UNIVERSIDAD MAYOR, REAL Y PONTIFICIA DE SAN FRANCISCO XAVIER DE CHUQUISACA

FACULTAD DE TECNOLOGÍA

Inteligencia Artificial

Carrera : Ingeniería en Ciencias de la Computación

Universitario : Lujan Renteria David Fernando

Git: [SIS-420/Laboratorios/LABORATORIO 02 at main · lujan-99/SIS-420 (github.com)]

Sucre – Bolivia

2024

Informe de Métodos de Regresión

1. Análisis y Descripción del Dataset

El dataset utilizado para este análisis contiene **12,844 filas** y **22 columnas**, con una combinación de datos numéricos y de texto. A continuación, se presentan las columnas más relevantes para una regresión lineal multivariable:

- **age (Edad)**: La edad del jugador en años.
- **player_height (Altura del jugador)**: La altura del jugador en centímetros.
- **player_weight (Peso del jugador)**: El peso del jugador en kilogramos.
- **gp (Juegos jugados)**: Número total de juegos en los que el jugador ha participado.
- **pts (Puntos)**: Total de puntos anotados por el jugador.
- **reb (Rebotes)**: Total de rebotes capturados por el jugador.
- **ast (Asistencias)**: Total de asistencias realizadas por el jugador.
- **net_rating (Calificación neta)**: Indicador del rendimiento neto del jugador.
- **oreb_pct (Porcentaje de rebotes ofensivos)**: Porcentaje de rebotes ofensivos capturados.
- **dreb_pct (Porcentaje de rebotes defensivos)**: Porcentaje de rebotes defensivos capturados.
- **usg_pct (Porcentaje de uso)**: Porcentaje de posesiones en las que el jugador está involucrado.
- **ts_pct (Porcentaje de tiros)**: Porcentaje de tiros realizados por el jugador.
- **ast_pct (Porcentaje de asistencias)**: Porcentaje de asistencias realizadas por el jugador.

Resultado del Análisis: Contamos con gran cantidad de datos numéricos que serán de gran utilidad para el entrenamiento pero con los jugadores no drafteados

```
RangeIndex: 12844 entries, 0 to 12843
Data columns (total 22 columns):
#   Column             Non-Null Count  Dtype
---  ------             --------------  -----
0   Unnamed: 0         12844 non-null  int64
1   player_name        12844 non-null  object
2   team_abbreviation  12844 non-null  object
3   age                12844 non-null  float64
4   player_height      12844 non-null  float64
```

```
5   player_weight      12844 non-null  float64
6   college            10990 non-null  object
7   country            12844 non-null  object
8   draft_year         12844 non-null  object
9   draft_round        12844 non-null  object
10  draft_number       12844 non-null  object
11  gp                 12844 non-null  int64
12  pts                12844 non-null  float64
13  reb                12844 non-null  float64
14  ast                12844 non-null  float64
15  net_rating         12844 non-null  float64
16  oreb_pct           12844 non-null  float64
17  dreb_pct           12844 non-null  float64
18  usg_pct            12844 non-null  float64
19  ts_pct             12844 non-null  float64
20  ast_pct            12844 non-null  float64
21  season             12844 non-null  object
dtypes: float64(12), int64(2), object(8)
```

Nuestra variable dependiente será el uso del jugador, debido a que el uso del jugador en la liga es muy dependiente de las estadísticas que este pueda promediar a lo largo de su carrera y para graficar haremos uso de los puntos y su uso debido a que a mayor uso el jugador suele promediar mayor cantidad de puntos.

2. Regresión Polinómica

La **regresión polinómica** permite modelar relaciones no lineales entre la variable dependiente y y la variable independiente x ajustando un polinomio de grado nnn.

