

## 머신러닝 적용 경륜 경주 순위 및 베팅방식별 결과 예측에 관한 연구

김 필 수(제1저자)  
한국스포츠경영전략연구원(원장)

전 성 삼(공동저자)  
에이치앤스포츠(연구원)

이 상 현(교신저자)  
아주대학교 일반대학원(박사과정)

## A Study on the Application of Machine Learning to Predict Keirin Competition Ranking and Sports Betting Methods

Kim, Philsoo(First Author)  
Korea Sports Management Research Institute (Director)

Jeon, Seong Sam(Co Author)  
HN Sports (Researcher)

Lee, Sang Hyun(Corresponding Author)  
Graduate School, Ajou University (Doctoral Student)

## Abstract

The The sports betting industry has proven to be one of the most influential areas in the service sector. Despite its significance, the Korean sports betting industry has been neglected which calls for research scrutiny. This research applies machine learning algorithms (Logistic Regression, Random Forest, AdaBoost, GradientBoost, Light-GBM, Multi-Layer Perceptron, Extra GradientBoost) to predict the results of Keirin competition along with sports betting methods. All of the race data generated in 「Gwangmyeong Speedome」 from 2016 to 2022 were collected and preprocessed for empirical analysis using Python. The results imply that the Logistic Regression had the highest accuracy performance among the machine learning algorithms, with an accuracy of 61.18% for the win prediction, 78.51% for perfecta, 42.37% for the quinella, 31.33% for the exacta, 31.63% for the trio, 22.10% for quinella place, and 14.30% for trifecta bet. Light-GBM and GradientBoost demonstrated the second-highest performance among the machine learning algorithms. In conclusion, this research provides an analysis of the machine learning application of Keirin competition based on sports betting methods. We believe this attempt may contribute to the service management research domain by providing actual prediction results of the sports game to consumers that may to sports betting industry expansion.

*Keywords* : machine learning, sports analytics, sports prediction, sports betting, sports service management

접수일(2023년 05월 07일), 수정일(2023년 05월 31일), 게재확정일(2023년 06월 07일)

## I. 서론

4차 산업의 시대를 본격적으로 맞이하여 현대사회가 경제 사회적으로 발전하고 기술 수준이 급격한 진보를 이루면서, 스포츠 서비스 산업은 지속해서 발전해왔다. 그중 가장 눈에 띄는 산업 중의 하나는 스포츠산업이다. 전 세계 스포츠산업의 규모는 2019년 기준 7,570억 달러로 나타난 바 있으며, 2025년에는 1조 달러를 넘어설 것으로 추정된다. 이미 스포츠산업에서 창조되고 파생되는 다양한 서비스 콘텐츠와 문화적 요소는 국가, 나이, 성별에 국한되지 않고 가장 많이 소비되는 것으로 나타났다(Milano & Chelladurai, 2011; PricewaterhouseCooper, 2021; Wang & Zeng, 2020; Yang, 2020; Zhang, Kim, Mastromartino, Qian, & Nauright, 2018). 스포츠산업 안에서도 무엇보다 주목할 것은 스포츠 서비스 산업 내 확장 속도가 가장 빠른 분야 중의 하나인 스포츠베팅(Sports Betting) 분야이다. 구체적으로 전 세계 스포츠베팅 산업 규모는 2022년 기준 9,360억 달러로 보고되고 있으며, 매년 10% 안팎의 성장을 기록하고 있다(Grand View Research, 2022; Lopez-Gonzalez & Griffiths, 2018; Lopez-Gonzalez, Guerrero-Solé, Estévez, & Griffiths, 2018; Vantage Market Research, 2022).

스포츠베팅을 운영하는 방식은 국가와 종목별로 상이할 수 있으나, 기본적으로 스포츠베팅은 기존에 존재하던 프로스포츠 종목의 경기 자체를 콘텐츠로 그대로 활용하여 확장할 수 있는 산업이라고 볼 수 있다. 스포츠베팅을 통해 국가 및 공익사업을 위한 각종 재원을 마련할 수 있으며, 새로운 일자리를 창출은 물론 스포츠에 관한 관심과 적극적 참여를 유도할 수 있다는 점에서 산업적인 성장잠재력과 가치가 큰 분야이다(노주원, 2021; 이원재, 황다빈, 신성연, 2020; 이장영, 조옥연, 김의진; 최창환, 윤지운, 2017). 하지만, 국내의 스포츠베팅 산업은 글로벌 패러다임과는 달리 제한적으로 발전되어왔다. 무엇보다 스포츠베팅 산업이 활발하게 전개되고 있는 미국이나 유럽 등과 비교하면 국내 스포츠베팅 산업의 경우 국가가 엄격하게 소비자들의 참여 방식을 통제하고 있다(김문영, 2012; 김진세, 김대희, 2020; 김현일, 2022; 정해상, 2011). 이러한 원인으로는 크게 스포츠베팅이 도박이라고 하는 부정적인 인식에 간혀 있다는 점과 과거 스포츠관계자들이 승부조작과 연루된 경험에서 오는 위험요인, 마지막으로 스포츠베팅에 대한 접근의 어려움을 들 수 있다(김경식, 한승백, 2018; 김필수, 이상현, 전성삼, 2023; 노주원, 2021; 이지용, 박재현, 윤지운, 윤효준, 2020). 따라서, 스포츠베팅과 스포츠 서비스 산업의 종합적인 발전을 위해서는 이러한 부정적 인식을 극복하는 것이 중요한 과제라 볼 수 있다.

한편, 최근 ChatGPT와 Bard로 대변되는 대화 생성형 인공지능의 발달은 학자들과 스포츠관계자들이 머신러닝을 통한 스포츠 데이터의 분석과 애널리틱스의 적용 및 예

측에 관한 관심을 증폭시켰다. 농구, 야구, 축구 등의 메이저 스포츠 종목에서는 머신러닝을 통한 경기결과의 예측에 대한 시도가 비교적 오래 반복되어 왔으며(김태훈, 임성원, 고진광, 이재학, 2020; Alonso & Babac, 2022; Baboota & Kaur, 2019; Xie, Xu, Nie, & Nie, 2016), 최근에는 다양한 머신러닝 알고리즘과 하이퍼 파라미터를 적용한 상당히 높은 수준의 예측이 가능해졌다(김필수, 이상현, 2023; 예원진, 이성노, 2022). 반면, 스포츠베팅 분야에서의 머신러닝 적용은 규칙의 복잡성으로 인해 비교적 늦게 태동하게 되었으며, 최근에서야 머신러닝의 적용 가능성에 대한 논의가 시작되었다(김필수, 이상현, 전성삼, 2023; Borowski & Chlebus, 2021; Bunker & Susnjak, 2022). 관련된 최근까지의 연구를 고찰하면 주로 경기 승/패의 결과 예측(Hubáček, Šourek, & Železný, 2019; Wilkens, 2021)과 방법론(Kollár, 2021; Sullivan, Myers, Harman, & Srivastava, 2023)에 대한 논의를 중심으로 진행되며, 주요 종목으로는 축구(Knoll & Stübinger, 2020; Stübinger, Mangold, & Knoll, 2019), 경마(Borowski & Chlebus, 2021; Chung, Chang, & Ko, 2017; Min, Lee, Park, & Lee, 2021), 테니스(Wilkens, 2021) 등으로 정리될 수 있다.

선행연구에서는 머신러닝을 스포츠베팅에 적용하면 기존 국내 스포츠베팅의 발전을 저해하였던 부정적인 ‘도박 이미지’나 ‘승부조작’과 연관될 수 있다는 측면에서의 관점을 탈피하고 스포츠 서비스 산업으로서의 베팅 범주에 신규로 진입하는 소비자들의 접근성을 높임으로써 전반적인 스포츠베팅 산업의 발전에 도움을 줄 수 있다고 본다(이지용, 박재현, 윤지운, 윤효준, 2020). 최근 이러한 점을 인지한 국민체육진흥공단은 대중이 건전한 방식으로 스포츠베팅과 관련한 서비스를 즐길 여건을 마련하기 위해 국가의 관리 아래 경륜 경주 결과를 예측하고 이를 활용할 수 있는 방법을 소비자들과 공유하고 교육하는 것의 중요성에 대해 발표하였다. 국민체육진흥공단은 경륜 종목에서 머신러닝을 통한 경기결과 예측과 분석 결과의 서비스화를 한국스포츠경영전략연구원과 본격적으로 시범운영 하여 본격화함으로써 새로운 서비스 산업의 성장동력을 마련하였다고 판단되며, 이를 다양한 종목으로 확장할 방법에 대해 강구하고 있는 것으로 확인된다(강석봉, 2023). 즉, 스포츠베팅 종목의 결과를 예측하고 이에 대한 서비스를 제공하는 것은 새로운 서비스 산업을 구축함으로써 국가의 경제에 긍정적인 효과를 미치는 것은 물론, 대중이 여가를 즐길 수 있는 선택지를 추가할 수 있다는 측면에서 큰 의미를 지닌다.

이러한 실무적 요구를 만족시키기 위해서 선행되어야 하는 것 중 하나는 서비스의 기반이 되는 기술 고도화와 학문적 검증이라고 볼 수 있다. 최근 이러한 관점을 확장하는 차원에서 국내에서는 머신러닝을 적용하여 스포츠베팅의 대표종목인 경륜 순위를 예측한 연구가 수행된 바 있다(김필수 외, 2023). 스포츠 애널리틱스에 적용될 수 있는 8가지 머신러닝 알고리즘의 성능을 비교하여 경륜 경주 결과를 예측함으로써 스

포츠베팅과 서비스 산업으로서의 경륜 경기 분야의 연구 발전에 이바지한 바 있지만, 순위의 다양한 조합을 이용한 실제 소비자들의 스포츠베팅 전략과 밀접한 세부적인 7가지 승식을 충분히 반영하지 못했다는 한계를 지닌다. 본 연구는 서비스 산업으로서의 스포츠베팅 발전과 소비자들의 의사결정에 실질적인 도움을 주기 위해 경륜 경주의 실제 베팅방식을 그대로 적용한 머신러닝 분석을 실증함으로써 솔루션을 제시하는 것을 목적으로 한다. 이를 위해 스포츠베팅과 관련된 대표적인 종목의 하나인 경륜 경주의 2016년부터 2022년 경기기록을 활용하여 실증분석하고, 실제 베팅이 이루어지는 승식의 예측 정확도를 제시하고자 한다. 보다 구체적으로 그라디언트 부스트, 다층신경망, 로지스틱 회귀, 랜덤 포레스트, 에이다 부스트, Light-GBM, XG 부스트 알고리즘을 사용하여 2016년부터 2022년까지의 학습한 경주기록을 바탕으로 2022년의 경주기록에 기반한 승식별 예측을 진행하고자 한다.

이러한 시도는 서비스 산업으로서의 스포츠베팅 분야의 대표적인 종목 중의 하나인 경륜 경주를 실제 승식을 예측하는 선도적인 연구의 노력으로서 해당 분야에 이론적으로 확장하는 바가 크다. 또한, 실무적으로 소비자들이 가지는 스포츠베팅의 부정적 이미지를 탈피하고 스포츠베팅을 이해하고 참여하는 데 도움이 되는 서비스를 실제로 제공할 수 있는 기초자료를 제시한다는 점에서도 의의가 있다고 볼 수 있다. 마지막으로, 머신러닝을 적용한 스포츠 애널리틱스와 세부적인 데이터 분석은 승부조작을 예방하는 시스템 구축을 가능하게 함으로써 스포츠 서비스 산업 전반의 발전에 영향을 미칠 것으로 본다.

## II. 이론적 배경

### 2.1 스포츠베팅 산업

전 세계적으로 스포츠베팅 산업이 활성화되고 발전하는 것에 비해, 국내 스포츠베팅 산업은 전체적인 규모나 역량 면에서 아직 국제표준(global standard)에 미치지 못한 상황이다. 스포츠베팅이 도박이라는 부정적인 인식으로 인해 베팅 참여자들도 주변에 쉽게 이를 표현하지 못하는 사회적 분위기나 스포츠베팅과 승부조작이 관련되어 있다는 인식이 존재한다. 이러한 이유로 국내에서 스포츠베팅은 제한된 방식으로만 허용되며, 프로스포츠계의 승부조작 사건으로 여론이 뜨거웠을 때나 사회적으로 도박이 문제가 되는 시점에 스포츠베팅이 중단되기도 했다. 또한, 엄격한 베팅참여 제한과 마케팅의 부재, 그리고 베팅전략 이해의 난해함, 참여자들의 노령화로 인해 스포츠베팅 산업의 배출은 감소하는 추세에 있다(김주연, 최현주, 안경모, 2016; 김진세,

2020; 노주원, 2021). 대표 분야인 경륜 매출은 코로나19의 여파를 고려하더라도 앞서 언급한 글로벌 패러다임의 흐름과는 정반대로 하락 추세에 있다. <표 1>은 스포츠베팅의 대표종목 중의 하나인 경륜의 매출 감소 추세를 나타낸다.

