APLICACIONES DE CIENCIA DE DATOS ORIENTADAS AL ANÁLISIS DEL FACTOR HUMANO:

CASO DE ESTUDIO EN EL ÁMBITO DEL DEPORTE Y ESTUDIO DE PARALELISMOS PARA SU APLICACIÓN A EQUIPOS HUMANOS EN EL ÁMBITO INDUSTRIAL

Universidad de Oviedo

Álvaro Ponte Blanco

uo233109@unoivi.es

Máster en Automatización e Informática Industrial

Tutor: Ignacio Diaz Blanco - Universidad de Oviedo

maiind
Máster en Ingeniería de Automatización e Informática Industrial

RESUMEN

El factor humano es un apartado clave en la industria, y predecir su comportamiento puede aportar un gran valor para estimar errores humanos y mejorar el rendimiento. Debido a la ausencia de información interesante sobre el factor humano en procesos de fabricación, se ha utilizado otra tipología de datos para predecir esta conducta: información deportiva. Este TFM pretende aplicar conocimientos y algoritmos de analítica avanzada de datos con el objetivo de estudiar las características del comportamiento humano ante unas situaciones determinadas relacionadas con el deporte, desarrollando en detalle el paralelismo con el mundo industrial. Se ha especificado el caso de uso en el fútbol, debido a la gran disponibilidad de datos en la web.

Se ha creado una base de datos propia, se han realizado análisis descriptivos para comprender la información, y se han diseñado modelos predictivos para estimar probabilidades de ciertos sucesos en el deporte. El resultado final es una aplicación en un entorno HTML que permite la visualziación los resultados de cada fase de una forma interactiva. Adicionalmente, el proyecto engloba otras funcionalidades con los datos como la generación automática de reportes que muestran los resultados de una temporada.

Palabras clave: **Big Data, Advanced Analytics, Data Science, IA, IoT, Human Factor, Boosting**

1. Objetivos y Estructura

El objetivo fundamental es la demostración del uso de la analítica avanzada de datos para **estimar el factor humano en la industria**.
Adicionalmente, se pretende exponer otras funcionalidades del Big Data en el contexto industrial : **analítica descriptiva**, generación de reportes…

Se resuelve el problema con datos sobre un deporte concreto:el fútbol. La programación se desarrolla con **R (RStudio)** y sus librerías.

El proyecto se integra dentro de la metodología para minería de datos **CRISP-DM** :

1. Estudio del Caso de Aplicación
2. Creación de la Base de Datos
3. Diseño de Análisis Descriptivos
4. Desarrollo de Modelos Predictivos
5. Resultados e Implementación

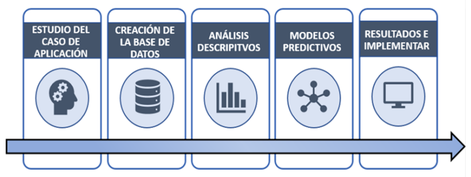

Fig. 1. 1 – Fases del Proyecto

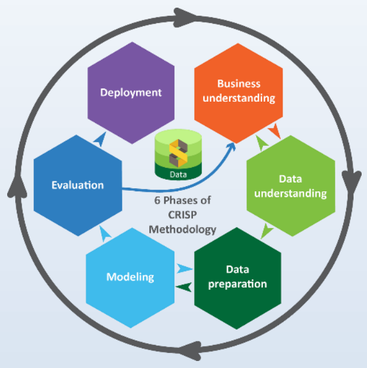

Fig. 1. 2: Metodología CRISP-DM – *http://www.sailotech.com*

2. Creación de la Base de Datos

Esta etapa consiste en la **creación de una nueva base de datos** sobre fútbol a través de procedimientos de **web scraping**. Se extrae información de cinco sitios webs distintos mediante algoritmos en R.

Además de la obtención de información, esta fase contempla otras tareas:

- **Diseño** de la base de datos
- **Extracción** de información
- **Procesamiento** inicial de los datos
- **Unión** de diferentes bases de datos
- **Definición de relaciones** entre variables

Fig. 2. 1 – Sitios web - fuentes de información

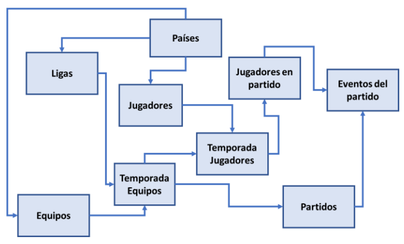

Fig. 2. 2 – Diseño de la base de datos generada

3. Diseño de Análisis Descriptivos

Esta fase tiene como objetivo el **desarrollo y la visualización de análisis descriptivos** utilizando la información almacenada en la base de datos generada. Esta etapa pretende ayudar al entendimiento del problema planteado, facilitar la comprensión de los datos y ofrecer una **visión gráfica e interactiva de la información**. Se distinguen cinco tipos de análisis principales:

ANÁLISIS DEPORTIVOS

Análisis de índole futbolístico que son específicos para el campo de aplicación seleccionado: el deporte.