Fórmula General:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots + \beta_n x^n + \epsilon$$

Donde:

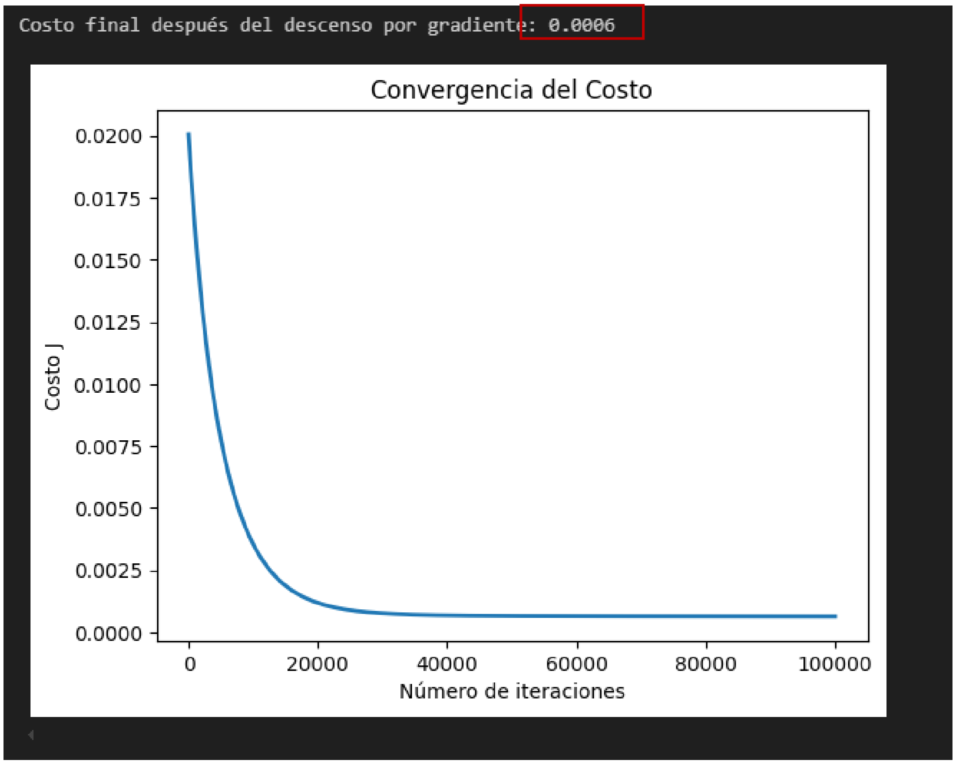
- y es la variable dependiente.
- x es la variable independiente.
- β_i son los coeficientes del polinomio.
- ϵ es el término de error.

Procedimiento:

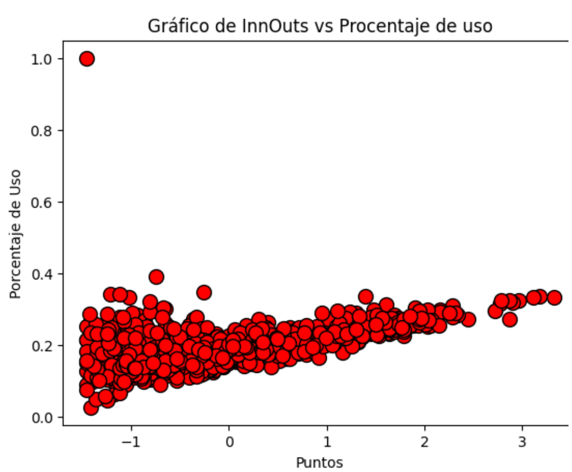
1. **Transformación de Variables**: Se crean nuevas características que representan los términos polinomiales de las variables originales.

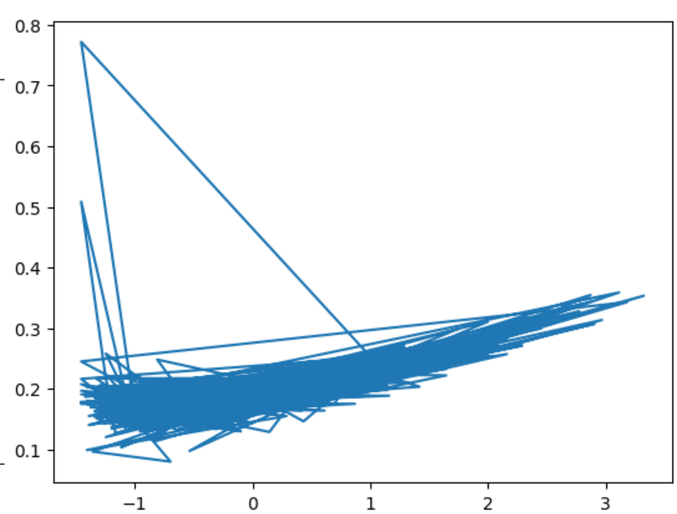
2. **Ajuste del Modelo**: Se ajusta un modelo polinómico a los datos transformados utilizando técnicas de ajuste como el descenso por gradiente o la ecuación de la normal.