<표 1> 경륜 매출 현황(단위: 억 원)

| 연도  | 2016   | 2017   | 2018   | 2019   | 2020  | 2021  |
|-----|--------|--------|--------|--------|-------|-------|
| 매출액 | 22,818 | 21,744 | 20,515 | 18,337 | 3,509 | 3,792 |

출처: 2022 사행산업 관련 통계

국가가 운영하는 스포츠베팅의 산업을 통해 마련된 재원은 국가의 공익적 목적과 더불어 스포츠산업의 발전을 위해 활용되기 때문에 산업의 활성화 그 자체가 국가 발전을 위해 중요하며, 거시적인 측면에서는 본연의 스포츠 경주 콘텐츠를 기반으로 새로운 산업 확장을 도모할 수 있다는 점에서 새로운 일자리 창출에 도움이 될 수 있다. 특히, 온라인이나 모바일 플랫폼을 활용한 스포츠베팅 산업은 이와 관련한 기술 분야, 마케팅, 운영, 전략 등과 연계한 긍정적인 가치를 창출할 수 있다고 본다(김현일, 2022; 이장영, 조옥연, 2016). 아울러, 스포츠베팅에 직·간접적으로 참여하여 경험의 쌓이게 된 소비자들은 이전보다 스포츠를 통한 즐거움을 더 많이 느낄 수 있으며, 스포츠 서비스 산업에 더 적극적으로 참여하여 개인의 복지와 산업적 확장에도 긍정적인 영향을 미칠 수 있다(Kollár, 2021).

스포츠베팅을 하나의 유망한 산업으로 받아들이고 활성화하는 방향에 대한 활발한 논의가 진행되는 세계적인 흐름과는 다르게(Killick & Griffiths, 2022; Vlastakis, Dotsis, & Markellos, 2009), 국내 대다수의 스포츠베팅 연구는 도박이나, 불법 사설 베팅 처벌 등과 관련한 주제에 한정되어 있다는 점에서 국내의 스포츠베팅 산업 발전의 명확한 논의 측면의 한계를 확인할 수 있다(김도우, 박경래, 이창무, 2012; 박준휘, 최성락, 오시영, 2014; 장승원, 2010). 따라서, 스포츠베팅 산업이 본격적으로 활성화되기 위해서는 소비자들이 스포츠베팅을 건전한 취미활동으로 인식할 수 있도록 체계화하는 과정과 절차적 정당성이 필수적이라고 본다. 기존의 선행연구에서는 서비스 산업으로서의 스포츠베팅의 인식 전환과 참여 확대를 위해 이에 관한 충분한 정보의 공개, 도박이 아닌 예측과 분석으로서의 이미지, 다양하고 접근하기 편한 서비스의 확대, 사회적 인식의 전환을 위한 마케팅 등이 중요하다는 것을 주장했다(김주연, 최현주, 안경모, 2016; 김필수 외, 2023; 노주원, 2021; Lopez-Gonzalez & Griffiths, 2018).

무엇보다 스포츠베팅에 대한 전반적인 인식을 개선하고 더 많은 소비자 확대가 이루어지기 위한 선행요인 중의 하나로 사회적인 인식의 개선을 들 수 있다. 스포츠베

팅이 아무런 정보나 분석 없이 결과만을 추측해서 금전적 이익을 추구하는 구조라면, 이는 도박의 영역에 들어갈 여지가 충분하다고 본다. 하지만, 베팅 의사결정을 내리기 위한 분석의 측면에서 기본적인 정보가 소비자들에게 주어지고 이를 활용하여 과학적이고도 체계적인 분석이 이루어질 수 있도록 자연스럽게 유도하여 경기에 더 관심을 두고 집중할 수 있는 구조가 만들어진다면, 이는 분석과 예측의 영역에 속한 스포츠 경기로 볼 수 있다. 이러한 맥락에서 스포츠베팅의 활성화는 믿을만한 정보의 공개와 스포츠베팅이 주변 사람들과 같이 할 수 있는 건전한 게임이라는 인식이 선행되어야 한다(Lopez-Gonzalez & Griffiths, 2018). 또한, 이를 넘어서는 다양한 상품개발과 마케팅 역시 중요한 요인 중의 하나로 볼 수 있다. 한 연구에 의하면 온라인 베팅 경험자들은 다양한 스포츠베팅 상품의 제공이 중요하며 이에 따라 소비자들은 좋아하는 스포츠 경기를 분석하고 성취감을 느끼는 등의 장점이 있다고 보고했다(김주연 외, 2016). 즉, 새로운 형태의 상품 혹은 정보제공과 스포츠베팅의 긍정적 효과에 대한 마케팅을 통한 사회적 인식변화가 스포츠베팅의 활성화를 위한 중요한 요인이 될 수 있다(김주연 외, 2016; 김혁, 2015; Lopez-Gonzalez & Griffiths, 2018).

## 2.2 스포츠베팅과 머신러닝

스포츠베팅은 그 자체가 스포츠의 결과에 대한 예측을 전제로 한다는 점에서 최근 급격하게 증가하고 있는 머신러닝을 활용한 스포츠 애널리틱스와 높은 정합성을 지닌다. 스포츠가 결과를 알 수 없다는 것이 중요한 매력 포인트이기는 하지만, 소비자들은 실시간으로 경기를 관람하면서 경기의 승패를 예측하고 예측이 적중했을 때 희열을 느낄 수 있다. 아울러, 스포츠관계자들은 경기결과를 예측하고 이에 적절한 전략의 수립이나, 구단 혹은 리그 운영에 참고하고자 한다. 이러한 수요에 따라서 스포츠 경기의 결과를 확인하고자 머신러닝을 적용하는 것은 학문적으로 실무적인 측면에서 더는 놀라운 일이 아니다.

Purucker(1996)를 시작으로 스포츠 분야에서 머신러닝을 적용하여 경기결과를 예측하고자 하는 노력은 꾸준히 진행되어왔다. 농구(안세환, 김영민, 2022; Horvat, Havaš, & Srpak, 2020; Thabtah, Zhang, & Abdelhamid, 2019), 야구(김태훈, 임성원, 고진광, 이재학, 2020; Huang & Li, 2021; Yaseen, Marhoon, & Saleem, 2022), 축구(Baboota & Kaur, 2019; Constantinou, 2019; Khan & Kirubanand, 2019) 등의 종목에서 머신러닝을 적용한 경기 예측에 관한 많은 연구를 확인할 수 있다.

경기 예측에 대한 사람들의 수요와 기술의 발달을 고려할 때, 경기결과에의 정확한 예측이 목적이 될 수 있다는 점에서 스포츠베팅 분야에서 머신러닝을 활용한 연구가 발전하는 것은 놀라운 일이 아니다. 하지만, 일반적인 스포츠 경기 승패 예

측과 비교해 스포츠베팅이 적용된 승자 예측에 관한 체계적인 실증연구는 많지 않다. 최근 축구(Knoll & Stübinger, 2020; Stübinger et al., 2019), 경마(Borowski & Chlebus, 2021; Chung, Chang, & Ko, 2017; Min, Lee, Park, & Lee, 2021), 테니스(Wilkens, 2021), 경륜(김필수 외, 2023) 등에서 산발적으로 스포츠베팅에 관한 연구가 수행되고 있으나 아직 충분한 연구가 이루어졌다고 보기 힘든 시점이다.

<표 2> 경륜/경정/경마 베팅방법

| 선수 수 | 베팅승식 | 내용                                              |
|------|------|-------------------------------------------------|
| 1    | 단승   | 1등 한 명 선택                                       |
|      | 연승   | 2등 안에 들어올 선수 한 명 선택                             |
| 2    | 복승   | 순서 관련 없이 2등 안에 들어올 선수 두 명 선택                    |
|      | 쌍승   | 1등과 2등을 순서대로 선택                                 |
| 3    | 삼복승  | 순서 관련 없이 3등 안에 들어올 선수 세 명 선택                    |
|      | 쌍복승  | 1등으로 들어오는 선수 한 명과 순서 관련 없이 2~3등으로 들어올 선수 두 명 선택 |
|      | 삼쌍복승 | 1·2·3등으로 들어올 선수 3명 순서대로 선택                      |

이는 스포츠베팅은 단순히 경기의 승패를 예측하거나 1위만을 예측하는 것보다는 복잡한 형태를 지닌다는 점에 기인하는 것으로 보인다. 스포츠베팅 산업의 대표적인 종목으로 분류될 수 있는 경륜 / 경정 / 경마의 경우, 국내에서는 <표 2>와 같이 다양한 조합을 통해 7가지 베팅승식을 가지고 있다. 이런 구조는 단순 승패나 1위를 예측하는 것보다 훨씬 복잡한 과정을 거치며 이에 따른 정밀한 결과를 예측하는 것은 매우 어려운 영역이라고 볼 수 있다.

하지만 지속적인 기술의 발달로 단순한 승패를 넘어선 여러 조합의 경기결과를 예측하는 것도 불가능하지는 않다. 경륜 출주표를 바탕으로 머신러닝으로 활용하여 단승, 연승, 삼복승의 승자를 예측한 국내의 실증분석은 눈여겨볼 필요가 있다(김필수 외, 2023). 해당 연구가 비교적 단순한 방식의 승식만으로 결과를 예측하였는데, 해당 실증연구는 머신러닝을 통한 스포츠베팅 종목의 승식 예측도 일정 수준 가능하다는 것을 보여주었다. 머신러닝을 통한 스포츠 경기결과 예측의 활용도는 스포츠베팅에 있어 세 가지 장점을 지닌다. 첫째, 분석 결과물을 가공하여 다양한 형태의 서비스 제공이 가능하다. 머신러닝을 통한 예측 결과는 여러 정보를 분석해서 승식별로 예상 확률을 제공한다는 점에서 이에 관한 정보를 가공하고 구체화하여 소비자들에 직접 제공할 수 있다. 스포츠 서비스의 측면에서 해당 경기 종목이나 선수에 관한 일정 수준의 지식이 있는 소비자는 머신러닝을 적용하여 도출된 실제 결과를 자신의 사전지식과 결합하여 더욱 깊이 있게 분석할 수 있으며, 이를 통해 스포츠베팅이 단순한 도

박에 불과하다는 인식에서 벗어나 예측과 분석의 영역이라고 하는 인식을 심어줄 수 있다.

둘째, 해당 경기 종목에 대한 충분한 지식이 없는데 스포츠베팅에 참여하고 싶은 소비자들에게 안전한 진입의 근거를 제시할 수 있으며 최소한의 길잡이가 될 수 있다. 특정 종목에 관한 관심이 베팅참여를 활발하게 할 수도 있지만, 거꾸로 베팅참여나 베팅참여를 통한 이익 획득이 해당 종목에 관한 관심을 불러일으키기도 한다. 이러한 현상은 인지적 사고의 결과가 행동을 일으키는 것으로 그치는 것이 아니라 행동이나 행동으로 인한 강화 체계가 사고나 관심의 방향을 바꿀 수 있다는 수많은 연구를 통해 지지가 되고 있다. 즉, 관심이 관여 정도를 높일 수도 있지만, 관여가 관심을 높이는 방향 역시 가능하다(권웅, 이현우, 2016; 김병률, 이진춘, 2021; Buzsáki, Peyrache, & Kubie, 2014).

셋째, 머신러닝을 통한 스포츠 경기의 분석은 부정행위 발견에 비교적 쉽다고 볼 수 있다. 머신러닝은 과거 자료를 바탕으로 새로운 정보의 패턴을 분석하는 데 장점을 지닌다. 이러한 속성 때문에 일부 연구자들은 머신러닝을 활용한 시뮬레이션을 통해 정상적인 경기와 비정상적인 경기결과를 구분할 가능성이 크다고 주장했다(이지용 외, 2020). 이는 스포츠베팅이 승부조작 등을 통해 건전한 스포츠의 활성화에 부정적인 영향력을 다소 완화할 수 있다는 점에서 시사하는 바가 크다. 오히려 스포츠베팅의 활성화로 많은 관찰자가 생기고 경기에서 생성되는 자료를 더욱 철저히 분석할 수 있는 계기를 만들게 됨으로써 현장에서 발견하기 힘든 비정상적인 플레이나 경기결과를 발견하게 될 가능성이 커질 수 있다.

