

컴퓨터공학 종합설계

상세 요구사항 명세서



컴퓨터공학 종합설계(202501-CSE4205-001)

890 팀
12192077 박륜
12180557 신성혁
12201798 정우성

목차

1. 소개

1-1. 개발 배경

1-2. 개발 목표

1-3. 기대 효과

2. 개요

2-1. 서비스 기능

2-2. 서비스 특징

3. 상세 요구사항

3-1. AI 모델 요구사항

3-2. 서비스 요구사항

1. 소개

1-1. 개발 배경

- 현대 스포츠에서 통계 기록(이하 스탯)을 데이터 사이언스 기반의 시스템으로 분석 및 활용하는 사례(야구 - 세이버메트릭스 등)가 활발히 이루어지고 있다. 이와 달리 데이터 기반의 예측이나 분석이 상대적으로 부진한 축구의 데이터 기반 분석 시스템을 구축하고, 경기와 관련된 다양한 정보를 제공하여 사용자의 흥미 및 정보 가치를 높이고자 서비스를 기획하였다.

1-2. 개발 목표

- 본 서비스의 목표는 각종 축구 통계 사이트에서 제공하는 통계 자료와 과거 경기 기록을 기반으로, 독자적인 승부 예측 모델을 개발하고 이를 서비스 형태로 배포하여, 사용자의 정보 탐색 및 소비 효율성을 높이고자 한다.
- 통계 외에도 유력 언론사의 기사, ITK 소스 기반 SNS 포스팅 등을 요약한 Match Fact를 시각화하여 예측 결과와 함께 제공하는 확장형 서비스를 구현하고자 한다.

1-3. 기대 효과

- **데이터 기반 승부 예측 서비스의 차별화**
기존의 단순 스코어 예측이나 배당률 중심 정보 제공을 넘어, 실제 축구 통계 데이터를 기반으로 한 예측 모델과 Match Fact 시각화를 결합함으로써, 사용자는 더욱 신뢰도 높은 정보를 직관적으로 받아볼 수 있다.
- **스포츠 데이터 활용 가능성 확대**
본 프로젝트는 단순 통계 분석을 넘어서, 기사·SNS 기반의 비정형 정보까지 통합적으로 분석·제공함으로써, 향후 다양한 스포츠 종목으로의 확장 가능성을 열 수 있다.

2. 개요

2-1. 서비스 기능

- **경기 결과 승부 예측**

팀 간의 과거 전적, 시즌 통계(xG, 점유율, 슈팅 수 등)를 바탕으로 머신러닝 모델이 경기의 결과(승/무/패)를 예측하여, 확률(예: 승 65%, 무 20%, 패 15%)의 형태로 제공되어 사용자가 직관적으로 해석할 수 있도록 구성된다.

- **경기 정보 및 통계 제공**

예측 대상 경기의 기본 정보(일정, 홈/어웨이, 팀 순위 등)와 함께, 각 팀의 최근 경기력, 주요 선수 스탯, 전술적 특징 등을 제공한다.

- **Match Fact 자동 요약**

오픈소스 LLM을 활용하여 해당 경기와 관련된 뉴스 기사, 팀/선수 관련 SNS, ITK 소스 등의 비정형 정보를 수집 및 요약한다.

요약된 정보는 간결한 문장과 함께 시각화 요소(예: 태그, 요약 카드)로 제공되어, 예측 결과의 보조 정보로 활용된다.

2-2. 서비스 특징

- 단순 수치 나열이 아닌, 그래프, 카드, 태그 등 시각적 요소를 중심으로 구성하여 사용자 이해도를 높이는 정보를 제공한다.

- 팀통계에 더하여 선수 개인 통계 자료를 데이터로 활용하여 학습시켜 더 높은 예측 정확도를 제공한다.

- 다양한 소스에서 정보를 수집할 필요 없이, 한 곳에서 통계적 예측과 경기 주요 정보 요약, 시각화 자료를 함께 제공받아 높은 밀도 정보를 제공한다.

3. 상세 요구사항

3-1. AI 모델 요구사항

3-1-1. 모델 목적

- 본 AI 모델은 축구 경기의 승/무/패를 예측하는 다중 클래스 분류 모델로, 최근 15시즌간의 리그/팀/선수 데이터를 학습하여 통계 기반의 예측 결과를 도출하는 데 목적이 있다. 해당 모델은 웹 서비스와 연동되어 사용자에게 정확하고 신뢰도 높은 예측 정보를 제공하는 핵심 엔진으로 사용된다.