Posicion	Equipo	Puntos	Victorias	Empates	Derrotas
1	Chelsea	87	26	9	3
2	Man City	79	24	7	7
3	Arsenal	75	22	9	7
4	Man United	70	20	10	8
5	Tottenham	64	19	7	12
6	Liverpool	62	18	8	12
7	Southampton	60	18	6	14
8	Swansea	56	16	8	14
9	Stoke	54	15	9	14
10	Crystal Palace	48	13	9	16
11	West Ham	47	12	11	15
12	Everton	47	12	11	15

Fig. 3. 1 – Ejemplo Análisis Deportivo

ANÁLISIS ESTADÍSTICOS

Análisis basados en patrones estadísticos como el promedio, frecuencia de sucesos o filtrados.

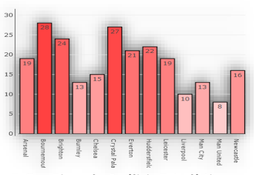

Fig. 3. 2 – Ejemplo Análisis Estadístico

ANÁLISIS MACHINE LEARNING

Análisis de mayor complejidad que utilizan técnicas más avanzadas (*regresión lineal, clustering,..*)

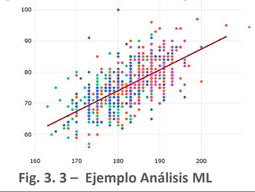

Fig. 3. 3 – Ejemplo Análisis ML

ANÁLISIS ESPECÍFICOS DEL CASO

Análisis en mayor profundidad sobre los casos de uso de los modelos predictivos (tarjetas)

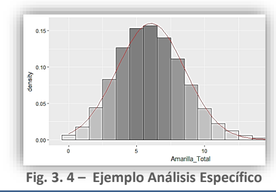

Fig. 3. 4 – Ejemplo Análisis Específico

ANÁLISIS ADICIONALES

Análisis sobre otros aspectos que proporcionan información útil adicional. Ej: Análisis Geoespacial.

Fig. 3. 5 – Ejemplo Análisis Geoespacial

4. Desarrollo de Modelos Predictivos

Las estimaciones se hacen a través de **modelos predictivos**. Se utiliza un algoritmo de **gradient boosting** implementado en R a través de la **librería XGBoost**. Se desarrolla un modelo de cada tipo (binario, multiclase, y de regresión) relacionado con el **comportamiento humano en el deporte.**

Etapas en el desarrollo de modelos predictivos:

1. Definición de la **variable objetivo**
2. Selección de **Información** para el modelo
3. Creación de **MasterTable**
4. Optimización de **Parámetros** y Ejecución
5. **Evaluación** del modelo y **Métricas** de resultado

MODELO DE CLASIFACIÓN BINARIO: ESTIMACIÓN DE LA PROBABILIDAD DE EXPULSIÓN

La variable objetivo es 1 **si hay expulsiones** y 0 en caso contrario.

La métrica de error es compleja debido a las clases desbalanceadas. Métricas basadas **en la captura del nº de expulsiones** en función del valor estimado en la predicción

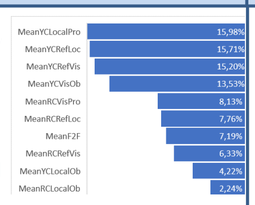

Fig. 4. 1– Importancia de las variables más significativas del modelo binario

MODELO DE CLASIFACIÓN MULTICLASE: CÁLCULO DE Nº TARJETAS AMARILLAS EN UN PARTIDO

La variable objetivo es el **número total de tarjetas amarillas** en un encuentro. Las clases son el número posible de tarjetas (0, 1, 2, 3, 4..)

El resultado para cada muestra es un **vector con la probabilidad** de pertenencia a cada una de las clases

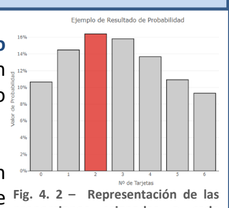

Fig. 4. 2 – Representación de las proporciones asignadas a cada clase para una muestra aleatoria del test

MODELO DE REGRESIÓN: ESTIMACIÓN DE LA ASISTENCIA DE PÚBLICO AL ESTADIO

La variable objetivo es el **porcentaje de ocupación de un estadio** para un encuentro determinado (rango de 0 a 1)

El resultado también es un valor continuo en el mismo rango. Para la métrica de error se ha utilizado el parámetro **RMSE**

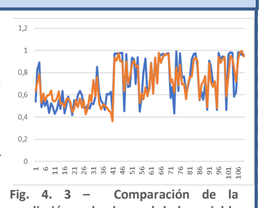

Fig. 4. 3 – Comparación de la predicción y el valor real de la variable para cien muestras aleatorias

ANALOGÍAS EN LA INDUSTRIA

Modelos de Estimación de Tarjetas

- Las tarjetas son el resultado de una **acción no reglamentaria** en el fútbol. La analogía de este caso es la predicción de errores humanos en un proceso de producción industrial
- **Modelo de Tarjetas Rojas**: Estimación de probabilidad de un error con consecuencias graves
- **Numero de Tarjetas Amarillas**: cantidad de errores leves (retrasos, poca calidad) en el procedimiento.