따라서, 머신러닝을 통한 스포츠베팅 승식에 관한 연구는 베팅 참여자들이 이 결과를 통해 제공되는 자료를 분석함으로써 스포츠베팅을 서비스 산업의 측면에서 바라보는 전체적인 시각을 전환할 기회를 제공한다. 아울러, 새롭게 진입하는 소비자들에게는 좋은 길잡이가 될 수 있으며 승부조작과 관련될 수 있는 비정상적인 경기를 가려낼 수 있다는 점에서 중요하게 작동할 수 있다고 본다. 하지만, 국내에서는 아직 머신러닝을 활용하여 스포츠베팅 결과를 면밀하게 분석한 연구가 부족한 실정이다. 대표적인 스포츠 서비스 산업 내 스포츠베팅 종목인 경륜의 경우에도 극소수의 연구만이 머신러닝을 활용하여 승자를 예측하였으며, 이마저도 실제 스포츠베팅과 밀접한 승식에 관한 연구는 아니었다고 볼 수 있다(김필수 외, 2023; 최창환, 윤지운, 2017).

인공지능 기술이 급격히 발전하고 스포츠 경기의 다양한 기록 축적이 비교적 수월해진 현대의 스포츠 서비스 산업은 스포츠베팅의 종목과 머신러닝과 관련성이 높은 스포츠 서비스 연구 천착이 필요하며 적절한 시점이라고 볼 수 있다. 이러한 측면에서 본 연구의 목적은 스포츠베팅의 대표적인 종목인 경륜 경주의 경기기록을 머신러닝을 통해 분석함으로써 스포츠베팅 산업의 발전을 위한 새로운 형태의 연구 발전 가

능성을 모색하는 것에 있다.

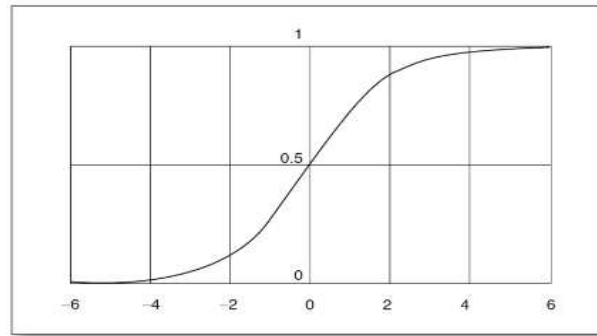
## 2.3 예측 모형에 대한 선행연구

본 연구는 경륜 경주의 1·2·3착을 예측하기에 적절한 머신러닝 알고리즘을 적용했다. 경륜 선수의 1·2·3착 이내 입상 가능성을 도출하는 것은 입상할 경우와 입상하지 못하는 경우로 분류하는 문제라는 점에서 분류(Classification) 모델을 적용하기에 적절하다. 경륜과 관련한 기존 연구(김필수 외, 2023)에서는 로지스틱 회귀, 랜덤 포레스트, 에이다 부스트 등이 경륜 순위를 예측하기에 적절한 알고리즘으로 나타났으며, 최근 XG Boost와 Light-GBM 등의 Gradient Boosting 기반의 알고리즘과 다층신경망 모델이 각종 데이터 분석 대회와 산업 분야에서의 활용이 늘어나고 있다는 점을 고려하여 이를 추가하였다. 따라서, 최종적으로 로지스틱 회귀, 랜덤 포레스트, 에이다 부스트, 그래디언트 부스트, Light-GBM, 다층신경망, XG Boost 7개의 모델을 사용하여 분석을 진행했다. 본 연구에서 사용된 머신러닝 알고리즘은 다음과 같다.

### 2.3.1 로지스틱 회귀(Logistic Regression)

로지스틱 회귀는 영국의 통계학자 D.R.Cox가 1958년에 제안한 확률 모델로 독립변수의 선형 결합을 이용하여 사건 발생 가능성을 예측하는 통계 기법이다. 독립변수의 선형 결합을 통하여 종속변수와 관계설명을 하고자 한다는 관점에서 선형 회귀분석과 유사하나, 종속변수가 범주형 데이터를 대상으로 하며 예측 결과가 특정 분류로 구분되기 때문에 일종의 분류 기법으로도 볼 수 있다. 일반적으로 로지스틱 회귀는 종속변수가 두 개인 이항 문제에 적용하며 종속변수가 두 개 이상이면 다항 로지스틱 회귀(multinomial logistic regression) 또는 분화 로지스틱 회귀(polytomous logistic regression)라고 한다(Hosmer, Lemeshow, & Sturdivant, 2013). 본 연구에서는 “1위 이내 입상”, “2위 이내 입상”, “3위 이내 입상”의 여부를 각각 1과 0으로 구분하는 이항 로지스틱 회귀모델(binomial logistic regression)을 적용하였다. <그림 1>은 로지스틱 회귀의 개념적 모형을 보여준다.

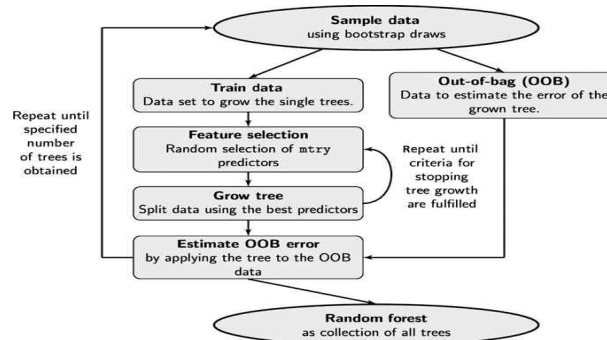
&lt;그림 1&gt; 로지스틱 회귀의 개념적 모형 (Osborne, 2014)



### 2.3.2 랜덤 포레스트(Random Forest)

랜덤 포레스트는 의사결정나무(Decision Tree)를 기반으로 하는 분류 알고리즘의 하나이다. 의사결정나무의 과적합 문제를 해결하기 위해 개발되었으며 원본 데이터에서 중복을 허용하여 무작위로 샘플을 추출하여 같은 크기의 샘플을 여러 개 생성하는 부트스트랩 샘플링(Bootstrap Sampling)과 각 의사결정나무의 특성을 무작위로 선택하는 무작위 특성 선택(Random Feature Selection)을 적용하였다. 생성된 각 의사결정나무는 서로 독립적으로 학습하고 예측하여 예측 결과의 평균을 최종 결과로 도출한다(Brieman, 2001; Migliorati, 2020). <그림 2>는 랜덤 포레스트의 개념적 모형을 보여준다.

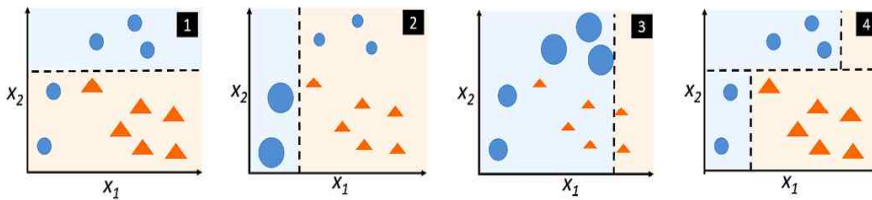
&lt;그림 2&gt; 랜덤 포레스트의 개념적 모형 (Janitza et al., 2012)



### 2.3.3 에이다 부스트(AdaBoost)

에이다 부스트는 앙상블(Ensemble) 기법의 하나로 여러 개의 약한 분류기를 결합하여 강력한 분류기를 만드는 기법이다. 모든 샘플에 같은 가중치를 할당하고 학습 결과를 바탕으로 모형의 오차를 계산하여 오차가 낮은 샘플에 대한 가중치를 감소시키고 오차가 높은 샘플에 대한 가중치를 증가시킨다. 이를 통해 다음 분류기에서 오분류된 샘플의 중요도를 높여 학습을 진행한다. 위 과정을 반복하여 최종적으로 모든 분류기의 결과를 결합하고 각 분류기의 가중치를 고려하여 최종 예측 결과를 도출한다. 적은 수의 오분류 샘플의 가중치를 높기 고려하기 때문에 데이터세트의 불균형 문제를 해결할 수 있으나 이상값(Outlier)에 민감하게 반응하고 복잡한 데이터세트에서 과적합(Overfitting) 우려가 존재한다. <그림 3>은 에이다 부스트의 개념적 모형을 보여준다(Rojas, 2009).

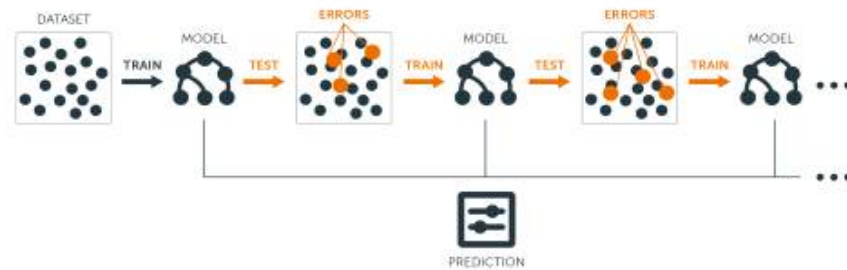
<그림 3> 에이다 부스트의 개념적 모형 (김필수 외, 2023)



### 2.3.4 그래디언트 부스트(GradientBoost)

그래디언트 부스트는 에이다 부스트와 같은 부스팅 알고리즘의 하나로 경사 하강법을 이용하여 오차를 최소화한다는 점에서 에이다 부스트와의 차이가 존재한다. 초기 분류기를 생성하고 예측 결과와 실제 결과의 차이를 계산하여 다음 분류기에 적용한다. 위 과정을 사용자가 지정한 조건에 만족할 때까지 반복하여 최종 분류기와 예측 결과를 도출한다. 다른 부스팅 알고리즘과 달리 잔차를 학습하여 모형을 개선하기 때문에 학습 데이터세트에 대한 예측 정확도와 일반화 성능을 모두 끌어올릴 수 있다(Arik & Pfister, 2021). <그림 4>는 그래디언트 부스트의 개념적 모형을 보여준다.

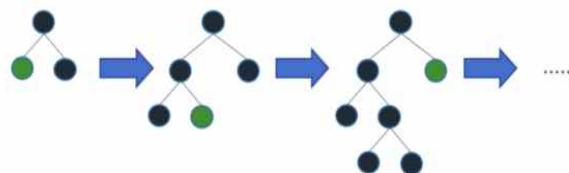
<그림 4> 그래디언트 부스트의 개념적 모형 (Boehmke & Greenwell, 2020)



### 2.3.5 Light-GBM

Light-GBM은 그래디언트 부스트를 발전시킨 알고리즘의 하나로 그래디언트 부스트의 단점인 대용량 데이터셋에서의 속도 저하를 해결하기 위해 기존의 알고리즘과 다르게 트리를 세분화하여 작은 데이터셋에서 빠르게 학습할 수 있도록 설계되었다. 이전 트리의 잔차를 이용하여 새로운 트리를 만들며, 그림자 트리(Shadow Tree)를 도입하여 효과적으로 계산을 처리한다. 또한, 병렬 계산이 가능하도록 설계되어 대용량 데이터셋에서도 빠른 학습이 가능하다. 다만, 깊은 트리를 생성할 가능성이 크기 때문에 과적합 위험이 있으며 데이터의 불균형이 심한 경우 예측 성능이 저하될 수 있다(Ke et al., 2017). <그림 5>은 Light-GBM의 개념적 모형을 보여준다.

<그림 5> Light-GBM의 개념적 모형 (Daoud, 2019)

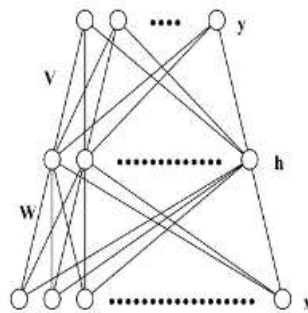


### 2.3.6 다층신경망(Multi-Layer Perceptron)

다층신경망은 인공신경망(Artificial Neural Network)의 기본적인 형태 중 하나로, 여러 개의 은닉층(hidden layer)을 가진 인공신경망이다. 각 층은 여러 개의 뉴런(neuron)으로 구성되며 입력층(input layer)에서 입력된 데이터는 은닉층을 거쳐 출력층(out layer)으로 전달되어 예측 결과를 출력한다. 다층신경망의 학습 과정은 예측

결과와 실제값의 차이를 최소화하기 위해 가중치(weight)를 업데이트하는 과정을 반복하는 것이다. 이때 역전파(Backpropagation) 알고리즘을 사용하여 각 뉴런에서의 오차를 계산하고 이를 기반으로 가중치를 업데이트한다. 다층신경망은 다양한 문제에 적용할 수 있으며 그림이나 음성 등의 복잡한 데이터의 처리에도 적합하지만, 학습 시간이 긴 편이고 과적합 문제가 발생할 수 있다는 단점을 지닌다(Agrawal, Girshick, & Malik, 2014). <그림 6>은 다층신경망의 개념적 모델을 보여준다.

