Diseño de una Base de Datos Para el Análisis de Estadísticas Deportivas de la Premier League

Roa Reina John Isidro[1], Martinez Chitiva Freddy Orlando[2], Sepulveda Garzón Jerson Joan[3], $LermaBarbosaPaulaAndrea$[4]

[1234]Facultad de Ingeniería y Ciencias Básicas
Universidad Central
Maestría en Analí tica de Datos
Curso de Bases de Datos
Bogotá, Colombia
{[1]$jroar1@ucentral.edu.co$,[2]$fmartinezc@ucentral.edu.co$,[3]jsepulvedag@ucentral.edu.co,[4]

May 3, 2024

Contents

1 Introducción

Las actividades deportivas han generado gran interés en la población, no solo en el ámbito de la práctica, sino que también en el entretenimiento y la ciencia. Uno de los deportes más populares en el mundo es el fútbol, el cual no solo se basa en la actividad física, sino que también se deriva en diferentes actividades, como la medicina, videojuegos, predicciones estadísticas, entre otros. El gran interés que ha despertado el fútbol exige ser tratado con el mayor profesionalismo y es aquí donde los datos comienzas a tener un papel fundamental, pues en cada partido de las ligas se recopila gran cantidad de datos que llegan a ser indicadores determinantes para predecir el éxito de un jugador y un equipo. De estos indicadores se pueden derivar nuevos modelos de negocios como son la casa de apuestas, los cuales dependen de los análisis estadísticos que están basados en los datos obtenidos en cada juego, con el fin de predecir el resultado del siguiente partido. Las bases de datos cobran gran importancia una vez se recopila información, pues debe ser almacenada y organizada, de tal forma que facilite el acceso rápido y eficiente para realizar el análisis de eventos deportivos.

2 Características del proyecto de investigación que hace uso de Bases de Datos

El proyecto presentado propone desarrollar una base de datos robusta y funcional destinada a recopilar, almacenar y analizar una v cantidad de datos deportivos. Este sistema permitirá medir el desempeño de los equipos de la Premier League mediante el uso de estadísticas avanzadas y refinadas. La base de datos relacional diseñada facilitará la evaluación de múltiples indicadores de rendimiento, tales como desempeño por equipo, local y visitante, goles marcados, tiros de esquina y tiros a puerta, entre otros. Estos datos no solo servirán para mejorar la preparación física y las estrategias de los equipos, sino que también proporcionarán insights valiosos para medios de comunicación, consultorías deportivas, desarrolladores tecnológicos y el sector de apuestas deportivas. El alcance del proyecto abarca desde el diseño estructural de la base de datos hasta la implementación de procesos de ETL y consultas complejas para el análisis estadístico, asegurando así la integración efectiva y el manejo óptimo de los datos. Con esto, el proyecto busca establecer un estándar en el análisis deportivo, ofreciendo una herramienta poderosa y competitiva en el mercado.

2.1 Titulo del proyecto de investigación

Diseño de una Base de Datos Para el Análisis de Estadísticas Deportivas de la Premier League.

2.2 Objetivo general

Diseñar una base de datos relacional que permita hacer análisis estadísticos orientados a medir el desempeño de los equipos que integran la Premier League.

2.2.1 Objetivos especificos

- Recopilar datos históricos de las últimas cinco temporadas de la Premier League con el fin de evaluar estadísticas de los diferentes eventos deportivos.
- Diseñar la estructura de la Base de Datos definiendo las entidades que le incorporarán, así como los atributos asociados a cada una, con el fin de garantizar el uso efectivo de los datos incorporados.
- Determinar las relaciones existentes entre cada una de las entidades definidas para la Base de Datos.
- Desarrollar procesos de extracción, transformación y carga (ETL, por sus siglas en inglés) para facilitar la recopilación, limpieza y carga de datos en la base de datos.
- Implementar consultas de los datos incorporados en la Base de Datos que permitan evaluar diferentes análisis estadísticos.

