

Proyecto de clase

Descripción

El dataset que se va a analizar contiene las estadísticas avanzadas de rendimiento de jugadores de baloncesto en la NBA durante la temporada 2017-2018, así como los salarios de los mismos. Se dispone del diccionario de datos siguiente:

- Player: nombre del jugador
- Salary: salario anual (en dolares)
- NBA_Country: país de procedencia
- NBA_DraftNumber: número de selección en el draft
- Age: edad
- Tm: equipo
- G: partidos
- MP: minutos jugados
- PER: Player Efficiency Rating
- TS%: True Shooting Percentage (porcentaje real de tiros)
- 3Par: tasa de intentos de tres puntos
- FTr: tasa de intentos de tiro libre
- ORB%: porcentaje de rebotes ofensivos
- DRB%: porcentaje de rebotes defensivos
- TRB%: porcentaje de rebotes totales
- AST%: porcentaje de asistencias
- STL%: porcentaje de robos
- BLK%: porcentaje de bloqueos
- TOV%: porcentaje de pérdidas
- USG%: porcentaje de uso
- OWS: Offensive Win Shares (número estimado de victorias contribuidas por un jugador dado su desempeño ofensivo)
- DWS: Defensive Win Shares (número estimado de victorias contribuidas por un jugador dado su desempeño defensivo)
- WS: Win Shares (número estimado de victorias contribuidas por un jugador)
- WS/48: Win Shares per 48 (número estimado de victorias contribuidas por un jugador por 48 minutos)
- OBPM: Offensive Box Plus/Minus (métrica para medir la contribución de un jugador al equipo mientras está en el juego dado su juego ofensivo)
- DBPM: Defensive Box Plus/Minus (métrica para medir la contribución de un jugador al equipo mientras está en el juego dado su juego defensivo)
- BPM: Box Plus/Minus (métrica para medir la contribución de un jugador al equipo mientras está en el juego)
- VORP: Value Over Replacement Player (métrica para medir la contribución total de un jugador al equipo)

La idea es poder predecir los salarios a partir de un modelo basado en los demás campos.

Puntos a desarrollar

1. **Limpieza y EDA:** Verifiquen si hay problemas de calidad de datos.
Se espera una primera sección de evaluación de la calidad de los datos y de entendimiento de la relación entre las variables predictivas y la variable objetivo (**OJO!** Solo poner gráficos y análisis de las relaciones importantes, **menos es más!**)
2. **Modelos predictivos:** Entrenen modelos predictivos (al menos 3 familias de modelos) que permitan estimar el salario de los jugadores de la NBA a partir de los valores de las demás variables. Escoja el mejor modelo, buscando sus parámetros óptimos.
Se espera una sección donde se establezca el protocolo de evaluación y los procesos de entrenamiento y evaluación de los modelos.
3. **Cambio de representación del dataset:** Considerando todas las variables (menos Salary), realice un análisis de componentes principales (PCA), escogiendo el número de componentes necesarios para conservar el 95% de la representación original.
4. **Caracterización de los jugadores:** Con los datos en su nueva representación de PCs, realice una segmentación, estableciendo el mejor número de clusters entre 3 y 5. Caracterice los clusters con respecto a las variables originales (incluyendo Salary).

Rúbricas de puntuación

1ª entrega: Sesión 9

Calidad de datos	Visualización de datos	Extracción de intuiciones de los datos	TOTAL PROYECTO PRIMERA ENTREGA
1.0	2.0	2.0	5.0

2ª entrega: Sesión 12

Entendimiento de los datos y limpieza	Protocolo de entrenamiento y evaluación de modelos	Entrenamiento de los 3 modelos	TOTAL PROYECTO SEGUNDA ENTREGA
1.0	1.0	3.0	5.0

3ª entrega: Sesión 15

Transformación de los datos por PCA		Caracterización de perfiles de basquetbolistas en la NBA	TOTAL PROYECTO TERCERA ENTREGA
2.0		3.0	5.0