważna jest siła postaci w danej aktualizacji, ponieważ wydawca gry (Riot Games) co dwa tygodnie wydaje aktualizacje, które wzmacniają „zbyt słabych” herosów i osłabiają tych „zbyt potężnych”, tym samym tworząc tak zwaną metę (ang. *most effective tactic available* – najlepszą dostępną w danym momencie strategię). Znaczenie wyboru odpowiedniej postaci podkreślają również autorzy publikacji „Victory prediction in League of Legends using Feature Selection and Ensemble methods”[[17]](#footnote-17). Ich analiza wykorzystuje zarówno dane przedmeczowe, takie jak wybrane i „zbanowane” (każda drużyna ma możliwość zablokowania przed meczem pięciu postaci, tak aby uniemożliwić przeciwnikowi ich wybranie, jednocześnie pozbawiając się możliwości ich użycia) postacie, jak i dane meczowe (np.: ilość zabójstw). Interesującym jest fakt, że uzyskana przez nich skuteczność predykcji na danych przedmeczowych jest niewiele niższa od tej uzyskanej na podstawie informacji z rozgrywki. Podważa to poniekąd tezę mówiącą o tym, że czas trwania meczu jest istotny, ponieważ skoro istnieje możliwość przewidzenia wyniku meczu przed jego rozpoczęciem, nie jest wymagana analiza w podziale na czas trwania pojedynku. Aczkolwiek Devan i współautorzy podkreślają, że najlepsza trafność predykcji przez nich osiągnięta, została uzyskana z wykorzystaniem zarówno danych przedmeczowych jak i meczowych. Ponadto wskazują oni na najważniejsze predyktory w analizie. W przypadku danych pozyskanych przed rozpoczęciem spotkania są to wybrane i zbanowane postacie, a podczas meczu: ilość zniszczonych wież, ilość zabójstw i śmierci oraz różnica w złocie między drużynami. Podobne podejście zostało zastosowane w artykule „League of Legends match outcome prediction”[[18]](#footnote-18). Również w tym wypadku autor zastosował zarówno dane przedmeczowe, jak i te pozyskane w trakcie meczu. Zgodnie z jego analiza najlepszymi algorytmami do predykcji meczy e-sportowych są Gradient Boosting oraz Gradient Boosting z wykorzystaniem regresji logistycznej. Inne algorytmy proponują jednak Costa i inni[[19]](#footnote-19). Ich zdaniem bardzo dobre rezultaty w przewidywaniu wyników meczy sportów elektronicznych dają zarówno regresja logistyczna, lasy losowe oraz algorytm maszyny wektorów nośnych (*ang. Support Vector Machines* - SVM). Jako najważniejsze zmienne objaśniające autorzy wskazują indywidualne cechy zawodników (rozumiane przez statystyki osiągane w meczu) oraz wybrane postacie. Wagę wyników graczy dla rozstrzygnięcia meczu zaznaczają również autorzy publikacji „ A Machine Learning based Analysis of e-Sports Player Performances in League of Legends for Winning Prediction based on Player Roles and Performances”. Stwierdzają oni, że statystyki osiągane przez graczy w danym spotkaniu są kluczowym aspektem w przewidywaniu wyników. W League of Legends oprócz graczy przeciwnych drużyn występują też obiekty neutralne (takie jak smoki, Baron Nashor czy Rift Herald), po pokonaniu których drużyna dostaje stałe lub czasowe premie do statystyk. Autor eseju „eSports Game Skill Analysis & Prediction: League of Legends”[[20]](#footnote-20) wskazuje, że oprócz statystyk takich jak złoto na konkretnych pozycjach, zabójstwa i asysty, istotną rolę odgrywają też obrażenia zadane w obiekty neutralne, co można powiązać ze zdobywaniem tych obiektów, a właściwie bonusów z nich płynących. Zdanie to wydaje się podzielać również Tian Wang. W swojej publikacji „Predictive analysis on eSports games: A case study on league of legends (LOL) eSports tournaments” stwierdza on, że najważniejszymi zmiennymi objaśniającymi w tego rodzaju analizie są przewagi danej drużyny w: zabójstwach, smokach, Rift Heraldach i Baronach Nashorach. Podkreśla on zarazem, że oprócz zmiennych płynących z rozgrywki powinniśmy brać pod uwagę również dobór postaci oraz preferencje konkretnych zawodników, ponieważ, pomimo że w rozgrywkach na najwyższym szczeblu gracze mają opanowane wszystkie postacie (bądź ich większość), to każdy gracz lepiej odnajduje się w konkretnym stylu gry i ważny jest również dobór postaci, pasujących do kompozycji drużyny. Jak wspomniano w rozdziale 2.1 gra League of Legends należy do gatunku gier MOBA. Gry z tego gatunku opierają się na podobnych zasadach i schematach. Autorzy artykułu „Real-time eSports Match Result Prediction”[[21]](#footnote-21) przeprowadzili predykcję wyników meczy w grze Dota 2, która również należy do tego samego gatunku. Wnioski płynące z ich analizy są bardzo podobne do tych, wynikających z publikacji dotyczących gry będącej przedmiotem badania tej pracy. Ponownie jako kluczowe predyktory zostały wskazane różnice drużyn w złocie, zabójstwach, asystach i zgonach, jednak w tym przypadku badacze stwierdzają istotny wpływ czasu trwania meczu. Podsumowując, zgodnie z dostępną literaturą naukową predykcja meczy e-sportowych w grze League of Legends jest możliwa. Różnorodność modeli i algorytmów wykorzystywanych przez badaczy sugeruje, że dobór odpowiedniej metody badawczej jest kwestią subiektywną i zależy od posiadanego zbioru danych oraz indywidualnych preferencji. Oczekuje się, że najważniejszymi predyktorami będą: różnice zasobów danej drużyny (w złocie, zabójstwach, zgonach i asystach), ilość zdobytych wież oraz obiektów neutralnych i dobór odpowiednich postaci. Warto też przeprowadzić analizę w podziale na różne fazy meczu, w celu zweryfikowania czy siła danej kompozycji w konkretnym momencie spotkania ma realny wpływ na jego wynik.

## **2.5 Jakie jest zdanie profesjonalnych zawodników na temat tej analizy**

Ważnym aspektem przy analizie wydarzeń sportowych jest poznanie zdania osób wywodzących się z danej branży. W tym celu skontaktowano się z profesjonalnym zawodnikiem Marcinem „bucu” Świechem[[22]](#footnote-22), który zgodził się udzielić wywiadu na potrzeby pracy magisterskiej. Poniżej umieszczona została transkrypcja tego wywiadu. Litery „Ł” i „M” oznaczają podział na odpowiednio prowadzącego oraz odpowiedzi Marcina. W przypisach wytłumaczono sformułowania potoczne oraz te występujące w obrębie gry.

**Ł:** Cześć, ja nazywam się Łukasz Makaruk, jest ze mną Marcin Świech pseudonim Bucu.

**M:** Witam bardzo serdecznie.

**Ł:** Witamy cię Marcin. Marcin jest weteranem polskiej sceny, jednym z najlepszych midlanerów [[23]](#footnote-23) w naszym kraju. Grał w takich dużych jak Devil’s One, Gentleman's Gaming, czy teraz w LODIS. O ile dobrze to wymawiam?

**M:** Dobrze, bardzo dobrze.

**Ł:** Dzięki wielkie ogólnie z góry, że zgodziłeś się udzielić mi wywiadu.

**M:** Nie ma problemu, ja zawsze chętny na takie akcje.

**Ł:** No to może bez przedłużania, pierwsze pytanko: **Czy e-sport jest gałęzią**, Twoim zdaniem oczywiście, **której analiza może dać nam jakąś wartość?**

**M:** Wydaje mi się, że e-sport jest jeszcze bardzo stosunkowo młodą formą rozrywki, więc wydaje mi się, że patrząc na ostatnie lata i na to jak ja zaczynałem, a w którym miejscu jest teraz obecnie e-sport to, że rozwija się bardzo dynamicznie. I uważam, że prowadzenie analiz na pewno może dużo dać. Czy to pod wartością tego, jak można inwestować w e-sport, ale również po tym, jak zawodnicy mogą się rozwijać oraz inne jakieś tam branże.

**Ł:** O, bardzo dobrze uderzyłeś w temat inwestowania, bo następne pytanie właśnie tego dotyczy. **Czy uważasz, że dodatkowe nakłady finansowe, inwestowane w branżę e-sportową dają nam jakąś wartość dodaną?**

**M:** Znaczy uważam, że patrząc na polski e-sport tak typowo, tylko na polski e-sport, to z pewnością, gdyby te pieniądze były większe i większe byłyby nakłady finansowe dla drużyn, to też poziom e-sportowy był rósł, bo wielu zawodników dobrych, bardzo dobrych, polskich na początku swojej kariery już wyjeżdża za granicę i to nie jest bez powodu. Głównym powodem, dlaczego zostajesz graczem e-sportowym i poświęcasz na to bardzo dużo czasu to jest tak jak każda inna praca czy każdy inny sport- pieniądze. I jak masz do wyboru grać za granicą w jakichś dobrych drużynach, a grać w Polsce za jakieś dużo mniejsze pieniądze, a robisz tą samą pracę to wiadomo, że wolisz sobie wyjechać, więc uważam, że w samej Polsce na pewno tak, za granicą już są dobre pieniądze, więc czy dają na pewno jakąś korzystną wartość dodatnią? Na pewno dają, ale w jakim stopniu to nie wiem.

**Ł:** Dobra, dzięki wielkie. No już poruszyłeś kolejny z tematów, które chciałem z Tobą przegadać[[24]](#footnote-24), a mianowicie kondycję polskiego e-sportu to rozumiem, że podtrzymuje swoje zdanie o tym, że te **pieniądze inwestowane u nas w kraju by się przydały**?

**M:** Tak, jak najbardziej i to zarówno powiem tak: jeżeli pieniądze byłyby od początku dobrze inwestowane i zajmowałyby się tym osoby, które się na tym e-sporcie znają, to na pewno by był zysk dla sponsorów i dla graczy. Ale często jest tak w polskim e-sporcie, że przychodzą inwestorzy, którzy nie mają pojęcia o e-sporcie, wykładają kasę, liczą na to, że samo wszystko się zrobi, że będą mieli zwrot. No tak to nie działa. Wydaje mi się, że e-sport jest na tyle specyficzną dziedziną, że trzeba trochę na to czasu, trzeba się na tym znać i na pewno zadbać o obsadę, która się na tym zna, i która zadba o to. Więc generalnie jakby miał porównać polskie e-sport do zagranicznych, no to uważam, że jest w tym momencie bardzo słabo. Jeszcze jakieś dwa/trzy lata temu widzieliśmy tendencję taką wzrostową, jeżeli chodzi nawet o samą oglądalność, w tym momencie jest ona raczej spadkowa, natomiast w krajach sąsiednich czy to Niemcy, czy trochę dalej Hiszpania, to cały czas ta widownia wzrasta, zainteresowanie wzrasta i też wiadomo, no pieniądze też są coraz większe wszędzie.

**Ł:** Widzę, że podzielasz opinie innych osób właśnie z branży, z którymi rozmawiałem. No i ja też się zgadzam, bo widzę te statystyki, i też przyszłość polskiego sportu nie napawa mnie optymizmem, i fajnie by było jakby jednak coś z tym zostało zrobione. Dobra, teraz może przejdźmy do pytań bardziej w temacie mojej pracy magisterskiej. Przed nagrywkami[[25]](#footnote-25) trochę gdzieś tam przybliżałem temat, no ale tak jeszcze dla przypomnienia mój temat dotyczy wykorzystania narzędzi Machine Learningowych i optycznego rozpoznawania znaków, żeby na podstawie streamów przewidywać mecze. Czy więc uważasz, że **zadanie predykcji meczy e-sportowych na podstawie nagrań meczowych jest wykonalne dla kogoś na moim etapie, czyli początkującego badacza?**

**M:** No powiem szczerze. Wydaje mi się, że dla takiego początkującego badacza jest to bardzo trudne zadanie. Tak naprawdę transmisje nie dają kompletu wszystkich informacji o grze, a ich pozyskanie może być raczej ciężkie. Wydaje mi się też, że z samego streamu[[26]](#footnote-26) jedna akcja może mieć wpływ na całą rozgrywkę, więc nie wiem, nie wiem. Wydaje mi się, że trzeba mieć trochę większy, jakby to powiedzieć…

**Ł:** Może zasoby?

**M:** Większe zasoby, większy dostęp do informacji niż tak sam stricte stream.

**Ł:** No właśnie, przygotowując się do tej magisterki widziałem, że większość osób w pracach badawczych wykorzystuje API stworzone przez Riot Games. Ja zdecydowałam się pójść inną drogą, może rzuciłem się na głęboką wodę, ale byłem po prostu sam ciekaw, czy to jest wykonalne. Powiedz mi, **czy twoim zdaniem istnieją jakieś decydujące akcje lub wydarzenia, które mają wpływ na wynik meczu?**

**M:** No to oczywiście, że tak. Często jest tak, że mecz jest przez czterdzieści minut bardzo wyrównany, nagle jedna walka, czy tam jedno złe ustawienie może wpłynąć na to, że w pół minuty przegrywasz całą grę. A gra do czterdzieści minuty była wyrównana. Tak samo na początku, czy w trakcie każdej minuty gry, jedno złe ustawienie drużyny może wpłynąć na całą rozgrywkę, więc na pewno tak.

**Ł:** Dobra. **Jakie są** takim Twoim prywatnym zdaniem **najważniejsze wskaźniki lub parametry, które można wyciągnąć z tych transmisji i wykorzystać do predykcji meczów? Czy są to na przykład statystyki dotyczące zabójstw, farmy, czy może coś innego?**

**M:** Do predykcji wyników meczów no to tak jak chyba w każdej dziedzinie czy to e-sport, czy nawet piłka nożna, czy cokolwiek, no to najważniejsze są wyniki. Kto najwięcej meczów wygrywa. Jeżeli stricte mówimy o LoL-u, no to wiadomo, że można przez sam stream patrzeć na to, w jaki sposób te drużyny wygrywają? Czy one wygrywają właśnie losowo przez jedną akcję, która im się uda? Czy wygrywają także od początku rozgrywki dominują cały mecz i zdobywają te dwadzieścia zabójstw, czy na przykład grają tak, że mają tylko trzy zabójstwa i raczej grają cały czas stabilnie- są blisko i później jedną akcją decydują. Złoto też na to wpływa, ale to idzie wszystko przez to właśnie jak grasz drużynowo, jak się ustawiasz czy grasz w tempo, czy szybciej bierzesz obiekty, jakby to wszystko wpływa na złoto, więc złoto to jest taki już skutek tego co robisz wcześniej. Więc tak naprawdę na pewno są czynniki i parametry, które można spokojnie wyciągnąć z takiej transmisji i jakieś tam predykcje swoje mieć.

**Ł:** Dobra. **A czy poza tymi statystykami, czy istnieją jakieś strategie lub podejścia taktyczne, które w tym modelu mógłbym uwzględnić? A jeśli kojarzysz jakieś, to jakbyś mógł przywołać ze splitu letniego LEC-a w 2022[[27]](#footnote-27)**, no bo w sumie to na nim opieram swoją pracę.

**M:** No to jedyne co mogę powiedzieć, to na pewno, że na sto procent duże znaczenie ma kompozycja, każda gra jest tak naprawdę inna, w każdej grze możesz wziąć inne postacie, ale często drużyny mają swój styl gry i trzymają się tego stylu. Czasem bywa tak, że jakiś przeciwnik jest dobry na jakiegoś, ale z innym sobie nie radzi. Z kolei tamten, z którym ma problem drużyna pierwsza to wygrywa sobie spokojnie z drużyną drugą, więc no myślę, że to jest tak trochę jak w sporcie. A jeżeli no pytasz tam o summer split[[28]](#footnote-28), to wydaje mi się, że największe znaczenie tutaj miały konkretne wybory postaci na linii duo, czyli na botlanie[[29]](#footnote-29). I to są takie typowe połączenia jak Lucian Nami, Senna Tahm Kench, Zeri Yummi. Tak naprawdę większość banów[[30]](#footnote-30), czyli wykluczeń tych postaci była kierowana w te postacie, żeby właśnie tych połączeń za dużo nie było. Ze względu na to, że często te zestawienia były po prostu zbyt mocne.

**Ł:** No właśnie też oglądając te mistrzostwa Europy zauważyłem, że w przypadku Lucian Nami to było pick or ban[[31]](#footnote-31), że albo postać, któraś z tych postaci była zbanowana albo brana. Dlatego też chciałabym poznać Twoją opinię na ten temat. **Jakie są Twoim zdaniem takie największe wyzwania związane z analizą transmisji meczowych w LoLu? Czy istnieją jakieś trudności techniczne, z którymi Twoim zdaniem mógłbym się spotkać w trakcie przygotowywania tej pracy?**

**M:** Wydaje mi się, że problematyczne może być wydobywanie poprawnych statystyk z nagrań. Generalnie oglądając streama można się opierać na tym co się widzi, ale nie widzi się, co się dzieje wewnątrz drużyny, o czym te drużyny rozmawiają i to ma największe jakby skutki na to co się dzieje później na transmisji. Wydaje mi się również, że na transmisjach często takie rzeczy jak animacje czy jakieś różne reklamy podczas streama mogą utrudniać zadanie takie typowo badanie statystyk podczas meczu. No i z tego co wiem to większość ludzi korzysta z gotowych baz, które zapewnia Riot oraz z Riot API. No i wydaje mi się, że to dużo ułatwia, ale generalnie i tak to nie jest łatwe zadanie.

**Ł:** Powiedz mi jeszcze, wracając do rozgrywki, **czy w meczu zdarzają się sytuacje, które są trudne do przewidzenia bądź wychwycenia z samej transmisji meczowej. Takie jak na przykład jakieś nieprzewidziane zagrania, zmiany taktyk, czy roamy[[32]](#footnote-32) chociażby. Czy to ma realny wpływ na wynik meczu?**

**M:** Tak, wszystkie te aspekty tak naprawdę, które wymieniłeś są bardzo ważne. Każda decyzja jest bardzo ważna. Często drużyny mają jakiś plan przed samym meczem i ten plan może ulec zmianie kilka, a nawet kilkadziesiąt razy podczas meczu. Po prostu trzeba się dostosowywać do taktyki też przeciwnika, której nie znamy wcześniej. Mogą czymś zaskoczyć, mogą tym razem inaczej się ustawić, inaczej iść. Często są wyciągane statystyki na podstawie tego, gdzie zaczynał przeciwny leśnik[[33]](#footnote-33), czyli tak jakby wolny elektron, który chodzi po mapie i pomaga. Może iść od dołu do góry, może iść z góry do dołu, a może w ogóle nie gankować[[34]](#footnote-34), tylko sobie farmić[[35]](#footnote-35) i zdobywać złoto. Często sami lanerzy[[36]](#footnote-36) zmieniają styl- raz grają agresywnie, raz mniej, jakby trzeba cały czas patrzeć na to co się dzieje na mapie, plan może się bardzo dużo razy zmieniać. Jedną z najważniejszych chyba takich rzeczy podczas profesjonalnych rozgrywek jest to nie samo stricte rozgrywanie początku i laning phase-u[[37]](#footnote-37), ale właśnie walki drużynowe. Trzeba się dogadać, kto jak ma zagrać. Wychwycić dobry moment, kiedy możesz wejść, zrobić dobry engage[[38]](#footnote-38). Pojedynczy dobry engage w odpowiednią osobę z drużyny przeciwnej może Ci wygrać grę. Więc wszystko to trzeba brać pod uwagę tak samo jak to, że np. cztery smoki dają Ci bardzo duży buff[[39]](#footnote-39) i na przykład jak przeciwnik ma już trzy te smoki i brakuje mu jednego do tego, żeby mieć ten właśnie duży buff, czyli takie duże wzmocnienie, no to drużyny poświęcają coś na drugiej stronie mapy po to, żeby walczyć o tego smoka. No i właśnie to są takie rzeczy, których przed meczem nie przewidzisz, musisz się dostosowywać.