2.3 Alcance

El alcance de este proyecto comprende desarrollar una base de datos refinada que nos permita evaluar diferentes indicadores de la Premier League, logrando que sea competitiva contra otras en el mercado para así obtener mejores análisis estadísticos, entre los que se encuentran: estadísticas por equipos, desempeño local y visitante, número de goles marcados en un partido, número de tiros de esquina, tiros a puerta. Estos datos podrían ser utilizados en diferentes áreas, como es la preparación física, medios de comunicación, consultoría deportiva para negocios, desarrollo tecnológico, apuestas deportivas y aficionados.

2.4 Pregunta de investigación

¿Cómo se puede diseñar e integrar una Base de Datos de la Premier League lo suficientemente robusta que pueda utilizarse para construir análisis estadísticos relacionados con los eventos deportivos?

2.5 Hipotesis

La implementación de una base de datos especializada y bien estructurada para el análisis de estadísticas deportivas de la Premier League incrementará significativamente la precisión y la profundidad de los análisis estadísticos realizados, lo cual resultará en predicciones más precisas del desempeño de los equipos. Esto, a su vez, mejorará la toma de decisiones en áreas como la estrategia de juego, la preparación física, y las apuestas deportivas.

3 Reflexiones sobre el origen de datos e información

Los datos no son simplemente números o estadísticas; son representaciones de eventos reales, desempeños y decisiones humanas, cada uno con su propio contexto y significado, lo que lo hace de mucho cuidado y tratamiento.

Primero, la calidad y confiabilidad de los datos comienzan con su origen. En el caso de un análisis deportivo, esto implica considerar cómo se recopilan los datos durante los partidos. ¿Se utilizan tecnologías de seguimiento avanzadas? ¿Cuán entrenados están los individuos que registran los eventos durante un juego? Estos factores determinan la precisión de los datos recopilados. Un origen de datos defectuoso o sesgado puede llevar a conclusiones incorrectas, afectando estrategias y decisiones basadas en estos análisis.

Además, es fundamental considerar la ética del origen de los datos. En el deporte, la privacidad y la integridad de los datos de los jugadores deben ser manejadas con el máximo cuidado. La recopilación de datos no debe comprometer la privacidad ni la integridad de las personas.

Finalmente, una reflexión profunda sobre el origen de los datos también abarca su actualidad y relevancia. En un campo tan sujeto a cambios como el deporte, los datos antiguos pueden no ser representativos de las condiciones actuales. Por lo tanto, mantener una base de datos actualizada y relevante es esencial para obtener análisis que sean verdaderamente útiles y aplicables en tiempo real.

Al considerar estos aspectos, no solo mejoramos la integridad de los análisis estadísticos, sino que también fortalecemos la confianza en las decisiones que se toman basadas en estos datos.

3.1 ¿Cual es el origen de los datos e información ?

Se utilizará una base histórica de las últimas cinco temporadas de la Premier League tomadas de la página FBREF, donde se extrajeron los datos en colaboración con aplicativos que hacen toma de datos en tiempo real a través de una combinación de anotaciones humanas, visión artificial y modelado de IA. Donde contiene variables como fecha del partido, equipo local, equipo visitante, goles anotados por el equipo local, goles anotados por el equipo visitante, entre otras variables.

Adicionalmente, mediante técnicas de web scraping utilizando Python, actualizaremos semanalmente los resultados de la liga en curso, para efectos de comparar el desempeño de los equipos en la actualidad.

3.2 ¿Cuales son las consideraciones legales o eticas del uso de la información?

Los datos utilizados para este proyecto corresponden a datos disponibles en la web, los cuales no están restringidos para su consulta, por lo que no estimamos la existencia de asuntos legales asociados al uso de estos.

3.3 ¿Cuales son los retos de la información y los datos que utilizara en la base de datos en terminos de la calidad y la consolidación?