Modelo de Estimación de la Asistencia

- Este modelo es análogo a un modelo **de predicción de la demanda** en la industria. Pretende estimar el comportamiento de espectadores, que cumplen el rol de potenciales clientes de un producto.

5. Generación Automática de Reportes

Se trata de una funcionalidad con los datos que permite **generar archivos** en formato de xlsx de **forma automática.**

Está basado en un **algoritmo en R** que recibe como entrada un csv con datos de los resultados de una temporada y genera un fichero de Excel con tres hojas:

- **Hoja 1**: Resumen de los resultados de una temporada por jornadas
- **Hoja 2**: Clasificación de una competición liguera
- **Hoja 3:** Tabla Cruzada de Resultados

	A	B	C	D	E	F	G	H	I	J	K
1		Alme	AthB	AthM	Barc	Espa	Geta	LaCo	Mala	Mall	Osas
2	Almeria		1-4	1-0	2-2	0-1	1-0	1-1	1-0	1-1	2-0
3	Ath Bilbao	4-1		1-0	1-1	1-0	2-2	2-0	1-1	1-3	2-0
4	Ath Madrid	2-2	2-0		2-1	4-0	0-3	3-0	0-2	1-1	1-0
5	Barcelona	1-0	4-1	5-2		1-0	2-1	3-0	2-1	4-2	2-0
6	Espanol	2-0	1-0	3-0	0-0		0-2	2-0	2-1	1-1	2-1
7	Getafe	2-2	2-0	1-0	0-2	1-1		0-2	2-1	3-0	2-1
8	La Coruna	0-0	3-1	2-1	1-3	2-3	1-3		1-0	1-0	1-0
9	Malaga	1-2	1-1	3-0	0-2	2-1	1-0	0-0		2-1	1-1
10	Mallorca	3-1	2-0	4-1	0-1	2-0	3-1	2-0	1-1		2-0
11	Osasuna	1-0	0-0	3-0	1-1	2-0	0-0	3-1	2-2	0-1	

Fig. 5.1 y 5.2 : Muestra de los reportes generados

	B	C	D	E	F	G	H	I	J	K	L	M
1		PRIMERA VUELTA							SEGUNDA VUELTA			
2												
3		Jornada: 1							Jornada: 20			
4	2007-08-11	Aston Villa	1	2	Liverpool			2007-12-29	Birmingham	1	1	Fulham
5	2007-08-11	Bolton	1	5	Newcastle			2007-12-29	Chelsea	2	1	Newcastle
6	2007-08-11	Derby	2	2	Portsmouth			2007-12-29	Everton	1	4	Arsenal
7	2007-08-11	Everton	2	1	Wigan			2007-12-29	Portsmouth	0	1	Middlesbrou
8	2007-08-11	Middlesbrou	1	2	Blackburn			2007-12-29	Sunderland	3	1	Bolton
9	2007-08-11	Sunderland	1	0	Tottenham			2007-12-29	Tottenham	6	4	Reading
10	2007-08-11	West Ham	0	2	Man. City			2007-12-29	West Ham	2	1	Man. Utd
11	2007-08-12	Arsenal	2	1	Fulham			2007-12-29	Wigan	1	2	Aston Villa
12	2007-08-12	Chelsea	3	2	Birmingham			2007-12-30	Derby	1	2	Blackburn
13	2007-08-12	Man. Utd	0	0	Reading			2007-12-30	Man. City	0	0	Liverpool

6. Aplicación Web

Se trata de una **aplicación en un entorno web** que muestra los resultados de las fases de este proyecto de **forma visual e interactiva.**

Está diseñada con el paquete **Shiny**, una librería de R que permite el desarrollo de aplicaciones **en lenguaje HTML** a través de la programación en RStudio. También se puede combinar con funciones de otros lenguajes.

Dentro de la aplicación, el usuario tiene la posibilidad de **visualizar resultados, interactuar** con los análisis descriptivos **o extraer información** de las tablas.

Figuras 6.1, 6.2, 6.3 y 6.4 : Aplicación en Entorno Web