**Ł:** Dobra. A jakie są twoje doświadczenia? Czy miałeś w ogóle kiedyś, czy spotkałeś się kiedyś z kimś, kto prowadził podobne badanie jak ja? **I czy uważasz, że przewidywanie na podstawie nagrań jest łatwiejsze bądź trudniejsze w porównaniu do bazy danych?** Wiem, że po części odpowiedziałeś na to pytanie. **Ale czy uważasz, że występuje aspekt losowości?** Na przykład, że jedna ma akcja zaważy o całym przebiegu danego spotkania?

**M:** No generalnie to powiem Ci, że w każdej drużynie, w której grałem i wszystkie drużyny, które są szanowane, i drużyny e-sportowe mają swoich analityków, więc tak naprawdę z dokładnymi analizami to ja praktycznie się spotykam w każdej drużynie, z meczu na mecz przygotowujemy się daną drużynę. Jest sprawdzane czym grali, jak grali, kiedy wygrali, w jakim momencie podejmują jakie decyzje. Analiza jest naprawdę bardzo obszerna zawsze przez analityków oczywiście prowadzona, później nam przedstawiana przed meczem. No i wiadomo, no to jest tak jak z piłką nożną, jak z każdym sportem, uważam, że można tutaj porównywać ze względu na to, że to jest też rywalizacja i dobrze, poruszamy się postaciami w grze, ale kierują tymi postaciami ludzie, więc jakby ludzie decydują koniec końców, co się wydarzy i w którą stronę dana postać pobiegnie. Więc nie da się w stu procentach wszystkiego przewidzieć i w stu procentach przygotować, i przewidzieć każdy ruch przeciwnika. Często, bardzo często, można powiedzieć, że w każdym meczu jest coś takiego jak czynnik losowy, że ktoś nagle zrobi coś, co nie miał w planach nawet on sam, bo po prostu przyszło mu to do głowy, zrobił to, wyjdzie albo nie wyjdzie. Później są wyciągane z tego wnioski czy to było dobre czy nie, więc no uważam, że da się przewidywać na pewno jakiś wynik, bo są gracze lepsi, są gracze gorsi i drużyny lepsze i gorsze, tak jak w każdym sporcie, ale nadal może być tak, że najgorsza drużyna będzie w stanie wygrać z najlepszą, jeżeli będą mniej więcej jakiś poziom trzymać i mogą czymś zaskoczyć, czy to jakimś niespecyficznym wyborem, znaczy specyficznym wyborem, ale nietuzinkowym. No i tak samo decyzjami. Jakimś losowym biciem się na dole, biciem się nagle na górze, przeciwnik się nie spodziewał, bo to mi się niby nie opłacało, ale koniec końców im się opłaca, więc losowość jest zawsze.

**Ł:** Dobra, no i na koniec takie pytanie może trochę na rozluźnienie, chociaż to dość kontrowersyjny temat. **Czy** Ty prywatnie **uważasz, że e-sport może być uznawany za sport?** Taki sport z krwi i kości.

**M:** Znaczy uważam, że może być uznawany za sport, ale czy chciałbym, żeby kiedyś sport był łączony ze e-sportem, to raczej nie. Wydaje mi się, że podobną dyscypliną, która jest uznana za sport, a głównie opiera się na myśleniu i graniu tak stricte też przez jakieś urządzenia to są na przykład szachy, warcaby i to jest też uznawane za sport. Jednak uważam, że fajnie jak e-sport rozwija się swoim torem, sport swoim, jest dużo zwolenników, jeszcze chyba więcej przeciwników, czyli osoby, które nie grają, nie wiedzą o co chodzi w tym i fajnie jak po prostu e-sport będzie szedł swoją drogą. Ja uważam, że można to uznać za sport, bo jednak nie pamiętam, czy byłem kiedykolwiek tak głodny jak po jakiejś BO5[[40]](#footnote-40), przed którą byłem najedzony, ale przez to jaki wysiłek umysłowy musiałem podczas całego meczu zrobić to naprawdę kończyłem wymęczony, bardzo głodny i na pewno to obciąża organizm. I uważam, że spokojnie można to uznawać za sport, ale lepiej, żeby nie był dodany jako sport, bo wtedy wiadomo dużo by było jakiś dymów[[41]](#footnote-41), a e-sport sam sobie może się rozwijać. Tyle.

**Ł:** Dobra, no to jeszcze raz, dzięki śliczne za udzielenie mi tego wywiadu. No i powodzenia w kolejnych meczach, i mam nadzieję, że za rok będę miał możliwość analizowania Twoich meczy w Mistrzostwach Europy, a może nawet na Worlds-y[[42]](#footnote-42) pojedziecie? No zobaczymy.

**M:** Dziękuję również. Wydaje mi się, że jest daleka droga do tego. No ale nigdy nic nie wiadomo, tak? Trzeba tak mówić, także za rok się widzimy na analizie. Nie ma problemu. Dzięki również za ciekawe pytania i za wywiad.

Rozmowa z profesjonalnym zawodnikiem z pewnością wniosła dużą wartość dodaną w temacie analizowanym w tej pracy. Marcin zwrócił uwagę na wiele aspektów, które należy wziąć pod uwagę w dalszej analizie. Potwierdził on również tezy podnoszone w omówionej w podrozdziale 2.4 literaturze naukowej oraz zwrócił uwagę na skomplikowanie podjętego zadania i potencjalne trudności, na które trzeba zwrócić uwagę.

# III. Omówienie danych wykorzystanych w analizie

## **3.1 Sposób pozyskania bazy danych**

Prace naukowe omówione w rozdziale 2.4 w większości bazowały na bazach danych utworzonych z wykorzystaniem oficjalnego Riot Games API- API stworzonego przez wydawców gry, dostępnego na stronie deweloperskiej tej gry[[43]](#footnote-43). W tej pracy wykorzystane zostało natomiast inne podejście. Do utworzenia bazy użyto zapisów nagrań meczowych kanału Polsat Games[[44]](#footnote-44) na platformie YouTube, które następnie zostały przetworzone i z wykorzystaniem optycznego rozpoznawania znaków oraz detekcji elementów pozyskano z nich wszystkie niezbędne informacje. Cały kod został napisany w języku Python. W pierwszym kroku za pomocą biblioteki pytube pobrano nagrania meczowe dotyczące rozgrywek League of Legends European Championship 2022 lato. Wykorzystano nagrania z ośmiu tygodni rozgrywek, po dwa dni w każdym tygodniu (oprócz tygodnia ósmego, w którym były trzy dni meczowe). Każdego dnia rozgrywanych było po pięć meczy co łącznie daje 85 spotkań. Następnie, aby pozyskać klatki z meczy, skorzystano z funkcji zamieniającej nagranie wideo na klatki (wykorzystującej bibliotekę cv2). Zważając na fakt, że nagrania mają częstotliwość wyświetlania klatek na poziomie 30 klatek na sekundę postanowiono ograniczyć ilość zapisywanych klatek do jednej na około minutę (sekundowe różnice mogą wynikać z czasu potrzebnego na pozyskanie klatki). Wynika to z faktu, że w ciągu mniej niż minuty statystyki meczowe z reguły nie ulegają znaczącej zmianie. Tym sposobem osiągnięto 6898 klatek. Jedynie część uzyskanych klatek przedstawia nagranie meczu, ponieważ do transmisji wliczają się również zdarzenia, które meczami nie są, takie jak studio przedmeczowe, wywiady czy przerwy między spotkaniami. Z tego powodu zdecydowano się podzielić klatki na meczowe i nie meczowe. Aby tego dokonać posłużono się funkcją wykorzystującą bibliotekę cv2. Kryterium decyzyjnym w tej kwestii było złoto danej drużyny (funkcja ta zostanie omówiona w następnym rozdziale). Jeśli funkcja zwracała wartość None to dana klatka trafiała do folderu klatek nie meczowych, w przeciwnym razie do meczowych. Niestety funkcja ta nie dawała stuprocentowej skuteczności (błędne odczyty funkcji), ale pojedyncze błędne przyporządkowania zostały poprawione ręcznie. W wyniku tej operacji uzyskano 4087 klatek niepochodzących z rozgrywki oraz 2811 klatek ze spotkań, które posłużyły do utworzenia bazy danych.

## **3.2 Wykorzystane zmienne**

Zmienne pozyskane do analizy zostały uzyskane z klatek meczowych. Schemat działania w większości przypadków był bardzo podobny. W pierwszym kroku ładowano obraz z wykorzystaniem funkcji **imread** z biblioteki cv2. Następnie przycinano obraz do obszaru zawierającego interesujące nas dane. Potem transformowano obraz na skalę szarości, aby umożliwić łatwiejsze rozpoznawanie znaków. W kolejnym kroku obraz zostawał wyostrzony z wykorzystaniem rozmycia gaussowskiego wykorzystującego jądro rozmycia o wymiarach 5 na 5, odchylenie standardowe rozmycia gaussowskiego w wysokości 1, wzmocnienie wyostrzenia na poziomie 1 oraz próg kontrastu wynoszący 0. Próbowano ponadto zaimplementować progowanie obrazu, jednak jego zastosowanie pogarszało rezultaty, przez co zdecydowano się z niego nie korzystać. Na tak przygotowanym obrazie stosowano funkcję **image\_to\_string** z biblioteki cv2. W celu zrozumienia, jakie informacje możemy wyciągnąć z transmisji meczy e-sportowych, należy przyjrzeć się danym, jakie zawarte są w klatkach ze spotkań. Rysunek 6 prezentuje przykładową klatkę z meczu.

Rys. 6. Przykładowa klatka z meczu.

Obraz zawierający Gra komputerowa, Strategiczna gra wideo, tekst, Oprogramowanie gier wideo

Opis wygenerowany automatycznie

Źródło: Opracowanie własne

Na rysunku 6 konkretne zmienne zostały oznaczone numerami, które zostaną wykorzystane w ich opisie.

1. Zmienna **gold**

Zmienna ilościowa obrazująca różnice w złocie pomiędzy drużynami. Dane były pobierane dla każdej drużyny oddzielnie, a następnie była liczona ich różnica. W funkcji rozpoznawania znaków dla tej zmiennej ograniczono ilość dopuszczalnych znaków do cyfr od 0 do 9, kropki oraz litery „k”, która jest popularnym skrótem słowa „tysiąc” (1k = 1000, 2k = 2000 itd.).

1. Zmienna **towers**

Obrazuje ona różnice w ilościach zniszczonych struktur między drużynami. Ponownie dane były pozyskiwane oddzielnie dla każdej z drużyn. Tym razem dopuszczalnymi znakami są tylko cyfry od 0 do 9, jednak, jeśli różnica wychodziła większa od 14 to funkcja zwracała wartość None. Zdecydowano się na taki warunek, ponieważ na każdej alei są 3 wieże i jeden inhibitor, więc doliczając dwie wieże broniące Nexusa (po jego zniszczeniu następuje koniec meczu) każda drużyna może zniszczyć maksymalnie 14 struktur.

1. Zmienne **blue\_team** i **red\_team**

Są to zmienne jakościowe określające drużyny, które uczestniczą w meczu. Zdecydowano się na zastosowanie tych zmiennych, ponieważ zauważono dominację w turnieju niektórych organizacji- np. drużyna G2 Esports zdobyła mistrzostwo Europy 9 razy, a Fnatic sięgnęło po tytuł siedmiokrotnie[[45]](#footnote-45). Do turnieju analizowanego w tej pracy zakwalifikowało się 10 drużyn. Tabela 2 prezentuje zestawienie drużyn i skrótów ich nazw, które używane są w transmisjach i pozwalają rozkodować uczestników spotkania.

Tabela 2. Drużyny i skróty ich nazw w turnieju LEC 2022 Lato.

|  |  |
| --- | --- |
| **Drużyna** | **Skrót** |
| Astralis | AST |
| Fnatic | FNC |
| G2 Esports | G2 |
| MAD Lions | MAD |
| Misfits | MSF |
| Rogue | RGE |
| SK Gaming | SK |
| BDS Esports | BDS |
| Team Vitality | VIT |
| Excel Esports | XL |

Źródło: Opracowanie własne na podstawie: https://pl.egamersworld.com/lol/event/lec-summer-2022-4J\_8wxAS3

1. Zmienna **kills**

Zmienna ilościowa określająca różnice w zabójstwach zdobytych przez każdą z drużyn. W tym przypadku również dane dla każdej drużyny były pobierane oddzielnie. Wprowadzono dodatkowy warunek stanowiący o tym, że gdy różnica w zabójstwach przyjmuje wartość większą od 14 to jej wartość ustawiana jest na None. Ograniczenie to wynika z faktu, że w profesjonalnych rozgrywkach rzadko zdarza się sytuacja, gdy ta różnica jest aż tak duża, więc różnice tego pokroju mają dużą szansę wynikać z niepoprawnego działania funkcji optycznego rozpoznawania znaków.

1. Zmienna **time**

Jest to zmienna, która przedstawia minuty i sekundy meczu. Pozwoli ona podzielić posiadane klatki meczowe na odpowiednie przedziały czasowe (zgodnie ze zaleceniami literatury o analizie w podziale na czas). Może być również pomocna przy ręcznej imputacji danych- pomoże namierzyć której klatki dotyczy dany rekord tabeli.

1. Zmienne dotyczące postaci

Jest to zestaw zmiennych określających postać wykorzystywaną przez konkretnego zawodnika w danym meczu. W spotkaniu biorą udział dwie drużyny (oznaczone jako red i blue), każda licząca pięciu zawodników różnych pozycjach (górna aleja, dżungla, środkowa aleja, strzelec i wspierający) co daje nam 10 zmiennych jakościowych. Jak wspominane jest w literaturze analiza wykorzystywanych postaci jest ważnym czynnikiem badania rzeczonego zjawiska. Ponadto, niektóre kompozycje postaci mają ze sobą lepszą synergię co sprawia, że ich połączenie charakteryzuje się większym współczynnikiem zwycięstw. Przykładem takiej pary postaci jest strzelec Lucian i wspierająca Nami[[46]](#footnote-46). Daje to możliwość utworzenia kombinacji zmiennych jakościowych i sprawdzenie czy ich wpływ na wygraną jest znaczący.

1. Zmienne dotyczące KDA

Opis tych zmiennych należy zacząć od omówienia czym jest KDA. Jest to skrót pochodzący z języka angielskiego oznaczającego *Kill Death Assist*, czyli zabójstwa, zgony i asysty. Nie jest zaskoczeniem, że aby wygrywać mecze należy eliminować przeciwników, pomagać sojusznikom oraz samemu nie umierać, co dokładnie reprezentuje ten współczynnik. Jest on liczony na podstawie wzoru[[47]](#footnote-47):

W teorii im wyższy wynik KDA, tym lepszy występ danego zawodnika. Współczynnik ten został obliczony dla każdej pozycji (górna aleja, dżungla, środkowa aleja, strzelec i wspierający) w obu drużynach. Następnie policzona została różnica między odpowiadającymi sobie pozycjami (czyli np. różnica między graczem z górnej alei drużyny niebieskiej oraz graczem górnej alei drużyny czerwonej). Tym sposobem uzyskaliśmy pięć zmiennych dotyczących KDA.

1. Zmienne dotyczące stworów

Stwory (ang. *minions*/*creeps)* są cennym źródłem złota dla zawodników, a ich wpływ na majętność drużyn w danej rozgrywce jest często niedoceniany przez graczy z niższych rang. Do tej grupy zalicza się pięć zmiennych, opisujących różnice w ich liczbie dla każdej pozycji. Ponownie sposób pozyskania polega na odczytaniu liczby stworów dla każdego gracza oddzielnie i wyliczeniu różnicy między odpowiadającymi sobie rolami w obu drużynach.

1. Zmienne dotyczące wizji oraz **herald**

W League of Legends występują tak zwane totemy wizji. Pozwalają one na odkrycie niewidocznych fragmentów mapy, a tym samym ujawnienie działań przeciwnika. Mimo, że metryka Vision Score[[48]](#footnote-48) z perspektywy gracza wydaje się ważnym czynnikiem wpływającym na zwycięstwo, to problemy i błędy podczas pozyskiwania tej zmiennej (zostaną one omówione w późniejszej części pracy) nakazały zrezygnować z wykorzystania jej. Jednym z tych problemów jest zmienna odpowiedzialna za pozyskanie informacji o zgładzeniu neutralnego obiektu jakim jest Rift Herald- w momencie, gdy dana drużyna zabije ten obiekt, na ziemi pojawia się tak zwane Oko Herolda[[49]](#footnote-49). Jest to wzmocnienie, które zapewnia szybszy powrót do bazy oraz po jego użyciu przywoływana jest kopia bestii, która zdecydowanie ułatwia zdobywanie struktur. Z tego powodu zdecydowano się na wykorzystanie tej zmiennej. Sposób jej pozyskania był zdecydowanie różny od omówionych wcześniej. W pierwszym kroku przygotowano szablon Oka Herolda, następnie wykorzystując funkcję **matchTemplate** z biblioteki cv2 porównano klatkę wejściową oraz szablon, w celu detekcji elementu Oka Herolda w ekwipunku graczy. Tym sposobem została utworzona zmienna mówiąca o tym, czy dana drużyna zrobiła Heralda. Zmienną tą należało jednak dopracować, o czym będzie mowa w dalszej części pracy.

1. Zmienne dotyczące smoków

W League of Legends występują obiekty neutralne jakimi są smoki żywiołów. Każdy rodzaj smoka daje drużynie inne bonusy, przedstawione w tabeli 3.

Tabela 3. Smoki i ich efekty w turnieju LEC 2022 Lato.

|  |  |
| --- | --- |
| **Rodzaj smoka** | **Efekt** |
| Smok Piekielny | +8 / 16 / 24% obrażeń od ataku i mocy umiejętności |
| Smok Górski | +10 / 20 / 30% obrażeń przeciwko potężnym potworom i wieżom w postaci nieuchronnych obrażeń |
| Smok Powietrzny | +25 / 50 / 75% jednostek prędkości ruchu poza walką |
| Smok Oceaniczny | Co 5 sekund otrzymujesz 4 / 8 / 12% brakujących punktów zdrowia i many, jeśli w ciągu 8 sekund wrodzy bohaterowie nie zadali ci obrażeń |
| Smok Hextechowy | Zmniejszony czas odnawiania umiejętności i zwiększona prędkość ataku |

Źródło: Opracowanie własne na podstawie: https://leagueoflegends.fandom.com/pl/wiki/Pogromca\_Smoków i https://boop.pl/league-of-legends/dwa-nowe-broken-smoki-w-lolu-beda-mialy-teleporty-tryb-zombie-przyspieszenia-i-nie-tylko

Zrobienie czterech smoków żywiołów skutkuje uzyskaniem tak zwanej duszy smoka danego żywiołu, co wiąże się z pozyskaniem jeszcze większego wzmocnienia. Ponadto po pokonaniu czterech smoków przez daną drużynę pojawia się Starszy Smok, którego zabicie daje danej drużynie wzmocnione nieuchronne obrażenia i możliwość natychmiastowego zabijania przeciwników mających poniżej 20% ich maksymalnego życia[[50]](#footnote-50). W przypadku pozyskania zmiennych dotyczących smoków ponownie wykorzystano funkcję **matchTemplate**. W przypadku smoków żywiołów zlicza ona również ich ilość.

1. Zmienne dotyczące różnicy w złocie na danej pozycji

Różnice w złocie w podziale na pozycje byłyby zmiennymi, które w ciekawy sposób uzupełniałyby analizę wpływu różnic majętności drużyn na przebieg rozgrywki. Jednak problemy z pozyskaniem tych danych (które dokładniej zostaną omówione w dalszej części pracy) oraz fakt, że są one wynikowymi pozostałych cech (zarówno graczy, jak i drużyn) sugeruje możliwość ich pominięcia w analizie. Z tego powodu te konkretne informacje nie zostały wykorzystane jako zmienne.