Es esencial asegurar que los datos estén completos, precisos, actualizados y consistentes. La completitud de los datos implica que no hayan vacíos críticos que puedan llevar a análisis erróneos. La precisión es vital para que las decisiones basadas en estos datos sean confiables. Además, la actualización constante de la base de datos asegura que los análisis reflejen las condiciones actuales.

Otro reto significativo es la consolidación efectiva de los datos, la cual debe realizarse de manera que se preserve la integridad de la información. Esto implica diseñar e implementar procesos de transformación y carga de datos (ETL) que no solo combinen eficientemente datos de diversas fuentes sino que también limpien y corrijan los datos sin comprometer su veracidad original. La habilidad para mantener la integridad de los datos a lo largo de este proceso es fundamental para asegurar la validez de los análisis estadísticos derivados de la base de datos.

3.4 ¿Que espera de la utilización de un sistema de Bases de Datos para su proyecto?

De la implementación de un sistema de bases de datos para nuestro proyecto, esperamos lograr una gestión eficiente y efectiva de grandes volúmenes de datos estadísticos de la Premier League. Esto permitirá realizar análisis profundos y precisos del desempeño de los equipos y jugadores, facilitando la toma de decisiones estratégicas basadas en datos confiables y actualizados. Además, buscamos que el sistema ofrezca una plataforma robusta y escalable que soporte la integración y consolidación de datos de diversas fuentes, garantizando así la integridad y la seguridad de la información recopilada..

4 Diseño del Modelo de Datos del SMBD (Sistema Manejador de Bases de Datos)

4.1 Características del SMBD (Sistema Manejador de Bases de Datos) para el proyecto

Para el proyecto en mención, hemos seleccionado Oracle Developer como nuestro SMBD. Las características que ofrece este SMBD son fundamentales para el éxito de nuestra implementación y gestión de datos, permitiéndonos optimizar el rendimiento y la eficiencia de nuestras operaciones de base de datos.

A continuación, se detallan algunas de las características clave de Oracle Developer que son particularmente relevantes para nuestro proyecto:

Gestión de Grandes Volumen de Datos: Oracle Developer está diseñado para manejar y procesar grandes volúmenes de datos de manera eficiente. Esto es esencial para nuestro proyecto, dado que se recopilan y analizan extensas cantidades de datos estadísticos cada temporada. Alto Rendimiento y Escalabilidad: Oracle ofrece una plataforma altamente escalable que puede crecer junto con las necesidades de nuestro proyecto. Su capacidad para manejar incrementos en la carga de datos sin degradar el rendimiento es crucial para mantener la fluidez de nuestras operaciones.

Seguridad Robusta: Oracle Developer proporciona amplias características de seguridad, incluyendo control de acceso, encriptación de datos y mecanismos de auditoría. Estas herramientas son imprescindibles para proteger la integridad y la privacidad de los datos sensibles del proyecto.

Herramientas de Desarrollo Integradas: Oracle Developer incluye herramientas integradas que facilitan el diseño, desarrollo, y mantenimiento de la base de datos. Esto simplifica la creación de esquemas, la gestión de objetos de base de datos y la depuración de aplicaciones.

S**oporte para Procedimientos Almacenados y Automatización:** El soporte para procedimientos almacenados permite encapsular la lógica del negocio directamente en la base de datos, mejorando el rendimiento y la seguridad. Además, las capacidades de automatización facilitan la ejecución de tareas de mantenimiento y actualización de datos.

Compatibilidad con Diversas Interfaces de Programación: Oracle soporta una amplia gama de interfaces de programación, lo que permite a los desarrolladores trabajar en un entorno flexible y con las herramientas que mejor se adapten a sus necesidades.

Estas características hacen de Oracle Developer una elección acertada para soportar el análisis estadístico detallado y en tiempo real requerido por nuestro

proyecto, asegurando que la infraestructura de datos sea tanto robusta como adaptable a los desafíos futuros.

