universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Data Usage, Visualization and Predictions in Esports

verfasst von | submitted by

Aleksandar Stojkov

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 921
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Informatik
Betreut von | Supervisor:	Univ.-Prof. Dipl.-Ing. Dr. Helmut Hlavacs

Abstract

In a world increasingly driven by data, it has become one of the most powerful tools shaping our decisions on a daily basis. As time goes on, data is steadily evolving into a more powerful tool, creating more opportunities in all work branches. Parallel to the rise of data, the growth of Esports is also increasing, especially in the youth. More and more people are getting involved into the professional competitive world of video games. To an ever increasing extend, enormous organizations are showing incredible interest and are constantly pushing the boundaries of Esports. As both of these are accelerating, maybe it is for data to find place in Esports as well. At the moment, with the assistance of the data that is accessible from each match, each player or organization can use the data and find the mistakes which are only visible by looking at the data. As the data which is being received from one individual match is enormous, it is also important to cherry-pick the things which are beneficial to the teams and players. While using data is a positive feature, using bad data can lead the teams towards a wrong path or even make them do mistakes in the overall preparations and strategies. The main goal of this research is to analyze already existing techniques which are used for gathering data. By processing the data, the worth of the players is being evaluated which is bringing competitive advantage over the players' opponent. One of the key aspects of the research is to predict the outcome of the future matches by using several machine learning techniques. With the help of these machine learning models, predictions are made for the results of the matches, based on gathering of historical data, the performance of the teams, as well as individual performance of the players. The goal is to train models that are precise enough and give realistic predictions, while helping the teams to see their potential outcomes and optimize their strategies moving forward. Additionally, in this research multiple machine learning models are analyzed in order to give a better picture on which models fit the best when it comes to Esports matches. Ultimately, the Master Thesis is contributing to a better understanding of Esports data and its crucial role in the modern Esports industry. With the help of analysis, the Master Thesis is giving the teams the possibility to develop more advanced techniques which should lead to better performance in the long run. Finally, the Master Thesis is serving as a foundation for further research in the field of usage of Esports data, allowing further and severely improved developments in the industry.

Zusammenfassung

In einer Welt, die zunehmend von Daten bestimmt wird, sind diese zu einem der mächtigsten Werkzeuge geworden, die unsere täglichen Entscheidungen beeinflussen. Im Laufe der Zeit entwickeln sich Daten zu einem immer leistungsfähigeren Werkzeug, das mehr Möglichkeiten in allen Arbeitsbereichen schafft. Parallel zum Anstieg der Daten nimmt auch das Wachstum des Esports zu, insbesondere bei Jugendlichen. Immer mehr Menschen engagieren sich in der professionellen, wettbewerbsorientierten Welt der Videospiele. In zunehmendem Maße zeigen riesige Organisationen ein unglaubliches Interesse und verschieben ständig die Grenzen des Esports. Da beides immer schneller voranschreitet, ist es vielleicht an der Zeit, dass auch Daten im Esports ihren Platz finden. Mit Hilfe der Daten, die von jedem Spiel zur Verfügung stehen, kann jeder Spieler oder jede Organisation die Daten nutzen und die Fehler finden, die nur durch das Betrachten der Daten sichtbar sind. Da die Datenmenge, die von einem einzelnen Spiel empfangen wird, enorm ist, ist es auch wichtig, die Dinge herauszupicken, die für die Teams und Spieler von Vorteil sind. Während die Nutzung von Daten eine positive Eigenschaft ist, kann die Verwendung schlechter Daten die Teams auf einen falschen Weg führen oder sie sogar zu Fehlern bei den allgemeinen Vorbereitungen und Strategien verleiten. Das Hauptziel dieser Untersuchung ist die Analyse bereits vorhandener Techniken, die zur Datenerfassung verwendet werden. Durch die Verarbeitung der Daten wird der Wert der Spieler bewertet, was einen Wettbewerbsvorteil gegenüber dem Gegner bedeutet. Einer der Hauptaspekte der Forschung ist die Vorhersage des Ergebnisses zukünftiger Spiele durch den Einsatz verschiedener maschineller Lerntechniken. Mit Hilfe dieser maschinellen Lernmodelle werden Prognosen für die Ergebnisse der Spiele erstellt, die auf der Erfassung historischer Daten, der Leistung der Mannschaften und der individuellen Leistung der Spieler beruhen. Ziel ist es, Modelle zu trainieren, die präzise genug sind, um realistische Vorhersagen zu machen und den Mannschaften dabei zu helfen, ihre potenziellen Ergebnisse zu erkennen und ihre Strategien für die Zukunft zu optimieren. Darüber hinaus werden in dieser Studie mehrere Modelle des maschinellen Lernens analysiert, um ein besseres Bild davon zu erhalten, welche Modelle sich am besten für Esports-Spiele eignen. Letztendlich trägt die Masterarbeit zu einem besseren Verständnis von Esports-Daten und ihrer entscheidenden Rolle in der modernen Esports-Industrie bei. Mit Hilfe der Analyse gibt die Masterarbeit den Teams die Möglichkeit, fortschrittlichere Techniken zu entwickeln, die langfristig zu einer besseren Leistung führen sollten. Schließlich dient die Masterarbeit als Grundlage für weitere Forschungen auf dem Gebiet der Nutzung von Esports-Daten, was weitere und stark verbesserte Entwicklungen in der Branche ermöglicht.

Acknowledgments

As someone living in a time where the digital era is expanding and who enjoys playing and watching video games, it brings me great joy to work on such a unique subject that connects these two interests. I hope this research serves as both a purpose and motivation for future studies to contribute to the ongoing evolution of digitalization and the visualization of Esports matches. I want to start by thanking Univ.-Prof. Dipl.-Ing. Dr. Helmut Hlavacs for his guidance and valuable insights throughout this journey. His support has helped me refine my ideas and stay on track and I truly appreciate the time and effort he dedicated to my work. A huge thank you also goes to my family, who have been my biggest source of encouragement. Their constant support by simply being there has made all the difference. I could not have done this without them. I am also especially grateful to my partner for her endless support throughout this process. Her encouragement has meant the world to me, and I truly appreciate everything she has done to keep me motivated. Lastly, I want to thank everyone who took the time to participate in the survey. Your input was essential for this research, and I truly appreciate your willingness to share your thoughts and experiences.

Contents

1 Introduction **6**

2 Related Work **9**

3 A Closer Look at the Process of the Preliminary Study **12**
3.1 Preliminary Study Structure 13
3.2 Preliminary Study Results 14

4 Technologies Used – Escore Web Application **20**
4.1 Backend 23
4.2 Database 27
4.3 Frontend 31
4.4 Connecting Machine Learning models with UI 33

5 Machine Learning Models **35**
5.1 Dataset 37
5.2 Logistic Regression 38
5.3 Random Forest 41
5.4 Gradient Boosting 43
5.5 Naive Bayes 47

6 Evaluation **51**
6.1 Data Usage in League of Legends 52
6.2 Data Visualization in League of Legends 56
6.3 Comparison Between Machine Learning Algorithms 59

7 Conclusion **60**

References **62**

1 Introduction

In the last decade, Esports has evolved from a niche phenomenon into a global industry with millions of players, spectators, and enthusiasts [3]. It can be said that Esports is becoming an important part of the daily lives of millions of individuals, especially individuals coming from the younger generations. Due to the raid and accelerated advancement of various technologies, as well as the broad availability of Internet connection, Esports is attracting a huge amount of public, no matter whether the public comes as viewers, competitive players or even investors [8]. Since Esports is dating as a relatively new sport, the amount of things that can contribute to an improvement of the performance of a certain team are yet unknown. With all that being said, the need of analysis and utilization of data which is being extracted from the competitive matches is growing daily.

The data plays a major and very powerful role in Esports, since each competitive match is generating an enormous amount of it, from individual statistics of the player to further on global trends of the selected game. Performing analysis on the data collecting is allowing overall better understanding of the team's competitive style of playing and it is uncovering the characteristics of the team, both strong and weak. Once the data is collected and processed, that same data can be used by the team to strengthen the weaker aspects and to accelerate the strong links. Additionally, the models and methods used for analysis can show the teams objectives which are not really noticeable just by watching or by playing directly the competitive match [25]. With the accumulated information and knowledge, the teams can start applying some methods which are directly given to them based on the processing of the data from the previous matches.

In addition to the data being used for analysis purposes, the usage of data visualization methods can play a major role in presenting a complex information into an intuitive and easy understandable way. Certain data visualization methods like intuitive graphs, heat-maps and movement patterns could be of value and provide help in order for a better and more efficient decision making to be done in real time. Methods like these bring value to the professional teams, to the individual players and also to the spectators who want to understand the game on a whole another level [51]. The usage of both data analysis and data visualization is not only limited for usage by the individual players and teams but it is also available to the industry itself. Organizations, corporations and companies which are organizing these types of competitions, as well as streaming platforms which stream the matches played in the competitions could also use the analyzed data in order to improve the viewing experience and to bring even bigger crowd on their respective tournaments.

Advances in the field of Esports data processing could not only help and improve the performance of the professional teams but it could also increase the interaction created with the audience, while making the industry even more dynamic and competitive.

In this research, special attention will be paid to the prediction of the results of an Esports competition, which is one of the most important aspect of data analysis performed. With the help of various machine learning algorithms which will be used, it would be possible to predict the results on a basis of historical data, momentary trends and certain factors such as individual and team performance, choice of tactics and even the psychological preparedness of the players on the field. Given the continued growth of the Esports industry and the increasing role of analytics, extensive research in this area is crucial for the future of the Esports industry itself. Additionally, as a part of the Master Thesis research a web application will be developed which could serve as a bridge between the viewers and the competitive scene. The spectators can easily view the upcoming and the previous matches. On top of that, with the help of the machine learning models used in the research, there would be a prediction given for each match with an individual certainty percentage. Interesting to note here would be that even for the previous matches there would be a prediction given, so the user could see how often the models presented in the web application are accurate compared to the results. If the users are curious how the teams are doing in the regional leagues, the web application would contain a complex graph which would depict the current standings and the score of each team in that particular regional league. Since Esports is a general term and it exists in many genres of games, our primary focus and target will be the game called 'League of Legends' which is a five versus five multiplayer online battle arena game.

The predictability of Esports matches is a subject of ongoing research and debate within data science, sports analytics and competitive gaming communities. In contrast to traditional sports, where physical capabilities are often the deciding factor, Esports matches are influenced by in-game mechanics or mechanics which are done in the game by the players. Throughout the research, the question *Are Esports matches predictable?* will be explored using multiple techniques and incorporating knowledge gathered from professionals during the preliminary study.

One of the primary methods which will be used in this Master Thesis involves usage of statistical modeling and machine learning models. By utilizing historical match data, game-specific attributes and player performance metrics, a concise machine learning model can be constructed which will properly predict the future Esports matches. The concern which is in regards to the game-specific attributes is that there are frequent updates to the game itself, which alters the game mechanics. A statistic that once clearly indicated a team's victory may no longer be accurate in the current

patch.

Another important factor worth mentioning is the team dynamics. In traditional sports, teams train over extended periods and usually executing the same execution plan. In Esports, the teams often undergo roster changes that significantly impact their performance. In addition, the mental state and adaptability of the players play a crucial role, as Esports competitions demand high levels of concentration and strategic thinking done under pressure.

A key factor which determines the success of the machine learning methods is the choice for a suitable data set. The quality of the data itself is directly affecting the precision and the applicability of the models. If the data set is containing a bias or a missing values, the model can learn wrong patters and therefore create incorrect predictions due to being fed the wrong data. With a careful selection of a data set, which should be properly labeled, the success of the machine learning model would go higher and the same model could apply the predictions successfully on competitive matches which are played in real-world. After the data set has been received, the data should be well cleaned and preprocessed, since the unnecessary noise could decrease the overall performance of the model itself, independent of which methodology was used in the process of training.

For the purpose of this research project, a dataset from Oracle's Elixir was chosen and therefore used [46]. Oracle's Elixir is offering high-quality data which is being collected from official Esports matches, meaning it ensures reliability and accuracy. This data set is covering various potential features which can further on be used, including player statistics, team compositions, gold differentials and objective control which could be of a great benefit for the models. Since it is offering a huge spectrum of potential features which can be used in the models, the data set can be seen as a perfect fit especially tailored for this research. By having a huge amount of potential features, with an incremental changes into the models, we can precisely measure the accuracy, precision and F1-score of each machine learning technique which is used in the research. This dataset provides granular information at both team and player levels such as: **Early-game indicators** (gold, experience, kills at 10/15 minutes), **Mid-to-late game performance** (teamfight success rate, objective control, vision score), **Champion and player-specific data** (win rates, synergy, counters), **Match metadata** (patch version, tournament stage, region), which are just a small number of features that the dataset offers.

To enhance readers' understanding of key terms used throughout this research paper, Table 1 provides a description of each keyword and its meaning. This ensures that the reader has additional information about the abbreviations used in the study.

At the end, absolute predictability remains elusive. The dynamic nature of games

Abbreviation	Description
XPDiff@10	Difference in experience at 10 minutes
GoldDiff@10	Difference in gold at 10 minutes
CSDiff@10	Difference in creep score at 10 minutes
WR	Win Rate
GD	Gold Difference
VS	Vision Score

Table 1: League of Legends Abbreviations and Their Descriptions

together with human decision-making and external variables, ensures that upsets and surprises remain an integral part of competitive gaming. The purpose of this Master Research is to bring the viewers closer to the competition itself. By having data visualization, the users would be able to create a better picture of the overall current game state. Additionally, by analyzing the available data and using it to build machine learning models, there would be development of a model with the highest possible success rate. The goal is to show to the interested parties that to some extent, predictability in Esports is indeed possible.

2 Related Work

For the last few years, it can be seen that the advancement of Analysis in Esports matches has accelerated drastically forward, especially by integrating machine learning models into the prediction of competitive Esports matches. The measurements of the game play of the competitive players has improved as well. Different studies have contributed to understanding of the statistical significance of game events, the influence of team composition and the predictive power of historical data. In this chapter, various related researches will be mentioned and how they are all relevant in regards to the Master Thesis. Each of the papers mentioned in the chapter have greatly contributed to the overall quality of the Master Thesis. In the current time, it is quite important to have a big amount of peers in order for the writer to be able to compare the results with various papers. Additionally, the writer could deduct what could be added in the future as future work in the research that it is performed.

By increasing the interest in Esports, the amount of research which was done in terms of competitive matches increased and using the data which was produced also increased. There are some research papers which provide a historical overview of competitive gaming and emphasizing the evolution of statistical methods and

their application in the competitive gaming environment [8] [3]. As both competitive games and statistical models continue to evolve and improve in quality, it can be said that the performance of machine learning models predicting Esports matches has also significantly improved. Additionally, some of them highlight the importance of data science in analysis of Esports, illustrating how predictive modeling and machine learning techniques can improve decision-making processes in Esports [25] [27]. By actively having these machine learning techniques, both observers and competitive players can benefit—enhancing the viewing experience for the audience and improve the play making strategies of the players in the game League of Legends.

Some studies investigate the validity and safety of the statistics of competitive matches in League of Legends, adding importance to the validity of the data itself [31]. Additionally, it is of great importance to mention the work that Tim Sevenhuysen has done throughout the years in regards to the data. With his work, the majority of the data for each competitive match of League of Legends is made available to the public [46]. By having this amount of data, researchers who are not affiliated or working with League of Legends, could use the exported data and create their own applications, whether that is visualization of the data, predictions of upcoming matches or purely usage. It is of great importance for the data to be available to the general public as well, in order for the development of these types of applications to evolve even further.

The usage of machine learning algorithms in order to predict Esports competitive matches has gone up tremendously in the previous years. A research paper explores statistical learning techniques for predicting Esports matches, using features such as gold differential, experience differential and team composition [7], similarly as it was done in this Master Research. In addition to that, others develop machine learning models to predict League of Legends competitive matches in real time, highlighting the effectiveness of Logistic Regression and Random Forest classifiers to determine the outcomes of the matches [24].

Some of the mentioned studies bring machine learning models which can help the teams even before the game starts. Some of them introduced a system which gives suggestions during the drafting process [11] [23], stating that machine learning models can be used even during the preparation stage. Several research efforts are exploring alternative machine learning techniques for prediction in Esports. Some use a Random Forest method to identify key metrics that influence the outcome of matches in Rocket League, another popular Esports game [45]. Furthermore, there are research papers that explore the predictive analytics in LoL tournaments, comparing different supervised learning models in terms of their accuracy in predicting match outcomes [51]. By analyzing these articles, it is clear that each supervised model has

its own advantages and disadvantages, depending on the type of data it is used on whether that varies by game type or the specific purpose of the model.

As already mentioned, visualization plays an important role in Esports as well. By having different types of visualizations, the teams and players can see some major potential mistakes which are happening often. There are research papers that explore data visualization methods in League of Legends, showing how in-game metrics can be effectively represented to improve strategic decisions [12].A key aspect of Esports analytics is understanding the player behavior and the performance trends. One of the most important features to be examined is the player behaviors patterns in online competitive games, highlighting the predictive value of combat statistics and in-game actions [38].

In order to have a fruitful application, some explore layered architectures for software development in competitive gaming applications, providing insights into scalable data processing methods [48]. The layers mentioned throughout the research and the external sources could be divided as: Presentation Layer, Business Layer, Persistence Layer and Database Layer. Microservices are quite powerful type of architecture in the present world by having flexibility and weakly associated connections between them [44]. In addition, there are researchers which investigate how a layered architecture can be safe with the combination of technologies used in this research [19]. In order to have a communication between the visual layer and the non-visual layer in the application implemented, the methods used are already investigated and analyzed from other researchers as well [4].

Choosing which type of database to use, many different researches were investigated and analyzed in depth. On one hand, we have the relational databases which are more popular and more broadly used with mentioned advantages and disadvantages [2] [36] [13]. On the other hand, we have the non-relational databases which are becoming more and more used throughout the industry [18] [13]. In order to decide which database to use in the Escore application, it was necessary to go through the mentioned research papers. With the knowledge gained and the advantages and disadvantages presented in [13] [30], a well-informed decision can be made to leverage the strengths of the two most prominent types of databases.

For the visual layer of the application or the layer which can be seen by the users a deep research was done in order to decide and asses which technology to be used. Many researches focused on which technology/library has advantages for which type of application a developer develops [22]. Many researches are leaning towards a library called ReactJS which is quite popular to the frontend community [9] [35]. Additionally, in order for the library to work properly and have additional features which could not be done purely by ReactJS, many additional dependencies are needed

and therefore, some researches focus solely on investigating components libraries which can be added to help ReactJS in order for the developer to develop the things which are needed [40].

The application of machine learning and data visualization lays a strong foundation for predictive modeling in competitive gaming. The integration of supervised learning techniques improves the ability to analyze and predict the outcome of matches with high accuracy. Building on these existing studies, this research aims to refine predictive models for League of Legends matches, leveraging structured datasets and machine learning methodologies to improve the accuracy of predictions and provide useful insights for both teams and analysts.