3-1-2. 모델 정의

항목	내용
모델 유형	다중 클래스 분류 모델 (Multi-class Classification)
클래스 구성	0: 패배, 1: 승리, 2: 무승부
예측 대상	PL 경기 승/무/패 결과

3-1-3. 알고리즘

모델	특징	장단점
XGBoost	<ul style="list-style-type: none">• Gradient Boosting 기반 앙상블 모델• 결측치 자동 처리 기능 내장• 정형 데이터에 최적화	장점: 안정적 성능, 결측치 처리, 성숙한 생태계 단점: 학습 속도 느릴 수 있음, 메모리 많이 사용
LightGBM	<ul style="list-style-type: none">• Microsoft 개발의 Gradient Boosting 프레임워크• Leaf-wise 트리 생성 방식으로 빠른 학습 속도• 대규모 데이터 처리에 효율적	장점: 빠른 학습, 대규모 데이터에 적합 단점: 과적합 가능성 있음, 복잡한 구조로 디버깅 어려움

※ 추후 학습 정확도 비교 후 선택 예정

3-1-4. 입력 데이터 명세

3-1-4 (a). 데이터 출처

- footystat.org

전 세계 345개 이상의 축구 리그와 컵 대회 통계 데이터를 제공하는 웹사이트로 경기 결과, 오버/언더, 코너킥, 전/후반 득점, 클린시트, 득점왕 등 다양한 통계를 상세하게 제공해준다.

3-1-4 (b). 입력 데이터

- 학습 데이터: EPL 2010/11 ~ 2020/21 시즌

- 검증 데이터: EPL 2021/22 ~ 2024/25 시즌

3-1-4 (c). 입력 데이터 항목 (상세)

1) 리그 전체 통계

항목	설명	예시	역할
평균 득점	경기당 평균 골 수	2.7	리그의 공격 성향 반영
홈 승률	홈팀 승리 비율	55%	홈 어드밴티지 반영
무승부 비율	PL 경기 승/무/패 결과	23%	무승부 보정
평균 파울 수	경기당 평균 파울 수	22.3	경기 흐름, 거칠기 반영
득점 편차	상위-하위 팀 평균 득점 차	1.8	리그 경쟁 균형성 판단

2) 경기별 상대 통계

항목	설명	예시	역할
최근 5경기 전적	맞대결 결과 이력	["D", "W", "L", "W", "D"]	상성 반영
상대전 점유율	팀 간 평균 점유율	58%	경기 지배력 비교
상대전 슈팅 수	평균 슈팅 수	12.5	공격 우위 반영
상대전 실점	평균 실점 수	1.7	수비력 반영
홈/원정 성적	홈팀/원정팀 최근 상대 전적	홈: 3승 1무 1패	홈 어드밴티지 보정

3) 팀 통계

항목	설명	예시	역할
최근 5경기 성적	현재 폼	["W", "W", "D", "L", "W"]	현재 경기력 반영
시즌 평균 득점	팀 공격력 지표	58%	공격력 판단
시즌 평균 실점	팀 수비력 지표	12.5	수비력 판단
슈팅 수 / 유효 슈팅률	공격 효율성	1.7	결정력 판단
점유율	경기 지배력	61%	주도권 반영
카드 수	거친 정도	2.3	경기 태도 반영
부상자 수	결장자 수	3	전력 누수 판단

4) 선수 개인 통계

항목	설명	예시	역할
출전 시간	최근 5경기 기준	420분	핵심 선수 여부 판단
득점 / 어시스트	시즌 누적	8 / 5	공격 기여도 판단
슈팅 / 유효 슈팅	결정력 지표	18 / 9	결정력 판단
패스 성공률	볼 연결 능력	87%	빌드업 관여도
인터셉트 / 태클	수비 기여도	3.1 / 2.4	수비적 기여도
최근 경기 평점	평균 평점 or 활약지수	7.4	컨디션 반영

※ 선수 통계는 팀 평균 또는 주요 선수 3~5인의 평균값으로 요약하여 입력

3-1-5. 출력 및 평가 지표

3-1-5 (a). 출력

항목	설명
클래스 예측	'W', 'D', 'L' 중 하나
클래스 확률	Softmax 결과 ([0.65, 0.20, 0.15])
Top-k 결과	확률 기준 상위 k개 클래스