4.2 Diagrama modelo de datos

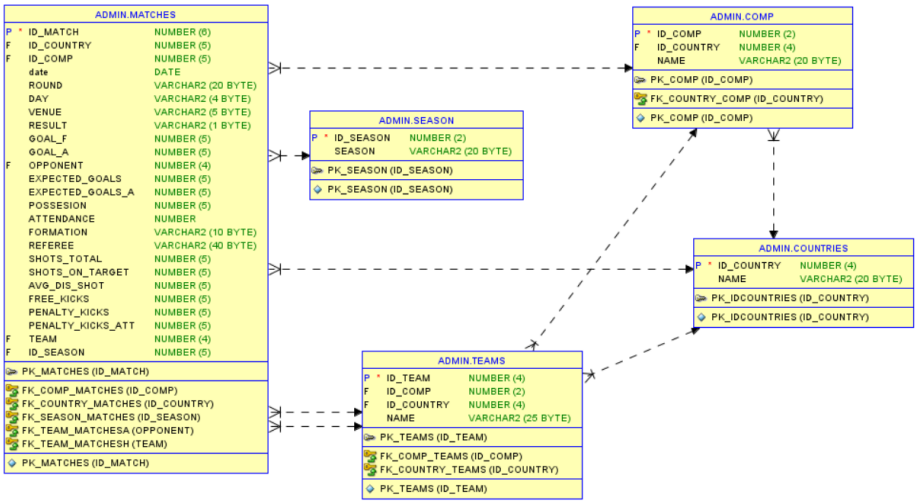

Figure 1: Diagrama Modelo de Datos

4.3 Imágenes de la Base de Datos

	ID_MATCH	ID_COUNTRY	ID_COMP	date	ROUND	DAY	VENUE	RESULT	GOAL_F	GOAL_A	OPPONENT	EXPECTED_GOALS	EXPECTED_GOALS_A	POSSESION	ATTENDANCE	FORMAT
1	101001	1001	1	23/11/19	Matchweek 13	Sat	Away	W	3	2	1026	21	7	52	599300	4-2-3-1
2	101002	1001	1	23/11/19	Matchweek 13	Sat	Home	L	2	3	1023	7	21	48	599300	4-4-1-1
3	101003	1001	1	23/11/19	Matchweek 13	Sat	Away	W	2	1	1009	14	10	60	254860	4-3-3
4	101004	1001	1	23/11/19	Matchweek 13	Sat	Away	W	2	0	1005	36	6	47	306400	4-1-4-1
5	101005	1001	1	23/11/19	Matchweek 13	Sat	Away	W	2	1	1003	11	6	53	105390	3-4-3
6	101006	1001	1	23/11/19	Matchweek 13	Sat	Home	D	2	2	1022	18	24	61	602950	3-4-1-2
7	101007	1001	1	23/11/19	Matchweek 13	Sat	Away	W	3	0	1024	19	10	38	197110	4-4-2
8	101008	1001	1	23/11/19	Matchweek 13	Sat	Away	D	2	2	1001	24	18	39	602950	4-4-2
9	101009	1001	1	23/11/19	Matchweek 13	Sat	Home	L	0	2	1019	9	17	57	392410	4-2-3-1
10	101010	1001	1	23/11/19	Matchweek 13	Sat	Home	L	1	2	1015	10	14	40	254860	4-5-1
11	101011	1001	1	23/11/19	Matchweek 13	Sat	Home	L	0	2	1014	6	36	53	306400	3-4-3
12	101012	1001	1	23/11/19	Matchweek 13	Sat	Home	L	1	2	1027	6	11	47	105390	3-4-3
13	101013	1001	1	23/11/19	Matchweek 13	Sat	Home	L	0	3	1006	10	19	62	197110	5-3-2
14	101014	1001	1	23/11/19	Matchweek 13	Sat	Away	W	2	0	1010	17	9	43	392410	4-2-3-1
15	101015	1001	1	23/11/19	Matchweek 13	Sat	Home	W	2	1	1008	10	9	47	544860	4-3-3