1. Zmienna celu **blue\_team\_win**

Z powodu niemożliwości pobrania danych do zmiennej zależnej bezpośrednio z nagrań meczowych, zostały one dodane na podstawie wyników zaczerpniętych ze strony Wikipedii League of Legends[[51]](#footnote-51).

## **3.3 Problemy ze zmiennymi**

W procesie pozyskiwania bazy danych wystąpiło wiele problemów. Można je podzielić na dwie główne grupy: problemy z transmisją oraz z pozyskiwaniem konkretnych zmiennych. Warto zaznaczyć, że problemy z transmisją przekładały się również na drugą grupę Pierwszym oczywistym problemem w przypadku transmisji meczowych są wszelkiego rodzaju animacje mające urozmaicać widzom rozgrywki. Pojawienie się tego rodzaju zmian w emisji powoduje niemożliwość pozyskania części zmiennych. Rysunek 7 prezentuje przykładową klatkę z meczu, w której powtórka uniemożliwiła odczytanie niektórych informacji.

Rys. 7. Powtórka meczowa.

Obraz zawierający Gra komputerowa, Strategiczna gra wideo, Oprogramowanie gier wideo, zrzut ekranu

Opis wygenerowany automatycznie

Źródło: Opracowanie własne

Jak można zauważyć na rysunku 7 przez zmianę scenerii transmisji tracimy informacje np. o liczbie stworów lub KDA. Kolejnym dość znaczącym problemem transmisji są wszelkiego rodzaju błędy interfejsu. Wielokrotnie zdarzają się nieprawidłowości w układu graficznego. W wyniku tego realizatorzy zmuszeni są podmieniać schemat transmisji na jeden z zapasowych. Rysunek 8 prezentuje porównanie klatek z różnymi układami.

Rys. 8. Porównanie klatek z różnymi układami graficznymi.



Źródło: Opracowanie własne (po lewej klatka z układem zapasowym, po prawej z głównym)

Różnice w układach graficznych są bardzo problematycznym aspektem pozyskiwania danych z klatek meczowych, ponieważ funkcje odwołują się do konkretnych pikseli, a zmiany w ich rozmieszczeniu skutkują brakami danych. Stworzenie bazy danych było trudnym zadaniem również z powodu problemów z pozyskaniem konkretnych zmiennych.

* Zmienna **time**

W przypadku tej zmiennej dużym wyzwaniem było rozpoznanie znaku „:”. Niemożliwym było dobranie takiej wartości powiększenia obrazu, aby w każdym z przypadków czas meczu był odczytany poprawnie. Po licznych testach mających na celu znalezienie odpowiednich wartości do skalowania zdecydowano na przygotowanie funkcji, która w pierwszym kroku ładuje obraz za pomocą funkcji **imread** z biblioteki cv2. Następnie obraz jest przycinany do kluczowego fragmentu, a jego wymiary powiększane o 160%. Tak przygotowana klatka jest konwertowana na skalę szarości i wyostrzana, po czym odczytywana jest wartość czasu. Jeśli przy próbie konwersji otrzymanego tekstu na format %M:%S funkcja zwraca błąd (np. w przypadku błędnego odczytania dwukropka) wykonywane są kroki z klauzuli except. Następuje wtedy konwersja klatki wejściowej na format obrazu Image z biblioteki PIL. Następne kroki są zbliżone do tych z pierwszej części klauzuli, jednak w tym przypadku wymiary obrazu powiększane są o 600%. Tym sposobem udało się zmniejszyć ilość braków danych tej zmiennej zaledwie do około 7% (wliczając również klatki, których dotykają omówione na początku tego podrozdziału problemy z transmisją).

* Zmienna **gold**

Przy pozyskiwaniu danych do tej zmiennej największe problemy ponownie stwarzał znak interpunkcyjny w postaci kropki, która oddziela wartości dziesiętne tysięcy sztuk złota gromadzonych przez daną drużynę. Jednak w tym przypadku, również litera „k” (odpowiednik tysięcy) generowała dodatkowe komplikacje, ponieważ funkcja odpowiedzialna za pozyskiwanie danych do tej zmiennej, po otrzymaniu tekstu odczytanego z obrazu, ucina ostatni znak (czyli właśnie tą literę) i otrzymaną wartość zmiennoprzecinkową mnoży razy tysiąc w celu otrzymania ostatecznej wartości liczbowej złota. Jeśli litera „k” nie zostanie w ogóle odczytana to otrzymana wartość będzie różna od tej rzeczywistej. Podobnie wygląda to w przypadku kropki. Jeśli dana drużyna będzie miała 22.1k złota, a znak kropki nie zostanie rozpoznany to otrzymana wartość będzie wynosiła 221 tysięcy. Z tego powodu zdecydowano się na ustalenie warunku sprawdzającego czy dana wartość nie przekracza 80 tysięcy (rzadko kiedy ma miejsce sytuacja, w której drużyna zdobędzie więcej złota). W przypadku przekroczenia tego poziomu ilość złota dzielona jest przez 10. Aby zmaksymalizować niezawodność funkcji pozyskującej tą zmienną zastosowano dwie wartości skalowania obrazu. W pierwszym podejściu wymiary zwiększane są o 150%, natomiast gdy ta wartość nie daje oczekiwanych wyników obraz wejściowy powiększany jest sześciokrotnie. Dodatkowo w przypadku liczenia różnicy w złocie między drużynami wprowadzono warunek sprawdzający czy wartość bezwzględna tej różnicy nie jest większa od 20 tysięcy- jest to spowodowane tym, że rzadko w profesjonalnych rozgrywkach ta różnica jest tak duża, więc jeśli ta sytuacja ma miejsce, to najprawdopodobniej wynika z faktu błędnego odczytu wartości złota jednej z drużyn.

* Zmienna **towers**

W przypadku tej zmiennej głównym problemem była czcionka wykorzystywana w transmisji meczu. Wielokrotnie cyfry oznaczające ilość zniszczonych struktur były odczytywane niepoprawnie (np. „0” było odczytywane jako „O”, więc w momencie dodania listy dopuszczalnych znaków otrzymywano wartość None). W celu poprawy działania tej funkcji zastosowano podobny sposób działania jak w przypadku zmiennej **time**, jednak przyjęte poziomy powiększenia były inne. Przy pierwszej próbie pozyskania zmiennej wymiary fragmentu klatki powiększane są o 120%, a w próbie drugiej 340%. Zdecydowano się na takie wartości powiększenia, ponieważ dawały one najlepsze wyniki podczas testów. Zdecydowano się również zastosować warunek dotyczący różnicy w ilości struktur omówiony w poprzednim podrozdziale.

* Zmienna **kills**

Podczas pozyskiwania wartości do zmiennej **kills** napotkano podobne problemy, jak w przypadku zmiennej **towers**. Jednak w tym przypadku przyjęte poziomy skalowania obrazu wynosiły odpowiednio 250% i 120%. Dodano również warunek sprawdzający, czy różnica w zabójstwach między drużynami nie przekracza 14, ponieważ takie różnice są nietypowe w meczach profesjonalnych drużyn.

* Zmienne **blue\_team** oraz **red\_team**

W przypadku zmiennych dotyczących nazwy grających drużyn główny problem wynikał ze specyfiki układu graficznego transmisji- w momencie, w którym dana drużyna zgładziła Barona Nashora (lub Starszego Smoka) informacja o tym, która drużyna ma jego wzmocnienie pojawia się w miejscu, gdzie normalnie znajduje się nazwa drużyny. Sprawia to, że niemożliwym staje się odczytanie tych zmiennych. Rysunek 9 przedstawia omawianą sytuację.

Rys. 9. Klatka prezentująca zmianę układu graficznego spowodowaną zgładzeniem Barona.

Obraz zawierający Gra komputerowa, Strategiczna gra wideo, zrzut ekranu, Oprogramowanie gier wideo

Opis wygenerowany automatycznie

Źródło: Opracowanie własne (Barona Nashora zgładziła drużyna czerwona)

* Zmienne: **kda\_top, kda\_jungle, kda\_mid, kda\_adc, kda\_supp**

Wszystkie zmienne dotyczące KDA cechowały się podobnymi trudnościami w pozyskaniu danych. Pierwszym problemem jest znak „/”. W przypadku gdy chociaż jeden z nich zostanie nieprawidłowo odczytany cała wartość wyliczanego współczynnika jest nieprawidłowa bądź brakująca. Do funkcji tworzących te zmienne dodano białą listę znaków, która częściowo radzi sobie z tym problemem, jednak utrudnienie w postaci czcionki sprawia, że duża część rekordów posiada w tych zmiennych braki danych. W celu odczytania danych dotyczących KDA wykorzystywana jest funkcja posiadająca podobne kroki do formuły pozyskującej czas. Zastosowano również warunek ograniczający ilość śmierci do 20, ponieważ często zwraca ich ilość była większa. W pierwszej próbie wykorzystano skalowanie o 120%, natomiast w następnym o 340%. Kolejnym dużym problemem jest tak zwany współczynnik „bounty”[[52]](#footnote-52). Pojawia się on w momencie, gdy dany gracz wyróżnia się na tle innych. Jest on problematyczny, ponieważ wyświetlany jest tuż ponad statystykami i przy maksymalnym możliwym przycięciu dalej jest w stanie zaburzyć odczyt. Jego czcionka jest również różna od czcionki zabójstw, asyst i zgonów. Rysunek 10 prezentuje wygląd współczynnika „bounty”.

Rys. 10. Współczynnik „bounty”.



Źródło: Opracowanie własne.

* Zmienne **minions\_top, minions\_jungle, minions\_mid, minions\_adc, minions\_supp**

W przypadku tych zmiennych głównym problemem był sam odczyt cyfr, szczególnie w przypadku gdy zmieniał się schemat transmisji. Zdecydowano się jednak wprowadzić ograniczenia na ilość stworów w zależności od pozycji (wiadomym jest, że wspierający zgromadzi mniej stworów od strzelca). Tabela 4 prezentuje zestawienie pozycji z nałożonymi restrykcjami.

Tabela 4. Ograniczenia nałożone na ilość stworów w zależności od pozycji.

|  |  |  |
| --- | --- | --- |
| **Pozycja** | **Ograniczenie na ilość stworów** | **Ograniczenie na wartość bezwzględną różnicy** |
| Górna Aleja | 5 < stwory < 400 | Najpierw 80, później 1000 |
| Dżungla | 5 < stwory < 400 | Najpierw 80, później 1000 |
| Środkowa Aleja | stwory < 400 | Najpierw 90, później 1000 |
| Strzelec | stwory < 400 | 90 |
| Wsparcie | stwory < 100 | 50 |

Źródło: Opracowanie własne.

Podobna różnorodność występuje również w wartości skalowania obrazu. Wartości wahają się od 120 do 600 w zależności od pozycji i klauzuli „try”.

* Zmienne **hextech\_dragon, infernal\_dragon, mountain\_dragon, ocean\_dragon, wind\_dragon**

W przypadku zmiennych dotyczących smoków głównym problemem był brak dostępnych obrazków wzorcowych. Aby móc zliczyć ilość wystąpień konkretnych elementów reprezentujących smoki należało pozyskać z klatki odpowiedni symbol, następnie obrobić go w programie graficznym GIMP. Tak przygotowane wzorce wyszukiwane są na odpowiednio przyciętej klatce meczowej. Jedyną zmienną wartością w przypadku tej funkcji był próg odcięcia, który był różny w zależności od rodzaju smoka i wahał się od 0,85 do 0,87. Następnie zliczano różnicę w ilości smoków między drużynami dla każdego smoka.

* Zmienna **baron**

Głównym problemem w pozyskaniu danych mówiących o tym, czy dana drużyna zgładziła Barona Nashora było wybranie odpowiedniej strategii ich pobierania. Gdy któraś z drużyn go pokona wyświetlany jest odpowiedni komunikat a wszyscy (żywi w momencie śmierci tego stwora) zawodnicy otrzymują świecącą na fioletowo poświatę. W pierwszej koncepcji pozyskania tej zmiennej próbowano pobrać informację bezpośrednio z komunikatu, jednak, aby móc to zrobić należy posiadać odpowiednią klatkę meczową- powiadomienie to jest widoczne jedynie przez krótką chwilę. Z uwagi na fakt pobierania klatki średnio co minutę, trafienie na odpowiedni moment było wręcz niemożliwe, dlatego koncepcja ta nie miała racji bytu. Kolejnym pomysłem na poradzenie sobie z tym problemem było wykorzystanie wyszukiwania elementu na obrazie do znalezienia wspomnianych fioletowych poświat wokół postaci. Niestety również i ten sposób nie był możliwy do zrealizowania. Powodem takiego stanu rzeczy była trudność z dobraniem progu odcięcia, tak aby wykrywał tylko iluminację symbolizującą wzmocnienie Barona Nashora. Ostatecznie zdecydowano się na zastosowanie odczytu tej informacji z górnego paska transmisji (rysunek 9). Wartym dodania jest, że metoda ta wymagała stworzenia dwóch wzorów do wyszukiwania elementów na obrazie, ponieważ różnią się one między sobą co generuje problemy z ustaleniem progu odcięcia (wzorzec dla drużyny czerwonej delikatnie różni się od wzorca dla drużyny niebieskiej). Zmienna ta przyjmuje wartości: -1 (gdy czerwona drużyna zrobi Barona), 1 (gdy niebieska drużyna zrobi Barona) i 0 (gdy żadna z drużyn nie zabiła Barona).

* Zmienna **elder**

Jest to zmienna mówiąca o tym, czy któraś z drużyn zabiła Starszego Smoka. W tym przypadku zastosowano to samo podejście co w przypadku Barona Nashora. Ponownie przygotowano dwa wzorce do rozpoznawania elementów na obrazie. Możliwe wartości tej zmiennej są takie same jak w przypadku zmiennej **baron**.

* Zmienna **herald**

Jednym z problemów przy pozyskaniu informacji dotyczącej zrobienia Rift Heralda był wskaźnik czasu pozostałego na użycie Oka Herolda (jak wspomniano w podrozdziale 3.2, właśnie detekcja tego przedmiotu została wykorzystany do stworzenia omawianej zmiennej). Gracz, który je podniesie ma 4 minuty na przyzwanie kopii w celu zniszczenia struktur, po tym czasie Oko znika. Wygląd Oka Herolda zmienia się w zależności od upływającego czasu. Rysunek 11 obrazuje tą sytuację.

Rys. 11. Różna animacja Oka Herolda w zależności od pozostałego na jego użycie czasu.



Źródło: Opracowanie własne.

Biała, poruszająca się linia wskazuje właśnie na pozostały na użycie czas. Dodatkowo utrudnia ona rozpoznanie szablonu Oka Heralda na obrazie. Jednak w przypadku tej zmiennej największym problemem nie było samo pozyskanie informacji o jego zgładzeniu, ale zliczenie ilości zrobionych Heraldów, ponieważ stwór ten w każdym meczu może być zgładzony dwukrotnie. Może więc mieć miejsce sytuacja, gdy jedna drużyna zabije tego stwora dwa razy lub każda z drużyn po razie, ale również, gdy żadna nie pokona go przez cały mecz. W celu zliczenia ilości jego wykonania wykorzystano informację o czasie na użycie Oka Herolda- w momencie, gdy od ostatniego rekordu posiadającego informację o zgładzeniu Rift Heralda minęło więcej niż 4 minuty, a informacja ta znowu się pojawi to ilość Heraldów zwiększana jest o 1 (dzieje się tak w przypadku, gdy zgładzi go niebieska drużyna) lub zmniejszana o 1 (gdy Heralda zgładzi drużyna czerwona).

* Zmienne: **champion\_red\_top**, **champion\_red\_jungle**, **champion\_red\_mid**, **champion\_red\_adc**, **champion\_red\_supp**, **champion\_blue\_top**, **champion\_blue\_jungle**, **champion\_blue\_mid**, **champion\_blue\_adc**, **champion\_blue\_supp**

Pozyskanie zmiennych dotyczących postaci było zadaniem trudnym. W pierwszej kolejności należało przygotować wzory oryginalnych ikon postaci w celu ich późniejszego rozpoznania. Do tego celu wykorzystano ekstrakcje danych ze strony internetowej Wikipedii League of Legends[[53]](#footnote-53). Następnie wykorzystując własną wiedzę usunięto te ikony, które nie są już wykorzystywane w grze (niektóre z nich uległy wizualnej zmianie na przestrzeni lat). W kolejnym kroku przygotowano model, w którym znajdowało się zestawienie nazwy postaci oraz zapisanej w skali szarości kopii obrazu. W ostatnim kroku stworzono funkcję, która porównuje ikony pozyskane z klatki meczowej ze wzorcowymi. Ponadto, aby zwiększyć skuteczność jej skuteczność zdefiniowano na podstawie własnej wiedzy listy postaci występujących na konkretnych pozycjach (w celu uniknięcia sytuacji, w której postać wspierająca zostanie rozpoznana na górnej alei). Główną wykorzystaną miarą porównawczą był błąd średniokwadratowy. Podczas testowania funkcji rozpoznającej bohaterów zauważono, że wykorzystanie jednej miary błędu nie daje najlepszych rezultatów. Postanowiono więc użyć również wskaźnika podobieństwa strukturalnego. Po jego dodaniu osiągane rezultaty testowe były dużo lepsze, zaobserwowano jednak pewne tendencje funkcji do mylenia konkretnych postaci. Aby temu zaradzić, na podstawie bazy testowej utworzono warunki, mające na celu zmaksymalizowanie poprawności działania funkcji. W dalszym ciągu funkcja ta nie osiąga stuprocentowej trafności, jednak największe błędy poprawiono ręcznie. Oprócz problemów bezpośrednio z transmisją oraz zmiennymi użytymi w bazie wartym wspomnienia są również problemy tych predyktorów, które nie zostały zawarte w końcowej bazie. W podrozdziale 3.2 wspomniano o dwóch informacjach, które przez trudności w implementacji nie zostały w bazie zawarte.

* Zmienne dotyczące wizji

Głównym problemem w ich pozyskaniu (tak jak w przypadku Rift Heralda) był wskaźnik czasu odnowienia. Animacja ruchu linii czasu odnowienia uniemożliwiała skuteczne odczytanie cyfr obrazujących ilość postawionych totemów wizji. Pomimo licznych prób ustawienia odpowiednich wartości parametrów, nie udało się osiągnąć satysfakcjonujących rezultatów, co może być spowodowane faktem, że wskaźnik czasu i vision score mają ten sam kolor. Z tego powodu zrezygnowano z wykorzystania tych zmiennych.

* Zmienne dotyczące różnicy w złocie na danej pozycji

W przypadku tych zmiennych głównym problemem było przyporządkowanie odczytanych różnic w złocie do konkretnej strony w spotkaniu. Wykorzystanie skali szarości, która zwiększa skuteczność rozpoznawania informacji o ilości złota, praktycznie uniemożliwiło wskazanie, który z zespołów ma przewagę na danej pozycji. Trudności w pozyskaniu tych danych i wspomniana wcześniej korelacja z innymi zmiennymi spowodowały decyzję o rezygnacji z wykorzystania tych zmiennych.

# IV. Analiza danych i modelowanie

## **4.1 Przygotowanie końcowej bazy danych**

Pozyskana z nagrań meczowych baza nie jest pozbawiona braków danych oraz innych problemów. Przed przystąpieniem do budowania modeli należy przygotować odpowiednio końcową ramkę danych. Pierwszym wykonanym krokiem było usunięcie rekordów, które mają braki danych w ponad 50% zmiennych. W tym celu stworzono odpowiednią funkcję. Okazało się jednak, że nie ma żadnego rekordu, który posiada mniej niż połowę cech. Następnie przygotowano zestawienie procentowej ilości braków danych w poszczególnych zmiennych. Wyniki tej analizy przedstawiono w tabeli 5.