3-1-5 (b). 목표 평가 지표

지표	목표 값	근거
Accuracy	$\geq 70\%$	- EPL의 승/무/패는 대체로 4050% / 2030% / 20~30% 수준으로 분포하며, 단순 무작위 예측 시 정확도는 약 33% 수준이다. 따라서 Accuracy 70%는 랜덤의 2배 이상에 해당하는 성능으로, 통계적으로도 충분히 유의미한 성과로 볼 수 있다. - 기존의 통계 기반 또는 배당률 기반 예측은 약 55~60% 수준의 정확도를 보이며, 머신러닝 모델이 이 보다 높은 성능인 70%를 목표로 하는 것은 합리적이다.
F1-score (macro)	≥ 0.60	- 0.60 이상이면 모든 클래스에서 균형 잡힌 성능을 내고 있다는 의미가 된다.
F1-score (weighted)	≥ 0.65	- Accuracy가 70%에 도달하면, weighted F1은 보통 0.65~0.7 정도로 따라 올라간다.
Log Loss	≤ 0.90	- 다중 클래스 분류에서 baseline 수준은 1.0 이상이고, 1.0 이하로 떨어뜨리면 의미 있는 신뢰도 개선이 이루어졌다고 본다. - 0.90 이하는 통계/머신러닝 기반 예측에서 실용적으로 관찰은 cutoff로 자주 쓰인다.
Top-2 Accuracy	$\geq 85\%$	- 축구처럼 이변이 많고 무승부 확률이 높은 게임에서는 정답을 딱 1순위로 못 맞춰도, Top 2 안에 포함시키는 게 더 현실적일 수 있다. - Accuracy가 70%면, 보통 Top-2 Accuracy는 그보다 10~15% 더 높게 나온다. 따라서 85%는 자연스럽고 달성 가능한 기준선이다.
Confusion Matrix	균형 잡힌 분포	- 특정 클래스(예: 무승부)에 너무 치우치지 않도록 점검하는 데 사용

3-1-6. 하이퍼 파라미터 튜닝

3-1-6 (a). XGBoost

파라미터	범위	근거
learning_rate	0.01 ~ 0.3	-학습률 -낮을수록 학습이 천천히 되 지만 성능이 안정적임 -일반적인 권장 범위
n_estimators	100 ~ 1000	-트리 개수 -많을수록 성능 향상 가능하 지만 과적합 위험 증가
max_depth	3 ~ 10	-트리 깊이 -복잡도 조절 -너무 깊으면 과적합 가능
subsample	0.5 ~ 1.0	-샘플링 비율 -과적합 방지 및 모델 일반 화에 도움
colsample_bytree	0.5 ~ 1.0	-트리마다 사용할 feature 비 율 -과적합 줄이는 데 도움
min_child_weight	1 ~ 10	-리프 노드가 가져야 할 최 소 가중치 합 -높을수록 과적합 방지
gamma	0 ~ 5	-분할에 필요한 최소 손실 감소 값 -클수록 보수적인 분할
scale_pos_weight	클래스 불균형 조정시 사용	-불균형 데이터셋에서 긍정 클래스에 가중치를 줄 때 사 용

3-1-6 (b). LightGBM

파라미터	범위	근거
learning_rate	0.01 ~ 0.3	-학습률 낮을수록 정밀하지 만 느림 -일반적인 추천 범위
num_leaves	20 ~ 150	-트리의 복잡도 조절 -leaf-wise 성장 특성상 조심 해서 튜닝 필요
n_estimators	100 ~ 1000	-트리 개수 -너무 크면 과적합, 적으면 성능 저하
max_depth	-1 or 3 ~ 10	- -1은 무제한 -과적합 방지를 위해 제한하 는 것이 일반적
min_data_in_leaf	10 ~ 100	-리프 노드가 포함할 최소 샘플 수 -과적합 방지 역할
feature_fraction	0.5 ~ 1.0	-모델이 학습 시 사용하는 feature 비율 -과적합 줄임

bagging_fraction	0.5 ~ 1.0	-데이터 샘플링 비율 -모델 일반화에 도움
bagging_freq	0 ~ 1.0	-배깅 실행 주기 -0이면 사용 안 함.
lambda_l1/l2	0 ~ 1.0	-정규화 파라미터 -과적합 방지 목적.
objective	'multiclass'	-다중 클래스 분류 문제임을 명시
num_class	3	-클래스 수: 승, 무, 패.

3-1-6 (c). 튜닝 도구

구분	도구	설명
자동	Optuna	-하이퍼파라미터 최적화를 위한 자동 튜닝 라이브러리 -베이지안 최적화 기반이며, 빠르고 효율적으로 최적값 탐색 가능 -탐색공간이 넓거나 복잡할 때 유리함
수동	GridSearchCV	-가능한 모든 파라미터 조합을 하나하나 시도하는 방식 -가장 단순하지만 계산량이 많음 -소규모 탐색에 적합
수동	RandomizedSearchCV	-지정된 횟수만큼 랜덤 조합을 시도하는 방식 -GridSearch보다 빠르지만, 최적값을 놓칠 가능성도 있음 -넓은 범위의 빠른 탐색에 유리함

3-1-7. 모델 해석 기능

기능	설명
Feature Importance	예측에 영향을 준 feature 순위를 시각화 (bar chart)
SHAP 분석	개별 예측에 기여한 feature 기여도를 정량화
실패 사례 분석	예측 실패 경기 수집 및 오류 원인 분석