3 A Closer Look at the Process of the Preliminary Study

In order to prepare and to fully understand what features should be selected for the training of the models, it was decided to conduct a research survey through a research questionnaire where a certain number of data engineers were asked about what features they use in their daily work. With this research questionnaire, a strong basis and a good feature comparison could be applied when constructing the machine learning models. Additionally, the participants in the questionnaire were given access to the web application in order to be able to give an answer to the questions which relate to the web application itself. With these types of questions, it can be seen whether the contestants are leaning towards a positive or a negative experience. Since the research evolves around data which is being used in the League of Legends competitive matches, the participants which were asked to answer the survey were selected based on their area of expertise and thinking that they have knowledge of working with both data and the game League of Legends.

To ensure a diverse dataset, the research survey was sent through various channels. Some of these channels were related to professional networking platforms, other to social media groups and to direct invitations being sent to experts in the industry. Additionally, the research questionnaire was structured in a way where it would minimize the bias. By minimization of the bias, it was ensured that the answers would be coming from a real-world usage and not from personal preference. With that approach, it was aimed that the received answers can truly help in improving the machine learning models and could help in the process of selection of used features. Furthermore, the research questionnaire contains questions which could give answers in regards to the design and the usability of the web application, which can help for

the improvement of the overall user experience. As Esports which became popular and famous in the last twenty years, it is important to note that the majority of the participants or exactly 71,4% answered that they are below thirty-five years old, which can also be seen on Figure 1.

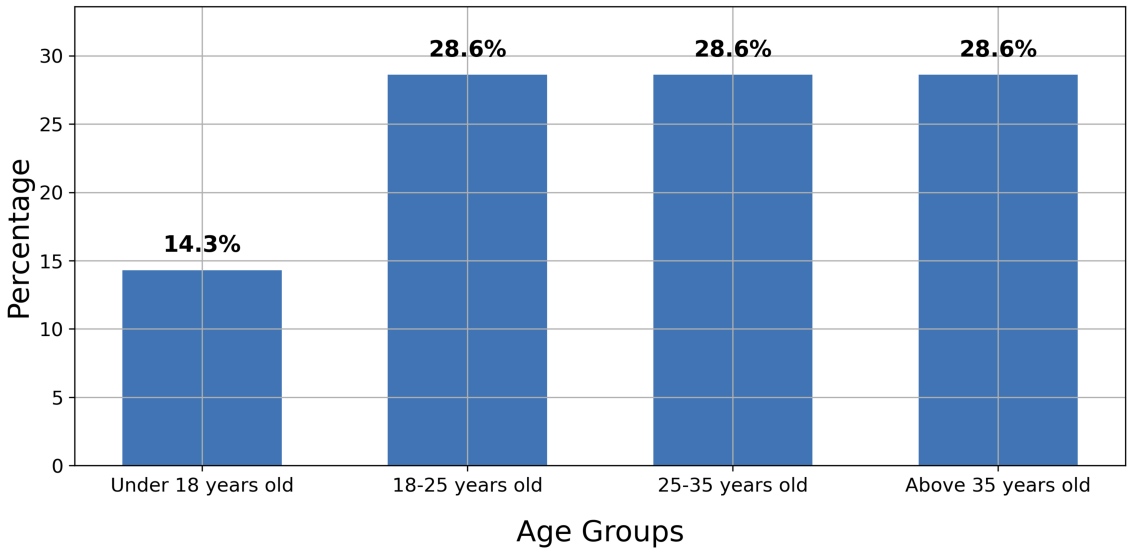


Figure 1: Research question regarding age of participants

3.1 Preliminary Study Structure

One of the most important phases before any implementation is to conduct a proper research of the topic which is going to be implemented. The research can be conducted as review of other literature research provided from other researchers on the topic and by the creation of research survey in a format of a research questionnaire. The conduct of the research survey is a must before the start of any development or implementation as it is important to understand whether the solution which is targeted would align with the results received from experts in the area and from future users. Such a research survey could only serve as a greater benefit for the development of the application, as based on the answers the positive features would be acknowledged and further used in the development phase.

In order to conduct a successful research survey, the questions provided in the research questionnaire should be concise, easy to comprehend and have one meaning. The questions should be short in order for the participant to give a straight answer

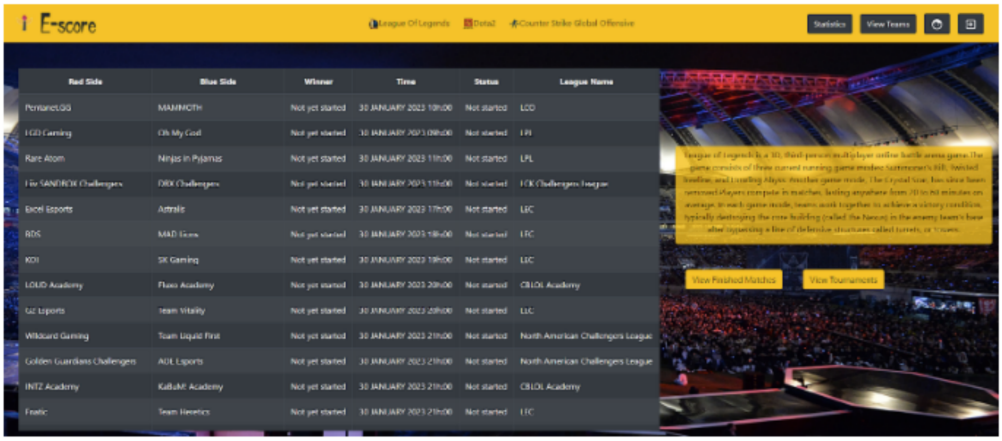
and the questions should be formulated clearly in order to not cause confusion. The research questionnaire in total consists of fifteen questions. Six of the questions are closed questions meaning the answers for the questions are given as already predefined options. The remaining questions, nine of them, are questions which are open, meaning that the participants had the chance to write their own answer as an answer to those questions. This mix of question types was intended as a possibility to keep the participant more engaged with the research questionnaire during the answering phase. Additionally, the purpose behind every question is made to be different in order to gain more significant data input from the participants. The questionnaire contains questions related towards demographic information like specification of the age, area od expertise in Esport and the work position of the participants, then specific questions relating to statistic methods and usage, followed by questions asking for opinion on planned features to be developed in the application and finally question relating to data visualization strategy. What was intended with the research questionnaire is that a diverse set of participants would be attracted, as the target audience in the research are passionate lovers and admirers of the Esports industry.

Furthermore, the questions were explained one after another to all of the participants before the start of the evaluation, in order for the participants to have a clear mind when writing or when ticking an answer. Four of the questions were marked as not mandatory and their intention was to be answered only based on the meaning of the previous question. For example, there is a question induced in the research questionnaire which is asking whether the participants think that the dynamic graph of showing the history of the teams in the regional championship leagues would bring a certain benefit to the participants when they are using the application. If the participant answered positively, then the participant could answer the next question which relates to what would the benefit be. In Figure 2, there can be seen one example of a conditional question from the research survey where the participant is being asked whether the implementation of a 'Predict' button located next to each match would be beneficial inside the Escore application.

3.2 Preliminary Study Results

Even though the research survey contains only fifteen questions, the questions were formulated to be concise and to bring the most value for the further research and development. The initial intention behind the research questionnaire was fulfilled, as the participants provided meaningful data which could be used for the future phases of development, implementation and research. Based on the answers provided, it can

Below, you have a picture of how a user sees the match results page. Would you be interested to see a 'Predict' button next to each match in order to see the prediction from our side? *



- Yes
- No
- Maybe

If you selected 'Yes', please state why.

Short answer text

Figure 2: Conditional Question used in the research questionnaire about visibility of predictions on the match overview

be understood that the research survey attracted participants which are coming from different working backgrounds which is meaning that the profile of the participants is very diverse. This was noticed from the answers on the question which was asked towards the profession which the participants currently do. As it can be seen on Figure 3, there is a mix of a lot of answers for this questions, where participants work as analysts, streamers, competitive as well as retired players, team managers, data engineers and many more positions.

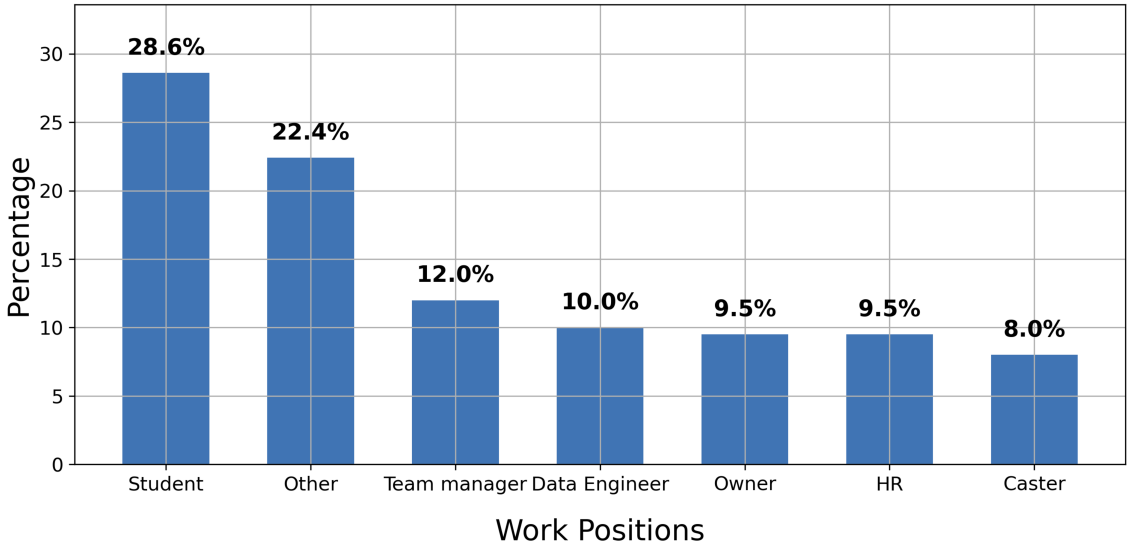


Figure 3: Current positions of work of the participants

As it was already expected and speculated towards the age question, most of the participants which were included in the research questionnaire were between twenty-five and thirty-five years old, while some of the participants were older than thirty-five and only a few of the participants were younger than eighteen years. This is positive insight, as it means that the age is not the most relevant factor and that the target audience should not be targeted based on the age, due to almost all ages being present as viewers of Esport matches. When the participants were asked which game statistics they considered most relevant, the majority of the participants mentioned the following statistics:

- **Gold difference at 10 minutes**
- **Experience difference at 10 minutes**

- **Team composition**

Some participants as an answer stated that players history and the players past performance can be an important factor for prediction calculation of match results. The opinions whether these parameters could be used as a basis for match predictions were mixed. Some participants felt that these parameters provided valuable insights, while other participants were skeptical about their reliability. Several participants expressed their interest in a 'Predict' button that could have a function to generate automated predictions for upcoming matches. There were mixed feelings on the topic of Data Visualization which could be answered based on the analysis of the answers. Many of the participants were not interested in the champion win/loss and pick/ban percentages, as it can be seen on Figure 4. The answers show that there is a division between the thinking, where there is also a percentage of the participants which answered with a 'Maybe' response, while those participants which answered with 'Yes' explained in the next question that it would be beneficial to include such percentage in order to identify a powerful champion.

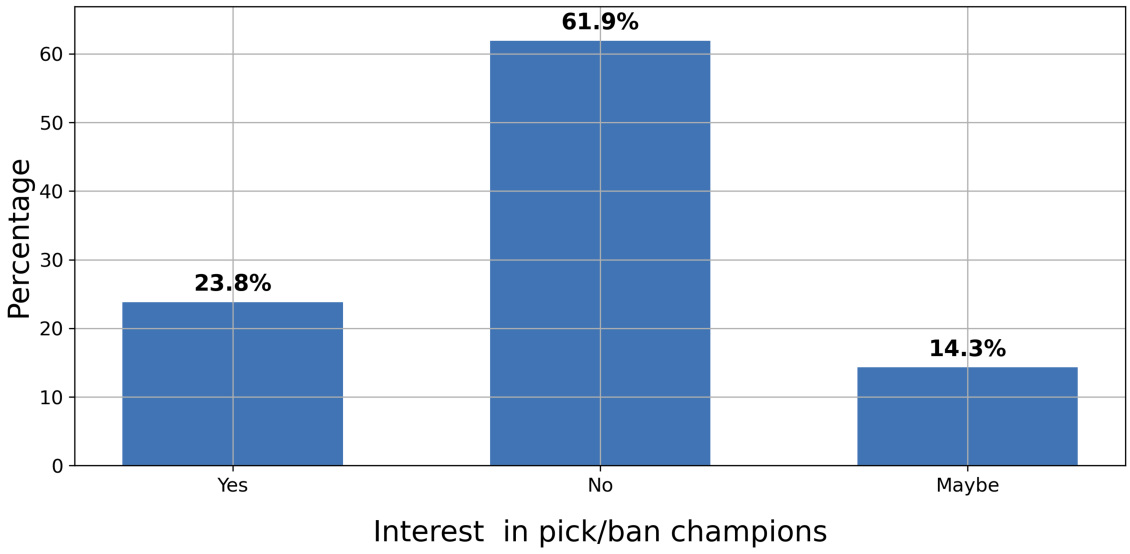


Figure 4: Question in regards to interest pick/ban champions

On the other hand, other participants found value in tracking previous match ups between certain teams, especially to be able to identify the potential and existing rivalries between teams. Some participants suggested some improvements in regards to the visualizations, such as tournament brackets, detailed user information, deeper

information about champions and more in-depth event breakdowns. When the participants were asked why they believe the statistics methods they already chose as part of another question would be very beneficial to be implemented in the Escore application, a lot of the participants gave an answer that they believe the statistics methods would be very helpful and that if you continuously win, the statistics would be positive. Some of the answers given for this open question can be seen on Figure 5.

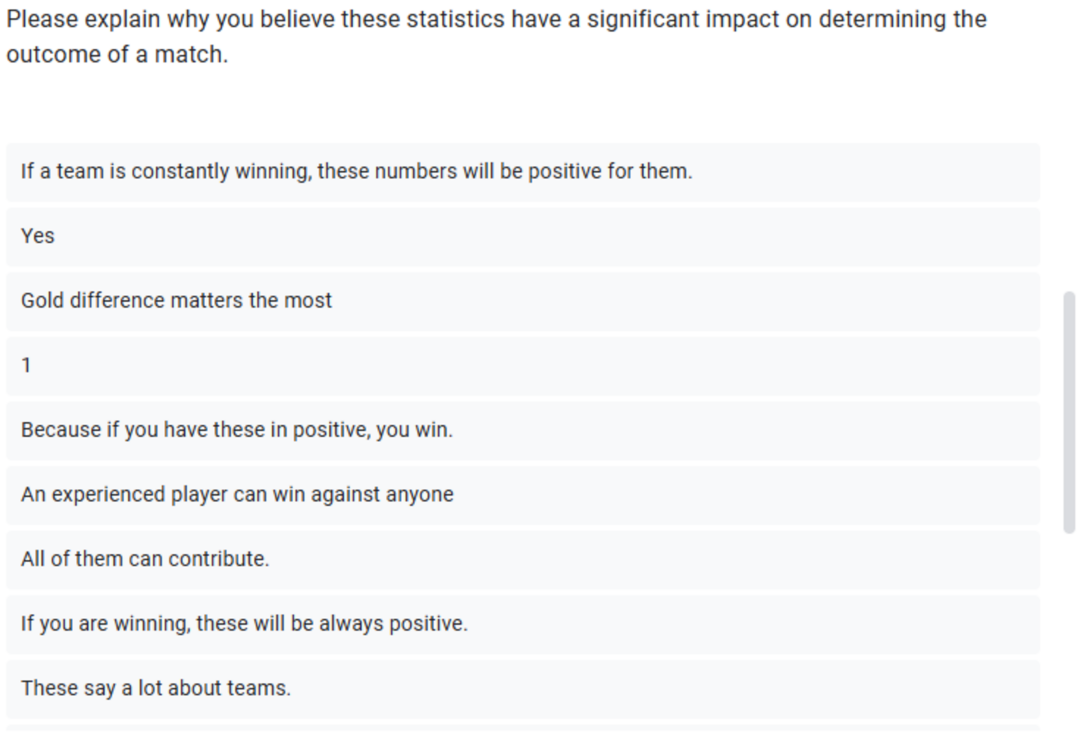


Figure 5: Answers on research question why statistics would be beneficial

Overall, the research survey results highlight a strong interest in the Esports data, particularly in match predictions and history of the teams performance. While there were different views on the accuracy of the predictive data, many participants saw potential in improving data visualization tools in order to offer more engaging insights for both Esports fans and industry professionals. The big interest for engagement in the research survey is suggesting a growing demand for advanced analytics that can provide a deeper understanding and precise analysis within the Esports field. Additionally, the participants included in the research survey emphasized the importance of user-friendly interfaces, which would mean that the accessibility and easy interpretation

of the data shown could further maximize the value of such data oriented Esports applications. By having more and more these type of applications, which are rich with data oriented features, Esports could become even more interesting topic for people who are not even included in Esports at that particular moment. Since, in general, competitive games have a lot of data for each game played, it could also draw people who are interested in the data itself and not the actual game. By having these people interested in the overall Esport, it could bring greater value to the overall health of Esports or competitive games in general.

Additionally, it can be seen on Figure 6 that the number of articles or researches done on this topic is increasing year by year. That means that the topic of Esports, whether the user is coming from a viewer perspective or the user is actively included in the games is becoming more and more popular.

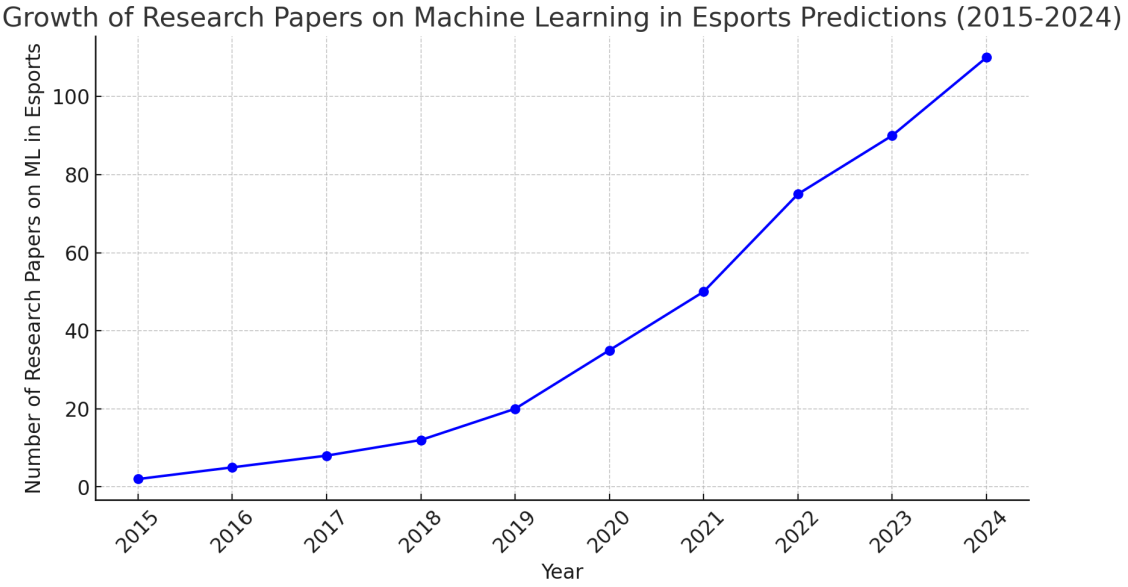


Figure 6: Growth of Research Papers on Machine Learning in Esports Predictions

By becoming more and more popular, the quality of the overall branch could be greatly improved even further in the upcoming years. Additionally, it means that the competitive players could extract some valuable information in order to become better. With this, the quality of one competitive match could increase as well.