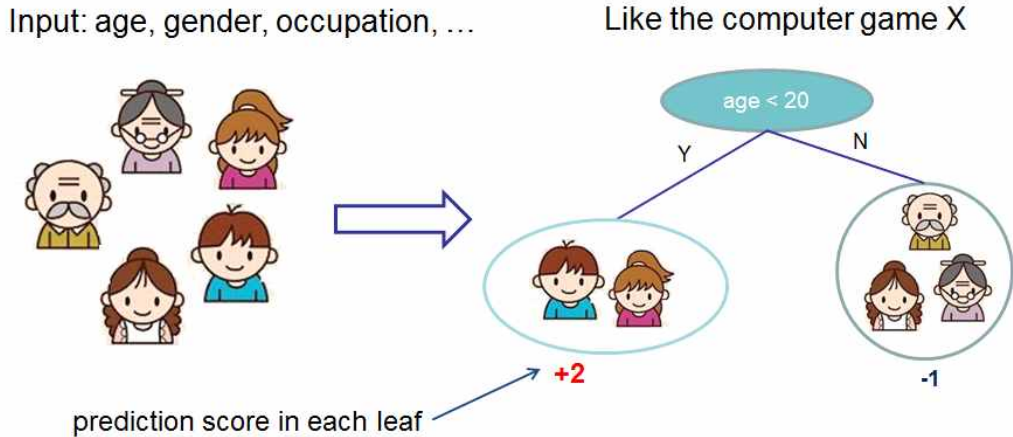
<그림 6> 다층신경망의 개념적 모형 (오상훈, 2009)



### 2.3.7 XG Boost (Extra GradientBoost)

XG Boost는 그래디언트 부스트를 발전시킨 또 다른 알고리즘으로 예측 성능이 높고 과적합 위험이 적어 대용량 데이터와 고차원 데이터의 처리가 비교적 쉽다고 볼 수 있다. XG Boost의 핵심 아이디어는 불순도를 최소화하며 트리를 구성하는 것이다. 이때 불순도란, 각 분기점에서 다른 클래스의 데이터가 섞여 있는 정도를 말하며 이 과정에서 다양한 기능을 사용한다. 예를 들어, 모델이 너무 복잡해지는 것을 예방하고 일반화 성능을 끌어올리기 위해 Regularization이라는 기능을 사용하며 이외에도 Tree Pruning, Custom Loss Function 등 다양한 기능을 사용하여 머신러닝 모델을 구성한다. 이러한 기능들로 인해 학습 및 예측을 보다 간편하고 효율적으로 수행할 수 있으며 그에 비해 우수한 예측 성능을 보여주어 대회 등 다양한 분야에서 널리 선호되는 알고리즘 중 하나이다(Ramraj, Uzir, Sunil, & Banerjee, 2016). <그림 7>은 XG Boost의 개념적 모형을 보여준다.

<그림 7> XGBoost의 개념적 모형 (XGBoost, 2022)



### Ⅲ. 연구방법

#### 3.1 연구 대상

본 연구에서는 머신러닝 알고리즘을 적용하여 경륜 경주 승자를 예측하는 것이다. 국민체육진흥공단의 「KCYCLE 홈페이지」와 「문화 빅데이터 플랫폼」을 통해 제공하는 “경륜 선수 출주표 데이터”와 “경륜 경주 결과 데이터”를 수집하여 2016년부터 2022년까지 진행된 경륜 경주에 대한 데이터를 확보하였다. 최근 경주에서 활용한 전법 코드를 변수로 활용하기 위해 해당 변수가 정밀하게 축적된 2016년도 경주 시점부터를 본 연구의 분석 샘플로 설정하였다. “경륜 선수 출주표 데이터”의 기어 배수, 200M, 승리율, 평균 득점 등 선수의 기록과 관련된 데이터와 “경륜 경주 결과 데이터”의 경주 별 1·2·3착 선수를 병합하기 위하여 각 선수가 1위, 2위, 3위 이내로 입상하는 경우를 각각 1로 코딩하였고 그렇지 않으면 0으로 코딩하였다.

경륜 경주는 결장 선수가 간혹 발생하여 6명의 선수로 경주가 진행되거나 공동으로 입상하여 1·2·3위가 한 명씩 존재하지 않는 경주가 발생할 수도 있다. 경륜 경주가 진행되는 금/토/일요일 이외에 공휴일이 있으면 4일 차까지 경주가 진행되는 사례도 발생한다. 큰 맥락에서 이러한 경주들은 예외적인 경우로 볼 수 있으며 예측 모델의 성능을 저하할 수 있으므로 본 연구의 분석대상에서는 제외하였다. 이러한 연구방법론의 절차적 타당성으로 본 연구에 실질적으로 적용된 최종 데이터 샘플은 2016년부터 2022년까지 10,164경주에 참가한 71,148명의 선수로 구성할 수 있으며 이를

대상으로 실증연구를 진행하였다. 이를 구체적으로 정리한 내용은 <표 3>과 같다.

<표 3> 연구 대상 자료 측정항목

| 구성개념       | 2016년  | 2017년  | 2018년  | 2019년  |
|------------|--------|--------|--------|--------|
| 선수 (단위:명)  | 13,650 | 12,880 | 13,146 | 8,288  |
| 경주 (단위:경주) | 1,950  | 1,840  | 1,878  | 1,184  |
| 구성개념       | 2020년  | 2021년  | 2022년  | 총계     |
| 선수 (단위:명)  | 2,170  | 4,858  | 16,156 | 71,148 |
| 경주 (단위:경주) | 310    | 694    | 2,308  | 10,164 |

### 3.2 연구 변수

출주표 데이터를 활용하여 경주 별 승자 예측을 진행하기 위해 79개의 독립변수와 3개의 종속변수를 사용하였다. <표 4>와 같이 79개의 독립변수는 3개 그룹으로 구분할 수 있다. 첫 번째 변수 그룹은 선수의 나이, 기어 배수, 연대율, 현재 등급, 평균 득점 등 출주표에서 확인할 수 있는 기본적인 기록 16개이며, 두 번째 변수 그룹은 진행 중인 회차의 1·2일 차와 최근 3회차 동안의 1·2·3일 차 경주에서의 네 가지 전법(선행, 찢히기, 추입, 마크)으로 입상하였는지를 1과 0으로 표시한 44개의 전법 관련 변수이다. 마지막은 선수가 참가한 경주에 대한 정보를 반영하기 위한 경주 데이터이다. 머신러닝의 특성상 선수 개개인의 기록과 입상 여부를 1:1로 연결하여 학습시키면 같은 경주에 출전한 선수들 간의 상대성이 고려되지 않는다는 한계가 있다. 이를 극복하기 위하여 경주에 참여한 선수 7명의 승리율, 평균 득점, 현재 등급 등의 평균과 해당 회차의 1일 차와 2일 차에 각 전법으로 입상한 선수의 수 등 19개의 변수를 같은 경주에 출전한 7명의 선수에게 같은 잣대로 부여하였다. 이를 통해 각 선수가 경주에 참여하는 7명의 선수 중 각 분야에서 어느 정도 수준에 위치하는지와 더불어 7명의 선수가 어떻게 구성되었는지를 예측에 반영할 수 있을 것으로 판단된다.

종속변수는 각 선수의 1·2·3착 이내 입상 여부이다. 1착으로 들어온 선수의 경우 “1위 이내 입상”, “2위 이내 입상”, “3위 이내 입상”을 모두 1로 표시하였으며 3착으로 들어온 선수는 “3위 이내 입상”은 1로 표시하고 “1위 이내 입상”과 “2위 이내 입상”은 0으로 표시하였다. 입상하지 못한 선수의 경우 세 변수 모두 0으로 표시하였다.

<표 4> 연구 변수 측정항목

| 독립변수_1 (16개)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 나이 수, 기어 배수 수, 기록 200M값, 승리율, 연대율, 3연대율, 입상 횟수, 경주 참가 횟수, 입상전법 선행 횟수, 입상전법 찢히기 횟수, 입상전법 추입 횟수, 입상전법 마크 횟수, 등급조정 현재명, 등급조정 이전 등급명, 평균 득점 종합값, 평균 득점 광명값                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
| 독립변수_2 (44개)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| 1일차 선행 전법, 1일차 추입 전법, 1일차 찢히기 전법, 1일차 마크 전법, 2일차 선행 전법, 2일차 추입 전법, 2일차 찢히기 전법, 2일차 마크 전법, 1전회차 1일차 선행 전법, 1전회차 1일차 추입 전법, 1전회차 1일차 찢히기 전법, 1전회차 1일차 마크 전법, 1전회차 2일차 선행 전법, 1전회차 2일차 추입 전법, 1전회차 2일차 찢히기 전법, 1전회차 2일차 마크 전법, 1전회차 3일차 선행 전법, 1전회차 3일차 추입 전법, 1전회차 3일차 찢히기 전법, 1전회차 3일차 마크 전법, 2전회차 1일차 선행 전법, 2전회차 1일차 추입 전법, 2전회차 1일차 찢히기 전법, 2전회차 1일차 마크 전법, 2전회차 2일차 선행 전법, 2전회차 2일차 추입 전법, 2전회차 2일차 찢히기 전법, 2전회차 2일차 마크 전법, 2전회차 3일차 선행 전법, 2전회차 3일차 추입 전법, 2전회차 3일차 찢히기 전법, 2전회차 3일차 마크 전법, 3전회차 1일차 선행 전법, 3전회차 1일차 추입 전법, 3전회차 1일차 찢히기 전법, 3전회차 1일차 마크 전법, 3전회차 2일차 선행 전법, 3전회차 2일차 추입 전법, 3전회차 2일차 찢히기 전법, 3전회차 2일차 마크 전법, 3전회차 3일차 선행 전법, 3전회차 3일차 추입 전법, 3전회차 3일차 찢히기 전법, 3전회차 3일차 마크 전법 |
| 독립변수_3 (19개)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| 경주 평균 기록200M값, 경주 평균 승리율, 경주 평균 연대율, 경주 평균 3연대율, 경주 평균 입상 횟수, 경주 평균 경주 참가 횟수, 경주 평균 입상전법 선행 횟수, 경주 평균 입상전법 찢히기 횟수, 경주 평균 입상전법 추입 횟수, 경주 평균 입상전법 마크 횟수, 경주 평균 등급조정 현재명, 경주 평균 등급조정 이전 등급명, 경주 평균 평균 득점 종합값, 경주 평균 평균 득점 광명값, 1일차 선행 전법 입상 선수, 1일차 추입 전법 입상 선수, 1일차 찢히기 전법 입상 선수, 1일차 마크 전법 입상 선수, 2일차 선행 전법 입상 선수, 2일차 추입 전법 입상 선수, 2일차 찢히기 전법 입상 선수, 2일차 마크 전법 입상 선수                                                                                                                                                                                                                                                                                                                              |
| 종속변수 (3개)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| 1위 이내 입상 여부 (단승식 승자), 2위 이내 입상 여부 (복승식 승자), 3위 이내 입상 여부 (삼복승 승자)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |

### 3.3 데이터 세트 분할

일반적으로 머신러닝을 활용한 분석에서 무작위로 학습데이터와 테스트데이터를 나누는 것이 보편적이지만, 본 연구에서는 연속된 7개의 데이터가 같은 경주에 참여한 7명의 선수에 해당하기 때문에 연도로 학습데이터와 테스트데이터를 구분하였다. <표 5>는 학습 및 테스트데이터의 구조를 보여준다. 2016년부터 2021년을 학습데이터로 2022년을 테스트데이터로 설정하였으며 학습데이터와 테스트데이터의 비율은

약 7.7:2.3에 해당한다. 6년간 축적된 과거의 경주 데이터를 기반으로 최근 경주 승자를 예측한다는 의미를 갖기도 한다.

<표 5> 데이터 분할

| 구성개념   | 기간         | 경주 수    | 선수 수    |
|--------|------------|---------|---------|
| 학습데이터  | 2016~2021년 | 7,856경주 | 54,992명 |
| 테스트데이터 | 2022년      | 2,308경주 | 16,156명 |

아울러, 방법론적인 측면에서 본 연구의 예측 모델 성능을 고도화하기 위하여 학습 데이터와 테스트데이터를 일차에 따라 구분하였다. 경륜 경기는 일반적으로 금·토·일 3일에 걸쳐 진행된다. 금요일과 토요일은 일반 경주가 진행되며, 일요일에는 일반 경주와 등급별 결승 경주가 진행된다. 회차에 참여한 선수 중 대부분이 3일 모두 경주를 진행하기 때문에 일차에 따라 경기 운영방식에 차이가 있을 수 있다는 가설을 기반으로 테스트하였다. 따라서, 학습데이터와 테스트데이터를 일차에 따라 3개로 나누어 학습과 예측을 진행하였다. <표 6>은 일차를 고려하여 학습데이터와 테스트데이터를 분할한 모습을 보여준다.

