Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Estado de México

Imagen que contiene Texto

Descripción generada automáticamente

Inteligencia artificial avanzada para la ciencia de datos I (Gpo 101)

TC3006C.101

Profesores:

Jorge Adolfo Ramírez Uresti

**Momento de Retroalimentación: Módulo 2 Uso de framework o biblioteca de aprendizaje máquina para la implementación de una solución.**

Abner Maximiliano Lecona Nieves | A01753179

09 de Septiembre 2024

Contenido

[Documentación de Random Forest Classifier 3](#_Toc176951787)

[Justificación del dataset 3](#_Toc176951788)

[Descripción de las Variables 3](#_Toc176951789)

[Importancia del dataset 4](#_Toc176951790)

[Separación y evaluación del modelo con el dataset 4](#_Toc176951791)

[Análisis del modelo 6](#_Toc176951792)

[Diagnostico 6](#_Toc176951793)

[Sesgo (Bias) 6](#_Toc176951794)

[Varianza 7](#_Toc176951795)

[Nivel de Ajuste 7](#_Toc176951796)

# Documentación de Random Forest Classifier

En este documento explicaremos el funcionamiento del archivo “randomForest.py” que tiene como objetivo implementar un modelo de clasificación usando Random Forest para predecir una variable objetivo en un conjunto de datos.

# Justificación del dataset

El presente análisis utiliza un conjunto de datos que contiene estadísticas detalladas de jugadores de béisbol con el propósito de predecir si un jugador entra o no al Salón de la Fama, representado por la variable objetivo HOF.

## Descripción de las Variables

El dataset incluye 465 registros de jugadores con diversas métricas. Aunque no todas las variables han sido descritas en detalle, se basan en convenciones comunes de estadísticas de béisbol y se describen a continuación:

* PLAYER: Nombre del jugador.
* YRS: Número de años jugados en las ligas mayores.
* G: Número de juegos jugados (Games).
* AB: Número de veces al bate (At Bats).
* R: Número de carreras anotadas (Runs).
* H: Número de hits (veces que el jugador llegó a la base mediante un golpe válido).
* 2B: Número de dobles (hits de dos bases).
* 3B: Número de triples (hits de tres bases).
* HR: Número de home runs (jonrones).
* RBI: Número de carreras impulsadas (Runs Batted In).
* BB: Número de bases por bolas (Walks).
* SO: Número de ponches (Strikeouts).
* SB: Número de bases robadas (Stolen Bases).
* CS: Número de intentos de robo fallidos (Caught Stealing).
* BA: Promedio de bateo (Batting Average), calculado como H/AB.
* HOF: Indicador de ingreso al Salón de la Fama (1 si el jugador está en el Salón de la Fama, 0 si no).

## Importancia del dataset

El dataset es adecuado para el modelo de Random Forest debido a la relevancia y diversidad de las variables, el volumen de datos suficiente, y la naturaleza numérica de las características, lo cual es perfectamente compatible con los requerimientos del modelo y los objetivos del análisis.

# Separación y evaluación del modelo con el dataset

Esta imagen representa los primeros datos del dataset, se puede apreciar como están divididos y la forma en la que están acomodados. Siendo un shape de “(465, 16)” (465 filas, 16 columnas).

Pantalla de computadora con letras

Descripción generada automáticamente con confianza media

Utilizaremos una gráfica para mostrar la distribución de la variable objetivo (HOF). Como hay un balance adecuado entre las clases, el data ser parece ser adecuado para el entrenamiento del modelo

Gráfico, Gráfico de barras

Descripción generada automáticamente

Dividiremos el data set en “training” y en “validation”, 80% a 20% respectivamente.

* Tamaño del conjunto de entrenamiento: (348, 13)
* Tamaño del conjunto de prueba: (117, 13)

Gráfico, Histograma

Descripción generada automáticamente

# Análisis del modelo

Utilizaremos learning\_curve de scikit-learn para calcular las puntuaciones de precisión en función del tamaño del conjunto de entrenamiento.

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza baja

Imagen de la pantalla de un celular con letras

Descripción generada automáticamente con confianza media

Gráfico, Gráfico de líneas

Descripción generada automáticamente

## Diagnostico

### Sesgo (Bias)

El Sesgo es bajo ya que el modelo se ajusta muy bien a los datos del entrenamiento

### Varianza

La varianza es alta, ya que el desempenio del entrenamiento es mucho mejor que el de testing.

### Nivel de Ajuste

El modelo está sobreajustado a los datos de entrenamiento, dado que la diferencia entre la precisión en entrenamiento y validación es considerable. Esto se debe a que el modelo aprende los detalles y el ruido de los datos de entrenamiento, pero no puede replicar este desempeño en datos nuevos.

# Mejora del modelo

## Grid Search

Se implemento un Randomized Search como una técnica inicial de exploración de hiperparámetros para mi modelo de Random Forest. La búsqueda aleatoria me permitió explorar eficientemente un rango amplio de combinaciones de hiperparámetros sin el elevado costo computacional de probar todas las combinaciones posibles, como lo haría un Grid Search.

## Randomized Search

Tras la exploración inicial con Randomized Search, utilicé Grid Search para realizar un ajuste más fino de los hiperparámetros identificados previamente. Grid Search es una técnica sistemática que prueba todas las combinaciones posibles dentro de un grid de hiperparámetros definido. Para este proceso, configuré Grid Search con valores ajustados basados en los resultados de Randomized Search.

## Resultados del modelo

La combinación de Randomized Search y Grid Search resultó ser una estrategia efectiva para la optimización de los hiperparámetros del modelo de Random Forest. Randomized Search permitió identificar rápidamente un rango óptimo de hiperparámetros sin requerir un tiempo de cálculo extenso, mientras que Grid Search exploró en detalle estos rangos para encontrar la mejor configuración posible. El uso complementario de ambas técnicas no solo mejoró el rendimiento del modelo en términos de precisión tanto en entrenamiento como en prueba, sino que también ayudó a evitar el sobreajuste a los datos de entrenamiento, manteniendo un buen equilibrio entre sesgo y varianza.

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Una captura de pantalla de un celular

Descripción generada automáticamente con confianza media

Texto

Descripción generada automáticamente

Gráfico, Gráfico de líneas

Descripción generada automáticamente