4 Technologies Used – Escore Web Application

As part of the Master Thesis research, it was decided to enlarge with new features and possibilities the web application called 'Escore' which was created by the author [5]. The purpose of the enlargement would be to add new features which would show the power and effectiveness of the data usage and data visualization. The web application is serving as a bridge between the spectators and the ongoing tournaments which are happening across multiple games. Upon entering the application, the users are shown an entry screen on which they can choose which game they want to interact with, as it can be seen on Figure 7.

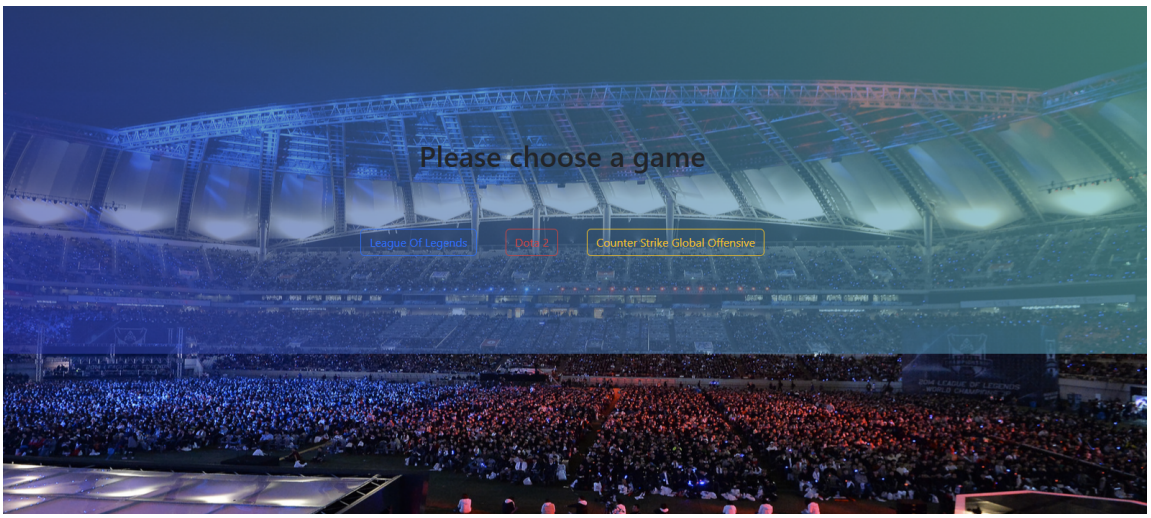


Figure 7: Entry screen in the Escore web application

Once a game has been chosen, the user has the ability to discover more about the competitive matches in the current game the user has chosen. One of the features of the application is to view future competitive matches in the selected game. With this feature, the user can constantly be up to date with the current matches played in the game that it is chosen. If a user misses a match which happened previously and the user wants to check the results of the missed match, there is a dedicated section for past competitive matches which is available inside the application. Additionally, the currently played tournaments can be viewed in the application as well, including all the information regarding the place, the date and the designated competitive prize pool. Furthermore, made for some newcomers, there is included a brief description and information in regards to the game that they are currently viewing on the screen in the application. This would allow the user to effectively plan if they have interest

in watching a specific tournament. For each competitive match which was played, there are several columns which are displayed a table which is visible in the screen shown in Figure 8.

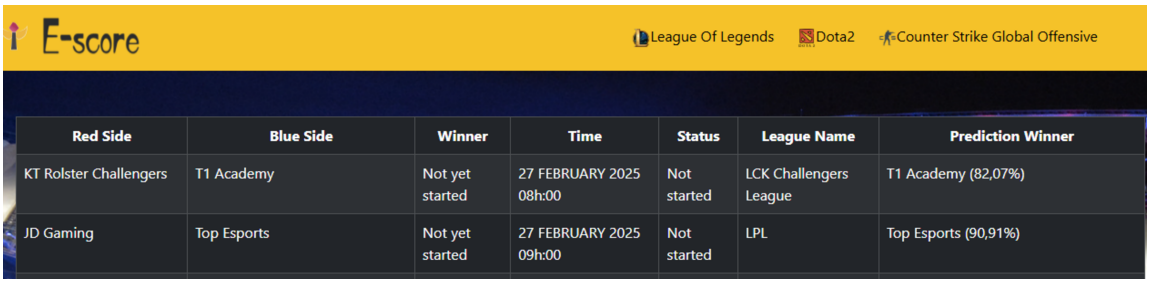


Figure 8: Web application feature for predictions of winners

One of the main features of the application is that one of the columns displayed in the table is the column called 'Prediction Winner', which fetches values from the machine learning model used in the web application, which would further on be explained in the research. With this feature, it could be interesting to the spectators or the users using the application, to see whether their favorite team is favored to win the upcoming game. On the other hand, maybe they did some analysis on teams and want to compare whether their intuition and analysis is correct and that the team they predicted to win, our models also predicted as well. Nevertheless, this feature brings a lot of value to the daily users of the application, whether they are just curious on the result itself or they have something other in mind. Additionally, the users can check whether our prediction model was correct for past matches. With that, they can see how accurate is the model itself.

On the other hand, data visualization seems to be important in every step of every application and for its success. Time after time, it was proven that data visualization plays a key part in Esports application as well. Data visualization has been shown to be essential and highly effective in enriching the experience of the spectators engaging in Esports [33]. With the help of data visualization methods, the spectators can feel like they are more included in the current state of the match and therefore making the applications and streaming services more user-friendly. By analyzing or inspecting the various graphs/numbers, the spectators can easily see which team is currently taking an advantage. Furthermore, by looking at the number presented, the spectators can have a hint and predict what the next objective would be for the team they want to see [24].

Nevertheless, data visualization methods can be helpful for the competitive organizations as well. In League of Legends, usually two teams are playing best-of-five

matches, which means whoever gets first to three wins, wins the match. In this scenario, the picking draft of the champions and the logic of the draft is done directly on the spot. The coaches of the teams are well equipped with the knowledge required to process the data which is extracted from the previous games in that particular series. In the pauses between the games played, the coaches are using the data which is prepared for them in order to gain insight and knowledge for the next match. With the usage of this data, the coaches can decide which pick of champions was a correct one and which one was not. The drafting of the champions is going in a scenario which is called snake draft scenario. Snake drafting is a type of drafting where each coach has one pick per round, and the picks go in a specific predetermined order. After a round is over, the following round is the reverse order of the previous round. So after one coach selects a champion, the other coach needs to select two champions and afterwards it is reversed, the first coach which selected one champion in the first round, now selects two champions. One of the key aspects of the data visualizations comes here into play, when the coaches need to draft the last champions. Since they know the majority of the opponent's champions, they can do a smart decision and choose a champion which counters them. This type of champions can be purely found by using statistics [23].

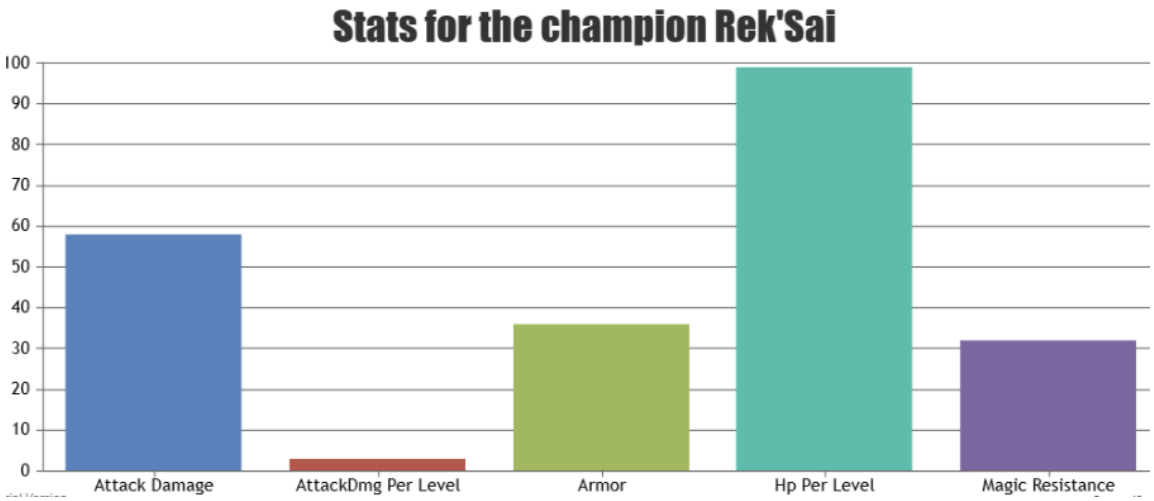


Figure 9: The individual statistics for champion Rek'sai in Escore

To assist the coaches, a feature has been developed inside the web application which would allow the selection of a champion and the display of the statistics of the champion. With this feature, the coach can select a counter champion purely based on the statistics of the enemy champion. By choosing a counter-champion, the

coach can gain a competitive advantage even before the game starts. Additionally, each team can ban five champions before the draft begins. This feature can help in the discovering of champions that are too powerful at the moment. Therefore, if the coach thinks that a champion is overpowered based purely on the visualization of the data of the champion, the coach can decide to ban that champion. The visualizations of this type of data can be found in the Figure 9. The famous statistics in regards to a champion in the game League of Legends are Attack Damage, AttackDmg Per Level, Armor, Hp Per Level and Magic Resistance which can be found in the Escore application.

As an another important thing, the coaches would like to know the previous history of the team they are competing against. The coaches would additionally like to know information regarding their current win streak, their current position on the ranking table and other relevant performance metrics. With all these data, the coaches can decide which type of draft they want to take part in. For example, if the team is currently not performing well, they can take more aggressive champions and finish the game faster. On the other hand, if a team is performing well, the coaches can opt to draft more safer picks, in order to wait for the opponent to make mistakes and later on to punish them. With numerous strategic paths available to the coach, this feature provides valuable insights to support the coaches throughout the draft process. Furthermore, the players playing the game can also be interested in the previous performance of the team that they are competing against next. By having this knowledge, the players can also gain competitive advantage throughout the game. On Figure 10, we can see the network of history of a regional league in Europe called LEC.

4.1 Backend

The backend is the backbone of the each web application. The backend manages the data, the processes the business logic and it connects with the frontend to make everything work as it should. In this subsection, an overview will be presented at how the backend is built, the technologies which are used for the development and how it was implemented in the Escore web application. The backend of the Escore application was developed by using Java 17, which has overall better performances, offers modern characteristics and a long-term support [39]. For a framework which eases the development, Spring Boot was chosen which is a Java-based open source microservices framework. Microservices are a type of architecture pattern which is allowing the build of services which can function independently of one other, meaning that they can be deploy independently. Each deployed service has its own process,

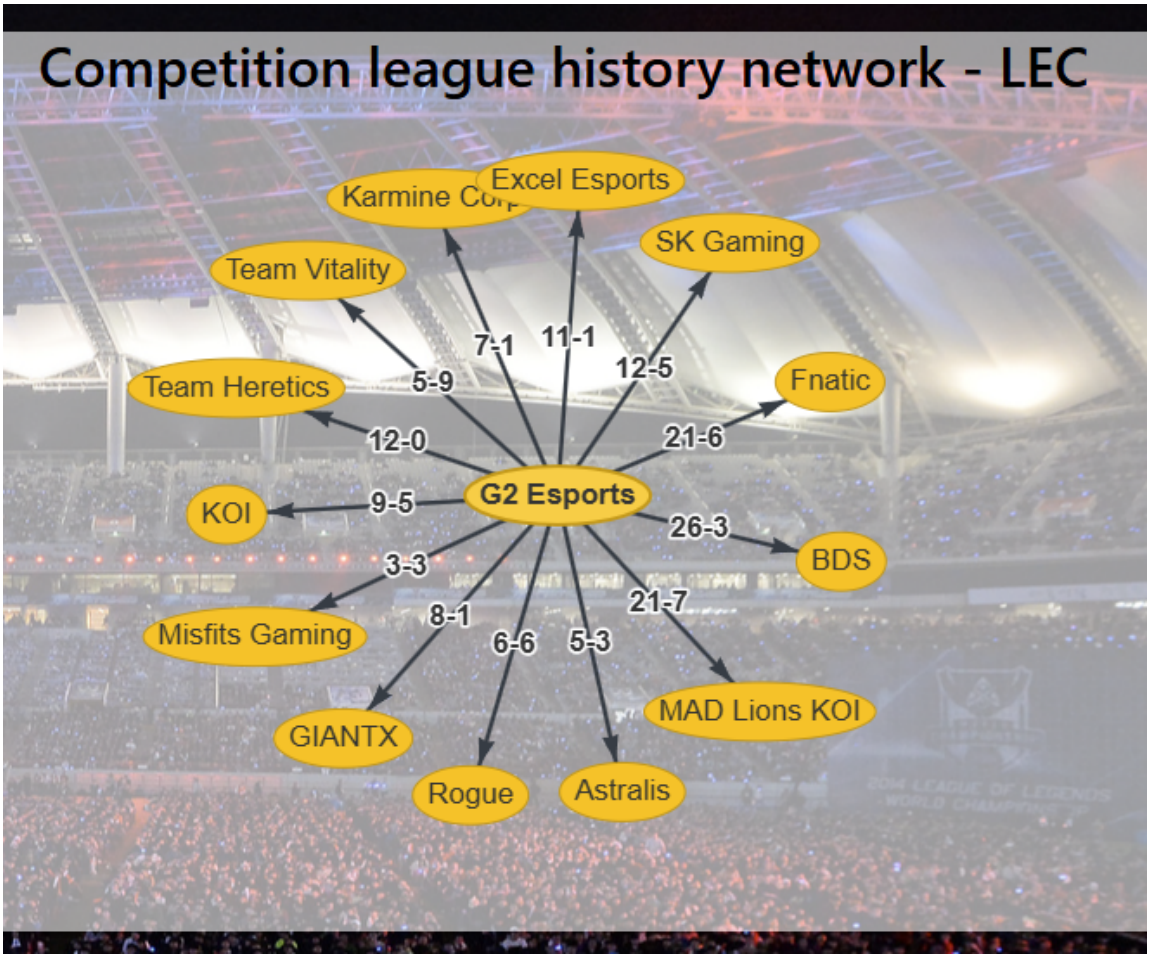


Figure 10: Historical graph for regional league in Europe, LEC, Escore application

which gives a feeling of having a lightweight application.

In this project, a **MVC (Model-View-Controller)** framework was used in order to help in the process of creation of flexible and weakly associated microservices [44]. The web application is structured as a layered architecture, meaning that it contains multiple Controller classes which handle the HTTP requests and responses. As a next layer in the layered architecture, there are the Services which serve as a bridge between the controllers and the repositories. At the end or the bottom layer of the layered architecture, there are the Repository interfaces which serve as handlers for the various database transactions. The layered architecture has several layers and it is structured in presentation layer, business layer, persistence layer and database

layer [48]. In Listings 1 2 3, the components included in the layers in the layered architecture are shown from the perspective in the web application Escore. With this examples, we can clearly see the distinction in the layered architecture.

```
@Transactional
@Repository
public interface Head2HeadRepository extends JpaRepository<Head2Head
    ,Long> {

    List<Head2Head> findAllByFirstOpponent(String firstOpponent);
    Head2Head findByFirstOpponentAndSecondOpponent(String
        firstOpponent, String secondOpponent);

    List<Head2Head> findAllByLeagueName(String leagueName);
}
```

Listing 1: Repository class, part of the persistence layer

```
@Service
public class BasicHTTPApiServiceImpl implements BasicHTTPApiService
    {

    @Autowired
    Head2HeadRepository head2HeadRepository;

    private String getPrediction(String opponentOne, String
        opponentTwo) {
        try {
            String encodedTeam1 = URLEncoder.encode(opponentOne,
                StandardCharsets.UTF_8);
            String encodedTeam2 = URLEncoder.encode(opponentTwo,
                StandardCharsets.UTF_8);
        ...
```

Listing 2: Service class, part of the persistence layer

```
@RestController
@RequestMapping("/lol")
@CrossOrigin(origins = {"http://localhost:3000","http://localhost
    :4200"}, maxAge = 36000)
public class LoLController {

  final
  BasicHTTPApiService basicHTTPApiService;

  public LoLController(BasicHTTPApiService basicHTTPApiService) {
    this.basicHTTPApiService = basicHTTPApiService;
  }

  @GetMapping("/upcomingMatches")
    public ResponseEntity<List<Match>> showUpcomingMatches() throws
        IOException {
      return new ResponseEntity<>(basicHTTPApiService.getMatches("
          upcoming","lol"), HttpStatus.OK);
    }}
```

Listing 3: Controller class, part of the persistence layer

As for the communication design between the resources, RESTful APIs were implemented on the backend side. These API calls in form of requests and responses are allowing the communication between the frontend and backend, by usage of standard HTTP methods, like GET for fetching or getting information, POST for creation of objects, PUT for updating of objects and delete for deletion of objects [4]. Every API endpoint is following the RESTful principles which is offering scalability and safety. On Listing 4, a GET HTTP request is shown which returns a HTTP response with all the upcoming matches of League of Legends game. This endpoint is called from the frontend in order to populate the table with all the necessary information.

```
  @GetMapping("/upcomingMatches")
    public ResponseEntity<List<Match>> showUpcomingMatches() throws
        IOException {
      return new ResponseEntity<>(basicHTTPApiService.getMatches("
          upcoming","lol"), HttpStatus.OK);
    }
```

Listing 4: Controller endpoint which is returning all upcoming matches of League of Legends

The safety is one of the main concerns of the architecture in backend development.

Therefore, in the Escore web application, security and safety is implemented with an Oauth 2.0 authentication, which allows the users to safely access the application. Oauth 2.0 allows authentication based on an access token, denying any access which is not containing the token [19]. In order to receive an access token to therefore use the APIs, first the token needs to be requested by using the Oauth endpoint. In the Escore API, the grant type 'password' is used. The password grant type is a way where there is an exchange of the user's credentials in order to receive an access token in return. Additionally, it is required that information for a trusted client username and a secret are provided. There are three steps for the Oauth 2.0 authentication which are the following:

1. **Endpoint with user credentials**
2. **Basic authentication needed, with the current username and the secret**
3. **Response from the API**

Once the access token has been retrieved, the user gains access to various API endpoints tailored for their use case. Of course, there are layers of privileges which are additionally implemented in the application as well. For example, an user with a role **ADMINISTRATORS** can query the active users of the application, while a user with a role **WATCHERS** can only see the Esports related data. Some of the endpoints which are related directly with the frontend are not available to the normal user. On Figure 11, all the necessary steps which are needed for a user to be able to successfully receive a token back from the API could be seen.