<표 6> 일차를 고려한 데이터 분할

| 구성개념   | 기간   | 경주 수    | 선수 수    |
|--------|------|---------|---------|
| 학습데이터  | 1일 차 | 2,632경주 | 18,424명 |
|        | 2일 차 | 2,627경주 | 18,389명 |
|        | 3일 차 | 2,597경주 | 18,179명 |
| 테스트데이터 | 1일 차 | 776경주   | 5,432명  |
|        | 2일 차 | 780경주   | 5,460명  |
|        | 3일 차 | 752경주   | 5,264명  |

### 3.4 예측 프로세스

경륜은 가장 빠른 선수를 가리는 기록 경주가 아니며, 능력과 기량이 비슷한 선수들이 각자 전법을 구사하여 결승선에 도착하는 순위 경주이다(KCYCLE 홈페이지). 이러한 맥락에서 종속변수로는 기록이 아닌 각 순위 내 입상 여부를 선택하였다. 그러나 실제 경륜에 적용하기 위하여 승식별 예측이 필요하며, 그에 앞서 1·2·3착으로 들어올 선수를 예측할 수 있어야 한다. 이를 파악하기 위해 모델의 예측 결과인 각

순위 내 입상 확률을 아래 프로세스를 통해 1·2·3착 선수로 치환하였다. <그림 1>은 해당 프로세스를 시각화한 것을 보여준다.

<그림 1> 경륜 경주 착순 예측 프로세스 모델

| 1위 선수 예측 |    |     |     |  |
|----------|----|-----|-----|--|
| 선수       | 1위 | 2위내 | 3위내 |  |
| A        | 16 | 30  | 45  |  |
| 1위 B     | 60 | 72  | 80  |  |
| C        | 14 | 25  | 40  |  |
| D        | 30 | 40  | 45  |  |
| E        | 20 | 30  | 50  |  |
| F        | 45 | 60  | 70  |  |
| G        | 20 | 25  | 35  |  |
| 단위: %    |    |     |     |  |

| 2위 선수 예측 |    |     |     |  |
|----------|----|-----|-----|--|
| 선수       | 1위 | 2위내 | 3위내 |  |
| A        | 16 | 30  | 45  |  |
| C        | 14 | 25  | 40  |  |
| D        | 30 | 40  | 45  |  |
| E        | 20 | 30  | 50  |  |
| 2위 F     | 45 | 60  | 70  |  |
| G        | 20 | 25  | 35  |  |
| 단위: %    |    |     |     |  |

| 3위 선수 예측 |    |     |     |  |
|----------|----|-----|-----|--|
| 선수       | 1위 | 2위내 | 3위내 |  |
| A        | 16 | 30  | 45  |  |
| C        | 14 | 25  | 40  |  |
| D        | 30 | 40  | 45  |  |
| 3위 E     | 20 | 30  | 50  |  |
| G        | 20 | 25  | 35  |  |
| 단위: %    |    |     |     |  |

우선 일차별로 학습데이터를 학습시킨 머신러닝 모델에 2022년 경주에 참여한 선수에 대한 기록 데이터, 전법 데이터, 경주 데이터를 7개 AI 머신러닝 알고리즘 모델에 입력하여 “1위 입상 확률”, “2위 이내 입상 확률”, “3위 이내 입상 확률”을 예측하였다. 이후 예측된 3가지 종속변수를 한 경주에 참여한 7명의 선수로 그룹화했다. 경주의 1·2·3착 선수를 도출하는 프로세스는 실제 경주에서 1·2·3착 선수가 결정되는 방식과 같은 순서로 구축했다. 가장 먼저 결승선을 통과하는 1착 선수는 7명 중 “1위 입상 확률”이 가장 높은 선수를 선택하고, 해당 선수를 제외한 6명의 선수 중 “2위 이내 입상 확률”이 가장 높은 선수를 2착 선수로 선택했다. 2착 선수까지 제외하고 남은 5명의 선수 중 “3위 이내 입상 확률”이 가장 높은 선수를 3착 선수로 선택하였다. 선택된 1·2·3착 선수를 바탕으로 승식별 예측을 진행했다. 단승식과 연승식 승자는 1착으로 예측된 선수로 설정하였고 복승식과 쌍승식은 1착과 2착 선수, 삼복승식·쌍복승식·삼쌍승식은 1·2·3착 선수로 선택했다.

이후 각 AI 머신러닝 알고리즘의 1·2·3착 선수와 7개 승식(단승식, 연승식, 복승식, 삼복승식, 쌍승식, 쌍복승식, 삼쌍승식)을 예측한 결과와 실제 결과를 비교하여 알고리즘별로 예측 정확도를 산출했다.

### 3.5 평가 지표

#### 3.5.1 머신러닝 모델 성능 평가

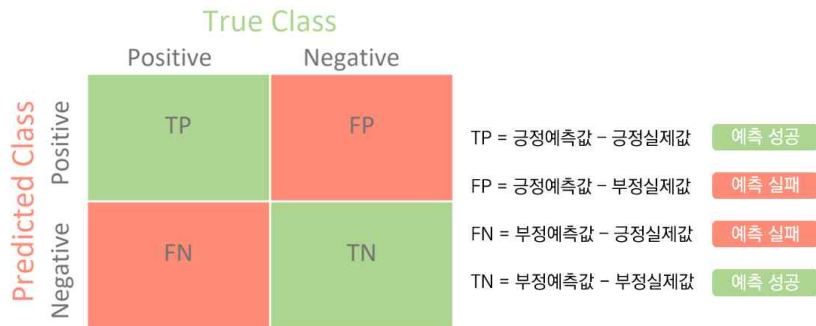
각 머신러닝 모델의 성능을 평가하기 위하여 학습 정확도, 테스트 정확도, 평균 제곱근 오차(RMSE), 결정계수(R-squared), 정밀도(Precision), 재현율(Recall), F1 점수(F1-Score)를 활용했다.

분류 모델을 평가하기 위해서 사용되는 정확도, 정밀도, 재현율, F1 점수는 <그림 9>에 표현된 Confusion Matrix를 통해 도출된다. Confusion Matrix는 실제 True를 True로 예측(TP), 실제 False인 것을 True로 예측(FP), 실제 True를 False로 예측(FN), 실제 False인 것을 False로 예측(TN)의 네 가지로 구분된다. 정확도는 True와 False와 관계없이 정확하게 예측한 비율로  $TP + TN / TP + FN + FP + TN$ 에 해당한다. 정밀도는  $TP / TP + FP$ 로 모델이 True라고 분류한 것 중 실제 True인 것의 비율이다. 재현율은  $TP / TP + FN$ 으로 실제 True인 것 중 모델이 True로 예측한 비율이다. 정밀도와 재현율은 1에 가까울수록 모델 성능이 좋다고 볼 수 있으나 각자 한계점을 가지며 서로 반비례하는 경향이 있다. 이를 보완하기 위해 정밀도와 재현율의 조화평균을 사용하는 F1 점수가 등장하였다. 즉 두 지표를 균형 있게 반영하여 모델의 성능을 직관적으로 평가할 수 있는 지표이다.

학습 정확도와 테스트 정확도는 각각 학습데이터와 테스트데이터를 적용하였을 때의 정확도를 의미한다. 모델이 학습데이터에 지나치게 최적화되어 학습데이터에 대해서는 높은 설명력을 갖지만 다른 데이터셋이 주어졌을 때 설명력이 떨어지는 것을 과적합이라고 한다. 학습 정확도와 테스트 정확도의 차이가 큰 경우에는 모델이 과적합 되었다고 볼 수 있다. 이를 극복하기 위하여 과적합 위험이 적은 알고리즘을 선택하거나 하이퍼 파라미터 조정을 통해 일반화 성능의 향상을 시도하기도 한다.

평균 제곱근 오차와 결정계수는 주로 회귀모델의 성능을 평가하기 위해 사용되지만, 본 연구에서는 각 선수의 입상확률을 예측하기 때문에 회귀모델과 유사한 특성이 있다. 따라서 결정계수와 평균 제곱근 오차를 적용하여 모델의 설명력을 평가했다. 평균 제곱근 오차는 실제값과 예측값 제곱의 평균을 구하는 MSE의 제곱근으로 오류 지표를 실제값과 유사한 단위로 나타내어 해석을 쉽게 한다. 예측 모델의 성능이 좋을수록 평균 제곱근 오차의 값은 감소한다. 결정계수는 모델이 데이터를 얼마나 정확하게 설명할 수 있는가를 직관적으로 수치화한 것으로 0에서 1 사이의 값을 가지며 1에 가까울수록 설명력이 크다고 판단한다(Haghighi, Jasemi, Hessabi, & Zolnvari, 2018; Liang, 2022).

&lt;그림 9&gt; Confusion Matrix



### 3.5.2 승식별 예측 정확도 평가

머신러닝 모델을 통해 예측할 수 있는 결과는 각 선수의 1·2·3위 이내 입상 여부이다. 경주에 따라 1위 입상 선수가 한 명도 존재하지 않을 수도 있고 반대로 2명 이상의 선수가 1위 이내 입상 선수로 예측될 수도 있다. 이는 현실에 적용하기 어려울 뿐만 아니라 경륜 경주 배팅에도 활용하기 어렵다.

따라서 3.4의 프로세스를 적용하여 한 경주에 참여하는 7명의 선수의 입상확률을 비교해 경주 별로 1·2·3착 선수를 한 명씩 도출하였다. 4.2에서 제공하는 예측 정확도는 2022년에 진행된 2,038경주에 대하여 모델을 통해 예측한 1·2·3착 선수와 실제 경기결과의 1·2·3착 선수를 각각 비교하고 착 순을 적용하여 승식별 예측에 성공한 비율을 나타낸다. 각 승식에 대한 단순한 산술적 확률은 단승(1/7), 연승(2/7), 복승(1/21), 삼복승(1/35), 쌍승(1/42), 쌍복승(1/105), 삼쌍승(1/210)에 해당한다.

## IV. 연구 결과

### 4.1 머신러닝 예측 모델의 성능 지표

7개 머신러닝 알고리즘을 활용하여 일차별로 1·2·3위 이내 입상 여부를 예측하는 63개의 머신러닝 예측 모델을 생성하여 각 일차에 적합한 학습데이터를 학습시킨 후 2022년에 경주에 참여한 선수들의 경주 결과를 예측하였다. 아래에 제시된 <표 6>에서부터 <표 14>까지는 63개의 머신러닝 모델의 학습데이터 정확도, 테스트데이터 정확도, RMSE, 결정계수, 정밀도, 재현율, F1 score를 나타낸다고 볼 수 있다. 로지스틱 회귀모델이 1일 차 1위 이내 입상 여부 예측 모형을 제외한 모든 모형에서 테스트 정확도가 가장 높게 나타났으며 그래디언트 부스트와 Light-GBM이 그 뒤를 이었다. 모델의 성능을 종합적으로 평가하기 위하여 고려되는 결정계수와 F1 점수는 로지스틱 회귀모델이 모든 모형에서 가장 높게 도출되었다.

1·2·3일차 모두 1위 예측 모형에서 2위 이내 입상 예측 모형, 3위 이내 입상 여부 예측 모형으로 범위가 확장될수록 테스트 정확도는 하락하지만, 정밀도와 재현율이 모두 상승하며 F1 점수는 상승하였다. 일차별 테스트 정확도는 2일 차, 1일 차, 3일 차 순이었으며 2일 차 테스트 정확도는 3일 차에 비해 1.46~4.62% 정도 높았다.