Tabela 5. Początkowe ilości braków danych w poszczególnych zmiennych.

|  |  |
| --- | --- |
| **Zmienna** | **Procent braków danych** |
| time | 7,11% |
| gold | 9,57% |
| towers | 11,92% |
| kills | 6,94% |
| blue\_team | 15,72% |
| red\_team | 15,9% |
| kda\_top | 12,2% |
| kda\_jungle | 30,49% |
| kda\_mid | 19,07% |
| kda\_adc | 10,92% |
| kda\_supp | 16,08% |
| minions\_top | 9,53% |
| minions\_jungle | 9,36% |
| minions\_mid | 6,58% |
| minions\_adc | 9,28% |
| minions\_supp | 15,8% |
| hextech\_dragon | 0% |
| infernal\_dragon | 0% |
| mountain\_dragon | 0% |
| ocean\_dragon | 0% |
| wind\_dragon | 0% |
| baron | 0% |
| elder | 0% |
| herald | 0% |
| champion\_red\_top | 0% |
| champion\_red\_jungle | 0% |
| champion\_red\_mid | 0% |
| champion\_red\_adc | 0% |
| champion\_red\_supp | 0% |
| champion\_blue\_top | 0% |
| champion\_blue\_jungle | 0% |
| champion\_blue\_mid | 0% |
| champion\_blue\_adc | 0% |
| champion\_blue\_supp | 0% |

Źródło: Opracowanie własne.

Jak można odczytać z tabeli 5 ponad połowa pozyskanych zmiennych nie posiada żadnych braków danych, jednak niedobory informacji obserwowane w przypadku pozostałych mogą stanowić poważny problem w dalszej analizie. Głównymi zmiennymi wykorzystywanymi do analizy, w których występowały braki danych, są zmienne ilościowe obrazujące różnice poszczególnych drużyn i graczy w konkretnych statystykach. W tym przypadku dane wydają się stosunkowo łatwe do imputacji, ponieważ są to statystyki, które nie zmieniają się bardzo gwałtownie w czasie. Zdecydowano się zastosować imputację danych polegającą na uzupełnieniu wartości na podstawie rekordów sąsiadujących. W tym celu wykorzystano metodę interpolacji najbliższego sąsiada. Aby móc to wykonać należało wcześniej odpowiednio przygotować bazę. Kluczowymi zmiennymi służącymi uporządkowaniu bazy danych są: **time**, **blue\_team** i **red\_team**. Pomogą one ustalić z której minuty danego meczu pochodzi rekord charakteryzujący się brakami danych, dlatego na początku wybrano rekordy, w których nie ma braków danych dla tych 3 zmiennych. Tych rekordów było 2005. Następnie wykorzystano metodę **sort\_values** celem posortowania ramki danych kolejno po zmiennych **blue\_tem, red\_team** i **time**. Pozwoliło to na podział na mecze i uszeregowanie względem czasu jego trwania. Po zastosowaniu metody interpolacji najbliższego sąsiada braki danych zostały uzupełnione w 1998 rekordach. Pozostałe 7 rekordów postanowiono imputować ręcznie. Aby tego dokonać podzielono bazę danych (zawierającą uzupełnione zmienne **time**, **blue\_team** i **red\_team**) na bazę z brakami oraz zbiór ich pozbawiony. W przypadku pierwszego z rekordów posiadających braki sprawa była dość prosta, ponieważ brakujące dane dotyczą ilości pokonanych stworów na alejach, a czas odczytany z klatki wskazuj 47 sekundę meczu. Z uwagi na fakt, że pojawiają się one dopiero po upływie minuty i pięciu sekund od rozpoczęcia spotkania[[54]](#footnote-54), ich liczba w tym przypadku jest z pewnością równa zero. Następne rekordy wymagają jednak dokładniejszego sprawdzenia. W przypadku klatki z 1 minuty i 37 sekundy meczu Astralis kontra Excel Esports ponownie ilość stworów w brakujących zmiennych jest równa 0, ponadto zauważono również błędne odczytanie danych dla dżungli, co również skorygowano. Kolejnym meczem wziętym do analizy był mecz Astralis przeciwko Team BDS. W tym przypadku brakującą zmienną była ta informująca o różnicy w KDA na środkowej alei. Z tego powodu znaleziono odpowiadającą rekordowi klatkę i obliczono poprawną wartość tej zmiennej. Podobnie sytuacja wyglądała w przypadku pozostałych brakujących danych. Wszystkie dotyczyły braków statystyk KDA na różnych pozycjach, więc odnaleziono odpowiadające klatki i obliczono brakujące wartości. Na samym końcu połączono ze sobą obie ramki danych i uzyskano ramkę bez braków danych dla wszystkich rekordów posiadających zmienne **time**, **blue\_team** i **red\_team**. W następnym kroku utworzono bazę, która zawiera rekordy, w których brakuje którejś z trzech zmiennych wykorzystanych w poprzednim kroku. W tym przypadku sprawa jest bardziej skomplikowana, ponieważ bez kompletu informacji o czasie i drużynach bardzo ciężko jest powiązać odpowiedni rekord z konkretną klatką. W celu imputacji danych wydzielono rekordy, w których brakuje zmiennej **time** ale obie drużyny są uzupełnione. Następnie prześledzono konkretne mecze w celu uzupełnienia informacji o czasie. Można było to wykonać przeglądając luki czasowe w spotkaniach. Tym sposobem uzupełniono ręcznie brakujące rekordy zmiennej **time** części rekordów. Następnie sprawdzono ilość rekordów, dla których nie mamy informacji o minucie meczu i nie posiadamy nazwy którejś z drużyn. Klatek, w których brakowało informacji o czasie i jednej z drużyn było 13. W celu dopasowania rekordów do konkretnego spotkania postanowiono wykorzystać posiadaną nazwę drużyny oraz postacie na konkretnych pozycjach. Zadanie to było bardzo czasochłonne z powodu zdarzających się błędów funkcji rozpoznawania postaci, jednak ostatecznie udało się wykonać tą imputację. W kolejnym kroku zbadano rekordy, które posiadają informację o czasie, ale nie posiadają którejś z nazw drużyn. W tym celu rozpatrzono dwa przypadki- gdy posiadamy informację od drużynie niebieskiej oraz gdy posiadamy nazwę zespołu grającego po czerwonej stronie (informacje o czasie posiadane były w obu przypadkach). Na podstawie czasu oraz innych statystyk uzupełniono braki danych o nazwach drużyn w obu tych przypadkach. Najtrudniejszymi klatkami w analizie były te, które nie posiadały informację o czasie, ale nie posiadały nazw drużyn oraz te, dla których brakowało wszystkich trzech zmiennych (**time**, **blue\_team** i **red\_team**). W tym przypadku postanowiono ponownie wykorzystać informacje o postaciach i inne statystyki. Niestety dla 47 (bez drużyn i czasu) oraz 32 (bez drużyn) rekordów nie udało się dopasować klatki do odpowiedniego meczu, a tym samym pozyskać informacji o drużynach i czasie. Zdecydowano się na usunięcie tych wierszy. Kolejnym krokiem w przygotowaniu końcowej bazy było uzupełnienie danych dla rekordów, które na samym początku nie posiadały kompletu informacji o drużynach oraz czasie. W tym celu połączono ze są ramkę posiadającą komplet danych oraz tą, w której braki dalej występują i posortowano rekordy po nazwach zespołów oraz czasie. Ponownie zdecydowano się na zastosowanie interpolacji najbliższego sąsiada, ponieważ kolumny z niepełną informacją pokrywały się z tymi, na których użyto jej ostatnio. Tym sposobem otrzymano bazę bez braków danych, zawierającą 2732 rekordy. Podczas imputacji danych o drużynach i czasie zauważono dość duże nieścisłości między rozpoznanymi postaciami oraz tymi faktycznie występującymi. Ponadto występowały przypadki, gdzie dane postacie różniły się w obrębie jednego meczu. Aby ujednolicić bohaterów w każdym spotkaniu i osiągnąć jak największą poprawność rekordów postanowiono poprawić te zmienne za pomocą zewnętrznych źródeł. W tym celu posłużono się Wikipedią League of Legends[[55]](#footnote-55), która zawiera informacje o postaciach wybieranych w konkretnych spotkaniach. Dla każdego meczu rozpisano zestaw postaci dla każdej drużyny oraz każdej pozycji. Kolejnym krokiem w przygotowaniu końcowej bazy danych było utworzenie dodatkowych zmiennych. Jak wspomniano w podrozdziale 3.2, nie ma możliwości pozyskania wartości zmiennej celu bezpośrednio z nagrań meczowych. Dlatego wykorzystując historyczne wyniki podane na stronie Wikipedii League of Legends utworzono zmienną celu **blue\_team\_win**, której wartości uzupełniono na podstawie nazw drużyn. Dzięki poprawie jakości danych o postaciach można było również przeprowadzić analizę konkretnych synergii bohaterów. Na podstawie własnej wiedzy oraz słów Marcina Świecha zdecydowano się wprowadzić zmienne dotyczące czterech kombinacji, z których wszystkie dotyczyły pozycji strzelca oraz wspierającego: Lucian z Nami, Kog’Maw z Lulu, Zeri z Yummi oraz Senna z Tahm Kench-em. W przypadku tego ostatniego zestawienia należało rozpatrzyć dwa przypadki: Senna jako strzelec oraz Senna jako wspierający, ponieważ jej umiejętność pasywna zwiększa jej siłę gdy nie zdobywa ona złota z zabijanych stworów. Następne dodane zmienne zawierały informacje o wspomnianych wcześniej smoczych duszach, jednak zadanie to było utrudnione z powodu logiki zastosowanej w funkcji pobierającej dane o smokach. Funkcja ta pobiera różnice w liczbie zrobionych smoków zamiast faktycznej ich ilości. Uniemożliwia to zliczenie sumy bestii zrobionych przez daną drużynę. Z tego powodu zdecydowano się sprawdzić klatki z meczy, w których wartość bezwzględna różnicy w liczbie któregoś ze smoków była większa bądź równa dwa. Tym sposobem uzyskano informacje o tym czy którakolwiek z drużyn była w posiadaniu duszy. Podczas pracy nad zmiennymi dotyczącymi smoczych dusz zauważono pojedyncze przypadki błędnego rozpoznania ilości pokonanych obiektów neutralnych (Baron-ów Nashor-ów, smoków, Rift Herald-ów oraz Starszych Smoków) - funkcja rozpoznała, że dana drużyna zdobyła obiekt przed czasem jego pojawienia się. W tym przypadku zdecydowano się na założenie warunków na czas pojawienia się danego obiektu. Wykorzystano do tego dane z artykułu ze strony esports.com[[56]](#footnote-56). Zgodnie z tym artykułem pierwszy smok pojawia się w piątej minucie każdego spotkania. Dlatego wszystkie rekordy obejmujące klatki sprzed piątej minuty mają ustawioną wartość zmiennych dotyczących smoków na 0 (żadna z drużyn nie miała możliwości zgładzić tej bestii wcześniej, niż zakłada warunek). Z uwagi na fakt, że do duszy potrzeba zabić aż cztery te potwory, a czas ich odnowienia wynosi pięć minut to minimalny czas, gdy któraś z drużyn może zdobyć duszę przypada na 20 minutę. Z tego powodu wartości zmiennych dotyczących smoczej duszy, dla rekordów odnoszących się do klatek sprzed 20-tej minuty zostały ustawione na 0. Podobnie sytuacja wyglądała w przypadku zmiennej **baron**. Ten obiekt neutralny również pojawia się po raz pierwszy w 20-tej minucie każdego spotkania, dlatego poczynione kroki były identyczne jak te w przypadku duszy. Sytuacja w kwestii zmiennej **herald** była zdecydowanie bardziej skomplikowana. Pierwszy Rift Herald pojawia się w ósmej minucie oraz znika, jeśli nie zostanie zrobiony do 20-tej minuty, dlatego też w tym przypadku wprowadzono te dwa warunki. Jednak, gdy dana drużyna zrobi Rift Herald-a to pojawia się on ponownie po upływie 6-ciu minut, co sprawia, że każda z drużyn ma możliwość zrobienia maksymalnie dwóch tych obiektów w ciągu meczu. Stworzono więc funkcję, która dla danej kombinacji nazw drużyn sprawdza, czy od pierwszej klatki, w której Herald został zgładzony do następnej, posiadającej taką samą informację minęło 6 minut. Jeśli tak jest to liczba Rift Herald-ów zostaje zwiększona bądź zmniejszona o 1 (zmienna **herald** jest różnicą między Heraldami drużyny niebieskiej i czerwonej, więc jest zwiększana, gdy to niebieska drużyna zrobi Heralda i zmniejszana, gdy dokona tego drużyna czerwona). W ten sposób otrzymano końcową bazę zawierającą 2732 rekordy bez braków danych oraz 48 kolumn. Ta ramka danych posłuży dalszej analizie, modelowaniu oraz dokonaniu predykcji.

## **4.2. Analiza danych**

Po stworzeniu bazy, która będzie wykorzystana do modelowania należy przeanalizować posiadane dane. W pierwszym kroku postanowiono przeanalizować podsumowanie podstawowych statystyk opisowych zmiennych ilościowych (razem ze zmienną celu) z wykorzystaniem funkcji **describe**. Tabela 6 prezentuje otrzymane rezultaty.

Tabela 6. Statystyki opisowe zmiennych ilościowych.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| **gold** | 2732 | 580,6435 | 4760,406 | -19000 | -1800 | 100 | 2900 | 18700 |
| **towers** | 2732 | 0,258053 | 1,601797 | -7 | 0 | 0 | 0 | 8 |
| **kills** | 2732 | 0,04246 | 4,669097 | -114 | -1 | 0 | 2 | 14 |
| **kda\_top** | 2732 | 0,260702 | 4,571044 | -89 | 0 | 0 | 1 | 70 |
| **kda\_jungle** | 2732 | -0,27606 | 7,369271 | -92 | -1 | 0 | 1,833333 | 90 |
| **kda\_mid** | 2732 | 0,29729 | 8,313671 | -92,6667 | -1 | 0 | 1 | 93,66667 |
| **kda\_adc** | 2732 | 0,45329 | 7,460995 | -88 | -1 | 0 | 1,5 | 90 |
| **kda\_supp** | 2732 | 0,450774 | 6,59734 | -87 | -1 | 0 | 2 | 71 |
| **minions\_top** | 2732 | 4,782943 | 53,05788 | -292 | -11 | 2 | 15 | 986 |
| **minions\_jungle** | 2732 | 3,436676 | 99,54886 | -3982 | -11,25 | 0 | 11 | 954 |
| **minions\_mid** | 2732 | 4,415081 | 51,71126 | -365 | -10 | 1 | 13 | 970 |
| **minions\_adc** | 2732 | -4,13836 | 107,388 | -538 | -18 | -1 | 11 | 2563 |
| **minions\_supp** | 2732 | 8,75 | 70,27379 | -243 | -4 | 0 | 9 | 1612 |
| **hextech\_dragon** | 2732 | 0,006589 | 0,171026 | -1 | 0 | 0 | 0 | 1 |
| **infernal\_dragon** | 2732 | 0,007687 | 0,222201 | -3 | 0 | 0 | 0 | 2 |
| **mountain\_dragon** | 2732 | -0,00549 | 0,312628 | -2 | 0 | 0 | 0 | 4 |
| **ocean\_dragon** | 2732 | -0,04136 | 0,306362 | -3 | 0 | 0 | 0 | 1 |
| **wind\_dragon** | 2732 | 0,010981 | 0,166457 | -1 | 0 | 0 | 0 | 3 |
| **baron** | 2732 | 0,22511 | 0,499929 | -1 | 0 | 0 | 1 | 1 |
| **elder** | 2732 | 0,036237 | 0,188863 | -1 | 0 | 0 | 0 | 1 |
| **herald** | 2732 | 0,06552 | 0,271476 | -1 | 0 | 0 | 0 | 1 |
| **blue\_team\_win** | 2732 | 0,675329 | 0,468337 | 0 | 0 | 1 | 1 | 1 |
| **zeri\_yummi\_blue** | 2732 | 0,012079 | 0,109259 | 0 | 0 | 0 | 0 | 1 |
| **zeri\_yummi\_red** | 2732 | 0,010615 | 0,102499 | 0 | 0 | 0 | 0 | 1 |
| **lucian\_nami\_blue** | 2732 | 0,065886 | 0,248128 | 0 | 0 | 0 | 0 | 1 |
| **lucian\_nami\_red** | 2732 | 0,028917 | 0,167603 | 0 | 0 | 0 | 0 | 1 |
| **senna\_tahm\_blue** | 2732 | 0,028551 | 0,16657 | 0 | 0 | 0 | 0 | 1 |
| **senna\_tahm\_red** | 2732 | 0,04429 | 0,205776 | 0 | 0 | 0 | 0 | 1 |
| **kog\_lulu\_blue** | 2732 | 0,008419 | 0,091383 | 0 | 0 | 0 | 0 | 1 |
| **kog\_lulu\_red** | 2732 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **hextech\_soul** | 2732 | -0,00549 | 0,131071 | -1 | 0 | 0 | 0 | 1 |
| **infernal\_soul** | 2732 | -0,00183 | 0,212215 | -1 | 0 | 0 | 0 | 1 |
| **mountain\_soul** | 2732 | 0,003294 | 0,128322 | -1 | 0 | 0 | 0 | 1 |
| **ocean\_soul** | 2732 | -0,00586 | 0,135181 | -1 | 0 | 0 | 0 | 1 |
| **wind\_soul** | 2732 | -0,03551 | 0,221142 | -1 | 0 | 0 | 0 | 1 |

Źródło: Opracowanie własne.

Pierwszą rzeczą, którą można odczytać z tabeli 6 jest fakt, że zgodnie z oczekiwaniami baza pozbawiona jest braków danych, co wskazuje na to, że imputacja przebiegła pomyślnie. W następnym kroku postanowiono dokładniej przyjrzeć się każdej zmiennej.

* Zmienna **gold**

W przypadku zmiennej obrazującej różnice w złocie pomiędzy drużynami można stwierdzić, że różnice w złocie w danej próbce mają szeroki zakres, od mocno ujemnych wartości (-19 000) do wysokich wartości dodatnich (18 700). Wartości te są możliwe, jednak mogą one również być obserwacjami odstającymi bądź błędami w danych. Średnia wartość sugeruje, że wartości zbliżone są do 580,6435, jednak duże odchylenie standardowe wskazuje na znaczną zmienność danych. Wartości kwartyli pokazuje, że większość obserwacji koncentruje się w przedziale od -1800 do 2900. Zarówno wartość średniej, jak i mediany mogą sugerować wyrównany pod względem złota poziom rozgrywek. Postanowiono również zbadać rozkład tej zmiennej. Rysunek 12 prezentuje odpowiedni histogram.

Rys. 12. Rozkład zmiennej gold.

Obraz zawierający piksel, sylwetka

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

Histogram prezentujący rozkład zmiennej **gold** nie wskazuje na żadne nieprawidłowości, a najczęściej pojawiające się wartości ponownie mogą sugerować, że rozgrywki w tym turnieju były wyrównane. Następnie zbadano wartości odstające z wykorzystaniem trzech metod: na podstawie odchylenia standardowego (przyjęto odległość trzech odchyleń standardowych), na podstawie procentyli (5% i 95%) oraz na podstawie z-score. Do weryfikacji wybrano rekordy, które zostały wskazane jako odstające przez wszystkie te metody. Ostatecznie zdecydowano się usunąć 14 wierszy, gdzie głównym kryterium decyzyjnym była wartość zmiennej celu (np. sytuacja, w której różnica w złocie wynosi 18 800 na korzyść drużyny czerwonej a wygrywa drużyna niebieska, ponieważ wygranie meczu z taką różnicą złota jest praktycznie niewykonalne).