The backend development in the web application plays a key role in the data processing, the security and the seamless communication realized through the usage APIs. The use of Java 17, REST API sand OAuth 2.0 is allowing the application to be robust, scalable and strongly secure. These technologies collectively improve the performance of the application while providing a seamless user experience.

4.2 Database

For the purpose of the web application, a **database** used for storage was also needed in order to have all the necessary information available for the users to view. Currently, there are two leading type of database storage components in the industry [30]. On one side there are the relation databases, such as MySQL, PostgreSQL, Oracle Database or Microsoft SQL Server. Then on the another side, there are the non-relational databases, such as MongoDB, Redis, CouchDB or Amazon DynamoDB. The biggest

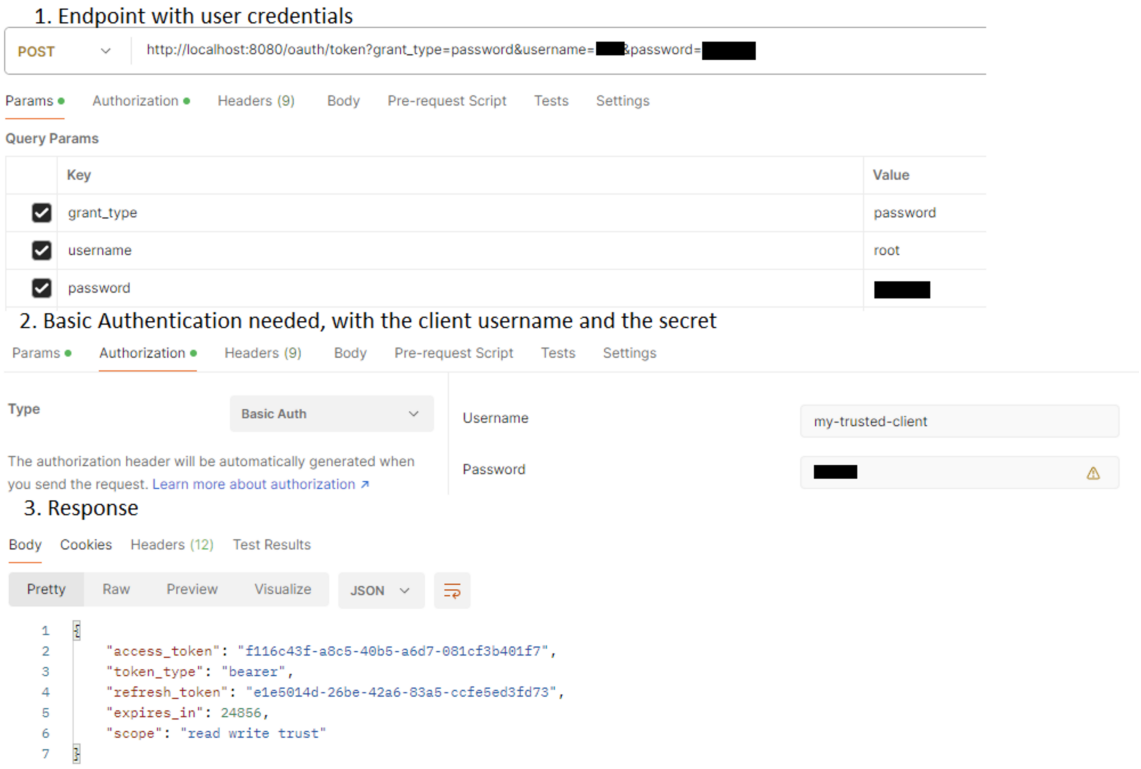


Figure 11: Necessary steps for successful Oauth 2.0 authentication

difference between these two types of databases is how the structure the data looks inside of them.

Relational databases structure their data in a way where everything is structured in tables, rows, and columns. A table is an entity which can have many columns, where each column contributes to the description of that entity. The columns in the relational database specify the data type which will be used and each row is containing the value of that particular data type. Nearly all tables have a column which is called 'primary key', and this key serves as a unique identifier of each row stored inside a table. In a relational database, connections or relations can be created between tables with something which is called 'foreign key' and it serves as a reference to the primary key of another table. With this 'feature', the developer can create relations between tables, which would help and ease the process of querying data from these tables. Listed below are some of the advantages and disadvantages of using relational database management systems in web development.

Of course, each type of database has advantages and disadvantages. For the

relational database, some of the advantages are: **Data Integrity and Consistency** – Relational database management systems enforce data integrity with constraints such as primary keys and foreign keys. The keys ensure that the data will remain accurate [2], **ACID Compliance** – Relational databases follow the **ACID** (Atomicity, Consistency, Isolation, Durability) principles which ensure reliable transactions and **Structured Query Language (SQL)** – SQL is providing a standardized method for querying and managing data, allowing for complex queries, indexing and optimization, which is intended to improve the database performance [36].

On the other hand, some of the disadvantages are: **Scalability Challenges** – Relational databases are often difficult to scale horizontally due to their strict consistency requirements [1], **Complex Schema Design** – As nearly everything is set in stone in a relational database, the use case to fulfill dynamic data requirements can lead quite often to schema modifications [13] and **Performance Bottlenecks** – Complex queries which involve multiple joins can lead to performance overhead, therefore reducing efficiency.

Non-relational databases are storage models which do not structure the data in a way like the relational databases do. In the non-relational database management systems, the data is not stored in a tabular structure as it was the case in the relational database management systems. In this case, non-relational databases use data models in order to store the data such as: **Key-value pairs** where the key is what defines the data and the value is the actual data, **Documents** where the data is stored in a document, usually in a JSON format, **Graphs** where the data is stored as nodes, it has relationships and properties, **Wide-Column** uses tables, rows and columns as the relational databases do, but the names and the format of the columns can vary from row to row in the same table, so it is more dynamic.

Some of the advantages by using Non-relational databases are: **Scalability** – Opposite the of relational databases, NoSQL databases are designed to scale horizontally, making them suitable for large-scale distributed systems [1], **Flexibility** – Non-relational databases allow the design dynamic schemas, allowing the developers to store and retrieve unstructured data efficiently and **High Performance** – Non-relational databases optimize reading and writing operations by reducing the complex joins, which results in faster query performance for large datasets [13].

On the other hand, the disadvantages are: **Lack of ACID Compliance** – Many non-relational databases prioritize scalability over consistency, leading to potential data differentiation in some cases [18], **Limited Standardization** – Unlike the relational databases language SQL, which is widely adopted and standardized, non-relational databases use different languages, which makes the use of each non-relational database much harder and **Complex Data Consistency Management** – Handling

data consistency across distributed nodes can be challenging and may require a bigger amount of work.

Since 'Escore' uses relatively small amount of data and the structure of the data is set in stone, it was decided to use a relational database management system called MySQL. In total there are fourteen tables in the database. The division of the entities is done into three groups, based on their usage in the web application. The first group which is used for the privileges of the web application and the users itself, the second group which is used for the Oauth 2.0 authentication process and the last group which is used for business related data. Some of the business related entities are: *head2head* which contains the data for the data visualization of the history in their competitive regional league and *teams* which contains data for the teams competing in Esports. All of these tables were developed in a schema called 'e_score'. On Figure 12, the ER diagram which is tailored for the web application 'Escore' can be seen.

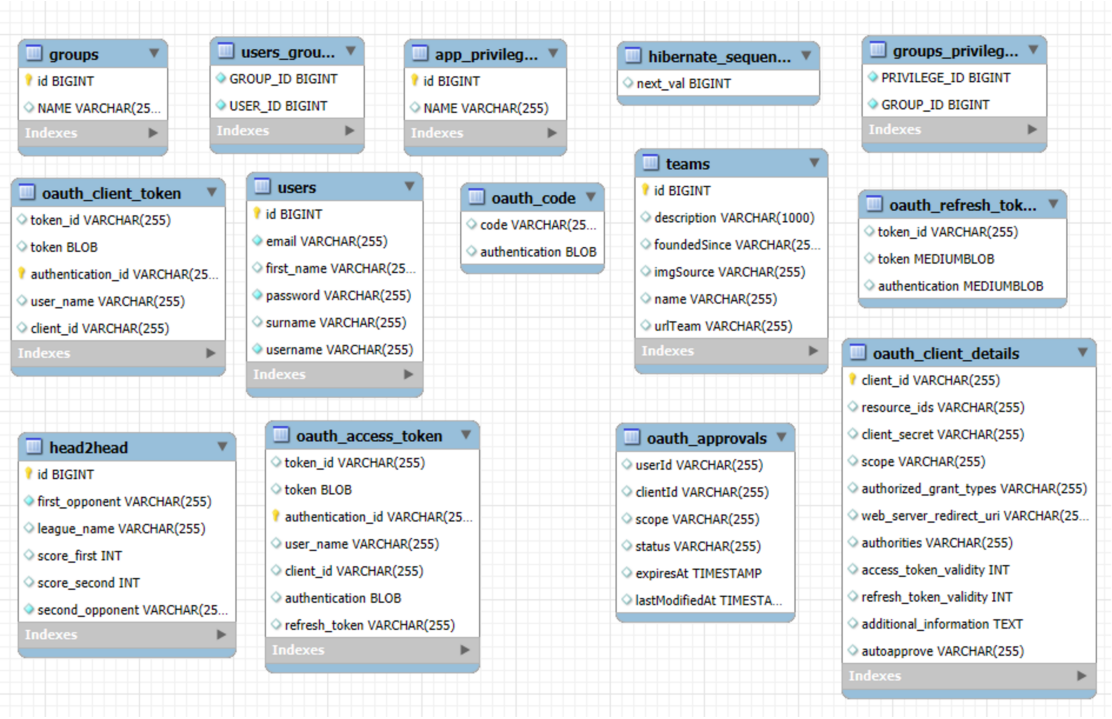


Figure 12: ER diagram for the database design of 'EScore' web application

4.3 Frontend

The frontend is the visual part included in every web application or the part of the application with which the user interacts. Since the spectrum of technologies which can be used to develop the frontend is enormous, there was a small research done in order to decide which one to use for the research project. Many of the technologies have similarities with each other and the majority of the popular ones are using **JavaScript** with a combination of **TypeScript**. JavaScript is a high-level programming language which has a primary use in the building of interactive and very versatile web applications. When the programming language first emerged, it was used as a scripting language and therefore the name contains the word Script [22]. From a scripting language, JavaScript further evolved into a language which is offering both backend and frontend possibilities by the usage of various libraries and frameworks. It is an object-oriented programming language like Java and it supports event-driven development. The key features of JavaScript are asynchronous programming with the usage of async and await function, a lot of support from the development community and dynamic typing. It is a dominant force in web development, which is offering the possibility to develop from simple to complex web applications. JavaScript is very powerful because it offers usage of popular frameworks and libraries like React, Angular and Vue.js [14].

At the end, the library chosen for the thesis was **ReactJS**. ReactJS is a widely used JavaScript library which can be used for building dynamic and interactive user interface and primarily single-page application. It was developed by Facebook, which quickly made its way in the development community to be known as a library which is component-based, offers clean structure and usage of reusable user-interface elements [9]. ReactJS is using a virtual DOM – Document Object Model which can optimize the performance done during rendering of the components and it will ensure fast updates by doing minimal interactions with the actual DOM. It is offering the support for one-way data binding and it empowers predictability and debugging. The framework introduces a feature called React Hooks which is easing the handling of side effects during state management. The library is widely used and it is one of the first picks in web development for the development of scalable applications because it offers flexibility, high performance and a developer friendly approach [35]. ReactJS is a modern web-development library as it has a strong relationship with MaterialUI. MaterialUI is a collection of customization components which follow the guidelines and principles of Google's Material Design. It offers pre-styled components which can be used to build very responsive user interfaces and designs which are visually appealing to the eye of the user. Since ReactJS is built based on component based architecture, the usage of Material UI as a library will just enhance the development

by the usage of already developed components like buttons, cards, forms or dialogs, which can further on be adjusted and designed depending on the need on the web application. MaterialUI is also allowing styling solution with CSS approach and several built in functions. The combination of MaterialUI with ReactJS will allow the developers to build very elegant and catchy web applications in a relatively fast time pace [35] [40].

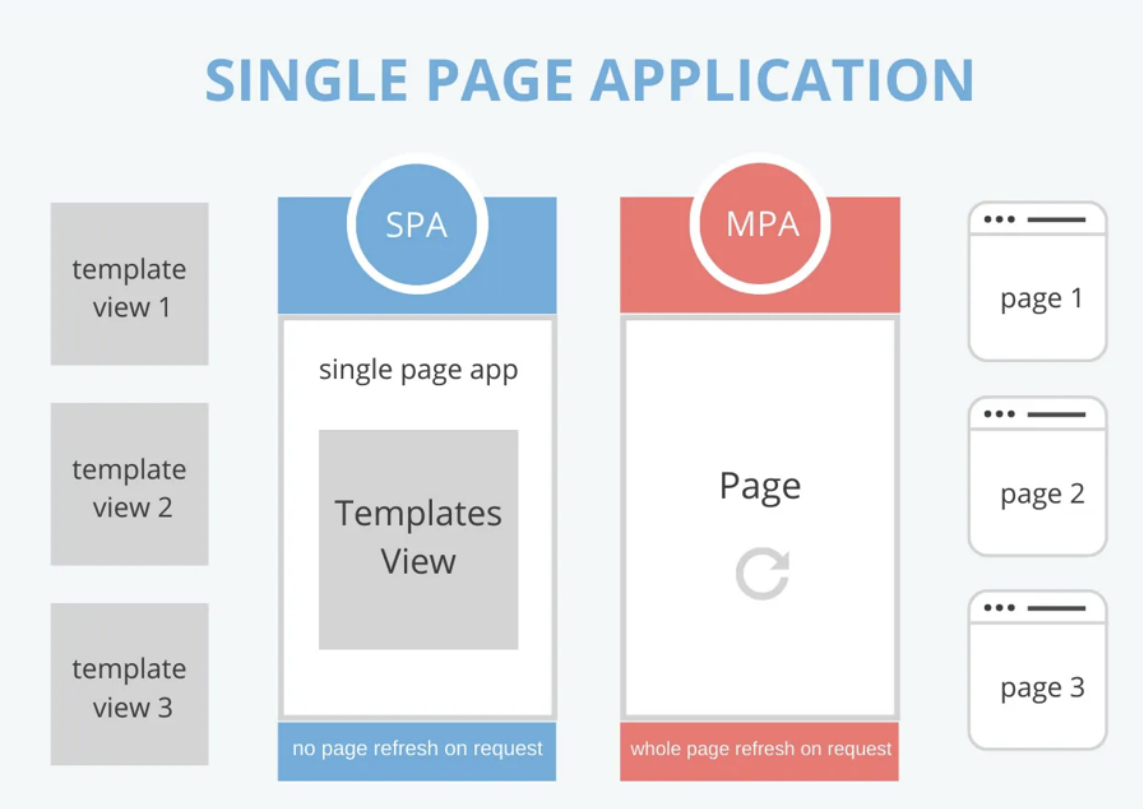


Figure 13: Difference between single-page and multi-page applications [29]

ReactJS is a library which is open-sourced and used to develop singe-page applications [34]. A single-page application loads only the requested content or a template. This is quite helpful in order to accelerate the time needed for communication between and for content reloading. On the other hand, in multiple-page applications the content is always reloaded, even though the majority of the data on the page is still the same. As it can be seen on the comparison Figure 13, the single-page application has only one page. Then the single-page application changes the template views which are shown to the user. The advantage of using single-page applications is that

whenever a view is changed, the previous view is moved and the new requested view is added. However there is not a page reload performed, so the time that the user needs to wait is significantly lower than in the multi-page applications. With this, the user will have a feeling that everyone from the developer's side is going smoothly and everything is loaded pretty fast, but in theory everything was loaded beforehand and only the template views are changed. This is one of the reason why ReactJS was chosen to be used in order to develop the frontend of the web application.

4.4 Connecting Machine Learning models with UI

Since the idea was to show predictions in the UI of the web application - Escore, a endpoint was exposed in the Python project as well, as it can be seen on Listing 5.

```
@app.route('/predict', methods=['GET'])
def predict():
    team1 = request.args.get('team1')
    team2 = request.args.get('team2')
    model_name = request.args.get('model', 'Logistic Regression')

    if not team1 or not team2:
        return jsonify({'error': 'Please provide both team1 and
            team2 as query parameters'}), 400

    if model_name not in models:
        return jsonify({'error': f'Model {model_name} not available.
            Choose from {list(models.keys())}'}), 400

    result, probabilities = predict_match_result(team1, team2,
        model_name)
    response = {
        'team1': team1,
        'team2': team2,
        'model': model_name,
        'prediction': 'win' if result == 1 else 'lose',
        'probability': {
            'team1_win': probabilities[1],
            'team2_win': probabilities[0]
        }
    }

    return jsonify(response)
```

Listing 5: Endpoint exposed in Python project

This endpoint serves as a bridge between the web application and the machine learning models developed. With this, the web application can show real time predictions generated by the machine learning algorithm. It accepts three arguments:

- team1 – which is the first team with which the user wants to compare
- team2 – which is the second team with which the user wants to compare
- model – which model the user wants to be used to predict a League of Legends match

The frontend will do a HTTP call to the backend to retrieve the upcoming matches for the game League of Legends. At the start, there is an HTTP call to an external API which will return the desired matches whether they are upcoming or previous. Afterwards, the backend calls with an HTTP call the python backend which is called for every League of Legends match. Once everything is aggregated, the backend sends the response with all the necessary data to the frontend. On Figure 14, it can be seen exactly how the communication flows, where each step is described in order to be able to fully understand.