<표 6> 1일 차 1위 이내 입상 여부 예측 모형 평가 지표

|           | 학습<br>정확도 | 테스트<br>정확도 | RMSE | 결정계수 | 정밀도  | 재현율  | F1 점수 |
|-----------|-----------|------------|------|------|------|------|-------|
| 로지스틱 회귀   | 89.43%    | 89.30%     | 0.33 | 0.13 | 0.73 | 0.40 | 0.52  |
| 랜덤포레스트    | 100.00%   | 88.81%     | 0.33 | 0.09 | 0.79 | 0.30 | 0.43  |
| 에이다부스트    | 88.82%    | 88.66%     | 0.34 | 0.07 | 0.74 | 0.32 | 0.45  |
| 그래디언트부스트  | 90.03%    | 89.38%     | 0.33 | 0.13 | 0.76 | 0.38 | 0.50  |
| Light-GBM | 93.34%    | 89.10%     | 0.33 | 0.11 | 0.72 | 0.39 | 0.51  |
| 다층신경망     | 91.34%    | 87.87%     | 0.35 | 0.01 | 0.61 | 0.41 | 0.49  |
| XG Boost  | 95.22%    | 88.62%     | 0.34 | 0.07 | 0.69 | 0.37 | 0.48  |

&lt;표 7&gt; 1일 차 2위 이내 입상 여부 예측 모형 평가 지표

|           | 학습<br>정확도 | 테스트<br>정확도 | RMSE | 결정계수 | 정밀도  | 재현율  | F1 점수 |
|-----------|-----------|------------|------|------|------|------|-------|
| 로지스틱 회귀   | 82.66%    | 81.94%     | 0.42 | 0.12 | 0.75 | 0.55 | 0.64  |
| 랜덤 포레스트   | 100.00%   | 81.44%     | 0.43 | 0.09 | 0.75 | 0.52 | 0.62  |
| 에이다 부스트   | 81.66%    | 80.45%     | 0.44 | 0.04 | 0.72 | 0.52 | 0.60  |
| 그래디언트 부스트 | 83.46%    | 81.44%     | 0.43 | 0.09 | 0.75 | 0.52 | 0.62  |
| Light-GBM | 87.47%    | 81.15%     | 0.43 | 0.08 | 0.74 | 0.52 | 0.61  |
| 다층신경망     | 85.10%    | 80.49%     | 0.44 | 0.04 | 0.69 | 0.58 | 0.63  |
| XG Boost  | 90.27%    | 81.37%     | 0.43 | 0.09 | 0.73 | 0.56 | 0.63  |

&lt;표 8&gt; 1일 차 3위 이내 입상 여부 예측 모형 평가 지표

|           | 학습<br>정확도 | 테스트<br>정확도 | RMSE | 결정계수 | 정밀도  | 재현율  | F1 점수 |
|-----------|-----------|------------|------|------|------|------|-------|
| 로지스틱 회귀   | 78.28%    | 77.25%     | 0.48 | 0.07 | 0.77 | 0.67 | 0.72  |
| 랜덤 포레스트   | 100.00%   | 76.73%     | 0.48 | 0.05 | 0.76 | 0.67 | 0.71  |
| 에이다 부스트   | 76.68%    | 75.42%     | 0.50 | 0.00 | 0.75 | 0.64 | 0.69  |
| 그래디언트 부스트 | 78.74%    | 76.73%     | 0.48 | 0.05 | 0.77 | 0.65 | 0.71  |
| Light-GBM | 83.05%    | 76.14%     | 0.49 | 0.03 | 0.76 | 0.65 | 0.70  |
| 다층신경망     | 80.25%    | 75.68%     | 0.49 | 0.01 | 0.74 | 0.67 | 0.70  |
| XG Boost  | 86.27%    | 75.81%     | 0.49 | 0.01 | 0.75 | 0.65 | 0.70  |

&lt;표 9&gt; 2일 차 1위 이내 입상 여부 예측 모형 평가 지표

|           | 학습<br>정확도 | 테스트<br>정확도 | RMSE | 결정계수  | 정밀도  | 재현율  | F1 점수 |
|-----------|-----------|------------|------|-------|------|------|-------|
| 로지스틱 회귀   | 89.78%    | 89.74%     | 0.32 | 0.16  | 0.71 | 0.47 | 0.57  |
| 랜덤 포레스트   | 100.00%   | 89.03%     | 0.33 | 0.10  | 0.73 | 0.36 | 0.49  |
| 에이다 부스트   | 89.21%    | 89.27%     | 0.33 | 0.12  | 0.71 | 0.42 | 0.53  |
| 그래디언트 부스트 | 90.30%    | 89.34%     | 0.33 | 0.13  | 0.71 | 0.43 | 0.54  |
| Light-GBM | 93.97%    | 89.73%     | 0.32 | 0.16  | 0.70 | 0.48 | 0.57  |
| 다층신경망     | 91.76%    | 87.67%     | 0.35 | -0.01 | 0.59 | 0.46 | 0.52  |
| XG Boost  | 95.88%    | 89.38%     | 0.33 | 0.13  | 0.69 | 0.47 | 0.56  |

<표 10> 2일 차 2위 이내 입상 여부 예측 모형 평가 지표

|           | 학습<br>정확도 | 테스트<br>정확도 | RMSE | 결정계수 | 정밀도  | 재현율  | F1 점수 |
|-----------|-----------|------------|------|------|------|------|-------|
| 로지스틱 회귀   | 83.69%    | 82.82%     | 0.41 | 0.16 | 0.74 | 0.61 | 0.67  |
| 랜덤 포레스트   | 100.00%   | 82.29%     | 0.42 | 0.13 | 0.76 | 0.56 | 0.64  |
| 에이다 부스트   | 82.41%    | 81.39%     | 0.43 | 0.09 | 0.73 | 0.56 | 0.63  |
| 그래디언트 부스트 | 84.28%    | 82.25%     | 0.42 | 0.13 | 0.75 | 0.57 | 0.65  |
| Light-GBM | 88.09%    | 82.29%     | 0.42 | 0.13 | 0.75 | 0.58 | 0.65  |
| 다층신경망     | 85.93%    | 81.32%     | 0.43 | 0.08 | 0.70 | 0.60 | 0.65  |
| XG Boost  | 91.32%    | 81.83%     | 0.43 | 0.11 | 0.73 | 0.57 | 0.64  |

<표 11> 2일 차 3위 이내 입상 여부 예측 모형 평가 지표

|           | 학습<br>정확도 | 테스트<br>정확도 | RMSE | 결정계수 | 정밀도  | 재현율  | F1 점수 |
|-----------|-----------|------------|------|------|------|------|-------|
| 로지스틱 회귀   | 78.18%    | 77.64%     | 0.47 | 0.09 | 0.78 | 0.67 | 0.72  |
| 랜덤 포레스트   | 100.00%   | 76.54%     | 0.48 | 0.04 | 0.76 | 0.66 | 0.71  |
| 에이다 부스트   | 76.76%    | 76.23%     | 0.49 | 0.03 | 0.75 | 0.66 | 0.70  |
| 그래디언트 부스트 | 79.00%    | 76.90%     | 0.48 | 0.06 | 0.77 | 0.65 | 0.71  |
| Light-GBM | 83.12%    | 76.52%     | 0.48 | 0.04 | 0.77 | 0.65 | 0.70  |
| 다층신경망     | 81.21%    | 76.12%     | 0.49 | 0.02 | 0.75 | 0.67 | 0.71  |
| XG Boost  | 87.56%    | 75.95%     | 0.49 | 0.02 | 0.76 | 0.64 | 0.70  |

<표 12> 3일 차 1위 이내 입상 여부 예측 모형 평가 지표 측정항목

|           | 학습<br>정확도 | 테스트<br>정확도 | RMSE | 결정계수  | 정밀도  | 재현율  | F1 점수 |
|-----------|-----------|------------|------|-------|------|------|-------|
| 로지스틱 회귀   | 87.64%    | 88.03%     | 0.35 | 0.02  | 0.67 | 0.32 | 0.44  |
| 랜덤 포레스트   | 100.00%   | 87.48%     | 0.35 | -0.02 | 0.72 | 0.20 | 0.32  |
| 에이다 부스트   | 86.95%    | 87.23%     | 0.36 | -0.04 | 0.65 | 0.23 | 0.34  |
| 그래디언트 부스트 | 88.23%    | 87.86%     | 0.35 | 0.01  | 0.69 | 0.28 | 0.39  |
| Light-GBM | 92.18%    | 87.71%     | 0.35 | 0.00  | 0.65 | 0.30 | 0.41  |
| 다층신경망     | 90.28%    | 86.15%     | 0.37 | -0.13 | 0.52 | 0.33 | 0.41  |
| XG Boost  | 93.66%    | 87.92%     | 0.35 | 0.01  | 0.64 | 0.35 | 0.45  |

&lt;표 13&gt; 3일 차 2위 이내 입상 여부 예측 모형 평가 지표

|           | 학습<br>정확도 | 테스트<br>정확도 | RMSE | 결정계수  | 정밀도  | 재현율  | F1 점수 |
|-----------|-----------|------------|------|-------|------|------|-------|
| 로지스틱 회귀   | 79.60%    | 80.26%     | 0.44 | 0.03  | 0.72 | 0.51 | 0.60  |
| 랜덤 포레스트   | 100.00%   | 79.18%     | 0.46 | -0.02 | 0.73 | 0.44 | 0.55  |
| 에이다 부스트   | 77.69%    | 78.17%     | 0.47 | -0.07 | 0.69 | 0.42 | 0.53  |
| 그래디언트 부스트 | 80.02%    | 79.26%     | 0.46 | -0.02 | 0.72 | 0.45 | 0.56  |
| Light-GBM | 84.89%    | 79.88%     | 0.45 | 0.01  | 0.71 | 0.50 | 0.59  |
| 다층신경망     | 82.84%    | 78.10%     | 0.47 | -0.07 | 0.65 | 0.51 | 0.57  |
| XG Boost  | 87.98%    | 79.67%     | 0.45 | 0.00  | 0.70 | 0.51 | 0.59  |

&lt;표 14&gt; 3일 차 3위 이내 입상 여부 예측 모형 평가 지표

|           | 학습<br>정확도 | 테스트<br>정확도 | RMSE | 결정계수  | 정밀도  | 재현율  | F1 점수 |
|-----------|-----------|------------|------|-------|------|------|-------|
| 로지스틱 회귀   | 73.78%    | 74.01%     | 0.51 | -0.06 | 0.74 | 0.61 | 0.67  |
| 랜덤 포레스트   | 100.00%   | 72.74%     | 0.52 | -0.11 | 0.73 | 0.58 | 0.65  |
| 에이다 부스트   | 71.67%    | 71.88%     | 0.53 | -0.15 | 0.71 | 0.58 | 0.64  |
| 그래디언트 부스트 | 74.27%    | 72.28%     | 0.52 | -0.11 | 0.72 | 0.59 | 0.65  |
| Light-GBM | 79.95%    | 73.14%     | 0.52 | -0.10 | 0.72 | 0.61 | 0.66  |
| 다층신경망     | 77.04%    | 71.88%     | 0.53 | -0.15 | 0.71 | 0.59 | 0.64  |
| XG Boost  | 83.72%    | 72.57%     | 0.52 | -0.12 | 0.71 | 0.61 | 0.66  |

## 4.2 승식 예측 정확도

본 연구에서는 2022년에 진행된 2,308경주의 결승선 도착순과 각 승식에 따른 승자를 예측한다. 위에서 평가된 머신러닝 모델을 기반으로 각 순위 이내의 입상 여부를 예측할 수 있다. 이를 3.4의 예측 프로세스에서 전술한 프로세스에 적용하여 각 경주의 1·2·3착 선수를 예측하였고 이를 실제 결과와 비교하여 승식별 예측 정확도를 확인하였다.

<표 15> 승식별 예측 정확도 (1)

| 순위/승식 | 로지스틱 회귀 | 랜덤 포레스트 | 에이다 부스트 | 그래디언트 부스팅 |
|-------|---------|---------|---------|-----------|
| 1차 선수 | 61.18%  | 58.45%  | 58.67%  | 59.92%    |
| 2차 선수 | 36.70%  | 33.80%  | 33.02%  | 34.92%    |
| 3차 선수 | 26.13%  | 25.26%  | 22.75%  | 25.26%    |
| 단승식   | 61.18%  | 58.45%  | 58.67%  | 59.92%    |
| 연승식   | 78.51%  | 76.60%  | 76.43%  | 78.21%    |
| 복승식   | 42.37%  | 40.16%  | 37.82%  | 40.68%    |
| 삼복승식  | 31.33%  | 30.46%  | 28.47%  | 30.72%    |
| 쌍승식   | 31.63%  | 29.42%  | 27.51%  | 29.90%    |
| 쌍복승식  | 22.10%  | 21.27%  | 19.15%  | 21.14%    |
| 삼쌍승식  | 14.30%  | 14.08%  | 11.53%  | 13.60%    |

로지스틱 회귀모델은 1·2·3차 선수에 대하여 61.18%, 36.70%, 26.13% 확률로 예측에 성공하였고 단승식 61.18%, 연승식 78.51%, 복승식 42.37%, 삼복승식 31.33%, 쌍승식 31.63%, 쌍복승식 22.10%, 삼쌍승식 14.30%로 모든 승식에서 7개 머신러닝 알고리즘 중 가장 높은 예측 정확도를 보였다. 랜덤 포레스트는 각 승식에 따라 단승식 58.45%, 연승식 76.60%, 복승식 40.61%, 삼복승식 30.46%, 쌍승식 29.42%, 쌍복승식 21.27%, 삼쌍승식 14.08%로 예측에 성공하였으며 쌍복승식과 삼쌍승식은 7개 알고리즘 중 세 번째로 잘 예측하였다. 에이다 부스트는 7가지 승식(단승식, 연승식, 복승식, 삼복승식, 쌍승식, 쌍복승식, 삼쌍승식)에 대하여 각각 58.67%, 76.43%, 37.82%, 28.47%, 27.51%, 19.15%, 11.53%의 예측 정확도를 보였다. 그래디언트 부스트는 7가지 승식에 대하여 59.52%, 78.21%, 40.68%, 30.72%, 29.90%, 21.14%, 13.60%의 예측 정확도를 도출하였다. 연승식과 복승식에서는 로지스틱 회귀 다음으로 높은 예측 정확도를 보였으며 나머지 승식에 대하여는 7개 머신러닝 알고리즘 중 세 번째로 정확하게 예측하였다.