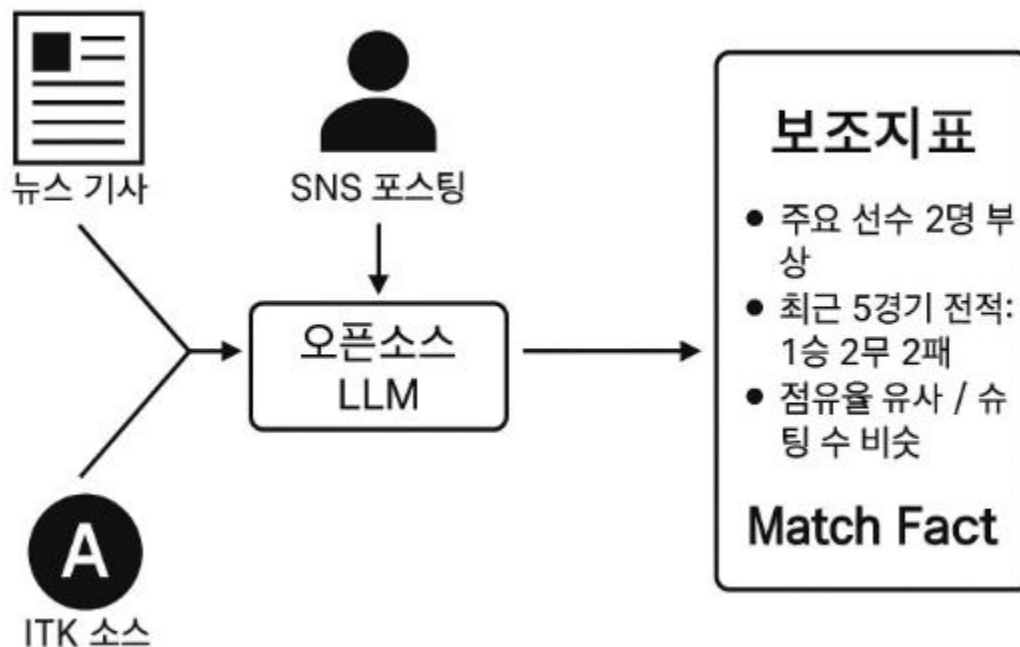
3-1-8. 유지보수 및 확장 고려사항

항목	설명
모델 재학습 주기	매주 정기 재학습
리그 확장	라리가, 세리에 A 등 타 리그로 확장 가능

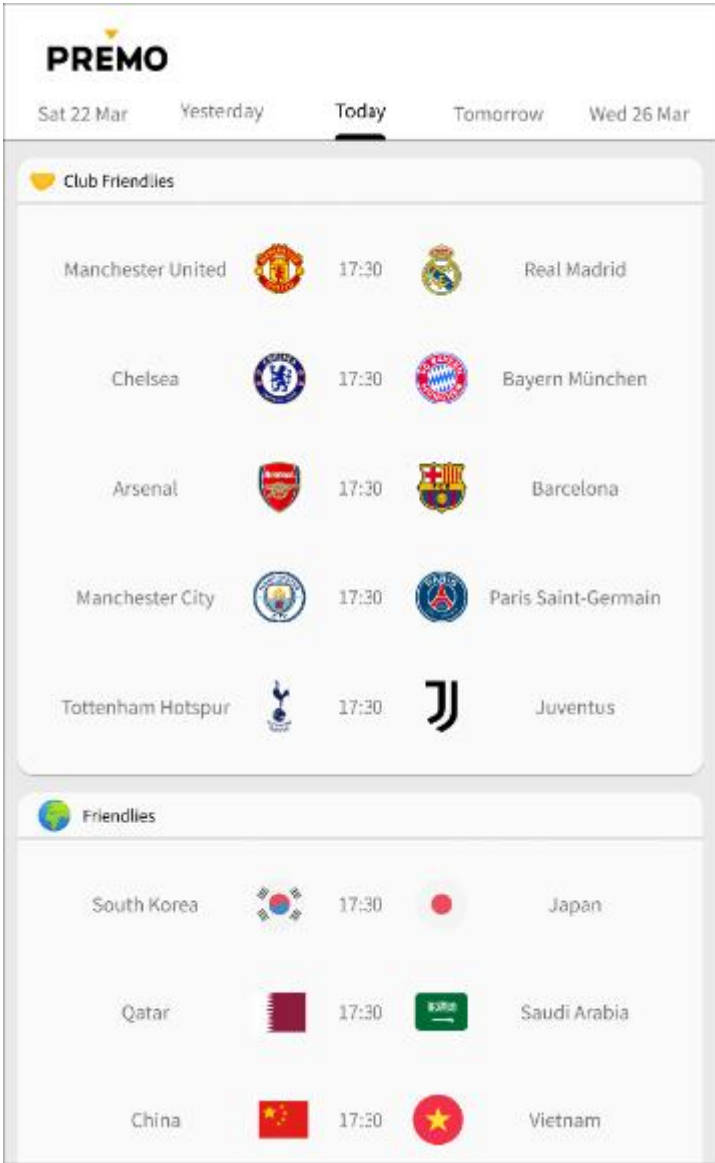
3-2. 서비스 요구사항

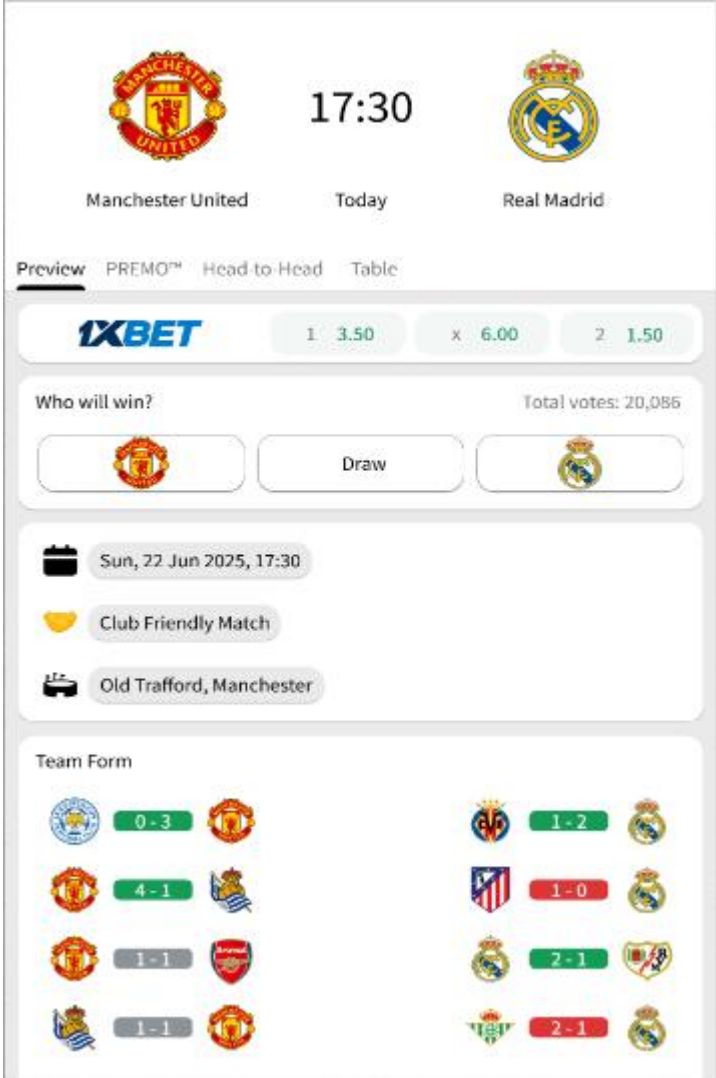
3-2-1. Match Fact

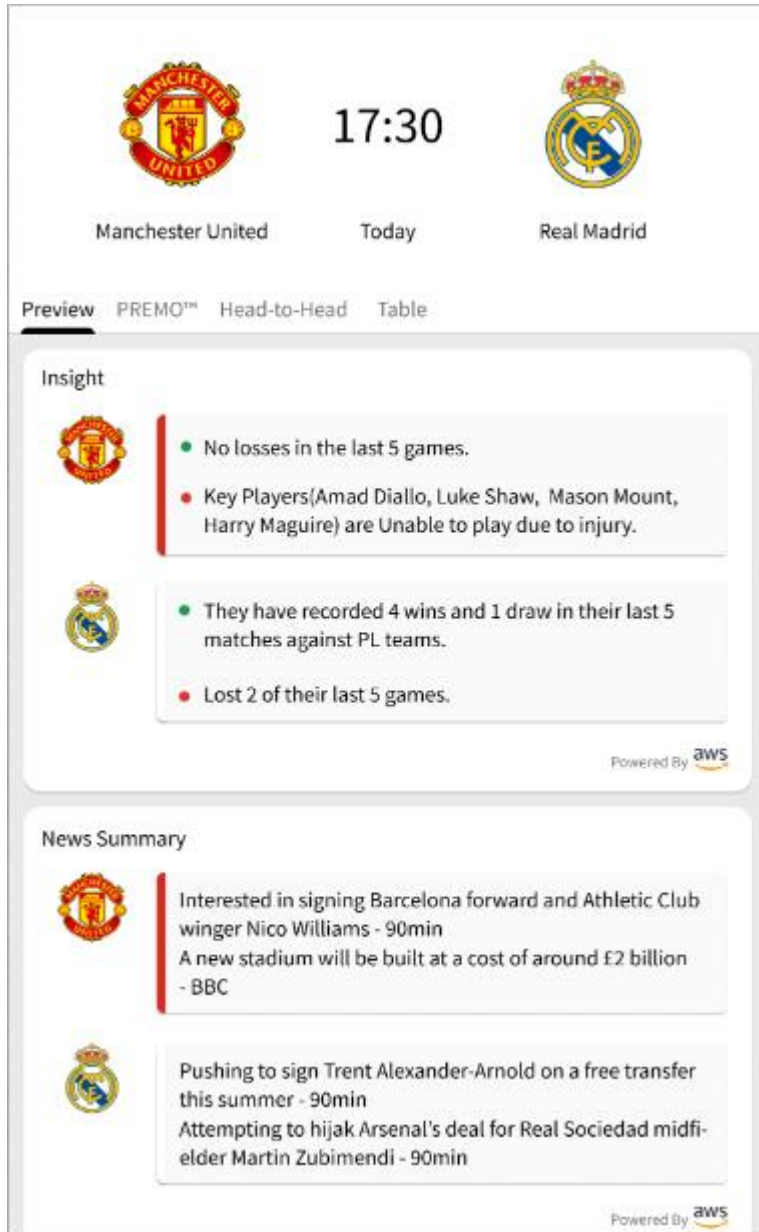
구성 단계	목적	기술 스택
① 데이터 수집	뉴스, SNS, ITK 소스 수집	BeautifulSoup, Selenium, Newspaper3k, SNS API (Twitter/X API)
② 텍스트 요약	핵심 정보 추출	KoBART, T5, GPT API, Huggingface Transformers, KeyBERT (키워드 추출)
③ 태그/카드화	시각적 요소로 가공	Jinja2, pandas, JSON schema, custom HTML template
④ 시각화	사용자에게 보기 좋게 보여 줌	React, Chart.js, Tagify, Card UI, Tailwind CSS, WordCloud
⑤ 연동 API	Match Fact 전달	FastAPI, Flask, JSON API
⑥ 자동화 / 스케줄링	주기적 수집/요약 수행	cron, Airflow, Python Script, Docker
⑦ 저장소	요약/태그 데이터 저장	MongoDB (비정형 JSON에 유리), PostgreSQL (정형 관리 시)