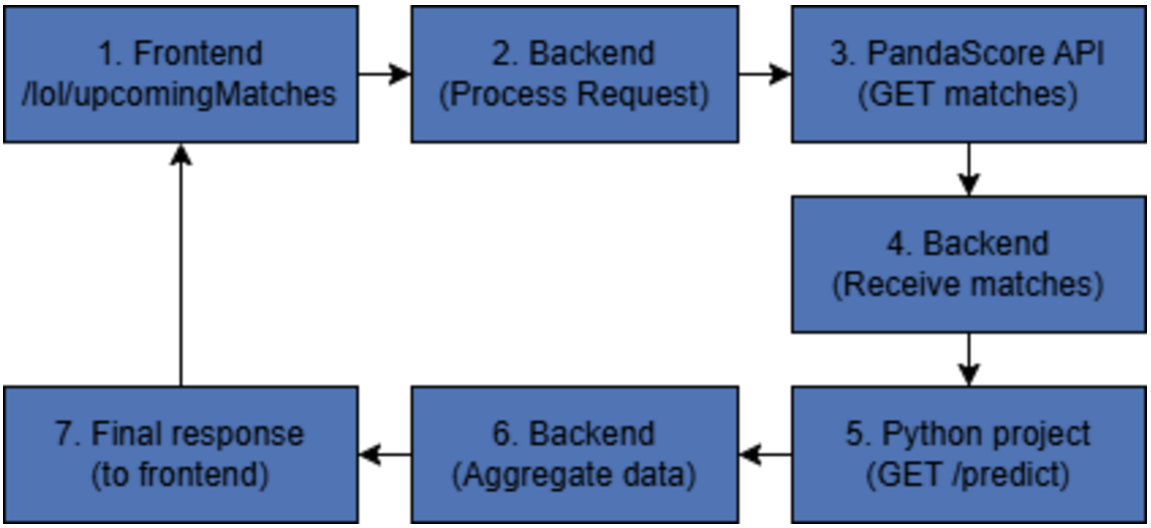


Figure 14: Architecture flow in fetching League of Legends matches

5 Machine Learning Models

One of the main features of the research conducted in the Master Thesis is the machine learning algorithms and models which are used in order to predict Esport matches in the game called League of Legends. Machine learning models are programs which can make decisions or find patterns in a certain dataset which is unknown to them at the moment. By previously training these models, the models gather knowledge in order to make a decision on something similar, but not familiar to them. For example, in natural language processing, the machine learning models can correctly recognize a sentences which were not known to them prior to the sentence being shown or in the context of the Thesis research, predictions of the competitive matches in League of Legends. Predictions in the competitive matches played in League of Legends, would be to predict an outcome or a winner from one match, based on the previous data gathered from the previous matches. In general, the machine learning techniques can be divided into four groups [41]:

- **Supervised technique** is using a labeled training dataset in order to understand the relationships between the input and the output data. The people training these models are manually creating datasets, where the input data corresponds with some labels and outputs. This technique trains the models in a way where the models can apply the correct output to new input data in a different use-case scenario.

- **Unsupervised technique** as a technique is quite the opposite of the supervised learning technique. It is a type of a machine learning model which learns from the data without any human interaction. In this scenario, the data which is given to the model is completely unlabeled and the idea is that the machine learning model will discover patterns on the data without any explicit counsel from the data engineers.

- **Semi-supervised technique** is a hybrid approach of the two techniques mentioned above. This machine learning methodology operates on both labeled and non-labeled data. The ultimate goal of a semi-supervised learning model is to provide a better outcome for prediction than that produced using the labeled data alone from the model [41].

- **Reinforcement technique** is based mostly on rewards and penalties. This machine learning model is making some decisions and based on that, the idea is for the model to take action to increase the reward or minimize the risk.

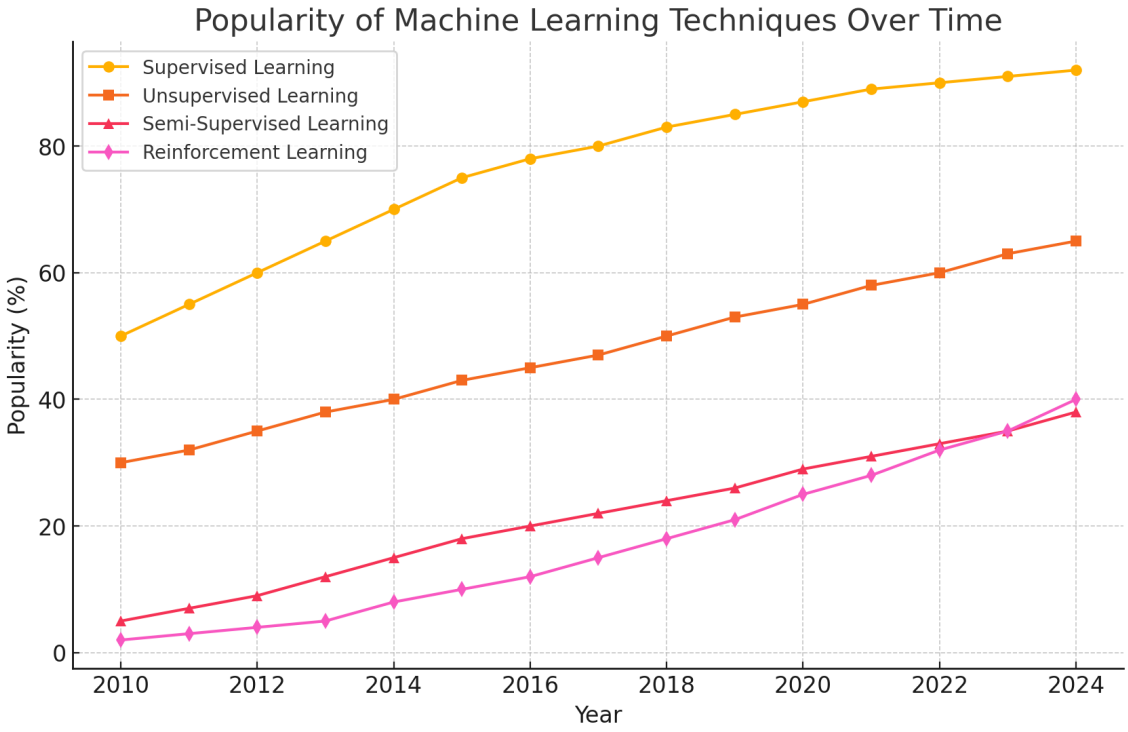


Figure 15: Popularity of Machine Learning techniques over time

Each machine learning methodology excels better in some fields than the others methodologies. Therefore, it is important to choose the technique wisely in order for the machine learning models to be efficient. As it can be seen from Figure 15, the most widely used currently and throughout the years is the supervised learning technique.

The percentages in Figure 15 do not represent exact usage numbers but instead show how the popularity of each machine learning technique has changed over the course of time. They are based on characteristics like: How often these techniques are mentioned in research papers, their usage in the industry and real-world projects and mentions in online discussions, forums and articles. For the purpose of this Master Thesis research, it was decided to use machine learning models which are using the supervised techniques. Since the choice of supervised machine learning models is big at the moment, it was decided to choose four models in order to be able to compare their performance through various different metrics. These models are: **Logistic Regression**, **Random forest**, **Gradient boosting** and **Naive Bayes**.

5.1 Dataset

As already mentioned, the data set that is going to be used in the training process of the machine learning models is taken from Oracle Elixir [46]. It consists of competitive matches which are played over a three-year time period, which can be seen on Listing 6.

```
file_paths = [
    'C:/Users/Alek/Desktop/Projects/data/2024_LoL.csv',
    'C:/Users/Alek/Desktop/Projects/data/2023_LoL.csv',
    'C:/Users/Alek/Desktop/Projects/data/2022_LoL.csv'
]
```

Listing 6: Import of competitive League of Legends matches

The dataset consists of many different statistics retrieved from every match, but some of the most important statistics are the following:

- **teamname** – Team name
- **xpdiffat15** – Experience difference at fifteen minutes
- **golddiffat15** – Gold difference at fifteen minutes
- **csdiffat15** – Creaps difference at fifteen minutes
- **dragons** – Number of times the epic monster Dragon was killed
- **firstblood** – Whether the team made the first kill

The major features which were used for the models were **xpdiffat15**, **golddiffat15**, **csdiffat15** and **dragons**. These features were chosen based on the answers received on the research questionnaire. Since the Master Thesis revolves around how much the experts in the Esports fields have the knowledge in order to understand the major statistics which are needed for a certain team to win, it was decided to use the previously mentioned features. As many machine learning algorithms do not handle missing values really well, there would need to be a pre-processing of the data itself.

```
data = pd.concat([
    pd.read_csv(file, low_memory=False) for file in file_paths
], ignore_index=True)

team_data = data[data['playername'].isna()]

X = team_data[['xpdiffat15', 'golddiffat15', 'csdiffat15', 'dragons'
    ]].fillna(team_data[['xpdiffat15', 'golddiffat15', 'csdiffat15',
    'dragons']].mean())
y = team_data['result']
```

Listing 7: Loading and Processing Team Data

The code shown in Listing 7 is showing some crucial steps which are needed in order to work with a fully functional data set. At the start, all three files are concatenated, so the files can be worked with as part of one big data set. Furthermore, the key aspect into the dataset is the team entries and the player specific rows are removed from the data set. Additionally, since it is not advisable to work with missing values, the mean of the three parameters is taken into account and it is assigned to the empty values of the above mentioned parameters. By having the missing values removed, we are ensuring that there would not be any over-fitting or under-fitting in the machine learning models.

5.2 Logistic Regression

Logistic regression is a popular machine learning technique that is used to solve classification problems. Unlike linear regression, Logistic Regression estimates the probability that a certain input will be classified in a specific class. Because of that characteristic, Linear Regression is one of the most suitable techniques to be used in order to predict matches in the competitive matches of League of Legends.

Logistic Regression uses a mathematical function called the **Sigmoid Function** which serves to transform numerical inputs into probabilities. The Sigmoid Function is given by the following:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{1}$$

where z is a linear combination of input variables:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \tag{2}$$

β_0 is the intercept and $\beta_1, \beta_2, \beta_3, ..., \beta_n$ are the coefficients that represent the impact of each characteristic.

Instead of minimizing the errors using the least squares like linear regression does, Logistic Regression optimizes its parameters using **Maximum Likelihood Estimation (MLE)**. The objective is to find the best values for β that maximize the probability of correctly classifying the data points. The model is trained using an optimization technique like **Gradient Descent**, which iteratively adjusts the parameters to minimize the **log loss function**:

$$J(\beta) = -\frac{1}{m}\sum_{i=1}^{m}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \tag{3}$$

where y_i represents the actual class and $\hat{y}_i$ is the predicted probability. Furthermore, m represents the total number of training samples that the training data set has.

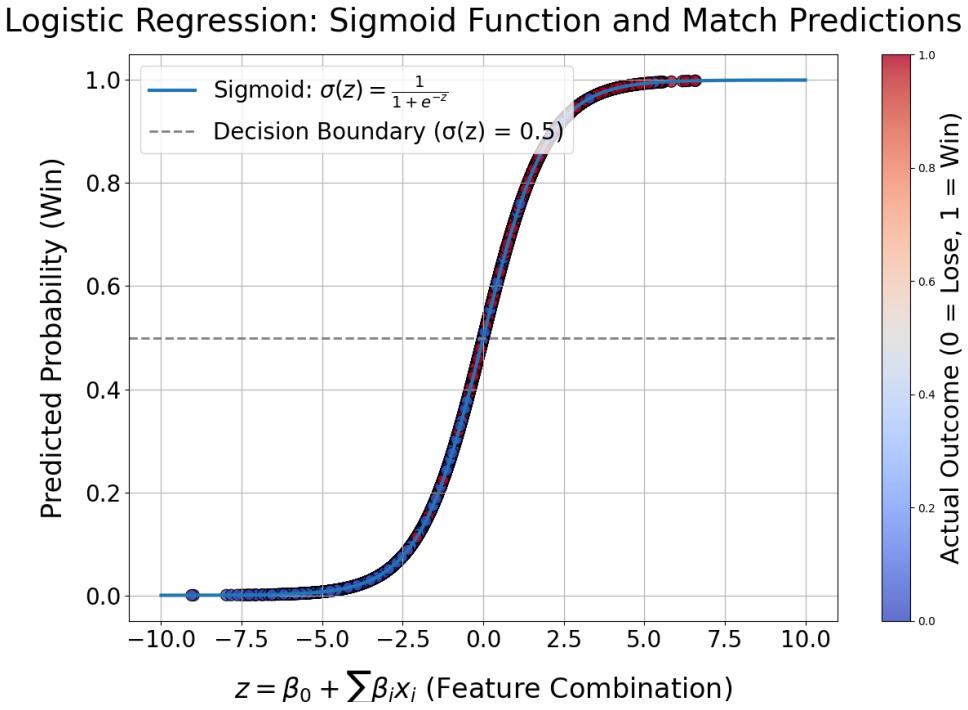


Figure 16: Logistic Regression sigmoid function

From the Figure 16, it can be observed that the blue dots which are the actual matches where the team lost, are mostly below the expected decision boundary. On

the other hand, the red dots which are the actual matches where the team won are mostly above the decision boundary. The accuracy for the Logistic Regression was put at 0.8, which indicates a pretty good overall fit [32]. Especially, since in this scenario the data set used is quite complex and it can contain possible noises. It means that the model can explain 80% of the variations in the target variable. Additionally, it also shows that the machine learning model is capturing important patterns without any overfitting. Of course, by adding more advanced techniques, the overall accuracy of this machine learning model could be improved.

Confusion Matrix was also created as an additional evaluation tool for the Logistic Regression. A Confusion Matrix summarizes the performance of the machine learning model in regards to the test data. It is a two-dimensional matrix, where

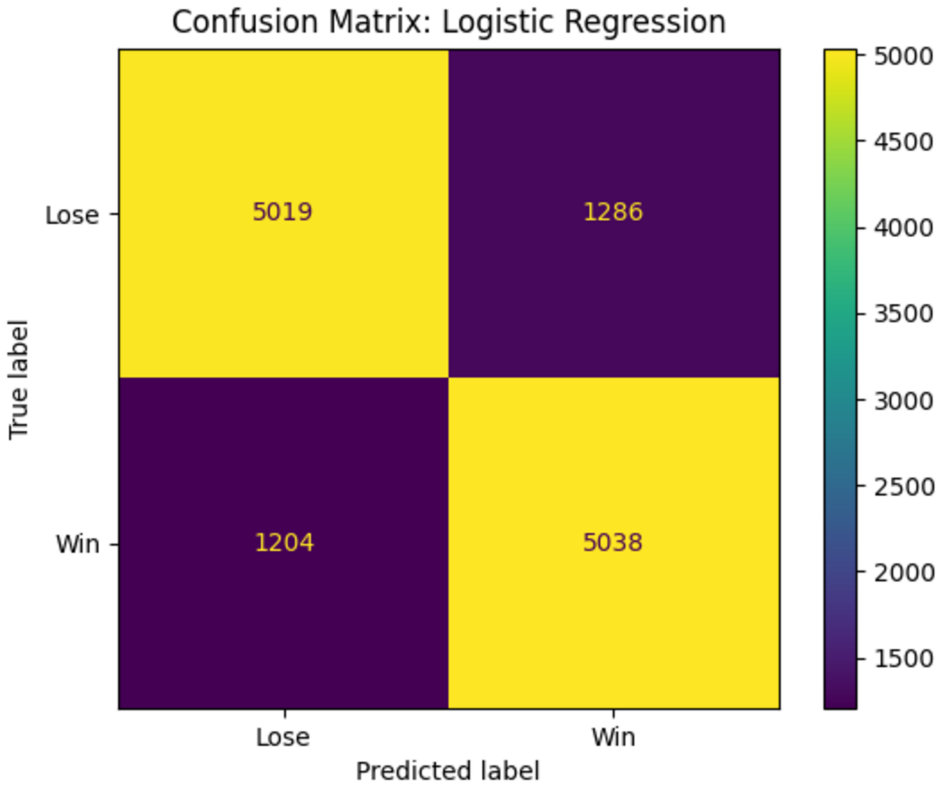


Figure 17: Confusion Matrix Logistic Regression

one dimension is the actual outcomes of a League of Legends match and the other

represents the predicted outcome from the machine learning model [47]. With the help of the Confusion Matrix, it can be seen where a machine learning model behaves quite well and where it behaves poorly. By using these observations, the machine learning models could be improved even further in order to eliminate the use case scenarios where the model predicts wrong outcome of a competitive match.

This Confusion Matrix 17 represents the performance of a Logistic Regression model which is predicting whether a team will win or lose a League of Legends match. Based on the matrix shown, it is known that there are **true positives** (Win predicted correctly): 5038 times, the model correctly predicted a win, **true negatives** (Lose predicted correctly): 5019 times, the model correctly predicted a loss, **false positives** (Predicted win but actually lost): 1286 times, the model thought the team would win, but they lost and **false negatives** (Predicted loss but actually won): 1204 times, the model thought the team would lose, but they actually won.

5.3 Random Forest

Random Forest is a machine learning method which works primarily with decision trees. The name "Random Forest" comes from the fact that it creates a "forest" of decision trees, each trained on a random subset of the data, making the model robust [49].

Random Forest works in a way where it uses a process called Bootstrap Aggregation or bagging. This process usually contains two steps. First step is called Bootstrapping, which is a process which selects random subsets of the training data. The second step is called Aggregating which is the process which lets the models which were selected from the bootstrapping vote on the final result or which class should be selected.

Bootstrap Aggregation works for Esports predictions as follows [45]:

1. **Bootstrapping the Data**
    - Multiple random datasets are created from the primary data set which has features like (xpdiffat15, golddiffat15, csdiffat15...).
    - It can happen that some data points will be repeated in a subset in order to ensure that there is diversity in the training samples.
2. **Training the decision trees**
    - Now the model has multiple decision trees, so each tree behaves as an independent decision tree model.
    - The decisions trees in Random Forest are also not pruned.

3. **Aggregating Predictions**

   - Once the above mentioned steps are finished, the predictions for each individual decision tree are aggregated.
   - For deciding whether a match will be won or lost, majority voting is used, where each decision tree classifies the match as win or loss and the majority of the votes will determine the final class.

There are some advantages with using Random Forest for Esports predictions. Bagging can help since match outcomes in Esports can be highly volatile, as it depends on many different factors such as: team composition, patch updates and the current form of players. This process can help by considering these factors and smooth out possible extreme predictions. Since it has many decision trees, the chances of overfitting are lower than using only an individual decision tree. Additionally, as

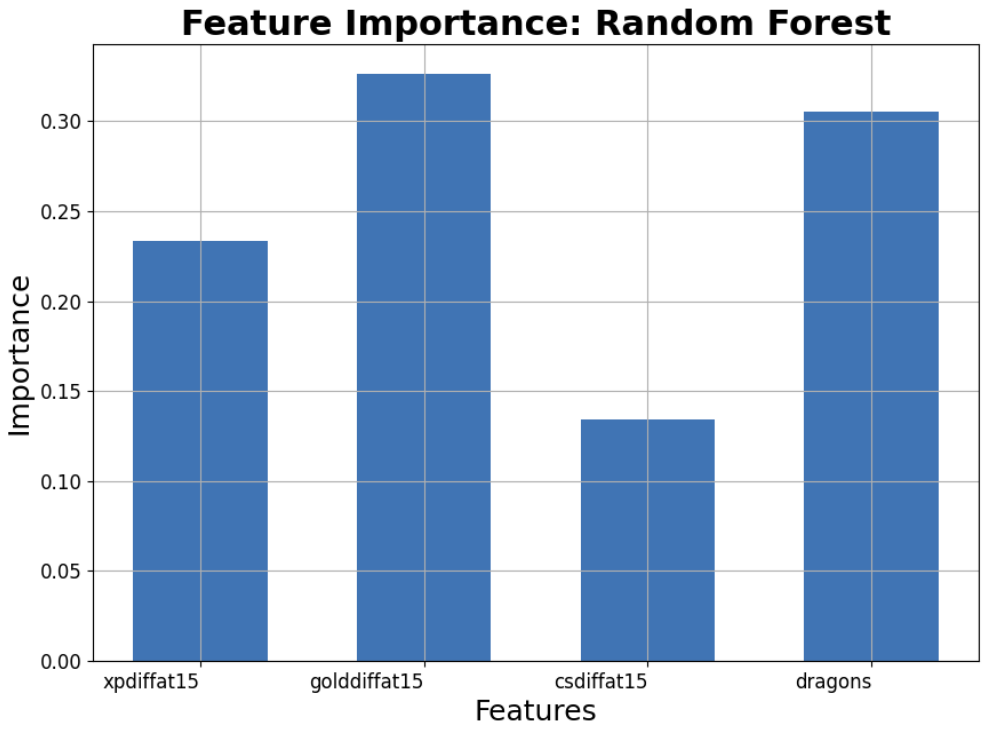


Figure 18: Feature Importance graph for Random Forest

already mentioned, by having multiple random subsets of the data set, it helps that the model is robust across different scenarios.