* Zmienna **towers**

W przypadku zmiennej dotyczącej różnicy w zniszczonych strukturach możemy zauważyć, że w większości analizowanych danych różnice te są zerowe bądź marginalne, jednak zdarzają się przypadki, w których występuje znacząca ich przewaga dla którejś z drużyn. Ponownie postanowiono zbadać rozkład tej zmiennej.

Rys. 13. Rozkład zmiennej towers.

Obraz zawierający symbol, zrzut ekranu, piksel, sylwetka

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

Rysunek 13 potwierdza tezę o braku różnicy w ilości wież w większości przypadków. Może być to spowodowane specyfiką rozgrywki i sugerować, że mecze były dość wyrównane- gdy konkurują ze sobą drużyny na podobnym poziomie to różnica w strukturach rzadko kiedy jest duża. Ponownie zdecydowano się na sprawdzenie wartości odstających na podstawie wspomnianych trzech metod, jednak w tym przypadku, na podstawie danych wydaje się, że żadna obserwacja nie jest obserwacją odstającą, więc zdecydowano się nie usuwać żadnego rekordu.

* Zmienna **kills**

Główną rzeczą, która rzuca się w oczy w przypadku zmiennej **kills** jest bardzo niska wartość minimalna, już na tym etapie analizy można stwierdzić, że najprawdopodobniej jest to błąd w danych, ponieważ w profesjonalnych rozgrywkach zdobycie tylu zabójstw jest praktycznie niewykonalne, jedna postanowiono podjąć dalsze decyzje na podstawie rozkładu tej zmiennej oraz wartości odstających.

Rys. 14. Rozkład zmiennej kills.

Obraz zawierający sylwetka, piksel

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

Jak wynika z histogramu zaprezentowanego na rysunku 14 wartość zmiennej **kills** na poziomie -114 faktycznie wydaje się błędna. Na tym etapie zdecydowano się przyjąć założenie o usunięciu rekordów, dla których różnica w zabójstwach wynosi więcej niż 20 (zmienna przyjmuje wartości większe od 20 lub mniejsze od -20), jednak zbadano również wartości odstające, w celu upewnienia się o słuszności tej decyzji. Po przeglądzie obserwacji odstających postanowiono usunąć rekordy zgodnie z przyjętym założeniem, w efekcie czego z dalszej analizy wykluczono 3 wiersze.

* Zmienne dotyczące statystyki KDA

W przypadku zmiennych **kda\_top, kda\_jungle, kda\_mid, kda\_adc, kda\_supp** główną rzeczą rzucającą się w oczy są wartości minimalne i maksymalne, z tego powodu postanowiono w pierwszym kroku zbadać rozkłady tych zmiennych.

Rys. 15. Rozkłady zmiennych dotyczących KDA

Obraz zawierający zrzut ekranu, linia, Prostokąt, Wielobarwność

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

Z wykresów na rysunku 15 ponownie wynika, że różnice w wielkości współczynnika KDA w większości przypadków były równe zero. Na podstawie znajomości tej gry zdecydowano się wprowadzić progi akceptowalnej różnicy KDA dla wszystkich pozycji. W przypadku większości pozycji (wszystkich oprócz wspierającego) przyjęto, że różnica nie może być większa od 20, natomiast w przypadku wspierającego od 15. Jest to spowodowane faktem, że wspierający najczęściej oddają zasoby, takie jak zabójstwa, osobom z innych pozycji, a ich statystyka KDA opiera się głównie na asystach, przez co jej wartość zazwyczaj jest niższa w porównaniu do innych ról. Przyjęcie tego ograniczenia spowodowało, że z bazy danych usunięto 144 rekordy.

* Zmienne dotyczące stworów

Zgodnie ze statystykami z tabeli 6 zmienne **minions\_top, minions\_jungle, minions\_mid, minions\_adc, minions\_supp** mogą charakteryzować się podobnymi problemami, do tych występujących w przypadku różnicy w statystyce KDA. Ponownie zbadano więc ich rozkłady.

Rys. 16. Rozkłady zmiennych dotyczących KDA

Obraz zawierający zrzut ekranu, linia, Prostokąt, Wielobarwność

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

Jak można zauważyć na rysunku 16 różnice w stworach również w większości wynoszą zero, jednak przesunięcie wykresów może wskazywać na duże wartości bezwzględne obserwacji odstających. Ponownie zdecydowano o ustaleniu progu akceptowalnej różnicy. W tym przypadku zastosowano trzy progi: 120 stworów różnicy (dla górnej alei i dżungli), 100 stworów różnicy (dla środkowej alei i wspierającego) oraz 150 stworów różnicy (dla strzelców). Po wprowadzeniu tych progów z dalszej analizy wykluczono 198 rekordów.

* Zmienne dotyczące smoków

Na podstawie informacji zawartych w tabeli 6 nie dostrzeżono żadnych nieprawidłowości w zmiennych **hextech\_dragon, infernal\_dragon, mountain\_dragon, ocean\_dragon, wind\_dragon**, postanowiono jednak zbadać ich rozkłady w celu prześledzenia częstości występowania konkretnych wartości różnicy w smokach.

Rys. 17. Rozkłady zmiennych dotyczących smoków

Obraz zawierający zrzut ekranu, Prostokąt, linia, Wielobarwność

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

Jak można odczytać z rysunku 17 w większości przypadków różnice w poszczególnych smokach były równe 0, jednak sytuacja ta wydaje się normalna, ponieważ w rozgrywkach profesjonalnych walki skoncentrowane są wokół prób zdobycia obiektów neutralnych i rzadko ma miejsce sytuacja, gdy jakiś obiekt oddawany jest przeciwnikom bez walki. W przypadku zdecydowano o nieusuwaniu żadnych rekordów. Wartym zaznaczenia jest również fakt, że do zmiennych dotyczących smoków możemy zaliczyć również zmienne opisujące zdobycie duszy. Z uwagi na fakt, że są to zmienne przygotowane ręcznie to nie powinny występować w nich błędy. Postanowiono jednak sprawdzić ich rozkłady.

Rys. 18. Rozkłady zmiennych dotyczących zdobycia smoczej duszy.

Obraz zawierający zrzut ekranu, Prostokąt, kwadrat, linia

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

Z rysunku 18 można zauważyć, że w większości rekordów żadna z drużyn nie była w posiadaniu wzmocnienia duszy smoczej.

* Zmienne **baron**, **elder** i **herald**

W podrozdziale 4.1 omówiono częściowo pracę analityczną nad tymi zmiennymi, natomiast dane z tabeli 6 nie wskazują na jakiekolwiek nieprawidłowości, dlatego w tym przypadku zdecydowano o nieusuwaniu żadnych rekordów z powodu tych zmiennych. Ponownie jedynym podjętym krokiem było sprawdzenie ich rozkładów.

Rys. 19. Rozkłady zmiennych baron, herald i elder.

Obraz zawierający zrzut ekranu, Prostokąt, linia, Grafika

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

Z rysunku 19 można odczytać kilka interesujących informacji. Ponownie wykresy rozkładu wskazują, że różnice w obiektach neutralnych są w większości rekordów równe 0. Ponadto zauważono, że najczęściej robionym obiektem jest Baron Nashor, a najrzadziej Starszy Smok. Wydawać by się mogło, że obiektem, który zdobywany jest w największej ilości, powinien być Rift Herald (ponieważ pojawia się najwcześniej oraz może on zostać zdobyty dwukrotnie w każdym meczu), jednak pokonanie go traci na znaczeniu, gdy pierwsza struktura zostanie zniszczona przed jego zdobyciem (drużyna, która zniszczy pierwszą strukturę dostaje dodatkową ilość złota). Wartym zaznaczenia jest również fakt, że w analizowanym zbiorze danych nie została żadna obserwacja, dla której drużyna czerwona jest w posiadaniu starszego smoka. Jak wynika z tabeli 6 obserwacje takie występowały, jednak zostały usunięte w poprzednich krokach.

* Zmienne dotyczące kompozycji

W przypadku zmiennych binarnych określających występowanie danej kompozycji w drużynach można zauważyć, że w większości przypadków zestawienia te nie występowały, na co wskazują wartości średnich z tabeli 6. Nawet po zsumowaniu średnich dla obu drużyn (np. dla zmiennych **lucian\_nami\_blue** i **lucian\_nami\_red**) wartość tej sumy nie przekracza 0,5, co jasno pokazuje, że dana kompozycja występowała w mniej niż połowie meczy. Kolejnym ważnym wnioskiem możliwym do odczytania jest fakt, że najczęściej występującą kompozycją z tych badanych było połączenie strzelca Luciana ze wspierającą Nami, a najrzadziej Kog’Mawa z Lulu. Ponadto kombinacja ta nie została ani razu wybrana przez stronę czerwoną.

* Zmienna celu **blue\_team\_win**

Analizując dane dotyczące zmiennej celu z tabeli 6 można zauważyć dysproporcję w ilości meczy wygranych przez drużynę czerwoną i niebieską. Średnia na poziomie około 0,675 wskazuje, że w blisko 68% rekordów to drużyna niebieska sięga po wygraną. Postanowiono zbadać rozkład tej zmiennej w celu upewnienia się co do słuszności tezy o niezbilansowaniu zmiennej celu.

Rys. 20. Rozkład zmiennej blue\_team\_win.

Obraz zawierający Prostokąt, zrzut ekranu, kwadrat, design

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

Rysunek 20 jasno wskazuje na dużą nierówność między ilością zer i jedynek w zmiennej **blue\_team\_win**. Wykorzystując funkcję **value\_counts** uzyskano dokładne liczby rekordów z konkretnymi wartościami tej zmiennej. Okazało się, że w 1845 rekordach to drużyna niebieska sięgała po zwycięstwo, a czerwona zaledwie 887 razy. Tak duże niezbilansowanie wartości zmiennej celu może prowadzić do sytuacji, gdy model będzie zdecydowanie lepiej rozpoznawał tylko sytuację wygranej drużyny niebieskiej. Postanowiono więc zrównoważyć ilość 0 i 1 w zbiorze danych. Metod wykonania tego dostępnych było kilka, np.: oversampling, undersampling lub algorytmy takie jak SMOTE. W tym konkretnym przypadku zdecydowano posłużyć się oversamplingiem. Decyzja ta spowodowana była między innymi faktem stosunkowo małej ilości obserwacji w zbiorze. Po analizie statystyk opisowych ramki danych, weryfikacji poprawności rekordów i usunięciu obserwacji błędnych oraz odstających postanowiono kontynuować analizę zmiennych ilościowych. Następną rzeczą do zbadania były możliwe zbyt wysokie korelacje między zmiennymi. W przypadku gdy sytuacja taka ma miejsce należy odrzucić jedną ze skorelowanych, ponieważ pozostawienie jej może skutkować błędnymi predykcjami. Postanowiono stworzyć macierz korelacji, która obejmuje wszystkie zmienne ilościowe z wyłączeniem zmiennej **kog\_lulu\_red**, ponieważ nie było żadnej obserwacji, która wskazywała na występowanie tej kompozycji po stronie czerwonej. Rysunek 21 prezentuje mapę cieplną analizowanego zbioru danych.

Rys. 21. Mapa cieplna korelacji zmiennych ilościowych.

Obraz zawierający zrzut ekranu, Wielobarwność, sztuka, kwadrat

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

Duża ilość zmiennych użytych w analizie sprawiła, że uzyskana na rysunku 21 mapa cieplna nie jest do końca przejrzysta, jednak już na pierwszy rzut oka widać, że występują pary zmiennych, które należy dokładniej zbadać. Postanowiono więc ograniczyć macierz korelacji na podstawie jej wartości- przyjęto próg na poziomie 20% skorelowania, więc wartości korelacji większe od 0,2 bądź mniejsze od -0,2.

Rys. 22. Ograniczona mapa cieplna korelacji zmiennych ilościowych.

Obraz zawierający zrzut ekranu, linia, Wielobarwność, Prostokąt

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

Niestety mapa cieplna nawet po ograniczeniu wartości korelacji branych pod uwagę nie jest w stanie przedstawić czytelnej prezentacji badanych korelacji, dlatego w celu dokładniejszego zaprezentowania możliwych kłopotliwych korelacji między zmiennymi postanowiono stworzyć tabelę prezentującą wszystkie korelacje zaliczające się do przyjętego warunku.

Tabela 7. Korelacje zmiennych ilościowych.

|  |  |  |
| --- | --- | --- |
| **Zmienna1** | **Zmienna2** | **Korelacja** |
| gold | towers | 0,393822 |
| gold | kills | 0,327893 |
| gold | kda\_top | 0,217354 |
| gold | kda\_supp | 0,211501 |
| towers | kills | 0,277898 |
| towers | kda\_top | 0,20178 |
| towers | baron | 0,23471 |
| towers | hextech\_soul | 0,223015 |
| kills | kda\_top | 0,206086 |
| kills | kda\_adc | 0,218103 |
| kills | kda\_supp | 0,217261 |
| minions\_top | minions\_mid | 0,210123 |
| minions\_jungle | minions\_mid | 0,262053 |
| minions\_mid | minions\_adc | 0,332144 |
| minions\_mid | minions\_supp | 0,204191 |
| hextech\_dragon | infernal\_dragon | 0,268458 |
| hextech\_dragon | mountain\_dragon | 0,3294 |
| infernal\_dragon | mountain\_dragon | 0,232538 |
| ocean\_dragon | zeri\_yummi\_red | 0,235539 |
| lucian\_nami\_blue | zeri\_yummi\_red | 0,390014 |
| lucian\_nami\_blue | senna\_tahm\_red | 0,480688 |
| lucian\_nami\_red | ocean\_soul | -0,20262 |

Źródło: Opracowanie własne.

Jak wynika z tabeli 7 w analizowanym zbiorze danych występują 22 pary zmiennych, których skorelowanie może być problematyczne w kontekście dalszego badania. W przypadku korelacji zmiennej **gold** ze zmiennymi **towers**, **kills**, **kda\_top** oraz **kda\_supp** korelacja na rzeczonych poziomach wydaje się być jak najbardziej zasadna- zarówno niszczenie struktur, jak i zdobywanie zabójstw, asyst i śmierci wiąże się z gromadzeniem złota (czy to przez drużynę niebieską, czy czerwoną). Można również bez problemu wytłumaczyć różnicę w wielkości korelacji w dwóch pierwszych wierszach tabeli, a tej obserwowanej w wierszu trzecim i czwartym. Najwyższy poziom korelacji dla zmiennej **towers** może wynikać z faktu dodatkowej premii złota za zniszczenie pierwszej struktury oraz z powodu, że złoto otrzymywane jest globalnie to znaczy, że wszyscy gracze drużyny niszczącej strukturę są beneficjentami jej zniszczenia, w odróżnieniu do zabójstw, gdzie osoba, która dokona egzekucji przeciwnika dostaje 300 sztuk, a osoby asystujące dostają tego złota mniej. Jest to oczywiście wartość bazowa nie uwzględniająca tak zwanej pierwszej krwi[[57]](#footnote-57), czy współczynnika „bounty”, który został omówiony w podrozdziale 3.3, jednak ich występowanie może być powodem drugiej najsilniejszej korelacji zmiennej **kills** ze zmienną **gold**. Jedynym zastanawiającym elementem może być fakt, że tak wysoka korelacja zmiennej obrazującej różnice we współczynniku KDA dotyczy tylko górnej alei i wsparcia, jednak w obu tych przypadkach można wyjaśnić to rolami tych pozycji pełnionymi w drużynie- gracze z górnej alei często pełnią rolę „split pusherów” (co zostanie omówione w późniejszej części tego podrozdziału), natomiast rolą graczy wsparcia jest pomoc całej drużynie, co sprzyja generowaniu asyst, a co za tym idzie złota. W przypadku korelacji zmiennej **towers** ze zmiennymi **kills**, **baron** oraz **kda\_top** ponownie można zgodzić się co do ich zasadności. Przewaga w zabójstwach powoduje szybsze pokonywanie struktur, tak samo jak wzmocnienie otrzymywane po zabiciu Barona Nashora. W kwestii korelacji statystyki KDA na górnej alei z różnicą w zniszczonych strukturach, kluczowe jest wspomniane wyżej pojęcie „split pushu[[58]](#footnote-58)”. Jest to strategia polegająca na wysłaniu gracza na przeciwną aleję mapy w celu niszczenia wież wroga. Najczęściej to właśnie gracz grający na górnej alei oddelegowywany jest do tego zadania, co jest powodem wysokiej wartości tej korelacji. Natomiast jeśli chodzi o korelację zmiennej **towers** ze zmienną **hextech\_soul** to trudno jest jednoznacznie określić jej pochodzenie, ponieważ bonus za zdobycie duszy Smoka Hextechowego nie ułatwia bezpośrednio zdobywania struktur, więc korelacja ta wydaje się losowa, jednak jej wartość nie jest na tyle krytyczna, aby należało usunąć tą zmienną. W przypadku korelacji zmiennej **kills** ze zmiennymi dotyczącymi różnicy we współczynniku KDA dla górnej alei, strzelca i wsparcia nie zauważono żadnych nieprawidłowości logicznych. Zwiększenie ilości zabójstw skutkuje w poprawie współczynnika KDA zarówno przez większą ilość eliminacji jak i asyst. Zastanawiającym jest jedynie fakt, dlaczego pozostałe pozycje (dżungla i środkowa aleja) nie znalazły się w tych korelacjach, jednak może to wynikać z faktu, że w początkowych i środkowych etapach rozgrywki gra rzadko kiedy skupia się na tych rolach. Korelacje zmiennych dotyczących stworów oraz zmiennych dotyczących smoków wydają się być całkowicie losowe. W przypadku stworów można spróbować to wytłumaczyć za pomocą pojęcia „roam-u[[59]](#footnote-59)”. Jest to strategia, która polega na tym, że wygrywając swoją linię gracze schodzą na aleje sąsiadujące w celu pomocy sojusznikom w uzyskaniu przewagi nad ich oponentami z alei. Natomiast co do korelacji zmiennych dotyczących smoków ciężko jednoznacznie wskazać co jest ich przyczyną, ponieważ to, jaki smok pojawi się następny jest albo losowe, albo ukryte w kodzie gry, jednak brak jest informacji pozwalających jednoznacznie to stwierdzić, więc może to po prostu wynikać ze specyfiki analizowanego turnieju. W przypadku zmiennych dotyczących kompozycji poziom korelacji wydaje się jak najbardziej uzasadniony. Jak wspomniano w rozdziale 4.1 zestawienie Lucian Nami jest jednym z wymienionych silnych duetów, więc naturalnym jest, że gdy drużyna niebieska je wybierze (zmienna **lucian\_nami\_blue**) to drużyna czerwona odpowie jednym z pozostałych zestawień (zmienne **sennna\_tahm\_red** oraz **zeri\_yummi\_red**). W przypadku korelacji zmiennych **ocean\_dragon** z **zeri\_yummi\_red** oraz **lucian\_nami\_red** z **ocean\_soul** trudno jest jednoznacznie wskazać przyczyny ich występowania, prawdopodobnie był to aspekt czysto losowy oraz specyfika turnieju. Po przeanalizowaniu i obróbce zmiennych ilościowych należy zająć się zmiennymi jakościowymi. W wykorzystywanej ramce danych występuje ich 13. Zmienne kategoryczne w niezmienionej formie nie będą możliwe do efektywnego wykorzystania w modelu predykcyjnym. Z tego powodu zdecydowano się je przekodować z wykorzystaniem funkcji **get\_dummies**, jednak wyłączono z tego zmienną **time**, ponieważ jest to pomocnicza zmienna, która w późniejszej fazie analizy będzie służyła podzieleniu rekordów na konkretne przedziały czasowe. Ostatecznie końcowa baza danych posiada 3168 wierszy oraz 233 kolumn, które zostaną użyte do modelowania i predykcji (z wyłączeniem zmiennej **time**).