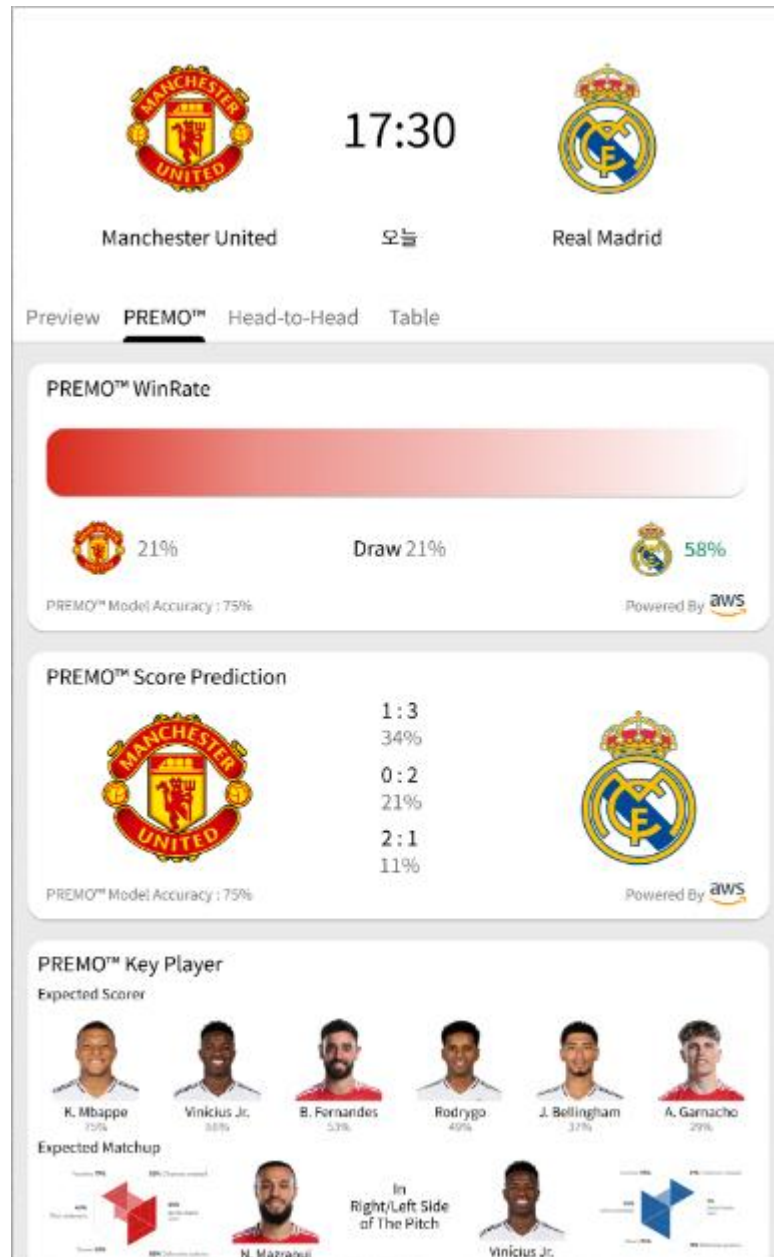
3-2-2. 서비스 화면

페이지 이름	Match Selection Page(Main Page)
페이지 설명	날짜 별, 리그 별 경기 확인 페이지
	
기능 상세	<ul style="list-style-type: none"> - 날짜별 경기 탐색 기능 제공. - 하단 스와이프를 통해 추가 경기 load. - 경기 리스트 카테고리별 섹션 구분.