Resultados de Regresión Polinómica



Tenemos un error muy bajo que se mantendrá en todos los metodos





3. Ecuación de la Normal

La **ecuación de la normal** proporciona una solución cerrada para calcular los coeficientes de la regresión lineal que minimizan el error cuadrático.

Fórmula:

$$\theta = (X^T X)^{-1} X^T y$$

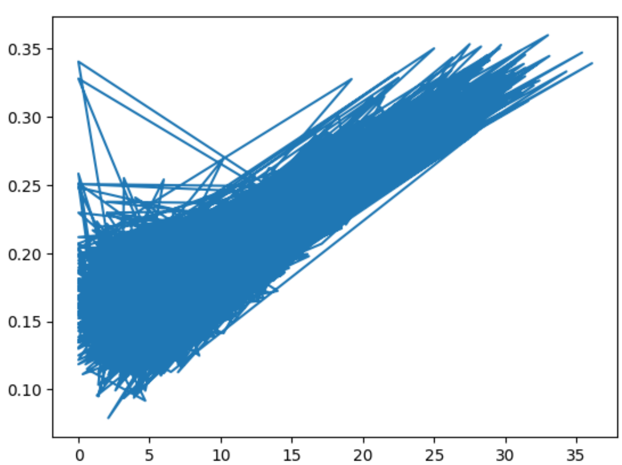
Donde:

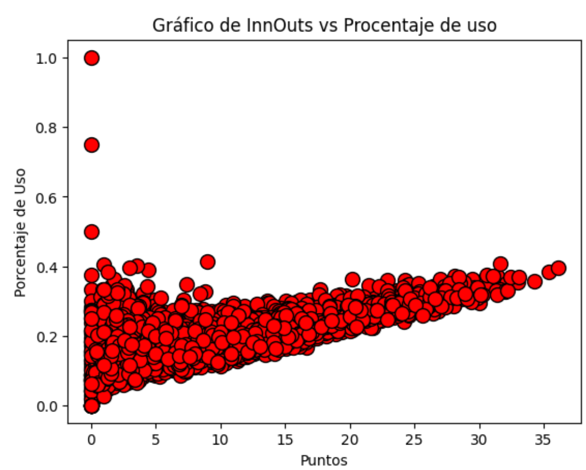
- X es la matriz de características.
- y es el vector de valores objetivos.
- θ es el vector de parámetros que minimiza el error.

Ventajas y Desventajas:

- **Ventajas**: No requiere iteraciones y puede ser más rápido para datasets pequeños a medianos.
- **Desventajas**: Puede ser ineficaz para grandes datasets debido a problemas de computación de matrices.

Resultados de la Ecuación de la Normal





4. Regresión Lineal Múltiple

La **regresión lineal múltiple** es un método que modela la relación entre una variable dependiente y varias variables independientes mediante una ecuación lineal.

