UNIVERSIDAD ICESI
ANALÍTICA DE DATOS
Javier Diaz Cely

Proyecto de clase

Descripción

El dataset que va a analizar contiene las estadísticas de rendimiento de jugadores de baseball en USA en 1986, así como los salarios de los mismos. Se dispone del diccionario de datos siguiente:

- Player: Nombre del jugador
- Atbat: número de veces que se ha presentado para batear en 1986
- Hits: número de veces que ha conectado un bateo sencillo y llegado a base en 1986
- HmRun: número de home runs en 1986
- Runs: número de carreras anotadas en 1986
- RBI: número de carreras impulsadas en 1986
- Walks: número de veces que ha pasado por bolas en 1986
- Years: número de años como profesional
- CAtbat: número de veces que se ha presentado para batear en su carrera
- CHits: número de veces que ha conectado un bateo sencillo y llegado a base en su carrera
- CHmRun: número de home runs en su carrera
- CRuns: número de carreras anotadas en su carrera
- CRBI: número de carreras impulsadas en su carrera
- CWalks: número de veces que ha pasado en su carrera
- League: categoría de la liga en la que jugaba en 1986
- Division: división en la que jugaba en 1986
- PutOuts: número de ponchadas que ha generado defensivamente en 1986
- Assists: número de veces que ha ayudado a ponchar a un jugador contrario indirectamente en 1986
- Errors: número de errores en 1986
- NewLeague: categoría de la liga en la que jugaba el día de apertura de la temporada 1987
- Salary: salario anual (en miles de dólares) el día de apertura de la temporada 1987

No se conocen los salarios de algunos de los jugadores (aparecen con NA). La idea es poder predecirlos 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 (diferentes a los NAs del campo Salary).
   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 mas!**)
2. **Modelos predictivos**: Entrenen modelos predictivos (al menos 3 familias de modelos) que permitan estimar el salario de los jugadores de baseball a partir de los valores de las demás variables (de los jugadores que tienen valor de Salary). 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: 5 de noviembre 2019

Calidad de datos	Visualización de datos	Extracción de intuiciones de los datos	Modelo predictivo preliminar	Análisis de resultados del modelo preliminar	TOTAL PROYECTO PRIMERA ENTREGA
1.0	1.0	1.0	1.0	1.0	5.0

2ª entrega: 12 de noviembre 2019

Entendimiento de los datos y limpieza	Protocolo de entrenamiento y evaluación de modelos	Entrenamiento de los 3 modelos	Transformación de los datos por PCA	Caracterización de perfiles de beisbolistas	TOTAL PROYECTO SEGUNDA ENTREGA
0.5	0.5	2.0	0.5	1.5	5.0