## **4.3. Modelowanie**

Następnym krokiem po ostatecznym przygotowaniu danych jest zbudowanie modeli predykcyjnych. W przypadku danych użytych w tej analizie zdecydowano o porównaniu ze sobą zdolności predykcyjnych pięciu algorytmów uczenia maszynowego: regresji logistycznej, drzew decyzyjnych, lasów losowych, metody wzmocnienia gradientowego oraz klasyfikatora XGBoost. Każdy z tych modeli sprawdza się inaczej w zależności od badanego problemu. Zdecydowano się zbudować modele na dwóch bazach: bazie ze zmiennymi dotyczącymi konkretnych postaci oraz bazie bez tych zmiennych. Powód takiej decyzji zostanie omówiony w późniejszej części pracy.

### **4.3.1 Modele zawierające zmienne dotyczących konkretnych postaci**

#### **4.3.1.1 Regresja logistyczna**

Pierwszą klasą modeli, których skuteczność postanowiono zbadać, były modele regresji logistycznej. Są to modele, które służą przewidywaniu binarnej zmiennej zależnej na podstawie zestawu zmiennych objaśniających. Model regresji logistycznej wykorzystuje logarytm stosunku szans do przekształcenia liniowej kombinacji zmiennych objaśniających. Algorytmy tej klasy przewidują prawdopodobieństwo przynależności do konkretnej klasy (np. w przypadku tej analizy są to: klasa meczy wygranych przez drużynę niebieską i klasa meczy wygranych przez drużynę czerwoną) na podstawie granicy decyzyjnej. W pierwszy kroku wykorzystano technikę optymalizacji Grid Search w celu znalezienie optymalnego zestawu parametrów modelu. Sprawdzono następujące parametry:

* Penalty- określa rodzaj regularyzacji zastosowanej w modelu. Regularyzacja pozwala zapobiec nadmiernemu dopasowaniu się do danych, przez co poprawia wydajność modelu. Możliwe wartości tego parametru to: „None” (brak regularyzacji), „L1” (nazywana inaczej „Lasso”, nakłada karę na absolut wartości), „L2” (nakłada karę na kwadraty wartości) oraz „Elasticnet” (jest niejako połączeniem regularyzacji L1 i L2). Sprawdzono wszystkie możliwe rodzaje regularyzacji z wyjątkiem jej braku. Ostatecznie Grid Search wskazał regularyzację L2 jako najlepszą.
* C- jest to parametr określający odwrotność siły regularyzacji. Im wyższe jest C tym nakładana kara jest mniejsza. Jako potencjalne wartości sprawdzono: 0,01; 0,1; 1; 10 oraz 100. W drodze optymalizacji wybrano 0,1 jako wartość tego parametru.
* Solver- parametr ten określa algorytm używany do optymalizacji parametrów modelu. Możliwe wartości tego parametru to: „newton-cg”, „sag”, „lbfgs", „liblinear" i „saga". Używana technika optymalizacyjna zasugerowała wykorzystanie „liblinear”.

Po wykonaniu optymalizacji przystąpiono do budowy modelu oraz dokonania predykcji. W celu weryfikacji skuteczności modelu posłużono się czterema miarami:

* Dokładność (*ang. accuracy*) - jest to procent poprawnie sklasyfikowanych próbek ze wszystkich. Mierzy ona ogólną skuteczność klasyfikacji.
* Precyzja (*ang. precision*) – mierzy stosunek prawdziwie pozytywnych predykcji (*ang. true positive*) do wszystkich pozytywnych predykcji (*ang. true positive and false positive*).
* Czułość (*ang. recall*) – mierzy proporcje prawdziwie pozytywnych predykcji (*ang. true positive*) do wszystkich istniejących prawdziwie pozytywnych przypadków (*ang. true positive and false negative*).
* Miara F1 – jest to harmoniczna średnia precyzji i czułości.
* AUC-ROC - krzywa ROC przedstawia wydajność modelu w zależności od wartości progowej, a AUC-ROC mierzy całkowitą wydajność klasyfikatora (gdzie 1 to klasyfikator idealny a 0,5 klasyfikator losowy).

W tabeli 8 przedstawiono wartości opisanych wyżej miar.

Tabela 8. Miary jakości klasyfikacji dla najlepszego modelu regresji logistycznej.

|  |  |
| --- | --- |
| Dokładność (accuracy) | 0,9517615176151761 |
| Precyzja (precision) | 0,9608062709966405 |
| Czułość (recall) | 0,9407894736842105 |
| Miara F1 (F1 score) | 0,9506925207756232 |
| AUC-ROC | 0,9516380380210978 |

Źródło: Opracowanie własne.

Analizując informacje zawarte w tabeli 8 można odnieść wrażenie, że stworzony model sprawdza się świetnie w predykcji badanego zjawiska, jednak sytuacja ta wydaje się niemożliwa do osiągnięcia. Z tego powodu postanowiono przeprowadzić walidację krzyżową z wykorzystaniem 5 foldów, ponieważ podejrzewano, że stworzony model może być niestabilny. Średnia dokładność osiągnięta w walidacji wynosi około 0,78. Jest to wielkość zdecydowanie mniejsza niż osiągnięta w modelu, z czego można wysnuć wniosek, że jest on faktycznie niestabilny. Jednym z powodów niestabilności modelu może być duży rozstrzał wartości występujący w zmiennych. Z tego powodu zdecydowano się wykorzystać moduł **StandardScaler()** w celu zmniejszenia rozstrzału wartości. Metoda ta polega na przekształceniu cech według następującego wzoru:

gdzie to oryginalna wartość zmiennej, a to przekształcona wartość. Temu procesowi poddano wszystkie zmienne, które posiadały więcej niż dwie unikalne wartości (oprócz zmiennej **time**, która nie była brana pod uwagę podczas modelowania). Następnie ponownie nauczono model oraz dokonano predykcji. Tabela 9 przedstawia wartości miar jakości klasyfikacji dla zmiennych z wystandaryzowanymi wartościami.

Tabela 9. Miary jakości klasyfikacji dla najlepszego modelu regresji logistycznej po standaryzacji.

|  |  |
| --- | --- |
| Dokładność (accuracy) | 0,964769647696477 |
| Precyzja (precision) | 0,9865900383141762 |
| Czułość (recall) | 0,9414990859232175 |
| Miara F1 (F1 score) | 0,9635173058933584 |
| AUC-ROC | 0,9644995429616088 |

Źródło: Opracowanie własne.

Jak wynika z tabeli 9, standaryzacja poprawiła wartości miar jakości klasyfikacji, jednak dalej wartości te mogą budzić wątpliwości, dlatego ponownie zdecydowano się na zastosowanie walidacji krzyżowej. Średnia dokładność dla wykorzystanych 5 foldów wyniosła około 0,776, co wskazuje, że poczynione kroki nie przyniosły oczekiwanej poprawy stabilności modelu. W ostatnim kroku pracy nad modelami tej klasy (z wykorzystaniem wszystkich pozyskanych zmiennych) postanowiono zbadać i wskazać 15 najbardziej istotnych dla predykcji zmiennych.

Rys. 23. Najbardziej istotne zmienne dla modelu regresji logistycznej.

Obraz zawierający zrzut ekranu, Prostokąt, Grafika, Czcionka

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

Rysunek 23 wskazuje na niepokojącą zależność w modelu- 11 spośród 15 najistotniejszych zmiennych to zmienne dotyczące konkretnych postaci. Nie jest to sytuacją pożądaną, ponieważ wybierane postacie nie wydają się być aż tak dominującym determinantem zwycięstwa bądź porażki. Przyczyny takiego rezultatu można upatrywać w specyfice gry. Jak wspomniano we wstępie analiza ta bazuje na jednym turnieju, a cały turniej rozgrywany jest na jednej aktualizacji, co sprawia, że zdarzają się postacie i kompozycje, które są po prostu zbyt silne (o czym wspominał Marcin Świech w udzielonym wywiadzie). Innym powodem takiego stanu rzeczy jest fakt, że baza użyta do modelowania zawiera stosunkowo mało rekordów, więc mogła zdarzyć się sytuacja, w której dane wybory postaci dawały stuprocentową szansę na wygraną danej drużyny. W takim przypadku powinno się rozważyć usunięcie zmiennych dotyczących konkretnych postaci (pozostawiając zmienne dotyczące ich kombinacji np. **zeri\_yummi\_blue**), jednak w celu podjęcia tej decyzji postanowiono zbudować pozostałe modele i zadecydować ich podstawie, ponieważ napotkana sytuacja może wynikać również z zastosowania zbyt prostego modelu mającego trudności z dopasowaniem się do skomplikowanych wzorców w danych.

#### **4.3.1.2 Drzewa decyzyjne**

Kolejną klasą modeli zbadaną w tej pracy były drzewa decyzyjne. Jest to popularny algorytm uczenia maszynowego wykorzystywany zarówno w klasyfikacji jak i regresji. Modele z tej klasy oparte są o hierarchiczną strukturę drzewa, w którym każdy węzeł reprezentuje test na jednej lub więcej cech, a każdy liść reprezentuje wynik tego testu. Przed konstrukcją modeli postanowiono sprawdzić optymalną wielkość parametru ccp\_alpha, który odpowiedzialny jest za regularyzacje drzewa, poprzez kontrolowanie stopnia jego przycinania w celu uniknięcia overfittingu i zwiększenia zdolności modelu do generalizacji badanego zjawiska. W tym celu sporządzono wykres zależności pierwiastka błędu średniokwadratowego od konkretnych wartości ccp\_alpha.

Rys. 24. Wykres zależności RMSE od wartości parametru ccp\_alpha.

Obraz zawierający zrzut ekranu, Prostokąt, linia, diagram

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

Jak można odczytać z rysunku 24 optymalna wielkość parametru ccp\_alpha wynosi 0, dla którego RMSE osiąga wartość 0. W następnym kroku sprawdzono głębokość oraz ilość liści drzewa zbudowanego z odpowiednią wartością parametru regularyzacji. Głębokość drzewa wyniosła 16 a ilość liści 20. Otrzymane wymiary sprawiają, że czytelna prezentacja graficzna jego budowy jest praktycznie niemożliwa. Z tego powodu zdecydowano się stworzyć funkcję, która wypisze cechy będące węzłami drzewa. Praktycznie wszystkie zmienne brane pod uwagę przy budowie drzewa dotyczyły konkretnych postaci (dwie z nich dotyczyły drużyn), co zdaje się potwierdzać wnioski otrzymane w przypadku regresji logistycznej- potrzebę utworzenia modeli bez zmiennych dotyczących konkretnych postaci.

#### **4.3.1.3 Lasy losowe**

Kolejną wykorzystaną w badaniu klasą modeli były lasy losowe. Algorytm ten polega na wyliczeniu dużej ilości drzew decyzyjnych i podjęciu decyzji predykcyjnych na podstawie wszystkich wykorzystanych drzew. Lasy losowe posiadają wiele zalet takich jak odporność na overfitting oraz odporność na braki danych. Pierwszym krokiem przy budowie lasów losowych był wybór odpowiedniej ilości drzew w lesie. W tym celu ponownie posłużono się błędem RMSE.

Rys. 25. Zależność błędu RMSE od ilości drzew w lesie losowym.

Obraz zawierający linia, diagram, origami

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

Jak wynika z rysunku 25 optymalną ilością drzew w lesie losowym jest 400. W następnym etapie postanowiono ustalić optymalną ilość zmiennych wpływających na RMSE.

Rys. 26. Zależność błędu RMSE od ilości cech w lesie losowym.

Obraz zawierający linia, zrzut ekranu

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

Optymalna ilość cech wskazana przez rysunek 26 wynosi 231, więc ostateczny las losowy poddany badaniu składa się z 400 drzew oraz 231 zmiennych. Wielkość RMSE osiągana przez ten model wynosi około 0,0053. Kolejnym podjętym krokiem było zbadanie relatywnej istotności 20 najważniejszych według modelu zmiennych oraz stopnia w jakim redukują one homogeniczność. Odpowiednie wykresy zostały zaprezentowane na rysunkach 27 i 28.

Rys. 27. Relatywna istotność zmiennych w lesie losowym.

Obraz zawierający zrzut ekranu, Prostokąt, linia, design

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

Rys. 28. Stopień redukcji homogeniczności w lesie losowym.

Obraz zawierający zrzut ekranu, linia, Prostokąt, Grafika

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

Na podstawie rysunków 27 i 28 ponownie można zauważyć dużą zależność predykcji od zmiennych odpowiadających konkretnym postaciom. Spośród 20 najważniejszych cech w modelu 18 dotyczy właśnie wyboru konkretnej postaci, co zdaje się potwierdzać sytuację występującą w poprzednio użytych modelach.

#### **4.3.1.4 Algorytm wzmocnienia gradientowego**

Modele predykcyjne z rodziny wzmocnienia gradientowego są zespołowymi metodami uczenia maszynowego, które łączą wiele słabych modeli, w celu stworzenia silnego modelu predykcyjnego. Jest to jedna z najmocniejszych technik uczenia maszynowego, która dobrze sprawdza się w przypadku złożonych wzorców danych posiadających wiele cech. Modele z tej klasy cechuje również wysoka zdolność do generalizacji, jednak wymagają niekiedy bardzo dużej mocy obliczeniowej. Kroki wykonane przy budowie modeli klasy Gradient Boosting były bardzo podobne do tych użytych w przypadku lasów losowych, więc w pierwszym kroku zbadano optymalną ilość drzew wykorzystaną do predykcji.

Rys. 29. Zależność błędu RMSE od ilości drzew w algorytmie wzmocnienia gradientowego.

Obraz zawierający zrzut ekranu, Prostokąt, kwadrat, linia

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

Minimalna wartość RMSE została osiągnięta dla modelu składającego się z 1500 drzew, więc wartość ta została użyta jako parametr **n\_estimators** w modelu. Ponownie zdecydowano się zbadać relatywną istotność 20 najistotniejszych zmiennych modelu.

Rys. 30. Relatywna istotność zmiennych w algorytmie wzmocnienia gradientowego.

Obraz zawierający zrzut ekranu, linia, Prostokąt, diagram

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

W przypadku modelu tej klasy ponownie można zauważyć, że 17 na 20 najistotniejszych zmiennych dotyczy wyboru konkretnych postaci, co podtrzymuje wcześniej postawioną tezę.

#### **4.3.1.5 Algorytm XGBoost**

Algorytm XGBoost jest wzmocnioną wersją algorytmu wzmocnienia gradientowego. Jest on bardziej uregulowaną formą tego algorytmu, która wykorzystuje zaawansowaną regularyzację (L1 i L2), w celu poprawy zdolności generalizacji modelu. Zapewnia on wysoką wydajność w porównaniu do zwykłego wzmocnienia gradientowego. W przypadku tego rodzaju modeli wybór dostosowywanych parametrów jest dość szeroki. Z tego powodu zdecydowano się na wykorzystanie techniki optymalizacji Grid Search w celu znalezienia optymalnej wartości następujących parametrów:

* **max\_depth-** jest to parametr określający głębokość każdego drzewa w modelu. Służy on redukcji ryzyka przeuczenia się modelu.
* **learning\_rate-** określa szybkość uczenia się modelu. Kontroluje on minimalizację funkcji kosztu w modelu.
* **n\_estimators-** określa liczbę drzew decyzyjnych użytych w modelu.
* **subsample-** określa procent próbek użytych do uczenia każdego drzewa.
* **colsample\_bytree-** oznacza procent zmiennych używanych do uczenia każdego drzewa.
* **reg\_alpha-** określa siłę regularyzacji Lasso, przez co ogranicza przeuczenie modelu.
* **reg\_lambda-** określa siłę regularyzacji L2, przez co zmniejsza wpływ potencjalnych obserwacji odstających.

Optymalny zestaw parametrów wskazany przez metodę Grid Search wygląda następująco:

* colsample\_bytree= 0,6
* learning\_rate= 0,1
* max\_depth= 3
* n\_estimators= 100
* reg\_alpha= 0,0
* reg\_lambda= 0,0
* subsample= 0,7

W następnym kroku zbudowano więc model ze wskazanymi przez technikę optymalizacji parametrami oraz ponownie sprawdzono, jakie są najistotniejsze parametry modelu.

Rys. 31. Relatywna istotność zmiennych w algorytmie XGBoost.

Obraz zawierający zrzut ekranu, linia, design

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

Z wykresu przedstawionego na rysunku 31 ponownie można zauważyć, że większość najistotniejszych zmiennych dotyczy wyboru konkretnych postaci, jednak w przypadku algorytmu XGBoost można zauważyć, że wpływ na predykcję modelu ma również zmienna **lucian\_nami\_red**, która opisuje jedno z najsilniejszych zestawień wymienionych przez Marcina Świecha.

#### **4.3.1.6 Podsumowanie analizy modeli zawierających zmienne dotyczące konkretnych postaci**

Po zbudowaniu i przeanalizowaniu modeli zawierających wszystkie zmienne postanowiono porównać je ze sobą. Jako miarę porównawczą ponownie wykorzystano pierwiastek błędu średniokwadratowego.

Rys. 32. Porównanie RMSE na zbiorze testowym dla modeli z pełnym zestawem zmiennych.

Obraz zawierający zrzut ekranu, Prostokąt, kwadrat, Wielobarwność

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

Jak wynika z rysunku 32 najmniejsze wartości RMSE osiągane są dla drzew losowych oraz modelu XGBoost, a największe dla regresji logistycznej, jednak analiza stabilności modeli oraz najważniejszych zmiennych spowodowała uzasadnione podejrzenie o niemożliwości wykorzystania tych modeli do rzeczywistej predykcji. Sytuacja, w której wybierane postacie definiują, która drużyna wygra mecz, wydaje się niedopuszczalna. Z tego powodu zdecydowano się ponownie przeprowadzić modelowanie oraz analizę dla bazy danych pozbawionej zmiennych dotyczących pojedynczych postaci. Działanie to powinno skutkować zbudowaniem modeli, które będą dużo bardziej stabilne oraz będą lepiej prognozować badane zjawisko.

### **4.3.2 Modele pozbawione zmiennych dotyczących konkretnych postaci**

Z używanej w poprzednim podrozdziale bazy danych postanowiono usunąć zmienne, które oznaczały wybór jednej z postaci, ale pozostawiono te, które odpowiadały duetom postaci (np. **lucian\_nami\_blue**), jednak w trakcie analizy danych i modeli zauważono, że zmienne dotyczące drużyn prowadziły do podobnych wyników, jak te otrzymane w podrozdziale 4.3.1. Stwierdzono, że rozsądnym będzie usunięcie również tych cech. Decyzja ta spowodowana była faktem, że składy drużyn rzadko ulegają zmianie w trakcie jednego turnieju, przez co poziom danej drużyny prezentowany na analizowanym turnieju może nie stanowić najlepszego predyktora, więc wykorzystanie tych zmiennych może prowadzić do pogorszenia predykcji. Użyta w tym podrozdziale baza posiada 3168 wierszy oraz 36 kolumn. Postanowiono również poddać odpowiednie zmienne standaryzacji (w podobny sposób jak zrobiono to w podrozdziale 4.3.1).

#### **4.3.2.1 Regresja logistyczna**

Budowę modeli rozpoczęto od regresji logistycznej. Ponownie posłużono się techniką optymalizacji Grid Search w celu znalezienia najlepszego zestawu parametrów. Na podstawie tego kroku zbudowano model z następującymi parametrami:

* **C** = 0.1,
* **penalty** = l2,
* **solver** = liblinear
* **max\_iter =** 200

Otrzymany model uzyskał około 71,71% skuteczności na zbiorze testowym. Postanowiono również sprawdzić pozostałe miary jakości klasyfikacji.

Tabela 10. Miary jakości klasyfikacji dla najlepszego modelu regresji logistycznej (na bazie bez zmiennych dotyczących postaci i drużyn).

|  |  |
| --- | --- |
| Dokładność (accuracy) | 0,7171398527865405 |
| Precyzja (precision) | 0,731006160164271 |
| Czułość (recall) | 0,7206477732793523 |
| Miara F1 (F1 score) | 0,72579001019368 |
| AUC-ROC | 0,7169978472523675 |

Źródło: Opracowanie własne.