페이지 이름	Match Info Page										
페이지 설명	해당 경기 팀, 날짜, 위치 정보 전달										
 <p>The screenshot displays a match preview for Manchester United (home) vs Real Madrid (away). The match is scheduled for Sunday, June 22, 2025, at 17:30 at Old Trafford, Manchester. The betting odds are 1: 3.50, x: 6.00, and 2: 1.50. A poll asks 'Who will win?' with 20,086 total votes. The 'Team Form' section shows recent results for both teams.</p> <table border="1"> <caption>Team Form</caption> <thead> <tr> <th>Manchester United</th> <th>Real Madrid</th> </tr> </thead> <tbody> <tr> <td>0-3</td> <td>1-2</td> </tr> <tr> <td>4-1</td> <td>1-0</td> </tr> <tr> <td>1-1</td> <td>2-1</td> </tr> <tr> <td>1-1</td> <td>2-1</td> </tr> </tbody> </table>		Manchester United	Real Madrid	0-3	1-2	4-1	1-0	1-1	2-1	1-1	2-1
Manchester United	Real Madrid										
0-3	1-2										
4-1	1-0										
1-1	2-1										
1-1	2-1										
기능 상세	<ul style="list-style-type: none"> - 기본 경기 정보 제공. (팀, 경기 시간, 경기 날짜). - 배당률 정보 제공. (제공 업체, 1/X/2 방식, 배당률 표시). - 유저 투표 기능 제공. (투표 후 결과 비율 시각화) - 경기 참여 팀 최근 경기 결과 정보 제공. (4~5 경기 결과 나열, 경기 클릭 시 경기 상세 정보 페이지로 이동) 										

페이지 이름	Match Preview Insight Page
페이지 설명	해당 경기 정보 요약 및 인사이트 제공 페이지
<div data-bbox="395 389 1157 1619">  <p>The image shows a match preview page for Manchester United vs Real Madrid. At the top, the team crests are displayed with the time 17:30 and the word 'Today'. Below this are tabs for 'Preview', 'PREMO™', 'Head-to-Head', and 'Table'. The 'Insight' section features two columns of information: Manchester United's recent performance (no losses in last 5 games, but key players injured) and Real Madrid's performance (4 wins and 1 draw in last 5 matches, but lost 2 of the last 5). The 'News Summary' section lists transfer rumors for both teams, such as Manchester United's interest in Nico Williams and Real Madrid's push for Trent Alexander-Arnold. Both sections are powered by AWS.</p> </div>	
기능 상세	<ul style="list-style-type: none"> - 기본 경기 정보 제공. (팀, 경기 시간, 경기 날짜). - Match Fact 카드 제공. (최근 경기 흐름, 리스크 정보, 전력 비교) - 양 팀 관련 최신 이슈 주요 뉴스 정보 요약 제공. - 뉴스 정보 클릭 시 원문 기사 하이퍼링크 연결

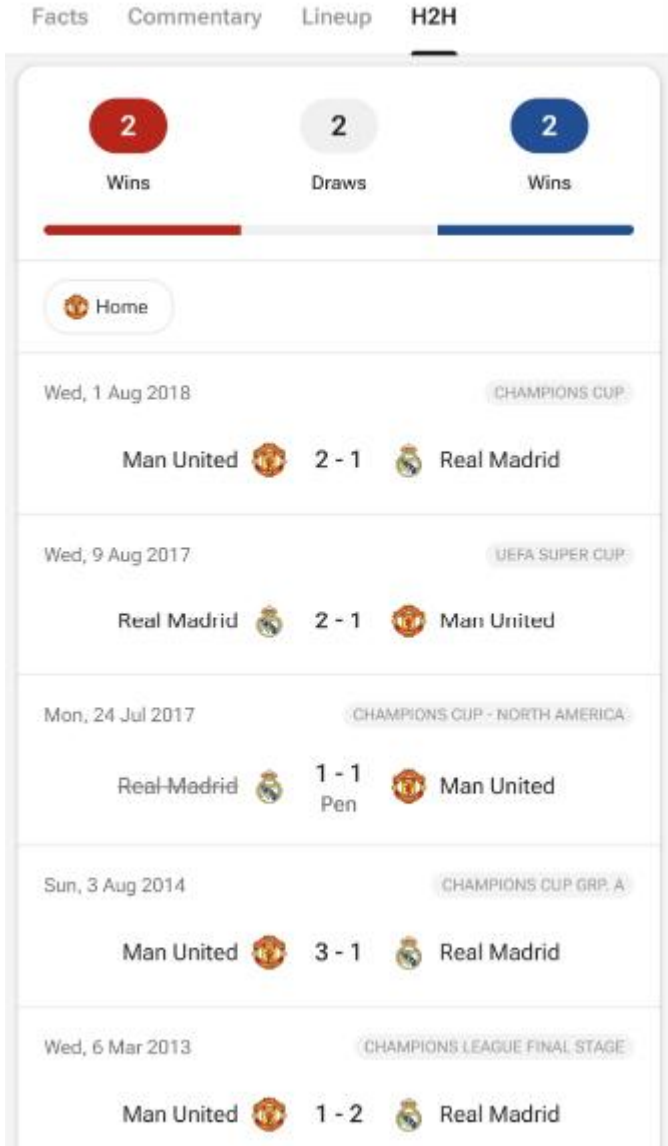
페이지 이름	Match Detail Page
페이지 설명	경기 예측 및 팀 데이터 분석 정보 전달