The accuracy for Random Forest was 0.79, which is also a decent accuracy number, since overall the dataset user in the Master Thesis is quite complex and concerning the predictions of the matches there is a big number of unpredictable factors such as: player form, meta changes and in-game strategies. In order to see which feature brings the greatest value, a **Feature Importance** graph was created. The graph can be seen on Figure 18.

From the Feature Importance the following can deducted. The feature with highest importance 33% is **Gold Difference at 15 Minutes (golddiffat15)**. It can be seen that the gold difference at fifteen minutes is a strong indicator and can be used as a dominant predictor. This means that an early advantage in gold correlates strongly with the match outcomes. The feature with second highest importance (31%) is **Dragons Secured - Significant Importance**. It influences long-term scaling (buffs, soul advantages) with the champions that the players are playing. Additionally, by killing dragons it provides early leads. The next feature with moderate importance (23%) is **Experience Difference at 15 Minutes (xpdiffat15)**. Experience difference is valuable but it has less impact when trying to determine a match outcome. At last it is the feature have **Creep Score Difference at 15 Minutes (csdiffat15)** with lowest importance (13%). A minor factor compared to the others. Likely due to the fact that CS indirectly contributes to gold and XP, which are already considered separately.

As an additional observation, a Confusion Matrix was generated for Random Forest as well, which can be seen on Figure 19.

Based on the Confusion Matrix shown on Figure 19, it can be concluded that the model has predicted a win correctly 4918 times, which is a pretty decent number. Additionally, the model has predicted a loss correctly 4971 times. On the other hand, the model has predicted the team would win, but they lost 1334 times and the model has predicted the team would lose, but they actually won 1324 times.

5.4 Gradient Boosting

Gradient Boosting is a machine learning algorithm which works in a similar way as the Random Forest, as Gradient Boosting also is using decision trees. The algorithm starts with a simple decision tree with basic predictions and usually the average values are used in the first step [28]. In Esports terms, it will predict that there is 50% chance of the team winning. Further, the algorithm checks how far were the predictions from the actual values. Instead of starting over from the start, a new

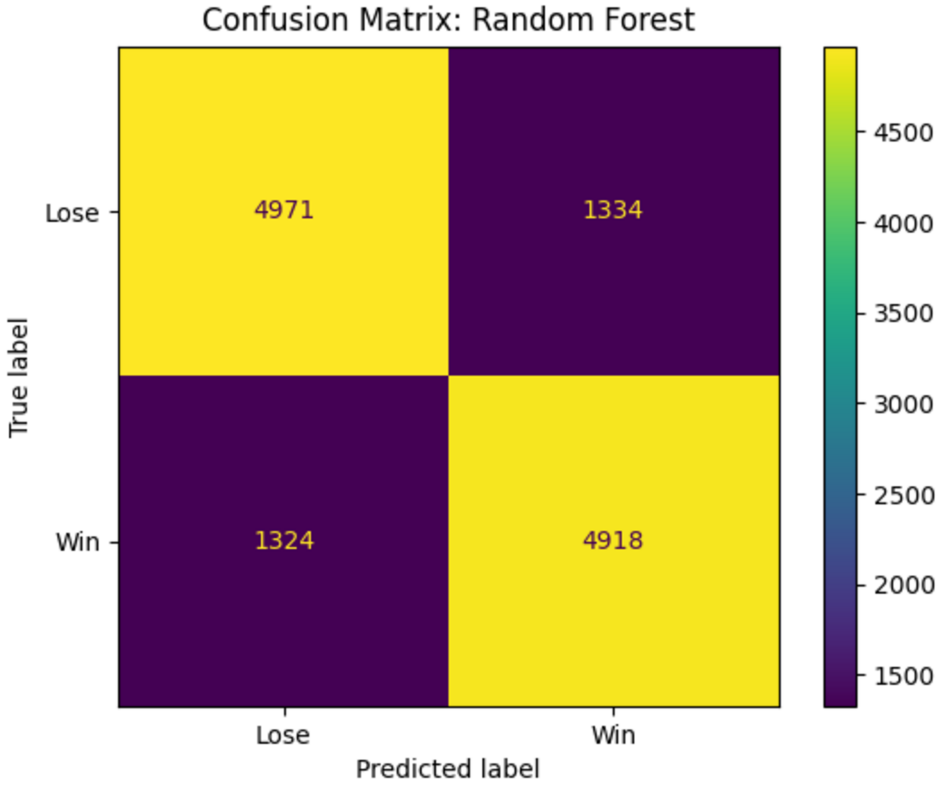


Figure 19: Confusion Matrix Random Forest

decision tree is trained to focus on what the first model got wrong. For example, if the team gold difference at fifteen minutes is really big and the team leading will win in 70% of the matches, it will adjust the predictions accordingly in the new decision tree.

At the exact moment, there are multiple Gradient Boosting implementations which can be used. In the case of Esports predictions, LightGBM was used because of the following reasons [28]:

- League of Legends competitive match prediction involves large datasets with player statistics, game metadata and historical match outcomes. LightGBM processes this efficiently, requiring less memory and computational power.

- LightGBM is optimized for big data, allowing it to train on millions of rows efficiently. Since the dataset that was used for this research contains many matches played throughout many years, LightGBM was a suitable choice.
- Since League of Legends match prediction relies on various game factors (xpdiffat15, csdiffat15 compositions, golddiffat15), LightGBM's leaf-wise growth leads to a more precise model, capturing important interactions in the data.
- Since predictions need to be made quickly before a match starts, LightGBM is the better option due to lower inference time.

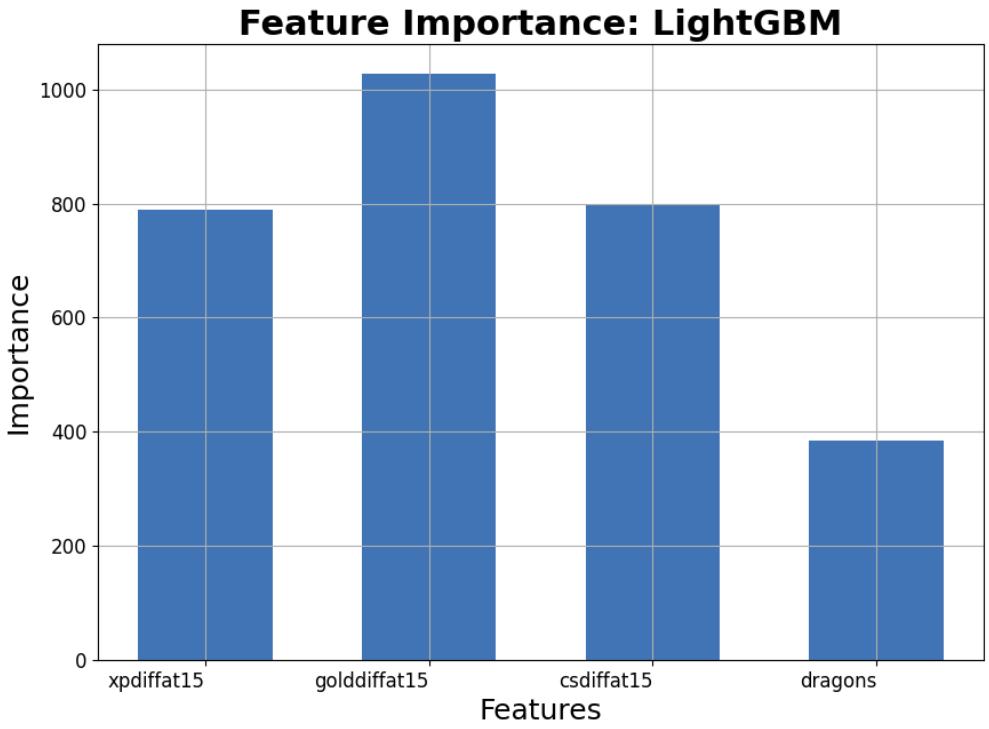


Figure 20: Feature Importance LightGBM

The accuracy score for LightGBM was 0.8 which is considered as a high accuracy score. This means that 80% of the predictions matches are predicted correctly. Overall, random guessing would yield around 0.5 accuracy, which means that this machine learning model can capture some meaningful patterns from the dataset. A 0.95–1.0

accuracy score indicates overfitting, which means the model memorized training data but would not generalize well. In the research case, 0.8 accuracy suggests a good balance between bias and variance, meaning the model performs well on unseen data. In other words, the model can perform quite well on predicting outcome of League of Legends competitive matches.

Based on that, LightGBM can be seen as a good machine learning model to predict League of Legends competitive matches.

Since Gradient Boosting works with minimizing errors in each decision tree iteration, feature importance is also quite valuable in order to know which features improve or worsen the performance of the model. With this observation, the Gradient Boosting machine learning model could be further improved by eliminating a feature which is not really important or adding a new feature which brings great value.

Based on the feature importance graph for the algorithm LightGBM 20, we can conclude that Gold Difference at fifteen Minutes (**golddiffat15**) is the most important feature in predicting of the League of Legends match outcomes, suggesting that early economic advantages lead to a crucial role in winning the match. Experience Difference (**xpdiffat15**) and CS Difference (**csdiffat15**) are also highly important, which means that teams with a stronger early game tend to perform better and win matches more often. Dragon Control (**dragons**) has the least impact in the models predictions, which means that taking dragons might not be as crucial as gold, XP or CS leads in the early game.

For an evaluation of Gradient Boosting, a Confusion Matrix was also generated 21. In regards to the League of Legends match predictions, a Confusion Matrix can help if the data is imbalanced (in League of Legends, predicting rare comeback wins vs. regular wins) and accuracy can be misleading.

In regards to LightGBM, the following can be deducted from the Confusion Matrix 21. The model correctly predicted most matches: 4984 losses and 5082 wins were predicted correctly. On the other hand, there were also mistakes such as: 1321 matches were incorrectly predicted as wins when they were actually losses and 1160 matches were incorrectly predicted as losses when they were actually wins.

The model does well at predicting the match outcomes, but sometimes it struggles with close games. The model overestimated wins in some cases, which means that the team lost, but the prediction was a win. The model underestimated winning chances in others, which means that the team won, even though the prediction was a loss.

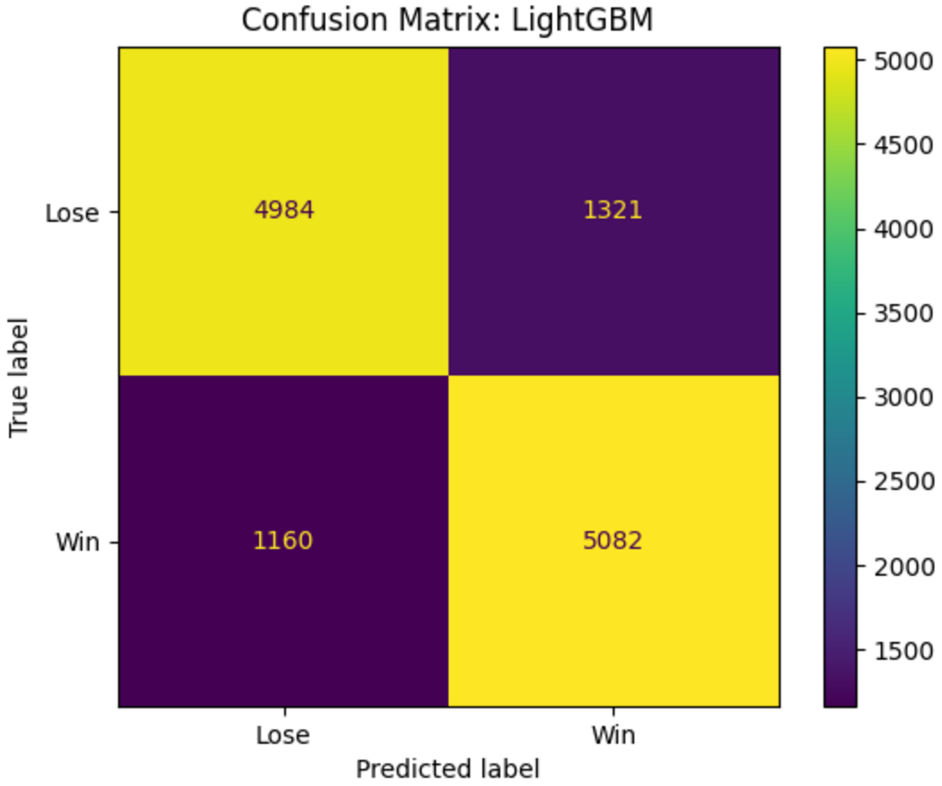


Figure 21: Confusion Matrix LightGBM

5.5 Naive Bayes

Naive Bayes is the last machine learning algorithm used in order to predict League of Legends competitive matches. For the implementation of this machine learning model, Gaussian Naive Bayes was used. Gaussian Naive Bayes is a classification algorithm based on the Bayes' Theorem, which predicts the probability of a certain class ("win" or "lose") given some input features. In the research case, the input features are csdiff15, xpdiffat15, golddiffat15 and the amount of dragons killed. For Gaussian Naive Bayes there are two assumptions [50]:

- All input features included in the classifier are independent. In some situations this is not true, but it simplifies the classification.
- Each feature follows a Gaussian (bell-curve) distribution, which is meaning the

probability of a certain feature value is calculated using a formula based on the mean and the standard deviation.

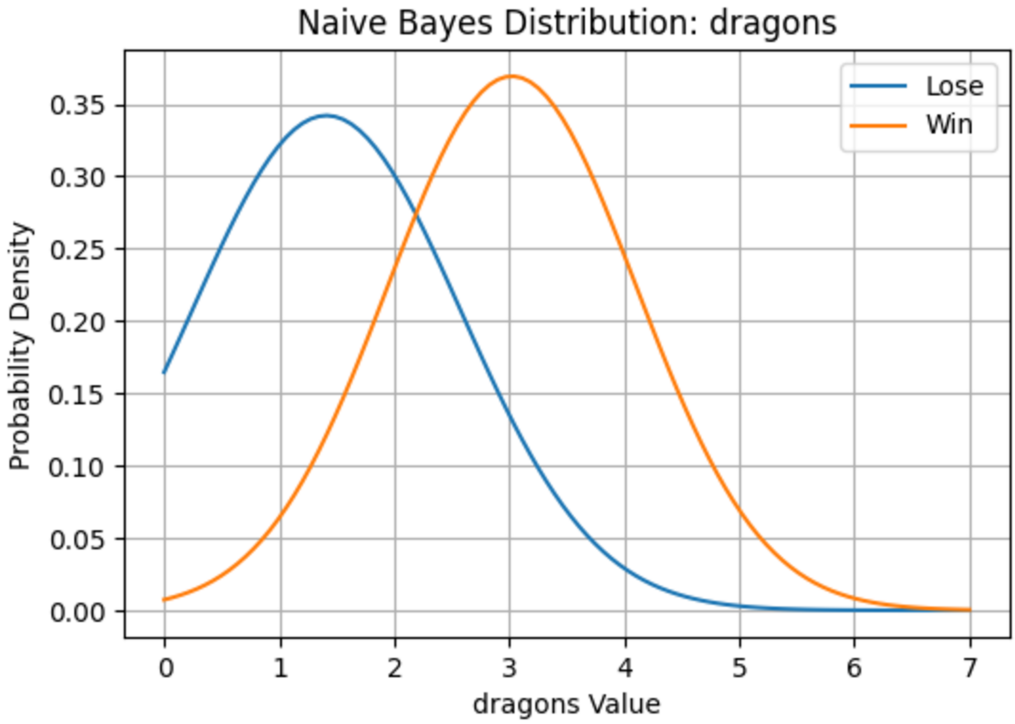


Figure 22: Naive Bayes Distribution for dragons killed

Since the input features are behaving as independent features, one example for number of dragons killed is taken into account. On Figure 22 there are two axis visible, which are the following:

- **X-Axis (Dragons Value):**
  - Represents the number of dragons taken by a team in a game.
  - Values range from zero to around seven, which means some teams take no dragons, while others secure up to seven.
- **Y-Axis (Probability Density):**
  - Represents how likely a specific number of dragons is for winning versus losing teams.

  - Higher values indicate that teams frequently take that number of dragons.

Additionally, on the graph it can be seen that there are two curves. The blue curve is the one which is describing the lose distribution or the characteristics for the teams that lose. The orange curve is the one describing the win distribution or the characteristics for teams that win.

- **Lose Distribution (Blue Curve):**
  - Peaks around one to two dragons.
  - Suggests that losing teams typically secure only a few dragons.
  - The probability density drops significantly after three plus dragons, meaning it is rare for losing teams to take that many.
- **Win Distribution (Orange Curve):**
  - Peaks around three dragons.
  - Winning teams are more likely to take three or four dragons.
  - The distribution extends further, meaning winning teams often take more dragons compared to losing teams.

Some key takeaways can be retrieved from the data shown in the graph. If a team takes less than two dragons, it means that the team will probably lose the game. Teams that are constantly getting more than three dragon means that they are likely winning more games. Securing dragons is strongly correlated with winning the competitive match. Additionally, just by looking at the graphs, teams can also retrieve some quality information. They should prioritize securing multiple dragons (minimum three) in order to improve their chances of winning. In case the team is taking less than three dragons constantly, they should rethink their strategy and develop a strategy which will end with taking more dragons. Additionally, teams can develop individual strategies for each individual team. If a team takes a small number of dragons, then the team can use that and force fights which will eventually bring them more dragons or develop a general strategy. A general strategy would be to play for dragons independently of the team that they are playing against.