W następnym kroku postanowiono upewnić się co do stabilność modelu zbudowanego na tym zestawie zmiennych. W tym celu posłużono się walidacją krzyżową dla 5 foldów. Uzyskana średnia dokładność predykcji wyniosła około 67%, co pozawala uznać model za stosunkowo stabilny. Ostatnim krokiem w pracy nad tą klasą modeli zbadano, które ze zmiennych były najistotniejsze dla predykcji. Wybrano 20 zmiennych z największym wpływem na wynik modelu i zaprezentowano je na rysunku 33.

Rys. 33. Najbardziej istotne zmienne dla modelu regresji logistycznej (bez zmiennych dotyczących postaci i drużyn).

Obraz zawierający zrzut ekranu, Grafika, Prostokąt, Czcionka

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

Z rysunku 3 możemy zauważyć, że największą istotność uzyskały zmienne dotyczące zestawień postaci, jednak zaskakującym jest fakt dodatniego wpływu kombinacji Senna z Tahm Kench-em w drużynie czerwonej na wygraną drużyny niebieskiej oraz ujemny wpływ zestawienia Zeri z Yummi dla drużyny niebieskiej, jednak wyniki te mogą być spowodowane wieloma czynnikami meczowymi, takimi jak wybór kontrującej kompozycji przez drużynę przeciwną lub faktem użycia oversamplingu w badaniu. Pozostałe z 20 najistotniejszych zmiennych wydają się być jak najbardziej uzasadnione.

#### **4.3.2.2 Drzewa decyzyjne**

Kroki poczynione przy budowie drzewa decyzyjnego były takie same jak w przypadku podrozdziału 4.3.1.2. Na początku zbadano optymalną wartość parametru regularyzacji cpp\_alpha dla danych wykorzystanych w tym rozdziale.

Rys. 34. Wykres zależności RMSE od wartości parametru ccp\_alpha (dla danych bez zmiennych dotyczących poszczególnych postaci oraz drużyn).

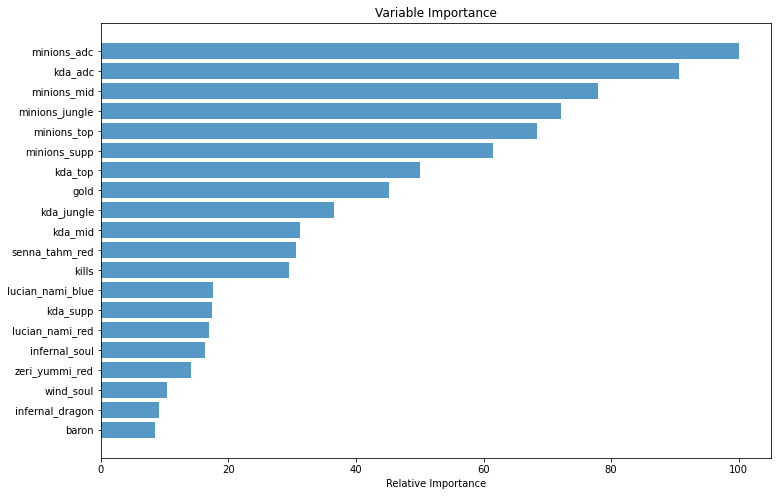
Obraz zawierający zrzut ekranu, linia, diagram, tekst

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

Jak można odczytać z rysunku 34, wartość tego parametru, dla której cpp\_alpha osiąga minimum to 0. Otrzymane drzewo ma głębokość 30 oraz posiada 292 liście, co sprawia, że niemożliwym jest pokazanie go na czytelnym wykresie. Z tego powodu postanowiono zbadać zmienne składające się na węzły tego modelu.

Rys. 35. Relatywna istotność zmiennych w drzewie decyzyjnym (dla danych bez zmiennych dotyczących poszczególnych postaci oraz drużyn).



Źródło: Opracowanie własne.

Po oszacowaniu modelu regresji spodziewano się, że najistotniejsza zmienna z poprzedniego podrozdziału (**lucian\_nami\_red**) będzie pierwszym kryterium podziału drzewa, jednak w tym przypadku była to zmienna **minions\_adc**. Większość najważniejszych kryteriów podziału w drzewie stanowiły zmienne dotyczące różnicy w statystyce KDA, stworach, zabójstwach i strukturach oraz zmienne dotyczące smoków i zdobycia Barona Nashora.

#### **4.3.2.3 Lasy losowe**

Następną klasą modeli, która będzie omówiona w przypadku bazy z usuniętymi zmiennymi są lasy losowe. Schemat analizy przyjęty w tym podrozdziale ponownie jest taki sam jak wykorzystany dla tej klasy modeli w rozdziale 4.3.1. W pierwszym kroku zbadano dla jakiej ilości drzew model osiąga najmniejszy pierwiastek błędu średniokwadratowego.

Rys. 36. Zależność błędu RMSE od ilości drzew w lesie losowym (dla danych bez zmiennych dotyczących poszczególnych postaci oraz drużyn).

Obraz zawierający linia, diagram, origami

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

Minimalna wartość błędu RMSE osiągana jest dla 500 drzew. Następnie zbadano jaka jest optymalna ilość cech w lesie, który składa się z takiej ilości drzew.

Rys. 37. Zależność błędu RMSE od ilości cech w lesie losowym (dla danych bez zmiennych dotyczących poszczególnych postaci oraz drużyn).

Obraz zawierający linia, diagram

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

Ilość zmiennych, dla których osiągane RMSE jest najmniejsze wynosi 1, jednak jego wartość nie wzrasta szybko wraz z dodaniem kolejnych cech. Postanowiono więc zbudować model składający się z 500 drzew oraz 20 zmiennych objaśniających. Dla tych wartości parametrów osiągnięto wielkość pierwiastka błędu średniokwadratowego na poziomie około 0,281. W kolejnym kroku postanowiono zbadać relatywną istotność zmiennych.

Rys. 38. Relatywna istotność zmiennych w lesie losowym (dla danych bez zmiennych dotyczących poszczególnych postaci oraz drużyn).

Obraz zawierający zrzut ekranu, linia, Prostokąt, design

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

W przypadku modeli tej klasy najważniejszą cechą okazała się różnica w stworach na środkowej alei, gdzie w przypadku drzew decyzyjnych była to różnica między strzelcami. Tym razem również największą istotność osiągały zmienne, które dotyczyły różnicy w statystyce KDA, stworach, zabójstwach i strukturach zdobyciu Barona Nashora, jednak w tym przypadku zmienna **gold** była relatywnie ważniejsza dla predykcji modelu, niż miało to miejsce w poprzednim podrozdziale. Ostatnim krokiem było zbadanie w jakim stopniu zmienne te redukują homogeniczność w tym modelu. W tym celu ponownie posłużono się wykresem wykorzystanym również w podrozdziale 4.3.1.3.

Rys. 39. Stopień redukcji homogeniczności w lesie losowym (dla danych bez zmiennych dotyczących poszczególnych postaci oraz drużyn).

Obraz zawierający zrzut ekranu, design

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

Wykres 39 oprócz prezentacji wizualnej stopnia redukcji homogeniczności przez zmienne prezentuje również informacje o zmienności ważności cech między drzewami. Największe wahania obserwujemy w przypadku zmiennych **gold**, **kda\_adc** oraz **kda\_top**.

#### **4.3.2.4 Algorytm wzmocnienia gradientowego**

W pierwszym kroku tego rozdziału zbadano optymalną wartość parametru **n\_estimators**. W tym celu ponownie wykorzystano RMSE.

Rys. 40. Zależność błędu RMSE od ilości drzew w algorytmie wzmocnienia gradientowego (dla danych bez zmiennych dotyczących poszczególnych postaci oraz drużyn).

Obraz zawierający linia, zrzut ekranu

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

Minimalna wartość RMSE jest osiągana dla wartości parametru na poziomie 3000. Zbudowano więc model wzmocnienia gradientowego z **n\_estimators** równym 3000. Następnie zbadano relatywną istotność parametrów.

Rys. 41. Relatywna istotność zmiennych w algorytmie wzmocnienia gradientowego (dla danych bez zmiennych dotyczących poszczególnych postaci oraz drużyn).

Obraz zawierający zrzut ekranu, linia, Prostokąt, design

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

Na podstawie rysunku 40 ponownie można zauważyć, że największy wpływ na decyzję modelu posiadały zmienne dotyczące statystyk takich jak KDA czy stwory oraz złota, jednak kolejność relatywnych istotności jest zupełnie inna niż w modelach wyżej omawianych klas modeli.

#### **4.3.2.5 Algorytm XGBoost**

W przypadku budowania modelu tej klasy, kroki postępowania były takie same jak w rozdziale 4.3.1.5. W pierwszym kroku ponownie wykorzystano technikę optymalizacyjną Grid Search w celu znalezienia optymalnego zestawu parametrów. W efekcie estymacji wybrano następujące wartości do budowy modelu:

* colsample\_bytree= 0,9
* learning\_rate= 0,1
* max\_depth= 9
* n\_estimators= 300
* reg\_alpha= 0,1
* reg\_lambda= 0,0
* subsample= 0,8

W następnym kroku ponownie sprawdzono relatywną istotność zmiennych. Wyniki zaprezentowano na rysunku 42.

Rys. 42. Relatywna istotność zmiennych w algorytmie XGBoost (dla danych bez zmiennych dotyczących poszczególnych postaci oraz drużyn).

Obraz zawierający zrzut ekranu, Prostokąt, linia, design

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

Na podstawie danych zaprezentowanych na wykresie można zauważyć ciekawą sytuację- najistotniejszą zmienną jest ta dotycząca różnicy w smokach piekielnych. Wartym zauważenia jest również fakt, że zmienne dotyczące kompozycji w przypadku tego modelu wykazują zdecydowanie większą relatywną istotność niż w modelach z innych klas.

#### **4.3.2.6 Porównanie modeli niezawierających zmiennych dotyczących konkretnych postaci i drużyn**

Po oszacowaniu modeli dla bazy danych pozbawionej zmiennych pogarszających stabilność modeli zdecydowano się dokonać ich porównania w celu wybrania najlepszego w tym przypadku modelu. Ponownie miarą porównawczą był pierwiastek błędu średniokwadratowego.

Rys. 43. Porównanie RMSE na zbiorze testowym dla modeli z niepełnym zestawem zmiennych.

Obraz zawierający Wielobarwność, Prostokąt, Grafika, zrzut ekranu

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

Jak wynika z rysunku najlepszy wynik pod względem RMSE otrzymał model lasu losowego, ale wartym zauważenia jest fakt, że różnice między wartością tego błędu dla trzech ostatnich modeli są stosunkowo niewielkie. Postanowiono więc wykorzystać również inną miarę oceny jakości modeli- dokładność predykcji. Wyniki przedstawiono w tabeli 11.

Tabela 11. Porównanie dokładności predykcji dla modeli z niepełnym zestawem zmiennych.

|  |  |
| --- | --- |
| **Model** | **Uzyskane accuracy** |
| Regresja logistyczna | 0.7171398527865405 |
| Drzewo decyzyjne | 0.8622502628811777 |
| Las losowy | 0.8969505783385909 |
| Algorytm wzmocnienia gradientowego | 0.8727655099894848 |
| Algorytm XGBoost | 0.886435331230284 |

Źródło: Opracowanie własne.

Z informacji w tabeli 11 możemy zauważyć, że model lasu losowego ponownie uzyskał najlepszy wynik- uzyskana dokładność predykcji wyniosła około 89,7%. Z tego powodu uznanie tego modelu za najlepszy w kontekście naszych danych wydaje się zasadne.

### **4.3.3 Modele w podziale na czas meczu**

Kierując się wskazaniami literatury postanowiono dokonać predykcji w podziale na konkretne fazy meczu, aby sprawdzić, w którym etapie spotkania wynik jest najłatwiejszy do przewidzenia oraz jakie zmienne mają wtedy największe znaczenie. Zdecydowano dokonać podziału na 3 grupy: do 10 minuty meczu, od 10 do 20 minuty meczu i po 20 minucie meczu. Ponownie posłużono się zestandaryzowanym zbiorem danych pozbawionym zmiennych o konkretnych postaciach i zespołach. Aby nie powtarzać kroków analizy z rozdziału 4.3.1 i 4.3.2 postanowiono przedstawić otrzymane rezultaty w postaci tabelarycznej.

#### **4.3.3.1 Obserwacje do 10 minuty meczu**

W pierwszym kroku postanowiono zbadać czy w efekcie podziału ramki danych pod względem zmiennej **time** nie wystąpiły zbyt duże nierówności w rozkładzie zmiennej celu. W przypadku obserwacji do 10 minuty spotkania „zer” jest 628 a „jedynek” 603. Nie jest to na tyle duża dysproporcja, aby trzeba było ją poprawiać.

Tabela 12. Porównanie modeli dla obserwacji do 10 minuty spotkania.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Parametry** | **Accuracy** | **RMSE** | **Najważniejsze zmienne** |
| Regresja logistyczna | penalty = 'l1',  max\_iter = 300,  C = 1,  solver = 'saga' | ≈ 0,5973 | ≈ 0,6346 | senna\_tahm\_red,  zeri\_yummi\_blue,  kog\_lulu\_blue,  lucian\_nami\_blue |
| Drzewo decyzyjne | ccp\_alpha ≈ 0,00099  depth = 25  leaves = 99 | ≈ 0,8135 | ≈ 0,4001 | minions\_adc,  minions\_mid,  minions\_top,  minions\_supp |
| **Las losowy** | **n\_estimators = 100**  **max\_features = 4** | **≈ 0,8351** | **≈ 0,3412** | **minions\_adc,**  **minions\_top,**  **minions\_mid,**  **gold** |
| Wzmocnienie gradientowe | n\_estimators = 1000 | ≈ 0,8135 | ≈ 0,3778 | gold,  minions\_top,  minions\_mid,  minions\_supp |
| XGBoost | colsample\_bytree = 0,8,  learning\_rate = 0,1,  max\_depth = 9,  n\_estimators = 200,  reg\_alpha = 0,5,  reg\_lambda = 0,1,  subsample = 0,9 | ≈ 0,827 | ≈ 0,4159 | senna\_tahm\_red,  senna\_tahm\_blue,  kog\_lulu\_blue,  lucian\_nami\_blue |

Źródło: Opracowanie własne.

Jak można odczytać z tabeli 12 zarówno pod względem RMSE jak i dokładności predykcji najlepszy okazał się model lasu losowego. Przewidział on poprawnie 83,51% rekordów. Najważniejszymi zmiennymi były różnice w stworach na różnych pozycjach i złoto, co wydaje się zgodne z logiką, ponieważ w początkowych fazach meczu najważniejsze jest gromadzenie zasobów, a zabójstwa nie padają zbyt często.

#### **4.3.3.2 Obserwacje od 10 do 20 minuty meczu**

W ramce danych wykorzystanej do budowy modelu dla obserwacji dotyczących środkowego etapu meczu rozłożenie zmiennej celu ponownie było dość wyrównane (478 „jedynek” oraz 431 „zer”).

Tabela 13. Porównanie modeli dla obserwacji od 10 do 20 minuty spotkania.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Parametry** | **Accuracy** | **RMSE** | **Najważniejsze zmienne** |
| Regresja logistyczna | penalty = 'l2',  max\_iter = 300,  C = 0,1,  solver = 'liblinear' | ≈ 0,7216 | ≈ 0,5276 | senna\_tahm\_red,  lucian\_nami\_blue,  zeri\_yummi\_red,  gold |
| Drzewo decyzyjne | ccp\_alpha ≈ 0,00351  depth = 9  leaves = 21 | ≈ 0,8135 | ≈ 0,3641 | gold,  minions\_adc,  kda\_supp,  minions\_supp |
| **Las losowy** | **n\_estimators = 100**  **max\_features = 4** | **≈ 0,9487** | **≈ 0,2468** | **gold,**  **minions\_top,**  **minions\_supp,**  **kda\_adc** |
| Wzmocnienie gradientowe | n\_estimators = 2000 | ≈ 0,9231 | ≈ 0,264 | gold,  minions\_supp,  kda\_adc,  kda\_top |
| XGBoost | colsample\_bytree = 0,8,  learning\_rate = 0,1,  max\_depth = 5,  n\_estimators = 100,  reg\_alpha = 0,0,  reg\_lambda = 0,0,  subsample = 0,9 | ≈ 0,9304 | ≈ 0,264 | senna\_tahm\_red,  senna\_tahm\_blue,  kog\_lulu\_blue,  lucian\_nami\_blue |

Źródło: Opracowanie własne.

W przypadku obserwacji od 10 do 20 minuty meczu ponownie najlepszy okazał się las losowy z **n\_estimators =** 100 i **max\_features =** 4. Przewidział on około 95% wyników. Najważniejszą cechą okazała się różnica w złocie, a następnymi różnice stworach. W przypadku tej fazy meczu znaczącą istotnością charakteryzował się również współczynnik KDA dla strzelca.

#### **4.3.3.3 Obserwacje po 20 minucie meczu**

W przypadku tej części danych również nie było potrzeby wykorzystania którejkolwiek z technik balansowania danych, ponieważ stosunek „zer” i „jedynek” wynosił 525 do 503.

Tabela 14. Porównanie modeli dla obserwacji po 20 minucie spotkania.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Parametry** | **Accuracy** | **RMSE** | **Najważniejsze zmienne** |
| Regresja logistyczna | penalty = 'l1',  max\_iter = 300,  C = 1,  solver = 'saga' | ≈ 0,8544 | ≈ 0,3816 | senna\_tahm\_red,  infernal\_soul,  mountain\_soul,  wind\_soul |
| Drzewo decyzyjne | ccp\_alpha ≈ 0,00131  depth = 12  leaves = 41 | ≈ 0,9061 | ≈ 0,2865 | kda\_adc,  minions\_mid,  kda\_top,  kda\_jungle |
| Las losowy | n\_estimators = 500  max\_features = 19 | ≈ 0,9288 | **≈ 0,239** | minions\_mid,  kda\_top,  kda\_adc,  kda\_jungle |
| Wzmocnienie gradientowe | n\_estimators = 2500 | ≈ 0,9353 | ≈ 0,2431 | minions\_mid,  kda\_top,  kda\_adc,  kda\_jungle |
| XGBoost | colsample\_bytree = 0,9,  learning\_rate = 0,1,  max\_depth = 5,  n\_estimators = 100,  reg\_alpha = 0,0,  reg\_lambda = 0,0,  subsample = 0,9 | **≈ 0,9385** | ≈ 0,248 | lucian\_nami\_red,  zeri\_yummi\_red,  lucian\_nami\_blue,  kda\_adc |

Źródło: Opracowanie własne.

W przypadku obserwacji z końcowych faz meczu można zauważyć bardzo ciekawą sytuację. Po raz pierwszy mamy sytuację, w której różne modele są najlepsze według konkretnych statystyk. Pod względem błędu RMSE najlepszy jest model lasu losowego a najlepszą dokładność predykcji otrzymuje algorytm XGBoost. Wartym zaznaczenia jest również fakt, że najistotniejsze zmienne w tych modelach zdecydowanie się różnią. Model lasu losowego wskazuje na zdecydowanie większe znaczenie statystyki KDA oraz liczby minionów, natomiast XGBoost premiuje zmienne dotyczące kompozycji i KDA strzelca.