Figure 2: Tabla Matches

	ID_SEASON	SEASON
1	1	2018-2019
2	2	2019-2020
3	3	2020-2021
4	4	2021-2022
5	5	2022-2023

Figure 3: Tabla Season

	ID_TEAM	ID_COMP	ID_COUNTRY	NAME
1	1001	1	1001	Arsenal
2	1002	1	1001	Aston Villa
3	1003	1	1001	Bournemouth
4	1004	1	1001	Brentford
5	1005	1	1001	Brighton and Hove Albion
6	1006	1	1001	Burnley
7	1007	1	1001	Cardiff City
8	1008	1	1001	Chelsea
9	1009	1	1001	Crystal Palace
10	1010	1	1001	Everton
11	1011	1	1001	Fulham
12	1012	1	1001	Huddersfield Town
13	1013	1	1001	Leeds United
14	1014	1	1001	Leicester City
15	1015	1	1001	Liverpool
16	1016	1	1001	Manchester City

Figure 4: Tabla Teams

	ID_COMP	ID_COUNTRY	NAME
1	1	1001	Premier League
2	2	1002	La Liga

Figure 5: Tabla Competition

	ID_COUNTRY	NAME
1	1001	Inglaterra
2	1002	EspaÃ±a

Figure 6: Tabla Country

4.4 Código SQL - lenguaje de definición de datos (DDL)

```
CREATE TABLE matches (
 id_match NUMBER(5),
 id_country NUMBER(5),
 id_comp NUMBER(5),
 "date" DATE,
 "time" INTERVAL DAY TO SECOND,
 round VARCHAR2(20),
 day VARCHAR2(4),
 venue VARCHAR2(5),
 result VARCHAR2(1),
 goal_f NUMBER(5),
 goal_a NUMBER(5),
 opponent VARCHAR2(25),
 expected_goals NUMBER(5),
 expected_goals_a NUMBER(5),
 possesion NUMBER(5),
 attendance NUMBER(5),
 formation VARCHAR2(10),
 referee VARCHAR2(25),
 shots_total NUMBER(5),
 shots_on_target NUMBER(5),
 avg_dis_shot NUMBER(5),
 free_kicks NUMBER(5),
 penalty_kicks NUMBER(5),
 penalty_kicks_att NUMBER(5),
 team VARCHAR2(25),
 id_season NUMBER(5)
```

```
CREATE TABLE season (
 id_season NUMBER(2),
 season VARCHAR2(20)
);
```

```
CREATE TABLE comp (
 id_comp NUMBER(2),
 id_country NUMBER(4),
 name VARCHAR2(20)
);
```

```
CREATE TABLE teams (
 id_team NUMBER (4),
 id_comp NUMBER (2),
 id_country NUMBER (4),
 name VARCHAR2(25)
);
```

```
CREATE TABLE countries (
 id_country NUMBER (4),
 name VARCHAR2(20)
);
```

Figure 7: Lenguaje Definición Datos DDL

4.5 Código SQL - Manipulación de datos (DML) (*Primera entrega*)

4.6 Código SQL + Resultados: Vistas (*Primera entrega*)

4.7 Código SQL + Resultados: Triggers (*Primera entrega*)

4.8 Código SQL + Resultados: Funciones (*Primera entrega*)

4.9 Código SQL + Resultados: procedimientos almacenados (*Primera entrega*)

5 Bases de Datos No-SQL (*Segunda entrega*)

5.1 Diagrama Bases de Datos No-SQL (*Segunda entrega*)

5.2 SMBD utilizado para la Base de Datos No-SQL (*Segunda entrega*)

6 Aplicación de ETL (Extract, Transform, Load) y Bodega de Datos (*Tercera entrega*)

6.1 Ejemplo de aplicación de ETL y Bodega de Datos (*Tercera entrega*)

6.2 Automatización de Datos (*Tercera entrega*)

6.3 Integración de Datos (*Tercera entrega*)

7 Proximos pasos (*Tercera entrega*)

8 Lecciones aprendidas (*Tercera entrega*)

9 Bibliografía