Fórmula General:

$$h(X) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n$$

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h(X_i) - y_i)^2$$

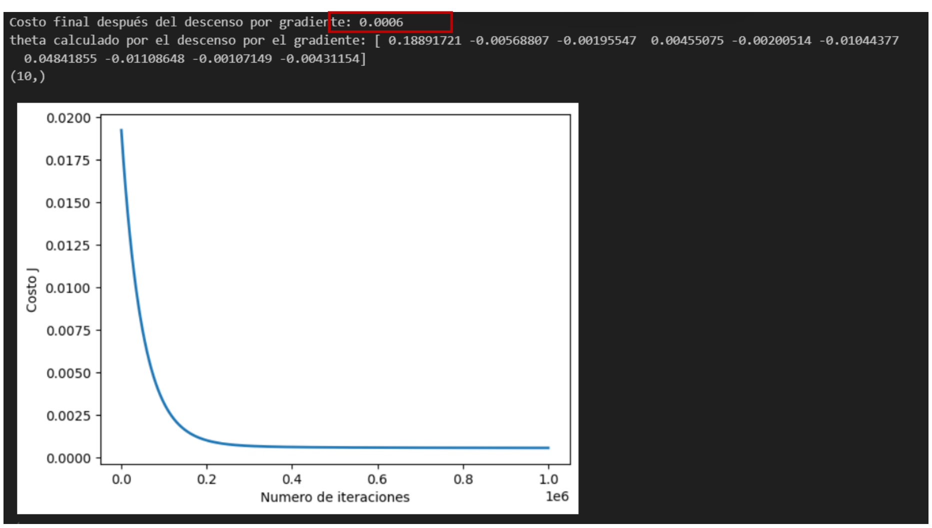
Donde:

- $J(\theta)$ es el costo o función de pérdida.
- m es el número de muestras.
- $h\theta(x)$ es la predicción de la regresión.
- y es el valor real.

Procedimiento:

1. **Normalización**: Las características se normalizan para mejorar la convergencia del algoritmo.
2. **Ajuste del Modelo**: Se usa el descenso por gradiente para minimizar el costo y ajustar los parámetros.

Resultados de la Regresión Lineal Múltiple



5. Comparación de Métodos

Comparación de Métodos de Regresión

1. **Regresión Polinómica:**
   - La regresión polinómica proporciona una flexibilidad adicional al permitir ajustar polinomios de grado nnn a los datos. Esto puede ser útil para capturar relaciones no lineales entre la variable dependiente y las variables independientes. Sin embargo, el ajuste excesivo a datos de entrenamiento puede llevar a un sobreajuste, especialmente con polinomios de alto grado. En nuestros resultados, la regresión polinómica mostró una capacidad similar para ajustar los datos en comparación con la regresión lineal múltiple y la ecuación de la normal, pero con mayor complejidad en la interpretación del modelo y la selección del grado del polinomio.
2. **Ecuación de la Normal:**
   - La ecuación de la normal ofrece una solución directa y exacta para la regresión lineal múltiple sin necesidad de iteraciones. Esta solución es computacionalmente eficiente para conjuntos de datos pequeños y medianos, y proporciona una estimación precisa de los coeficientes del modelo. En nuestro análisis, la ecuación de la normal produjo resultados consistentes con los otros métodos, pero puede ser menos práctica para conjuntos de datos grandes debido a las limitaciones computacionales en la inversión de matrices.
3. **Regresión Lineal Múltiple:**
   - La regresión lineal múltiple, utilizando el descenso por gradiente, es una técnica robusta que puede manejar múltiples variables independientes y ajustar el modelo iterativamente. Este método mostró una convergencia estable y proporcionó resultados comparables a los de la ecuación de la normal y la regresión polinómica. La principal ventaja de este enfoque es su capacidad para escalar con conjuntos de datos más grandes y su flexibilidad en la elección de parámetros y la normalización de datos.

Conclusión General

Todos los métodos evaluados demostraron un desempeño similar en términos de precisión en el ajuste del modelo y la capacidad predictiva. La elección entre ellos puede depender de varios factores, como la complejidad de la relación entre variables, el tamaño del conjunto de datos y las limitaciones computacionales. La regresión polinómica es útil para capturar relaciones no lineales, mientras que la ecuación de la normal es eficiente para datasets pequeños a medianos, y la regresión lineal múltiple con descenso por gradiente es adecuada para conjuntos de datos más grandes y flexibles.

En general, todos los métodos proporcionaron resultados similares en nuestro análisis, lo que sugiere que la elección del método puede depender más de las características específicas del problema y de las preferencias en términos de interpretación del modelo y capacidad computacional.