# V. Omówienie wyników najlepszych modeli

## **5.1 Modele z podziałem na etap spotkania**

Autorzy publikacji „Continuous outcome prediction of league of legends competitive matches using recurrent neural networks” zauważyli wahania trafności predykcji w zależności od minuty meczu. Z tego powodu, ciekawym aspektem było zbudowanie modeli dla różnych faz meczu. Analiza ta dostarczyła ciekawych informacji. Kluczowym pytaniem, na które odpowiedzi można udzielić, jest pytanie o możliwości dokonywania trafnych predykcji konkretnych fragmentów meczu na podstawie nagrań transmisji. Zgodnie z wynikami analizy, prognoza zwycięzcy meczu na podstawie odpowiedniej fazy jest możliwa, ale wartym zauważenia jest fakt, że stopień zaawansowania tego zadania jest różny. Najłatwiejsze jest dokonywanie predykcji dla środkowej fazy meczu (94,87% poprawnych prognoz), z kolei najtrudniejsze do przewidzenia są zgodnie z oczekiwaniami początkowe fazy meczu (85,31% dokładności), jednak w każdym z etapów otrzymywane wyniki były satysfakcjonujące. Na podstawie najważniejszych cech wskazanych przez modele w każdej fazie meczu istnieje możliwość omówienia najważniejszych czynników zwycięstwa dla każdego etapu spotkania.

Rys. 44. Relatywna istotność zmiennych dla najlepszego modelu do 10 minuty.

Obraz zawierający zrzut ekranu, linia, Prostokąt, design

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

Jak można odczytać z rysunku 44 najważniejszymi czynnikami we wczesnych fazach gry są zmienne dotyczące zasobów takich jak stwory czy złoto. Wartym zauważenia jest również fakt, że waga zdobytych stworów jest różna w zależności od pozycji. Co potwierdza poniekąd słowa autorów publikacji „E-Sports Player Performance Metrics for Predicting the Outcome of League of Legends Matches Considering Player Roles”, którzy stwierdzają, że waga statystyk graczy między pozycjami nie jest równa. Na podstawie najważniejszych cech dla tego modelu można wysunąć tezę, że aby w początkowym etapie meczu zwiększyć szansę na zwycięstwo, powinno skupić się na zdobywaniu zasobów w postaci stworów oraz możliwie jak największej ilości zabójstw i asyst. Trochę odmienną sytuację można zauważyć w przypadku obserwacji pomiędzy 10 a 20 minutą meczu.

Rys. 45. Relatywna istotność zmiennych dla najlepszego modelu od 10 do 20 minuty.

Obraz zawierający zrzut ekranu, linia, Prostokąt, design

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

Z rysunku 45 możemy odczytać zdecydowaną przewagę w relatywnej istotności dla zmiennej **gold** w porównaniu do pozostałych zmiennych. Statystyki dotyczące KDA oraz stworów w dalszym ciągu były dość istotnymi predyktorami zwycięstwa, ale ich rola była mniejsza niż gromadzonego złota. W tym etapie meczu możemy również zauważyć wzrost istotności zmiennych dotyczących neutralnych obiektów, a także kompozycji. Wyższa waga zmiennych dotyczących kompozycji Senna Tahm Kench oraz Lucian Nami nad kompozycją Zeri Yummi również wydaje się stać w zgodności z logiką, ponieważ Lucian Nami jest kompozycją skupioną na wczesne bądź środkowe etapy rozgrywek, a Senna Tahm Kench na środkowe bądź późne, gdzie zestawienie Zeri Yummi osiąga punkt mocy w późnych fazach rozgrywek. W przypadku obserwacji powyżej 20 minuty określenie najważniejszych składowych zwycięstwa jest bardziej skomplikowane niż w przypadku innych etapów rozgrywek. Tym razem należy wziąć pod uwagę wyniki otrzymane z dwóch modeli.

Rys. 46. Relatywna istotność zmiennych dla najlepszego lasu losowego (po 20 minucie).

Obraz zawierający zrzut ekranu, linia, Prostokąt, design

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

Rys. 47. Relatywna istotność zmiennych dla najlepszego modelu XGBoost (po 20 minucie).

Obraz zawierający zrzut ekranu, linia, Prostokąt, design

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

Model lasu losowego (rysunek 46) jako najważniejsze predyktory wskazuje zmienne dotyczące stworów oraz statystyki KDA, co może mieć podłoże w tym, że w raz z upływem gry, czas potrzebny do odrodzenia się po śmierci jest coraz dłuższy, a więc nawet jedna bezsensowna śmierć może doprowadzić do porażki. Z kolei podejście prezentowane przez model algorytmu XGBoost (rysunek 47) jest zupełnie inne. Jako kluczowe warunki zwycięstwa wskazuje głównie zmienne dotyczące kompozycji oraz smoków. To podejście również jest do wytłumaczenia, ponieważ synergie między postaciami są rzeczą, która bardzo często definiuje przebieg walk drużynowych, a smocze wzmocnienia (szczególnie te procentowe) przybierają na sile wraz z upływem kolejnych minut spotkania. Podsumowując analizę modeli z podziałem na fazę spotkania można stwierdzić, że satysfakcjonująca dokładność predykcji jest możliwa do osiągnięcia na podstawie transmisji meczowych. W początkowych etapach meczu najważniejszymi aspektami są statystyki, które w późniejszych etapach tracą na znaczeniu na rzecz zdobytych obiektów neutralnych oraz kompozycji postaci. Najlepszymi modelami dla użytych danych są bezdyskusyjnie lasy losowe oraz algorytm XGBoost.

## **5.2 Model ogólny dla całego spotkania**

W przypadku danych dla pełnych meczy ponownie okazało się, że las losowy składający się z 500 drzew oraz 20 zmiennych osiągał najlepsze wyniki. Uzyskał on pierwiastek błędu średniokwadratowego na poziomie około 0,2815 oraz dokładność predykcji równą blisko 89,7%. Można zauważyć, że jest to trochę gorszy rezultat niż w przypadku danych dla czasów powyżej 10 minuty meczu, jednak w dalszym ciągu wynik ten wydaje się być dość satysfakcjonujący. Na podstawie najbardziej istotnych zmiennych dla tego modelu można wytypować najważniejsze czynniki zwycięstwa w skali całego meczu.

Rys. 48. Relatywna istotność zmiennych dla najlepszego modelu w skali całego meczu.

Obraz zawierający zrzut ekranu, linia, Prostokąt, design

Opis wygenerowany automatycznie

Źródło: Opracowanie własne.

Na podstawie rysunku 48 można stwierdzić, że w przypadku danych obejmujących całe mecze najważniejszymi predyktorami są:

* Zmienne dotyczące różnicy w stworach

Wydaje się to dość logicznym, ponieważ za zabijanie stworów gracze dostają złoto oraz doświadczenie, które niewątpliwie pomaga w wygraniu meczu. Warto również zauważyć, że liczba stworów najważniejsza jest w przypadku gracza ze środkowej alei, a najmniej ważna w przypadku leśnika.

* Różnica w złocie

Obecność tej zmiennej również nie powinna dziwić, ponieważ została ona wspomniana zarówno w literaturze, jak i w wywiadzie z Marcinem Świechem. To, że zmienna ta nie ma największej relatywnej istotności może wynikać z faktu, o którym mówił Marcin- złoto jest tak naprawdę wynikową wcześniej poczynionych zagrań.

* Statystyki KDA

Ilość zabójstw, asyst i zgonów niewątpliwie jest kluczowym czynnikiem, który ułatwia wygranie meczu, dlatego fakt znacznej istotności tych zmiennych nie jest zaskakujący. Ponownie możemy jednak zauważyć, że waga tych statystyk różni się w zależności od pozycji- najważniejsza jest dla strzelca, z kolei najmniej ważna dla gracza ze środkowej alei. Może wynikać to ze specyfiki aktualizacji, na której rozgrywany był ten turniej, a w której to środkowa pozycja skupiała się nie na zdobywaniu zabójstw, a zdobywaniu priorytetu na linii, który ułatwiłby zdobycie obiektów neutralnych.

* Zabójstwa oraz zdobyte struktury

Te zmienne również wydają się być dość oczywistymi predyktorami zwycięstwa. Zdobywanie struktur pomaga uzyskać dostęp do bazy przeciwnika, co staje się jeszcze łatwiejsze w przypadku zabicia oponentów.

* Zmienne dotyczące kompozycji postaci

Wśród 20 najistotniejszych zmiennych znalazły się dwie dotyczące kompozycji postaci. Jedna dotyczyła kompozycji Senny i Tahm Kench-a, a druga Luciana z Nami. Oba te zestawienia zostały wspomniane przez Marcina Świecha w wywiadzie i ich istotność potwierdza rzeczywisty wpływ wyboru odpowiednich postaci na rezultaty meczu.

* Zmienne dotyczące neutralnych obiektów

W przypadku neutralnych obiektów można zauważyć, że jedynie ich część wpływa znacząco na sytuację meczową. Najmniej ważnym z istotnych obiektów jest Rift Herald. Fakt ten nie wydaje się sprzeczny z logiką, ponieważ wzmocnienie zdobywane po pokonaniu tego potwora nie jest ani na tyle znaczące jak wzmocnienie Barona Nashora, ani na tyle trwałe jak wzmocnienia zapewniane przez smoki. Jedynym co w przypadku tych zmiennych może być zastanawiające, jest przewaga relatywnej istotności zmiennej **baron** w porównaniu do smoczych dusz. Może to wynikać jednak ze specyfiki stylu gry preferowanego przez drużyny na tym turnieju, ponieważ konkretne wzmocnienia sprawdzają się różnie w zależności od kompozycji drużyny je zdobywające.

# VI. Podsumowanie

Zadanie predykcji meczy e-sportowych na podstawie transmisji meczowych z wykorzystaniem narzędzi uczenia maszynowego oraz optycznego rozpoznawania znaków jest niewątpliwie bardzo interesującym tematem, jednak stopień skomplikowania tego typu analizy jest duży. Podstawowym problemem było pozyskanie dobrej jakościowo bazy danych, co wymagało nie tylko dużej ilości mocy obliczeniowej, ale również znaczącej ilości czasu potrzebnego na ręczne korygowanie błędów w rekordach. Przedstawiona w tej pracy analiza sugeruje jednak, że przy odpowiednim nakładzie pracy nawet początkujący badacz ma możliwość zbudowania modeli, które z zadowalającymi rezultatami posłużą do predykcji meczy e-sportowych. W pracy tej zastosowano wiele algorytmów uczenia maszynowego. Niektóre były stosunkowo mało zaawansowane (tak jak regresja logistyczna czy drzewa decyzyjne), jednak pojawiły się również potężne narzędzia, takie jak algorytm XGBoost. Zarówno dla danych podzielonych na konkretne etapy meczu, jak i ogółu spotkań najskuteczniejszy okazał się algorytm lasów losowych. Na drodze analizy otrzymano też potwierdzenie dla tez stawianych w ramach tej pracy, takich jak:

* Zależność dokładności predykcji od minuty meczu,
* Nierówna waga konkretnych ról w zespole,
* Zależność wygranej od wybieranych postaci,
* Możliwość uzyskania zadowalających wyników predykcji dla danych pozyskanych z transmisji meczowych

Ponadto analiza podjęta w tej pracy ukazała stopień rozbudowania zjawiska jakim jest e-sport. Nie tylko pokazano, że sporty elektroniczne mogą być traktowane na równi ze sportami tradycyjnymi, a rynek e-sportowy jest bardzo złożoną gałęzią, ale również wskazano najważniejsze czynniki zwycięstwa w meczach e-sportowych (na przykładzie konkretnej gry). W przyszłych badaniach warto by wykorzystać większe nakłady mocy obliczeniowej, aby zbudować modele na danych z wielu turniejów (na różnych aktualizacjach, z różnych regionów itp.) lub stworzyć narzędzia, które pozwolą przewidywać procentową szansę na wygraną każdej drużyny w czasie rzeczywistym. Analiza ta wymaga olbrzymich zasobów, jednak posiada również olbrzymie spektrum potencjalnych zastosowań, takich jak zakłady bukmacherskie czy też analityka e-sportowa.

1. Wagner, M. G. (2006, June). On the Scientific Relevance of eSports. In International conference on internet computing (pp. 437-442). [↑](#footnote-ref-1)
2. https://cybersport.pl/310412/swieto-csgo-rekord-ogladalnosci-pobity-o-ponad-milion-widzow [↑](#footnote-ref-2)
3. https://online.maryville.edu/blog/different-types-of-esports/ [↑](#footnote-ref-3)
4. Mała encyklopedia sportu, tom 2, Wydawnictwo Sport i Turystyka, Warszawa 1986, s. 439. [↑](#footnote-ref-4)
5. M. Demel, A. Skład, Teoria wychowania fizycznego, PWN, Warszawa 1970, s. 18–19. [↑](#footnote-ref-5)
6. Stępnik, A. (2009). E-sport z perspektywy teorii sportu. Homo ludens, 1(1), 213-222. [↑](#footnote-ref-6)
7. Jenny, S. E., Keiper, M. C., Taylor, B. J., Williams, D. P., Gawrysiak, J., Manning, R. D., & Tutka, P. M. (2018). eSports venues: A new sport business opportunity. Journal of Applied Sport Management, 10(1), 8. [↑](#footnote-ref-7)
8. Kim, Y. H., Nauright, J., & Suveatwatanakul, C. (2020). The rise of E-Sports and potential for Post-COVID continued growth. Sport in Society, 23(11), 1861-1871. [↑](#footnote-ref-8)
9. Johannes, N., Vuorre, M., & Przybylski, A. K. (2021). Video game play is positively correlated with well-being. Royal Society Open Science, 8(2), 202049 [↑](#footnote-ref-9)
10. Źródło: https://www.who.int/news-room/q-a-detail/addictive-behaviours-gaming-disorder [↑](#footnote-ref-10)
11. Źródło: https://www.health.harvard.edu/blog/the-health-effects-of-too-much-gaming-2020122221645 [↑](#footnote-ref-11)
12. Newzoo 2022 Global Esports & Live Streaming Market Report [↑](#footnote-ref-12)
13. Bahrololloomi, F., Klonowski, F., Sauer, S., Horst, R., & Dörner, R. (2023). E-Sports Player Performance Metrics for Predicting the Outcome of League of Legends Matches Considering Player Roles. SN Computer Science, 4(3), 238. [↑](#footnote-ref-13)
14. Wielki słownik języka polskiego, https://wsjp.pl/haslo/podglad/11761/efekt-snieznej-kuli [↑](#footnote-ref-14)
15. Silva, A. L. C., Pappa, G. L., & Chaimowicz, L. (2018). Continuous outcome prediction of league of legends competitive matches using recurrent neural networks. SBC-Proceedings of SBCGames, 2179-2259. [↑](#footnote-ref-15)
16. Do, T. D., Wang, S. I., Yu, D. S., McMillian, M. G., & McMahan, R. P. (2021, August). Using machine learning to predict game outcomes based on player-champion experience in League of Legends. In Proceedings of the 16th International Conference on the Foundations of Digital Games (pp. 1-5). [↑](#footnote-ref-16)
17. Ani, R., Harikumar, V., Devan, A. K., & Deepa, O. S. (2019, May). Victory prediction in League of Legends using Feature Selection and Ensemble methods. In 2019 International Conference on Intelligent Computing and Control Systems (ICCS) (pp. 74-77). IEEE. [↑](#footnote-ref-17)
18. Lin, L. (2016). League of Legends match outcome prediction. Comput. Sci. Dept., Univ. Stanford, Stanford, CA, USA, Rep. [↑](#footnote-ref-18)
19. Costa, L. M., Mantovani, R. G., Souza, F. C. M., & Xexeo, G. (2021, August). Feature analysis to league of legends victory prediction on the picks and bans phase. In 2021 IEEE Conference on Games (CoG) (pp. 01-05). IEEE. [↑](#footnote-ref-19)
20. Zandaki, A. (2021). eSports Game Skill Analysis & Prediction: League of Legends. [↑](#footnote-ref-20)
21. Yang, Y., Qin, T., & Lei, Y. H. (2016). Real-time esports match result prediction. arXiv preprint arXiv:1701.03162. [↑](#footnote-ref-21)
22. https://lol.fandom.com/wiki/Bucu [↑](#footnote-ref-22)
23. Gracz w League of Legends grający na środkowej alei. [↑](#footnote-ref-23)
24. O którym chciałem porozmawiać [↑](#footnote-ref-24)
25. Przed wywiadem [↑](#footnote-ref-25)
26. Transmisja meczowa [↑](#footnote-ref-26)
27. Split letni LEC 2022 jest potocznym określeniem na letnie Mistrzostwa Europy w 2022 roku. [↑](#footnote-ref-27)
28. Inne określenie na Split letni LEC 2022 (patrz przypis 27). [↑](#footnote-ref-28)
29. Inaczej dolna aleja, na której gra strzelec i wspierający. [↑](#footnote-ref-29)
30. Przed każdą grą drużyny mogą wykluczyć z rozgrywki po 5 bohaterów. [↑](#footnote-ref-30)
31. Sytuacja, gdy postać jest albo wybierana, albo wykluczana w danej grze. [↑](#footnote-ref-31)
32. Zejście ze swojej linii w celu pomocy sojusznikom. Pojęcie to zostanie wyjaśnione w późniejszej części pracy. [↑](#footnote-ref-32)
33. Gracz grający w Dżungli. [↑](#footnote-ref-33)
34. Sytuacja, gdy gracz z Dżungli przychodzi na inne linie w celu pomocy sojusznikom. [↑](#footnote-ref-34)
35. Zabijanie stworów w celu pozyskania złota. [↑](#footnote-ref-35)
36. Gracze grający w alejach (górnej, dolnej, środkowej). [↑](#footnote-ref-36)
37. Faza gry na linii, gdzie gracze koncentrują się na zdobywaniu zasobów. [↑](#footnote-ref-37)
38. Inicjacja walki, najczęściej drużynowej. [↑](#footnote-ref-38)
39. Inaczej wzmocnienie. [↑](#footnote-ref-39)
40. Formuła pięciomeczowa. Drużyny rozgrywają maksymalnie 5 meczy, do czasu zdobycia przez daną drużynę 3 zwycięstw. [↑](#footnote-ref-40)
41. https://www.miejski.pl/slowo-Dymy [↑](#footnote-ref-41)
42. Inaczej Mistrzostwa Świata w League of Legends: https://pl.wikipedia.org/wiki/Mistrzostwa\_Świata\_w\_League\_of\_Legends [↑](#footnote-ref-42)
43. https://developer.riotgames.com [↑](#footnote-ref-43)
44. https://www.youtube.com/channel/UCrPE9FDBnSMX6W5K0-r7K-g [↑](#footnote-ref-44)
45. https://pl.wikipedia.org/wiki/League\_of\_Legends\_EMEA\_Championship [↑](#footnote-ref-45)
46. https://dotesports.com/league-of-legends/news/why-lucian-nami-is-one-of-the-most-powerful-bottom-lane-duos-in-competitive-league-of-legends [↑](#footnote-ref-46)
47. https://trucoteca.com/pl/kda-de-la-league-of-legends/ [↑](#footnote-ref-47)
48. https://leagueoflegends.fandom.com/wiki/Vision\_score [↑](#footnote-ref-48)
49. https://leagueoflegends.fandom.com/pl/wiki/Oko\_Herolda [↑](#footnote-ref-49)
50. https://leagueoflegends.fandom.com/wiki/Elder\_Dragon [↑](#footnote-ref-50)
51. https://lol.fandom.com/wiki/LEC/2022\_Season/Summer\_Season [↑](#footnote-ref-51)
52. https://leagueoflegends.fandom.com/wiki/Bounty [↑](#footnote-ref-52)
53. https://lol.fandom.com/wiki/Category:Champion\_Square\_Images [↑](#footnote-ref-53)
54. https://mobalytics.gg/blog/wave-management/ [↑](#footnote-ref-54)
55. https://lol.fandom.com/wiki/LEC/2022\_Season/Summer\_Season/Match\_History [↑](#footnote-ref-55)
56. https://www.esports.com/en/when-do-all-the-objectives-in-league-of-legends-spawn-277135 [↑](#footnote-ref-56)
57. Jest to dodatkowa premia w wysokości 100 sztuk złota za dokonanie pierwszego zabójstwa w grze. [↑](#footnote-ref-57)
58. https://how2play.pl/leagueoflegends/split-push-mini-poradnik/ [↑](#footnote-ref-58)
59. https://www.lolporadnik.pl/roaming/ [↑](#footnote-ref-59)