The accuracy score for Naive Bayes was 0.79, which means that 79% of the matches were correctly predicted. This would be considered as a good accuracy score, since League of Legends match prediction could be a quite complex task to predict. Additionally, for an outcome of a match there are many non-predictable factors. As

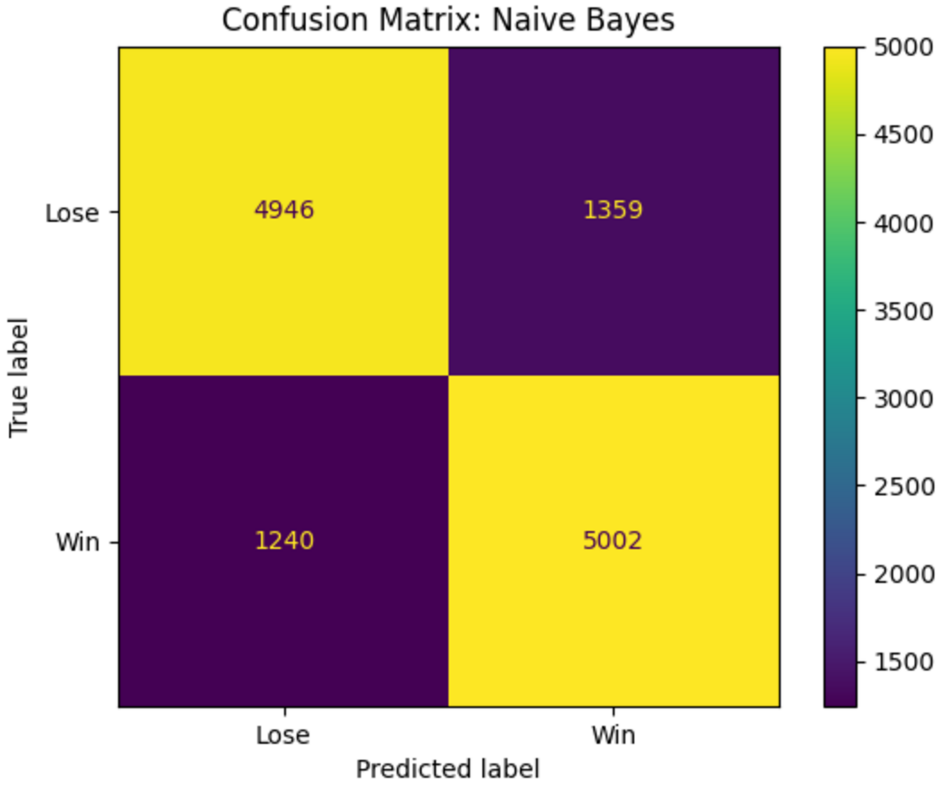


Figure 23: Confusion Matrix Naive Bayes

same for the other three machine learning models, a Confusion Matrix 23 was also generated for the Naive Bayes machine algorithm.

On the Confusion Matrix, it can be seen that there are True Negatives (4946): The model correctly predicted Lose when the actual outcome was Lose, False Positives (1359): The model incorrectly predicted Win when the actual outcome was Lose, False Negatives (1240): The model incorrectly predicted Lose when the actual outcome was Win and True Positives (5002): The model correctly predicted Win when the actual outcome was Win.

6 Evaluation

The evaluation of this research focuses on trying to understand how well the mentioned machine learning algorithms have performed in predicting competitive matches in Esports, specifically in the game League of Legends. In the assessment various methods were included such as: statistical analysis, accuracy measurements, visual representation of data that gets out of competitive matches and many more. By utilizing the multiple evaluation approaches, an extensive view was ensured of how effective and reliable the models are in real time.

One of the key aspects of the evaluation was measuring the accuracy, precision, recall and F1-score of the different machine learning models included in the research. The models included Logistic Regression, Random Forest, Gradient Boosting and Naive Bayes. Based on the results achieved, Gradient Boosting and Random Forest were the best machine learning algorithms with the highest accuracy of 80%. On the other hand, Naive Bayes has a slightly lower accuracy, which means that the assumptions in regards to the feature importance might not fully align with the dataset characteristics [32].

Additionally, for each machine learning algorithm, a Confusion Matrix was created. The confusion matrices can help to analyze how often the models made correct or incorrect predictions. For example, Random Forest performed quite well in minimizing false positives, which means that it rarely predicted a win when a loss was the actual outcome. On the other hand, Logistic Regression has a slight tendency to misclassify certain close games, which is expected since once a game is close in terms of gold, the game can flip on any team [51].

Another essential part of the evaluation was user feedback. A research survey was conducted with participants who have experience in Esports, especially working in Esports with data, since the features which were used in the machine learning models were actually decided based on the answers received from the survey. Additionally, the participants were asked questions in regards to the other aspects of the Master Thesis as well. They provided insights into the usability and effectiveness of the predictive system. Some participants pointed out that the models could benefit from additional real-time data points such as: player fatigue or in-game adaptations [25]. The participants also highlighted the importance of data visualization which is evaluated in the upcoming chapters.

A final part of the evaluation involved comparing the machine learning models with already existing ones in the Esports branch. Many of these systems rely on simple data, such as team rankings or past win rates which often is not enough since the in-game data is not considered at all. The developed machine learning models

demonstrated an improvement in accuracy by using a wider range of game metrics such as: gold difference at 10 minutes and objective controls [42].

6.1 Data Usage in League of Legends

In many cases, the usage of data has proven to be beneficial in the optimization of the performances and in the process of correct decision-making. By leveraging the insights received from the usage of the data, user experience can be guided and an overall improvement of efficiency can be achieved. One of the most notable examples is with the series of playing best out of five matches. In the best-of-five matches series, the winning team is determined by securing three wins out of the five possible matches. Within this very competitive scenario, the drafting and the selection of the champions is occurring right on the spot and this is requiring a lot of strategic decision-making which is done usually in very stressful and fast conditions. The team coaches play a crucial role in this process, where their expertise to analyze the data generated from the previous match within the series is put into practice. Between the pause of the match played and the match to be played, they are using the pre-prepared data in order to refine and further define the strategy which will be used for the drafting phase. The coaches can assess the effectiveness of a previous selection of a champion by actually analyzing the data driven evaluation prepared. This data driven evaluation can help the coach in making selections and adjustment to the team. One key metric which can be analyzed right on the spot in real time is the **Blind Pickability of Highest Presence Champions** [43], which is providing instant insight into the champion selection trends and how these trends actually have impact on the outcomes of the competitive matches.

The Figure 24 shows all of the champions with a **50%** or a higher presence based on the pick + ban rate in patch 10.16, in the top four Esports world leagues, compared to how often they are blind picked, meaning to be chosen before the enemy champion in their role or counter-picked, meaning to be chosen after the enemy champion in their role. Caitlyn is holding a 100% pick ban rate and a 100% blind picked rate. Graves and Renekton have also been blind picked in every game where they were being played. Nidalee, Tahm Kench and Gangplank have been seen a lot of times, but they were often picked after they know their matchup [43].

Guided by this data-driven approach, the coaches can refine their drafting strategies and give suggestions to the players while optimizing the team performance throughout the whole duration of the series. This research paper will present some of the most important and key statistical analysis. Among the most popular and widely used approaches for game analysis are those that leverage quantitative machine learning and

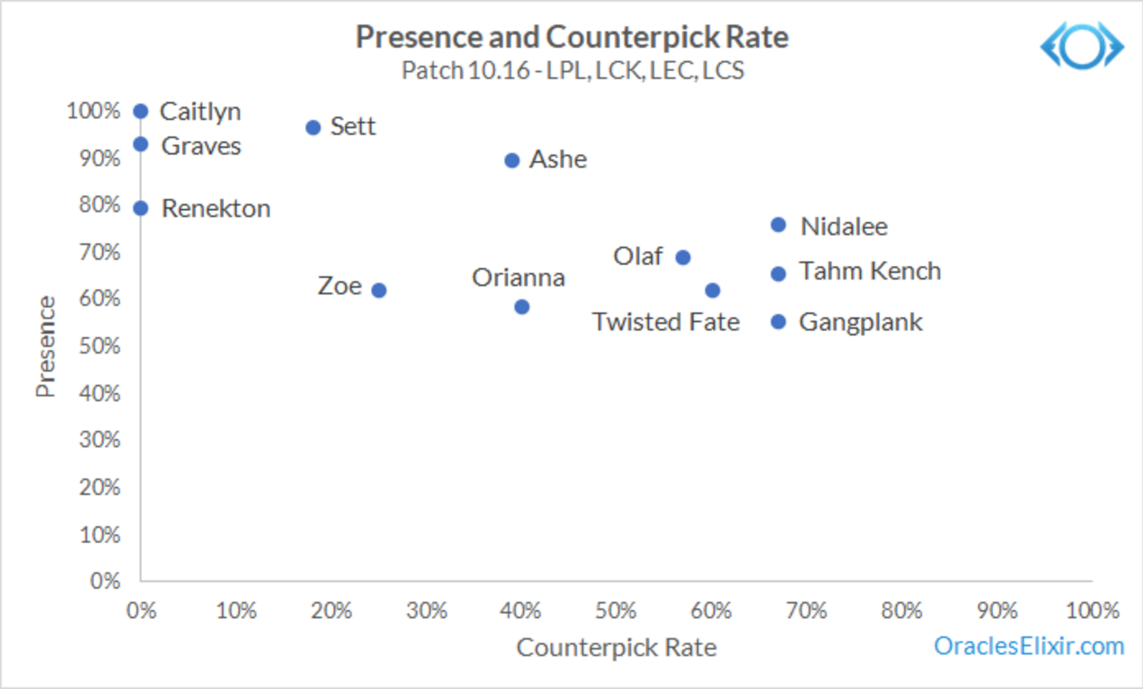


Figure 24: Blind Pickability of Highest Presence Champions [43]

statistical analysis methods which provide player behaviors [17] [37], predict match victories and outcomes [42] [52] and provide strategic recommendations [10] [11] [27]. However, the data which is processed and worked on is focused more towards the overall team performance, while integrating the statistics for each individual player in the team. One of the few inspirations for the decision of the selection of the types of data was the match statistics from the North American and European professional League of Legends [31]. The Table which contains the match statics can be seen on Figure 25.

As can additionally be noticed in Figure 25, there are some big discrepancies in the data between the North America matches and the Europe matches. For instance, the win rate for teams playing on the blue side is significantly higher than the ones in Europe. One assumption which can be extracted from this trend is that the win rate of the team is highly correlated with the amount of gold that they have accumulated throughout the match. Generally, the gold which is accumulated by the winning team is higher compared to the gold accumulated by the losing one.

Another approach which can be used to utilize the data provided by League of Legends is to group and categorize the data based on roles. Currently in the game of

	North America (n = 15)		Europe (n = 15)	
	Blue Side	Red Side	Blue Side	Red Side
Win Rate (%)	67	33	53	47
Match Duration (mm:ss)	32:41 ± 04:59		32:07 ± 06:50	
Level	75.6 ± 6.6	73.7 ± 6.8	75.1 ± 8.5	73.1 ± 9.9
Gold (in thousands)	59.1 ± 9.2	56.3 ± 9.9	59.8 ± 11.7	57.5 ± 15.4
Creep Score	1076.3 ± 158.8	1056.7 ± 175.6	1074.0 ± 240.3	1036.8 ± 244.9
Kills	12.0 ± 4.1	8.6 ± 5.9	14.6 ± 5.7	12.2 ± 7.4
Deaths	8.6 ± 5.9	12 ± 4.1	12.2 ± 7.4	14.6 ± 5.7
Assists	29.6 ± 11.4	20.9 ± 14.9	32.1 ± 15.0	26.9 ± 17.4
Wards Placed	119.3 ± 29.2	118.0 ± 27.4	105.4 ± 26.9	113.0 ± 39.0
Wards Destroyed	55.7 ± 15.3	49.9 ± 19.4	48.9 ± 18.8	42.1 ± 17.8
Towers Taken (Out of 11)	7.5 ± 3.5	5.2 ± 3.6	6.7 ± 3.7	5.6 ± 4.2
Inhibitors Taken (Out of 3)	1.6 ± 1.5	0.73 ± 1.0	1.1 ± 1.1	0.9 ± 1.1
Dragons	2.7 ± 1.3	1.6 ± 1.2	2.0 ± 1.1	1.9 ± 1.3
First Bloods (%)	33	67	73	27
First Blood Time (mm:ss)	05:25 ± 01:47		05:39 ± 02:32	
First Towers (%)	53	47	53	47
First Tower Time (mm:ss)	14:09 ± 02:48		14:26 ± 01:35	
Heralds (%)	47	53	53	47
First Barons (%)	53	47	40	53
Barons	0.67 ± 0.49	0.73 ± 0.70	0.53 ± 0.51	0.67 ± 0.72
First Inhibitors (%)	67	33	53	47
First Elder Dragons (%)	7	7	0	13
Elder Dragons	0.07 ± 0.26	0.07 ± 0.26	0	0.13 ± 0.35

Note: Creep Score is a combination of Minion and Monster kills.

Figure 25: Match statistics for North American and European professional leagues in League of Legends [31]

League of Legends, there are five different roles which are Top, Jungle, Mid, Attack Damage Carry and Support. With this approach by segmenting the data based on roles, a further deeper data analysis can be done on the performance of each individual player in specific positions [20]. Driven by this approach, there are three key characteristics which have been selected as primary indicators in assessment of the impact in each of the individual roles. These characteristics were chosen based on the relevance and in order to understand the influence on the overall performance of the team.

- **Gold difference at 15 minutes** – This statistic is one of the major indicators whether the game is going into the right desired direction or not. The explanation is that the following value will show the difference between the amount of gold acquired by the player and the amount of gold acquired by the opposing player playing in the same role. The statistic can be used in two directions where the first direction is in regards to how ahead is the player in the fifteen minute cap and the second way to use is in the direction of how ahead usually the players are when they are playing a specified champion [7] [15].

- **Creaps difference at 15 minutes** – This statistic is based usually on the individual performance against the designated opponent. If the player is mechanically performing better, then eventually the value will increase with a very steady rate. Worth to mention here is that this data can behave quite tricky. If the team intentionally picks a champion which at the start has performed worse than the designated opponent champion, this value will be nearly always negative, since the opponent has chosen a more dominant champion. By doing this choosing strategy intentionally, the team commits to a plan which means playing for a late game. In the late game, the particular champion can be more dominant than the opponent [20] [15].

- **KPW (Kills per Win) / DPL (Deaths per Loss)** – Since the advantages in the game are primarily gained from both kills and assists, it is important to have an idea of how many kills a team can reasonably be projected to have during a match. Plus, it is very rare for the losing teams to provide more value than the winning teams in League of Legends. So, whenever projecting kills within a given match-up, the team would want to know how many kills the team generally can get in a victory match and how many deaths the opposing team normally has when facing a defeating match [21] [38].

All of these statistics are tightly coupled and analyzed, in order for the teams to paint and form a better picture for the upcoming matches. The mentioned statistics can

be further used towards the comparison of different combinations pair for champion and opponent champion, to see whether those combinations bring a positive impact on the game play or sadly bring a negative one.

6.2 Data Visualization in League of Legends

All of the previously mentioned statistics are playing a huge part in deciding whether some choice was decided poorly or not. This data is represented only by numbers and therefore should be properly visualized in order for the people which are involved in the process of decision making to be able to acknowledge the data properly.

There are many possibilities how the data can be visualized. One approach would be to use Charts [12], in order for the person which is viewing the data to be able to compare different statistics based on roles or on champions. Another approach would be to use histograms and correlation tables [16] [20] [12]. Each of these data visualization types brings a unique characteristic when presenting the data. In the work ahead, the primary data visualization types which will be used are going to be elaborated further and those data visualization types are the graphs and the charts, since they brings a good representation in order for the reader to compare the values by himself. With this visualization, the data given will be easily explainable through charts and graphs. At the end, the most important thing to be achieved is for the readers to understand the idea behind the data.

For the purpose of this Master Thesis, the data given by **Oracle's Elixir** [46] will be used. The decision was based on the fact that this source of data consists of all of the competitive games which are played in the last few years. This data spans throughout the last three years of Spring and Summer splits done in all regions. This data has enormous amount of games which were played. For a simple experiment, the following five teams from the major regions were chosen: **Fnatic** in the League European Champions (LEC), **Edward Gaming** in the League of Legends Pro League (LPL), **DWG KIA** in the Legends Champions Korea (LCK), **100 Thieves** in the League Championship Series (LCS) and **DetonatioN FocusMe** in the League Japan League (LJL).

There are many comparisons used throughout the world for the gold difference at fifteen minutes [7] [15]. The bar chart would be one of the many data visualization type that can be used. With the bar chart, the discrepancy between the data is easily detectable. This feature can be used in order for the Esports teams to acknowledge both their strong and their weak points. Below, you can find a glimpse of what can be achieved with the usage of this type of data. As mentioned before, the gold difference at fifteen minutes is referring to how much a team or an individual

player is ahead against their designated opponent. On Figure 26, it can be seen that **FNATIC** (298.8262911), **DWG KIA** (296.5487528) and **DetonatioN FocusMe** (298.6923077) were more dominant than **Edwarg Gaming** (115.5882353) and **100 Thieves** (264.7715356) in their league. One of the advantages of using bar chart can be detected in Figure 26. As it can be seen from the Figure, the team Edward Gaming has quite a low gold difference. This can only mean some of the following things:

1. Edward Gaming were always playing for late game.
2. Edward Gaming's summer split was with ups and downs.
3. Edward Gaming's early game into a match was very poor.
4. Edward Gaming's drafted champions that do not thrive in early game.

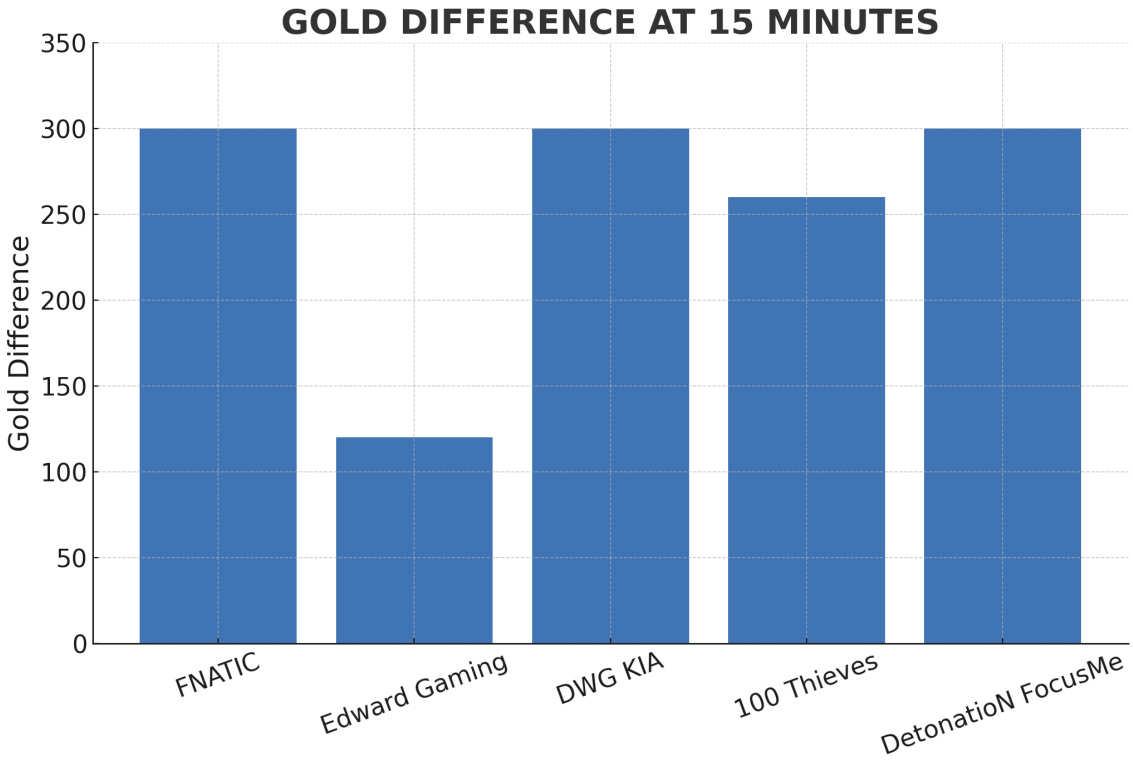


Figure 26: Gold difference at 15 minutes

Creaps difference is based on the individual performance. This statistic can also be seen as whether the teams are playing more into the late game composition of champions or into the early game composition of champions [20]. On Figure 27, it can be seen that **FNATIC** (3.338028169) are the most dominant team and they have a good creaps difference at 15 minutes. Behind them are **Edward Gaming** (2.667), **DWG KIA** (2.616780045), **100 Thieves** (2.355805243) and **DetonatioN FocusMe** (1.13). From this bar chart on Figure 27 it can be deducted that Detonation FocusMe are playing quite often for the late game composition of champions and their mindset into the game is to team fight. Additionally, for FNATIC it can be that they are always drafting champions which are really powerful in the early game, since a game can last up to one hour. On the other hand, it could be that the opponents who are against FNATIC in the League European Champions are quite weak so FNATIC as a team is always ahead in the early game of a match in League of Legends.