<표 16> 승식별 예측 정확도 (2)

| 순위/승식 | Light-GBM | 다층신경망  | XG Boost |
|-------|-----------|--------|----------|
| 1차 선수 | 59.97%    | 53.12% | 58.06%   |
| 2차 선수 | 34.84%    | 31.33% | 34.62%   |
| 3차 선수 | 25.04%    | 23.44% | 24.35%   |
| 단승식   | 59.97%    | 53.12% | 58.06%   |
| 연승식   | 77.90%    | 72.57% | 76.47%   |

|      |        |        |        |
|------|--------|--------|--------|
| 복승식  | 40.08% | 37.18% | 40.42% |
| 삼복승식 | 30.85% | 28.55% | 30.07% |
| 쌍승식  | 29.98% | 25.13% | 29.46% |
| 쌍복승식 | 20.80% | 17.68% | 19.93% |
| 삼쌍승식 | 13.43% | 11.27% | 12.09% |

Light-GBM 알고리즘은 7가지 승식에 대하여 59.97%, 77.90%, 40.08%, 30.85%, 29.98%, 20.80%, 13.43%의 정확도로 예측하였다. 단승식과 삼복승식, 쌍승식에서 로지스틱 회귀 다음으로 잘 예측하였으며 연승식에서는 세 번째로 높은 예측 정확도를 보였다. 다층신경망과 XG Boost를 기반으로 하는 예측 모델은 7가지 승식에서 각각 다층신경망 53.12%, 72.57%, 37.18%, 28.55%, 25.13%, 17.68%, 11.27%, XG Boost 58.06%, 76.47%, 40.42%, 30.07%, 29.46%, 19.93%, 12.09%의 예측 정확도를 도출했다. 두 머신러닝 알고리즘은 다른 알고리즘에 비해 상대적으로 예측 정확도가 낮은 편에 속했다.

로지스틱 회귀모델이 모든 결승선 도착순과 승식에서 가장 높은 예측 정확도를 보였으며 Light-GBM과 그래디언트 부스트 기반 모델이 뒤를 이었다.

## V. 논의 및 결론

본 연구는 스포츠배팅 산업의 대표종목인 경륜 경주의 순위 및 승식을 머신러닝을 활용하여 예측하였다. 경륜 경주 결과의 예측은 승식에 따라 경기별 1위부터 3위까지를 7가지 방식으로 예측한다는 점에서 프로스포츠의 두 팀의 경기 승패나 정규리그 순위를 예측하는 것보다 훨씬 더 복잡한 형태를 지닌다. 이러한 이유로 경륜 결과 예측에 관한 연구와 이러한 아이디어를 스포츠 서비스 산업에 적용하고자 시도하는 노력은 국내외를 막론하고 찾아보기 힘들다고 볼 수 있다. 본 연구는 기존의 선행연구가 내재한 한계점을 극복하고 스포츠배팅 산업의 발전을 도모하기 위해 실제 경륜 경주의 배팅이 진행되는 것과 유사한 방식으로 머신러닝 알고리즘을 적용하여 승식별 예측을 시행하였다는 점에서 그 의의가 있다. 국내 경륜 경주로 살펴본 실증결과, 본 연구에 적용된 7가지 머신러닝 알고리즘 중에서 로지스틱 회귀가 가장 높은 성능을 보이는 것으로 나타났으며, 단승식 59%, 연승식 77.90%, 복승식 40.08%, 삼복승식 30.85%, 쌍승식 29.98%, 쌍복승식 20.80%, 삼쌍승식 13.73%의 예측 정확도를 보였다.

본 연구는 다음과 같은 이론적 시사점을 지닌다. 첫째, 국내 최초로 경륜 경주의 승

식별 결과를 예측했다. 국내 스포츠 서비스 산업의 큰 축으로서 스포츠베팅 종목의 성적 예측과 관련한 연구는 매우 희소할 뿐만 아니라 존재하는 소수의 연구조차 1위 예측을 중심으로 한 순위 예측에 한정된다(김필수 외, 2023; 최창환, 윤지운, 2017; 최혜민, 황나영, 황찬경, 송종우, 2015). 하지만 앞서 기술하였듯이, 스포츠베팅의 경우 여러 조합을 통해 순수하게 승패나 1위 확률의 순위를 맞추는 것 이상의 다양한 승자 예측 방법이 존재한다. 경륜, 경정, 경마와 같은 국내 사행성 스포츠는 단승/연승/복승/쌍승/삼복승/삼쌍승/쌍복승식의 7가지 승식을 지닌다. 각 승식을 예측하기 위해서는 2위, 3위를 예상하는 능력과 함께 이를 조합하는 능력이 필요하며 이를 위해서 1위 가능성을 크기순으로 단순히 나열하는 것과는 다른 방식으로 접근하는 노력이 필요하다. 본 연구에서는 각 선수의 전략이나 상성 등을 고려하여 순위별 가능성을 별도로 모델링하고 디자인하였으며, 이를 바탕으로 승식별 예측을 진행하였다. 본 연구는 이러한 방식을 사용하여 승식을 분석한 국내 최초 연구라는 점에서 향후 이를 토대로 새로운 방식의 분석을 확장할 수 있는 실마리를 제공한다는 점에서 학문적 기여가 크다고 볼 수 있다.

둘째, 본 연구는 각 승식을 예측하기 위해 로지스틱 회귀, 랜덤 포레스트, 에이다 부스트, 그래디언트 부스트, Light-GBM, 다층신경망, XG Boost의 다양한 머신러닝 알고리즘을 비교하고 분석하였다. 머신러닝 알고리즘은 각각의 가정이나 계산 방법의 차이 때문에 스포츠의 특성에 따라 적절한 알고리즘이 모두 다를 수 있다. 따라서, 최근 머신러닝 알고리즘을 이용해서 스포츠 경기결과를 예측하는 연구들은 여러 머신러닝 알고리즘을 활용하여 이를 각각 비교하기도 한다(김필수 외, 2023, 김필수, 이상현, 2023; 예원진, 이성노, 2022). 본 연구에서는 이러한 선행연구를 고찰하여 아직 연구가 많이 이루어지지 않은 경륜 데이터에 7가지 알고리즘을 적용하였으며 그 성능을 각각 테스트하고 비교했다. 경륜 경주의 분석을 위해 현시점에서 가장 적절한 알고리즘을 탐색하였다는 점에서 향후 해당 분야의 연구 발전에 이바지하였다는 장점을 지닌다.

아울러, 본 연구는 세부적으로 다음과 같은 실무적 시사점을 지닌다. 첫째, 서비스 산업으로서의 스포츠베팅 영역이 제시할 수 있는 다양한 마케팅이나 상품개발의 기초 자료를 제공한다고 볼 수 있다. 스포츠베팅 산업은 공익산업과 스포츠 재원 마련에 중요한 수입원이지만, 도박이 가지는 불편한 이미지와 새로운 소비자가 이해하기 힘든 규칙 및 자료로 인해 이용자 수와 매출액이 급격하게 감소하는 추세에 있다. 따라서, 이를 반전시키는 방안의 강구가 필요하며 본 연구의 결과와 같이 베팅 참여자가 이해할 수 있는 방식으로 경기 예측 정보를 제공하는 것은 하나의 방안이 될 수 있을 것으로 판단된다. 스포츠베팅과 관련한 여러 국가기관에서는 이를 인지하고 스포츠베팅 예측 서비스를 기반으로 한 마케팅을 시행하여 대중의 접근성을 높이려고 노력 중

이지만, 이를 수행할 수 있는 시스템의 부재와 기술의 미비로 아직 체계적인 서비스를 시행하지 못하고 있다. 선수별 우승 가능성뿐만 아니라 승식별 적중 가능성을 신뢰할만한 정보를 활용해서 제공하는 것은 베팅 참여자들의 흥미를 불러일으킬 수 있다. 또한, 향후 이를 활용하여 소비자들이 자체적으로 분석할 수 있는 자료를 제공한다면 이 또한 스포츠베팅이 도박보다는 분석이 가능한 하나의 게임이 될 수 있다는 측면에서 긍정적인 마케팅 이미지를 확보할 수 있을 것이다. 스포츠베팅 사업의 활성화는 새로운 산업을 구축하고 경제 활성화에 도움을 줄 수 있다는 면에서 경제 사회적 측면에서 기여도가 높을 것으로 판단된다(김필수 외, 2023).

둘째, 각종 스포츠협회나 의사결정자 및 운영자들은 머신러닝을 활용한 분석의 고도화를 통해 승부조작, 경기 방해 등의 이상징후를 찾아낼 수 있을 것이다. 현재까지 유지되고 있는 스포츠베팅의 위험성 중의 하나는 승부조작의 우려이다. 국내 프로스포츠에서 발견된 승부조작의 상당 부분은 불법 온라인 베팅업체와 관련되어 있다고 알려져 있다. 이런 이유로 대중들은 불법도박과 스포츠베팅, 승부조작을 연관 지어 생각하기도 한다(이상현, 김필수, 2023). 하지만, 이지용 외(2020)의 견해를 살펴보면 머신러닝을 활용한 경기분석은 오히려 이상징후의 발견을 통해 승부조작과 같은 비정상적인 경기상황을 발견해내기에 좋은 도구가 될 수 있다. 이는 실제 Stats Perform이나 Opta Sports와 같은 해외 스포츠 데이터 분석 업체들의 비즈니스모델 중의 하나로 작동하고 있음을 주지할 필요가 있다. 특별히 스포츠베팅을 중심으로 이루어진 스포츠의 경우 이를 발견하는 것은 더욱 중요한 요인이 될 수 있다. 본 연구는 향후 국내에서도 본 연구와 비슷한 형태의 분석을 통해 스포츠 데이터 분석을 통한 이상징후의 발견을 현실화할 가능성을 보여준다. 농구, 야구, 축구 등과 같이 아직 스포츠베팅이 활성화되지는 않았지만 많은 관심을 받는 주요 프로스포츠 종목에서도 머신러닝을 적용한 스포츠 애널리틱스의 적용을 통해 이상징후를 발견하고 경각심을 불러일으키는 방안으로 활용 가능할 것이다.

본 연구는 이상과 같은 이론·실무적 기여를 통해 머신러닝을 활용한 서비스 산업의 중요한 스포츠베팅 연구의 발전에 도움이 되고자 노력하였다. 향후 연구에서는 본 연구에서 고려하지 못한 다음과 같은 사항을 포함하여 분석함으로써 해당 분야 연구의 발전에 더 크게 이바지할 수 있을 것으로 보인다. 첫째, 본 연구는 스포츠베팅 산업의 대표적인 종목인 경륜 경주의 경기결과를 분석함으로써 스포츠베팅 산업에서의 활용 가능성을 모색했다. 비록 비슷한 승식을 지니고 있다고 하지만, 본 연구에서 다루지 않은 경정이나 경마의 경우에도 유사한 결과가 나타나는지를 분석함으로써 본 연구에서 사용한 방법론의 일반화 가능성을 확인해 볼 수 있다. 또한, 사행성 스포츠로 명명되지는 않지만, 스포츠평타에서 이루어지는 베팅의 종류로서 프로스포츠 등은 본 연구가 이루어진 경륜과는 다른 형태의 데이터와 맥락을 지닌다. 스포츠평타와 같은 베팅

팅산업에서 다루고 있는 형태를 적용하여 분석하는 것은 스포츠베팅 산업의 현실을 반영한 연구의 발전에 도움이 될 것으로 보인다.

둘째, 본 연구는 머신러닝을 경기결과에 적용하여 실제 베팅이 이루어지는 승식별 적중 가능성을 분석함으로써 스포츠베팅 산업에서 활용할 수 있는 기초자료를 제공하여 해당 산업의 발전에 도움이 되고자 노력했다. 스포츠베팅 산업에 대한 국내 연구는 아직 걸음마 단계이며, 이를 위해서는 경기 자체뿐만 아니라 경기결과를 산업에 활용하기 위한 다양한 연구와 노력이 필요하다. 경기결과를 스포츠베팅 산업에 실제로 어떻게 활용하며 이것이 스포츠베팅 산업에 미치는 효과성에 대해 더욱 면밀하게 검증하는 연구가 이루어질 필요가 있다. 이러한 종합적인 노력이 실질적인 경기결과 예측을 위한 실증연구와 융합적 가치를 보편화할 때 서비스 산업으로서의 스포츠베팅이 더 발전하고 활성화될 수 있을 것으로 기대한다.