기능 상세

- 기본 경기 정보 제공. (팀, 경기 시간, 경기 날짜).
- 승부 확률 예측 정보 제공. (승리, 무승부 확률, 모델 정확도)
- 점수 예측 정보 제공. (예측 스코어 후보, 확률 수치, 모델 정확도)
- Key Player 분석 정보 제공. (예상 득점자, 기여도 수치)
- 경기 내 주요 포지션 매치업 예측 및 선수 능력 비교 레이더 차트 제공.

페이지 이름	Match Head to Head Page
페이지 설명	역대 상대전적 분석 정보 페이지

 <p>The screenshot displays the 'H2H' (Head-to-Head) section of a match page. At the top, there are three circular icons representing 'Wins' (red with '2'), 'Draws' (grey with '2'), and 'Wins' (blue with '2'). Below these is a horizontal bar chart. A 'Home' button with a house icon is visible. The main content is a list of matches:</p> <ul style="list-style-type: none"> Wed, 1 Aug 2018 (CHAMPIONS CUP): Man United 2 - 1 Real Madrid Wed, 9 Aug 2017 (UEFA SUPER CUP): Real Madrid 2 - 1 Man United Mon, 24 Jul 2017 (CHAMPIONS CUP - NORTH AMERICA): Real Madrid 1 - 1 Pen Man United Sun, 3 Aug 2014 (CHAMPIONS CUP GRP. A): Man United 3 - 1 Real Madrid Wed, 6 Mar 2013 (CHAMPIONS LEAGUE FINAL STAGE): Man United 1 - 2 Real Madrid 	
---	--

기능 상세

- 전적 요약 상단 표시.
- 경기별 H2H 리스트 표시 (경기 날짜, 대회명, 홈/원정 팀 이름과 로고, 스코어 및 결과, 승부차기 결과 표시 여부).
- 클릭 시 경기 상세 페이지 이동.

페이지 이름	League Standings Page
페이지 설명	리그 순위 정보 페이지

04:00

Preview

Table

H2H

Premier League

#	Team	Pl	W	D	L	+/-	GD	Pts
1	<div><div></div><div>Liverpool</div></div>	29	21	7	1	69-27	+42	70
2	<div><div></div><div>Arsenal</div></div>	29	16	10	3	53-24	+29	58
3	<div><div></div><div>Nottm Forest</div></div>	29	16	6	7	49-35	+14	54
4	<div><div></div><div>Chelsea</div></div>	29	14	7	8	53-37	+16	49
5	<div><div></div><div>Man City</div></div>	29	14	6	9	55-40	+15	48
6	<div><div></div><div>Newcastle</div></div>	28	14	5	9	47-38	+9	47
7	<div><div></div><div>Brighton</div></div>	29	12	11	6	48-42	+6	47
8	<div><div></div><div>Fulham</div></div>	29	12	9	8	43-38	+5	45
9	<div><div></div><div>Aston Villa</div></div>	29	12	9	8	41-45	-4	45
10	<div><div></div><div>Bournemouth</div></div>	29	12	8	9	48-36	+12	44
11	<div><div></div><div>Brentford</div></div>	29	12	5	12	50-45	+5	41
12	<div><div></div><div>Crystal Palace</div></div>	28	10	9	9	36-33	+3	39
13	<div><div></div><div>Man United</div></div>	29	10	7	12	37-40	-3	37

기능 상세
<ul style="list-style-type: none"> - 해당 경기 참여 팀 음영 처리. - 하단 스와이프 시 하위 순위 팀 load. - 리그 클릭 시 상단 드롭다운을 통해 리그 선택 가능.