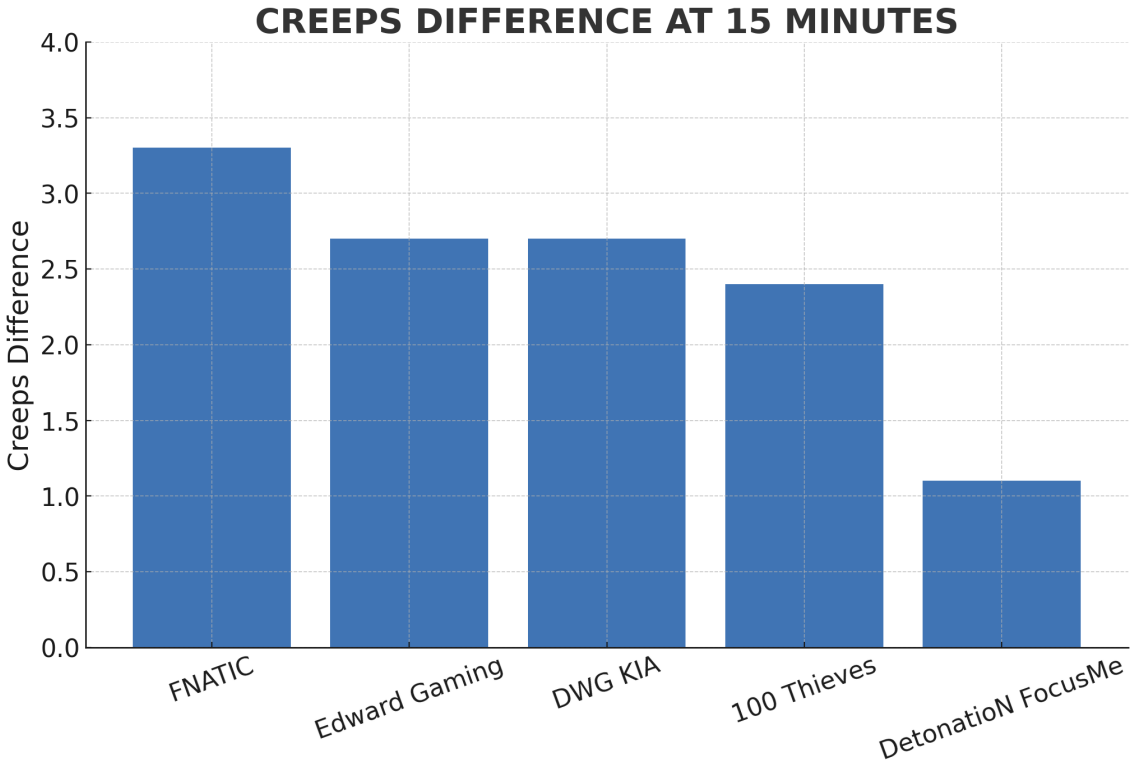


Figure 27: Creaps difference at 15 minutes

6.3 Comparison Between Machine Learning Algorithms

In every data-oriented branch, machine learning plays a crucial role in predictive analytics. By implementing various machine learning models, it can evaluate complex problems such as Esports match outcomes [41]. In the Master Research, four distinct machine learning algorithms were implemented and compared based on key performance metrics such as: accuracy, precision, recall and F1-score.The performance of these machine learning models is visually represented on Figure 28.

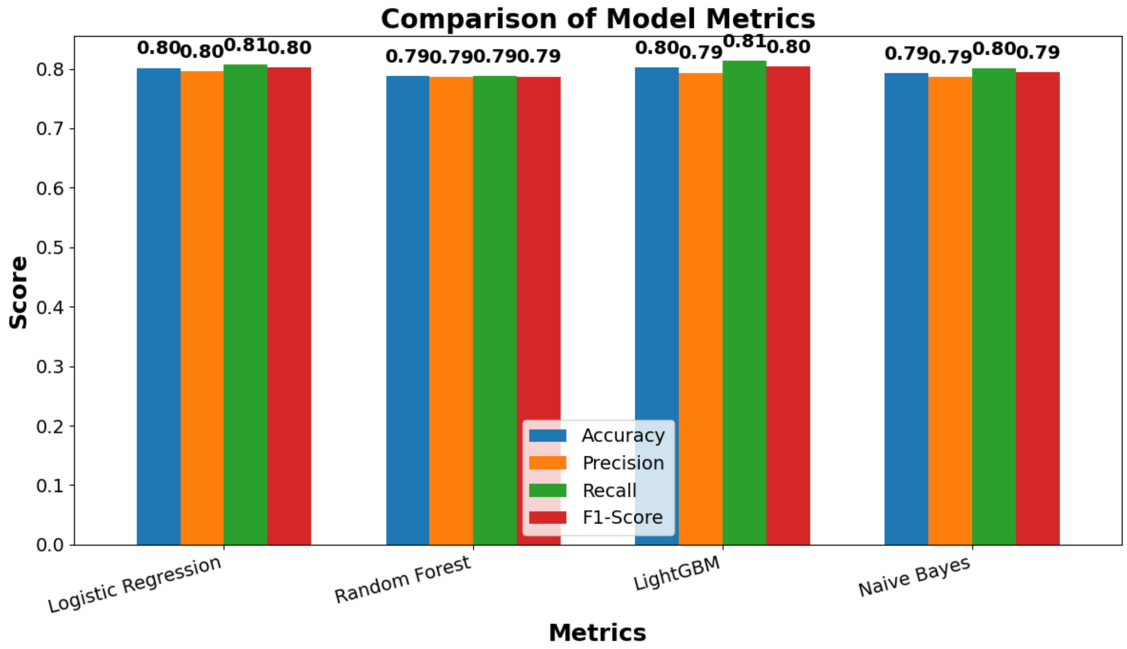


Figure 28: Performance of four machine learning models

All four models performed relatively well, with accuracies ranging between 0.79 and 0.8. Two models, Logistic Regression and LightGBM performed slightly better with accuracy score of 0.8. On the other hand, Random Forest and Naive Bayes had a slightly lower accuracy with 0.79. Precision and recall scores closely followed accuracy trends, suggesting consistent model performance across different evaluation metrics.

Logistic Regression achieved an accuracy of 0.8. The machine learning model is simple and interpretable, which means it is quite effective for linear decision boundaries, but it may struggle with some complex feature interactions. Random Forest had an accuracy of 0.79. This model can handle effectively feature importance and it can mitigate overfitting with the ensemble learning. On the other hand, it requires more computational power and is less interpretable. LightGBM also recorded an accuracy of 0.80. For LightGBM it can be said that it is quite scalable, which means it is particularly suitable for large datasets. However, it is sensitive to hyperparameter tuning and requires proper data pre-processing. Naive Bayes performed with an accuracy of 0.79. It is a fast and efficient model, especially suitable for categorical data. However, it assumes feature independence, which may not always hold true in complex datasets.

Above mentioned results indicate that there is a small variance in the accuracy between the four mentioned models. This means that the dataset allows for relatively stable predictions across each of them. Since Esports matches are predicted, factors as real-time inference speed and interpretability are crucial in order to predict correctly and increase the accuracy score [51]. Therefore, LightGBM and Random Forest might offer the best balance between accuracy and efficiency.

Recent studies have compared additional machine learning models for Esports predictions which are not included in this Master Thesis. Some of them are Support Vector Machines and Neural Networks, which can have additional advantages over traditional models. SVMs can demonstrate strong performance in regards to classification problems. They can be highly useful for capturing non-linear relationships in Esports Data [6]. Additionally, deep learning models such as Recurrent Neural Networks can be used in order to analyze sequential game data and time-dependent patterns. With these capabilities, they can achieve superior predictions in dynamic environments [26]. While these models can offer improved accuracy, they require extensive computational power and much larger datasets, which means they are not really suitable for real-time applications. The choice of which machine learning model to be used depends mostly on the balance between accuracy, interpretability and resource efficiency.

7 Conclusion

The findings of this Master Thesis contribute to researches who are working on Esports analytics, by helping with the potential of using machine learning models to predict competitive match outcomes. By levering historical data, player performances and team statistics, the machine learning models can give valuable insights to the

competitive space of the game League of Legends. The above mentioned findings align with previous researches in Esports. The traditional machine learning models can be used with high potential in order to analyze structured data [25].

Additionally, the written research helps by expressing how important is the data visualization in Esports and how much the research can contribute to the teams and the players with decision-making. The Escore web application serves as a practical implementation of the mentioned concepts. It provides the users with an interactive platform to analyze and predict match outcomes based on real-world data [15]. This Master Thesis research provides a strong foundation for Esports analytics, but several areas can be enhanced even further as part of the future work.

Integration of Advanced Machine Learning Techniques can help in order to capture temporal patterns in the competitive match data. These models could improve the accuracy by analyzing sequential game events and long-term trends. By implementing a real-time data processing algorithm, the usability of Esports analytics tools could further be improved as well. This could involve using streaming data pipelines and cloud-based machine learning models for real-time inference. As research is mostly based around the game League of Legends, including additional games can broaden the applicability of the predictive models. Introducing other Esport genres, such as first-person shooters may reveal new predictive features and modeling techniques.

In the end, this research demonstrates that predictions for Esports matches can be highly valuable to the teams and fans included in the process. With the advancements in data science and artificial intelligence happening at the moment, the field of Esports analytics will continue to develop and it will create new features which could be potentially used in the machine learning models. By expanding on the proposed future research directions, Esports analytics can bridge the gap between the data-driven strategies and the on-the-ground gameplay, which will contribute to the continued growth in the industry.

References

[1] M. Stonebraker and U. Cetintemel . One Size Fits All: An Idea Whose Time Has Come and Gone, 2010.

[2] A. Silberschatz, H. F. Korth, and S. Sudarshan. Database System Concepts, 2019.

[3] Al Ardha, M., Ridwan, M., Wijaya, A., Rohman, M., Putra, N., Bikalawan, S., Mubarak, J., Putra, K. P., and Yang, C. The development of esports research and technology in the last 3 decades. *TEM Journal* (05 2024), 1537–1547.

[4] Alam, S., Cartledge, C., and Nelson, M. Support for various http methods on the web, 2014.

[5] Aleksandar Stojkov. Data visualization through the lens of esport, 2021.

[6] Bailey, K. Statistical learning for esports match prediction, 202.

[7] Bailey, K. *Statistical Learning for Esports Match Prediction*. California State Polytechnic University, Pomona, 2020.

[8] Bousquet, J., and Ertz, M. esports: Historical review, current state, and future challenges. `https://www.researchgate.net/publication/352808190_eSports_Historical_Review_Current_State_and_Future_Challenges`, 2021.

[9] Chen, S., Thaduri, U., and Ballamudi, V. Front-end development in react: An overview. *Engineering International 7* (12 2019), 117–126.

[10] Chen, Z., Amato, C., Nguyen, T.-H. D., Cooper, S., and Sun, Y. *A fast deck recommendation system for collectible card games*. The 2018 Computational Intelligence and Games (CIG) conference, 2018.

[11] Chen, Z., Nguyen, T.-H. D., Xu, Y., Amato, C., Cooper, S., Sun, Y., and El-Nasr, M. S. *The art of drafting: a team oriented hero recommendation system for multiplayer online battle arena games*. Conference'17, July 2017, Washington, DC, USA, 2017.

[12] Cong, Y. *League of Legends Data Visualization*. School of Information, Pratt Institute, 2019.

[13] D. J. Abadi. Query Execution in Column-Oriented Database Systems, 2009.

[14] Delcev, S., and Drazen, D. Modern javascript frameworks: A survey study. In *2018 Zooming Innovation in Consumer Technologies Conference (ZINC)* (2018), pp. 106–109.

[15] Do, T. D., Wang, S. I., Yu, D. S., McMillian, M. G., and McMahan, R. P. *Using Machine Learning to Predict Game Outcomes Based on Player-Champion Experience in League of Legends*. International Conference on the Foundations of Digital Games (FDG) 2021, 2021.

[16] do Nascimento Junior, F. F., da Costa Melo, A. S., da Costa, I. B., and Marinho, L. B. *Profiling Successful Team Behaviors in League of Legends*. Brazillian Symposium on Multimedia and the Web, 2017.

[17] Drachen, A., Sifa, R., Bauckhage, C., and Thurau, C. *Guns, swords and data: Clustering of player behavior in computer games in the wild*. 2012 IEEE Conference on Computational Intelligence and Games (CIG), 2012.

[18] E. A. Brewer. Towards Robust Distributed Systems, 2000.

[19] Ferry, E., O'Raw, J., and Curran, K. Security evaluation of the oauth 2.0 framework, 2015.

[20] Gaina, R., and Nordmoen, C. *League of Legends: A Study of Early Game Impact*. Queen Mary University of London, UK, 2019.

[21] George, J. *Kills per Win and Deaths per Loss: The Most Important League of Legends Stats*. Fantasy Labs, 2018.

[22] Hari Om Pathak, Dr. Vishal Shrivastava, D. A. P., and Kumar, S. Research on analysis of java script, 2024.

[23] Horst, R., Meyer, F., and Dörner, R. Draftcompromise - on draft composition recommendations in league of legends. In *2024 IEEE Gaming, Entertainment, and Media Conference (GEM)* (2024), pp. 1–6.

[24] Junior, J., and Campelo, C. League of legends: Real-time result prediction, 2023.

[25] Karelia, Dhruv and Mehta, Dhaivat. The Role of Data Science in Esports Analytics And Performance Evaluation. https://www.researchgate.net/publication/369912153_The_Role_of_Data_Science_in_Esports_Analytics_And_Performance_Evaluation, 2023.

[26] Ke, C., Deng, H., Xu, C., Li, J., Gu, X., Yadamsuren, B., Klabjan, D., Sifa, R., Drachen, A., and Demediuk, S. Dota 2 match prediction through deep learning team fight models, 08 2022.

[27] Kleinman, E., and El-Nasr, M. S. *Using Data to "Git Gud": A Push for a Player-Centric approach to the Use of Data in Esports.* CHI 2021, May 08–13, Yokohama, Japan, 2021.

[28] Minami, S., Koyama, H., Watanabe, K., Saijo, N., and Kashino, M. Prediction of esports competition outcomes using eeg data from expert players. *Computers in Human Behavior 160* (2024), 108351.

[29] Natasha Ferguson. Single Page Applications (SPA), 2023.

[30] Nishtha Jatana, Sahil Puri, Mehak Ahuja, Ishita Kathuria, Dishant Gosain. A Survey and Comparison of Relational and Non-Relational Database, 2012.

[31] Novak, A. R., Bennett, K. J. M., Pluss, M. A., and Fransen, J. *Performance analysis in esports: part 1 - the validity and reliability of match statistics and notational analysis in League of Legends.* SportRxiv, 2019.

[32] Owusu-Adjei, M., Hayfron-Acquah, J. B., Frimpong, T., and Abdul-Salaam, G. A systematic review of prediction accuracy as an evaluation measure for determining machine learning model performance in healthcare systems. *medRxiv* (2023).

[33] Pedrassoli Chitayat, A., Block, F., Walker, J. A., and Drachen, A. Applying and visualising complex models in esport broadcast coverage. In *Proceedings of the 2024 ACM International Conference on Interactive Media Experiences* (2024), Association for Computing Machinery, p. 108–116.

[34] Prateek Rawat. ReactJS: A Modern Web Development Framework, 2020.

[35] Prateek Rawat, A. N. M. Reactjs: A modern web development framework, 2020.

[36] R. Elmasri and S. B. Navathe. Fundamentals of Database Systems, 2015.

[37] Ramirez-Cano, D., Colton, S., , and Baumgarten, R. *Player classification using a meta-clustering approach.* CGAT 2010 - Computer Games, Multimedia and Allied Technology, Proceedings, 2010.

[38] Ravari, Y. N., Spronck, P., Sifa, R., and Drachen, A. *Predicting Victory in a Hybrid Online Competitive Game: The Case of Destiny.* AIIDE, 2017.

[39] Reddy, K. Web applications with spring boot, 2017.

[40] Reddy, M. Analysis of component libraries for react js. *IARJSET 8* (06 2021), 43–46.

[41] Sarker, I. H. Machine learning: Algorithms, real-world applications and research directions, 2021.

[42] Schubert, M., Drachen, A., and Mahlmann, T. *Esports analytics through encounter detection.* MIT Sloan Sport Analytics Conference 2016, 2016.

[43] Sevenhuysen, T. *Blind Pickability of Highest Presence Champions in Pro – Patch 10.16.* Oracle's Elixir, 2020.

[44] Singha, R. Spring boot backend development, 2022.

[45] Smithies Tim D., Campbell Mark J., R. N. T., and J., A. A random forest approach to identify metrics that best predict match outcome and player ranking in the esport rocket league, 2021.

[46] Tim Sevenhuysen. Oracle's Elixir: League of Legends Statistics and Data.

[47] Ting, K. M. *Confusion Matrix.* Springer US, Boston, MA, 2010, pp. 209–209.

[48] Tu, Z. Research on the application of layered architecture in computer software development. *Journal of Computing and Electronic Information Management 11* (11 2023), 34–38.

[49] Vrushali Y Kulkarni, D. P. K. S. Random forest classifiers :a survey and future research directions, 2013.

[50] Wan, Z., and Sándor, B. Predicting game outcome in dota 2 with nlp and machine learning algorithms, 2023.

[51] Wang Tian. Predictive Analysis on eSports Games: A Case Study on League of Legends (LoL) eSports Tournaments. https://cdr.lib.unc.edu/concern/masters_papers/8s45qd54c, 2018.

[52] Yang, P., Harrison, B. E., and Roberts, D. L. *Identifying patterns incombat that are predictive of success in MOBA games*. Proceedings of the Foundations of Digital Games 2014 Conference (FDG 14), 2014.