## 참고문헌

- 강석봉(2023). '경륜' 경주사업 최초, AI 활용 경륜 승자예측 서비스 시범운영, 스포츠경향, 2023.5.28.
- 권웅, 이현우 (2016). 스포츠관여 수준에 따른 P300 뇌파반응의 차이: 체화된 인지 이론에 따른 소비자 행동과 EEG/ERP 측정법의 융합연구. *한국스포츠산업경영학회지*, 21(6), 45-56.
- 김경식, 한승백 (2018). 최근 10년간 보도된 스포츠-승부조작 신문기사 의미연결망 분석. *한국스포츠사회학회지*, 31(3), 21-41.
- 김도우, 박경래, 이창무 (2012). 불법 사행산업의 실태 및 대응전략에 관한 연구. *한국공안행정학회보*, 21(1), 9-42.
- 김문영 (2012). '사행산업통합감독위원회'의 명칭과 역할이 '불법사행행위감독위원회'로 바뀌어야 하는 이유. *동서언론*, 15, 23-66.
- 김병률, 이진축 (2021). 골퍼의 관여, 경험, 인지가치가 골프만족경험, 골프생활품질과 지속의도에 미치는 영향. *한국의사결정학연구*, 1-16.
- 김주연, 최현주, 안경모, 2016). 온라인베팅 확대허용 및 불법 온라인베팅 축소방안에 관한 인식 차이: 이용자, 중독자 및 일반인의 인식비교, *한국콘텐츠학회논문지*, 17(1), 426-435.
- 김진세, 김대희(2020). 경륜·경정 온라인 발매 도입을 위한 법령개정안 연구, *스포츠엔터테인먼트와법*, 23(1), 85-107.
- 김태훈, 임성원, 고진광, 이재학 (2020). 인공지능 모델에 따른 한국 프로야구의 승패 예측 분석에 관한 연구. *한국빅데이터학회지*, 5(2), 77-84.
- 김필수, 이상현(2023). 빅데이터 분석을 적용한 한국프로농구 리그 정규시즌 경기결과와 머신러닝 분류모형 예측성능 비교에 관한 연구. *한국체육학회지*, 62(2), 263-277.
- 김필수, 이상현, 전성삼 (2023), 머신러닝을 적용한 경륜 경기 순위 예측 및 평가에 관한 연구: 2016~2022년 출주표 정보 및 경주 결과 활용, *한국스포츠산업경영학회지*, 28(2), 1-19.
- 김헌일 (2022). 사행성 스포츠 베팅 산업의 현안과 정책 대안. *한국사회체육학회지*, 90, 73-86.
- 김혁 (2015). 불법스포츠도토 실태와 개선방안탐색. *한국체육과학회지*, 24(5), 81-98.
- 노주원 (2021). *국내스포츠베팅게임 도입의 비즈니스모델 연구*, 미간행 석사학위논문

- 문, 한양대학교 대학원.
- 박준휘, 최성락, 오시영 (2014). 스포츠 토트 불법 행위에 대한 처벌 강화 정책의 효과 분석. *한국사회와 행정연구*, 25(3), 35-57.
- 사행산업통합감독위원회(2022). *2021 사행산업 관련 통계*. 경기도 과천.
- 예원진, 이성노 (2022). 2022 FIBA 남자농구 아시안컵 경기결과를 활용한 머신러닝 분류 모형의 예측 성능 비교. *한국체육측정평가학회지*, 24(3), 53-69.
- 오상훈 (2009). 다층퍼셉트론에 의한 불균형 데이터의 학습 방법. *한국콘텐츠학회논문지*, 9(7), 141-148.
- 이상현, 김필수 (2023). 비윤리적 사건에 대응하는 프로스포츠팀의 전략: 사안의 속성에 따른 커뮤니케이션 패턴. *체육과학연구*, 34(1), 131-143.
- 이원재, 황다빈, 신성연 (2020). 스포츠서비스산업 활성화를 위한 스포츠산업 분류체계 개선 방안에 관한 연구. *한국스포츠산업경영학회지*, 25(5), 46-58.
- 이장영, 조옥연, 김의진 (2014). 스포츠 베팅산업의 긍정적 효과와 부정적 효과. *사회과학연구*, 27(1), 181-207.
- 이지용, 박재현, 윤지운, 윤효준 (2020). Monte-Carlo 시뮬레이션을 적용한 이상 배팅 징후 탐지: 남자프로농구승률과 배당률의 관계. *한국체육측정평가학회지*, 22(4), 55-65.
- 장승원 (2010). 경마베팅 참가자의 도박동기와 비합리적 도박신념 및 도박중독의 관계. *한국체육과학회지*, 19(2), 83-93.
- 정해상 (2011). 스포츠의 온라인베팅규제정책. *스포츠엔터테인먼트와 법*, 14(4), 199-219.
- 최창환, 윤지운(2017). 경륜 출주정보를 활용한 승자 예측모형 탐색: 데이터마이닝 기반 의사결정나무분석의 적용. *한국체육측정평가학회지*, 19(4), 15-26.
- 최혜민, 황나영, 황찬경, 송종우 (2015). 서울 경마 경기 우승마 예측 모형 연구. *응용통계연구*, 28(6), 1133-1146.
- Agrawal, P., Girshick, R., & Malik, J. (2014). Analyzing the performance of multilayer neural networks for object recognition. In *Computer Vision-ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII* 13 (pp. 329-344). Springer International Publishing.
- Al Daoud, E. (2019). Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset. *International Journal of Computer and Information Engineering*, 13(1), 6-10.

- Alonso, R. P., & Babac, M. B. (2022). Machine learning approach to predicting a basketball game outcome. *International Journal of Data Science*, 7(1), 60–77.
- Arik, S. Ö., & Pfister, T. (2021, May). Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, No. 8, pp. 6679–6687).
- Baboota, R., & Kaur, H. (2019). Predictive analysis and modelling football results using machine learning approach for English Premier League. *International Journal of Forecasting*, 35(2), 741–755.
- Borowski, P., & Chlebus, M. (2021). *Machine learning in the prediction of flat horse racing results in Poland*. University of Warsaw, Faculty of Economic Sciences.
- Bradley, B., & Brandon, G. (2020). *Hands-On Machine Learning with R*. CRS Press.
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24, 123–140.
- Boulesteix, A. L., Janitza, S., Kruppa, J., & König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(6), 493–507.
- Bunker, R., & Susnjak, T. (2022). The application of machine learning techniques for predicting match results in team sport: A review. *Journal of Artificial Intelligence Research*, 73, 1285–1322.
- Buzsáki, G., Peyrache, A., & Kubie, J. (2014). Emergence of cognition from action. In *Cold Spring Harbor symposia on quantitative biology* (Vol. 79, pp. 41–50). Cold Spring Harbor Laboratory Press.
- Chung, W. C., Chang, C. Y., & Ko, C. C. (2017, August). A svm-based committee machine for prediction of hong kong horse racing. In *2017 10th International Conference on Ubi-media Computing and Workshops (Ubi-Media)* (pp. 1–4). IEEE.
- Constantinou, A. C. (2019). Dolores: a model that predicts football match outcomes from all over the world. *Machine Learning*, 108(1), 49–75.
- Grand View Research (2022). *Sports etting market size, share & trends analysis report by platform, by betting type (fixed odds wagering, exchange betting, live/in-play betting, esports betting), by sprots type, by region, and*

- segment forecasts, 2023–2030*, Grand View Research.
- Haghighi, S., Jasemi, M., Hessabi, S., & Zolanvari, A. (2018). PyCM: Multiclass confusion matrix library in Python. *Journal of Open Source Software*, 3(25), 729.
- Horvat, T., Havaš, L., & Srpak, D. (2020). The impact of selecting a validation method in machine learning on predicting basketball game outcomes. *Symmetry*, 12(3), 431.
- Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.
- Huang, M. L., & Li, Y. Z. (2021). Use of machine learning and deep learning to predict the outcomes of major league baseball matches. *Applied Sciences*, 11(10), 4499.
- Hubáček, O., Šourek, G., & Železný, F. (2019). Exploiting sports–betting market using machine learning, *International Journal of Forecasting*, 35(2), 783–796.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.
- Khan, S., & Kirubanand, V. B. (2019). Comparing machine learning and ensemble learning in the field of football. *International Journal of Electrical and Computer Engineering*, 9(5), 4321–4325.
- Killick, E. A., & Griffiths, M. D. (2022). Sports betting advertising A systematic review of content analysis studies. *International Journal of Mental Health and Addiction*, 1–27.
- Knoll, J., & Stübinger, J. (2020). Machine–learning–based statistical arbitrage football betting. *KI–Künstliche Intelligenz*, 34(1), 69–80.
- Kollár, A. (2021). *Betting models using AI: A review on ANN, SVM, and Markov Chain*.
- Liang, J. (2022). Confusion Matrix: Machine Learning. *POGIL Activity Clearinghouse*, 3(4).
- Lopez–Gonzalez, H., & Griffiths, M. D. (2018). Understanding the convergence of markets in online sports betting. *International Review for the Sociology of Sport*, 53(7), 807–823.
- Lopez–Gonzalez, H., Guerrero–Solé, F., Estévez, A., & Griffiths, M. (2018).

- Betting is loving and bettors are predators: A conceptual metaphor approach to online sports betting advertising. *Journal of Gambling Studies*, 34, 709–726.
- Migliorati, M. (2020). Detecting drivers of basketball successful games: an exploratory study with machine learning algorithms. *Electronic Journal of Applied Statistical Analysis*, 13(2), 454–473.
- Milano, M., & Chelladurai, P. (2011). Gross domestic sport product: The size of the sport industry in the United States. *Journal of Sport Management*, 25(1), 24–35.
- Min, M., Lee, J. J., Park, H., & Lee, K. (2021). Detecting anomalous transactions via an iot based application: A machine learning approach for horse racing betting. *Sensors*, 21(6), 2039.
- Osborne, J. W. (2014). *Best practices in logistic regression*. Sage Publications.
- Ramraj, S., Uzir, N., Sunil, R., & Banerjee, S. (2016). Experimenting XG Boost algorithm for prediction and classification of different datasets. *International Journal of Control Theory and Applications*, 9(40), 651–662.
- PricewaterhouseCoopers (2021). *Sports outlook*, PricewaterhouseCoopers.
- Rojas, R. (2009). *AdaBoost and super bowl of classifiers a tutorial introduction to adaptive boosting*. Freie University, Berlin, Tech. Rep.
- Stübinger, J., Mangold, B., & Knoll, J. (2019). Machine learning in football betting: Prediction of match results based on player characteristics. *Applied Sciences*, 10(1), 46.
- Sullivan, K., Myers, D., Harman, A., & Srivastava, D. (2023). *Sports Betting Optimization with Machine Learning Algorithms*.
- Thabtah, F., Zhang, L., & Abdelhamid, N. (2019). NBA game result prediction using feature analysis and machine learning. *Annals of Data Science*, 6(1), 103–116.
- Vantage Market Research (2022). *Sports betting market size & share to surpass USD 12.9 Billion by 2028*, Vantage Market Research.
- Vlastakis, N., Dotsis, G., & Markellos, R. N. (2009). How efficient is the European football betting market\_ Evidence from arbitrage and trading strategies. *Journal of Forecasting*, 28(5), 426–444.
- Wang, Y., & Zeng, D. (2020). Development of sports industry under the influence of COVID–19 epidemic situation based on big data. *Journal of Intelligent &*

- Fuzzy Systems*, 39(6), 8867–8875.
- Wilkins, S. (2021). Sports prediction and betting models in the machine learning age: The case of tennis. *Journal of Sports Analytics*, 7(2), 99–117.
- Xie, J., Xu, J., Nie, C., & Nie, Q. (2016). Machine learning of swimming data via wisdom of crowd and regression analysis. *Mathematical Biosciences & Engineering*, 14(2), 511–527.
- Yang, K. (2020). The construction of sports culture industry growth forecast model based on big data. *Personal and Ubiquitous Computing*, 24(1), 5–17.
- Yaseen, A. S., Marhoon, A. F., & Saleem, S. A. (2022). Multimodal Machine Learning for Major League Baseball Playoff Prediction. *Informatica*, 46(6), 1–9.
- Zhang, J. J., Kim, E., Mastromartino, B., Qian, T. Y., & Nauright, J. (2018). The sport industry in growing economies: critical issues and challenges. *International Journal of Sports Marketing and Sponsorship*, 19(2), 110–126.
- XGBoost, XGBoost Tutorials, <https://xgboost.readthedocs.io/en/latest/tutorials/model.html>

## \* 저자소개 \*

### · 김 필 수(philsookim@naver.com)

한국스포츠경영전략연구원 원장으로 재직하고 있다. 주요 관심 분야는 스포츠 애널리틱스, 스포츠산업, 스포츠경영전략, 인공지능, 머신러닝, 빅데이터 분석이다.

### · 전 성 삼(jss709@naver.com)

에이치앤스포르츠에서 연구원으로 재직중이다. 주요 관심분야로는 인공지능, 머신러닝을 활용한 스포츠데이터 분석 등이 있다.

### · 이 상 현(ich45337@hanmail.net)

아주대학교 경영학과 인사조직 박사수료 상태이며, 스포츠 데이터 분석 전문가로서 활동 중이다. 주요 연구 분야로는 고성능작업관리시스템, 리더십, 스포츠팀 감독 역량, 스포츠클럽의 커뮤니케이션 전략 등이 있다.