Tarea N° 2
Imputación y Normalización

Nombre: Alejandra Verónica López Chiquito. **Profesor:** Dr. Marco Antonio Aceves
Fecha 2 oktober 2023 **Materia:** Machine Learning

Resumen

En el mundo de data science, no todos los conjunto de datos son perfectos. Es decir que se pueden presentar perdida de datos que pueden afectar al caso de estudio. De igual manera, se pueden ajustar algunos o todos los parámetros dentro de un rango para el fácil manejo de los datos.

En este documento se explicarán y aplicarán algunos métodos de imputación de datos, tratando de no afectar la distribución de los mismos. Además se aplicarán dos métodos de noramalizarán a 2 columnas del dataset con el objeivo de analizar el resultados de métodos diferentes.

Introducción

Es bastante común usar conjuntos de datos o bases de datos completos cuando se estudia ciencia de datos o machine learning, sin embargo, al enfrentar problemas reales los datos faltantes son más comunes de lo imaginado. Esto representa un gran problema, especificamente cuando se trata de realizar cualquier tiempo de predicción o entrenamiento con modelos de inteligencia artificial. Existen bastantes técnicas auxiliares para crear y completar estas bases de datos, a este proceso se le llama imputación de datos.

La imputación de datos es apliacada en la etapa de preprocesamiento de los datos, es decir, ayuda a preparar los datos de manera que sean lo más completos y útiles para el entrenamiento de algún modelo. Así como la imputación, existen otras técnicas de preprocesamiento de datos que son de gran ayuda para mejorar la calidad de los datos, una de ellas es la normalización. En esta técnica se considera un rango para

Objetivo

Con este documento se planea exponer y aplicar algunas técnicas de imputación y normalización para analizar qué técnicas son las más adecuadas para el conjunto de datos seleccionado.

En este caso se tienen 5 columnas, las cuales son ID, Age, Height, Weight, y Year. En este ejercicio se aplicarán técnicas de imputación y normalización a las columas Age, Height y Weight.

Grupo de estudio

El conjunto de datos seleccionado para este ejercicio es sobre atletas de diferentes ramas deportivas y años en los que compitieron. El data set tiene un total de 271,117 instancias y 15

atributos.

En este dataset se tienen 3 columnas de datos continuos en los que se pueden observar datos faltantes, marcados con un valor NaN. Estas columnas son Age, Height y Weight. Se tiene una columna más con datos faltantes, es la columna Medals, la cual tiene datos discretos, sin embargo no se tomará en cuenta para este ejercicio debido que el calculo del número de medallas para un atleta se podría considerar con un ejercicio de predicción u otro tipo de técnicas de machine learning.

Link de dataset: https://www.kaggle.com/code/chadalee/olympics-data-cleaning-exploration-prediction

Marco Teoríco

Existen bastantes técnicas auxiliares para crear y completar estas bases de datos, a este proceso se le llama imputación de datos.

Dentro de este proceso se usarán algunas ténicas o términos que es importante queden muy claros, ya que es el punto central del ejercicio. A continucación se redactan algunas definiciones:

- Definiciones

    1. Imputación: Proceso en el que, por medio de algúna técnica específica, se crearán datos sintéticos con el objetivo de evitar que un conjunto de datos tenga datos faltantes.
    2. Normalización: Esta técnica es utilizada para escalar datos de un atributo o varios dentro de un rango especifico.
    3. Datos sintéticos: Datos generados aritficialmente por medio de alguna técnica de imputación. Normalmente se crean por no tener suficiente datos reales.
    4. Imputación aleatoria: Se requiere encontrar el valor máximo y minimo de la columna. Una vez que se encuentran se aplican número aleatorios dentro de ese rango a los valores faltantes.
    5. Imputación por vecinos cercanos: Se requiere encontrar otro atirbuto relacionado y se infirere el valor faltante basado en el atributo que si se tiene. Para esto se puede utilizar la correlación de Pearson u otro método de relación.
    6. Imputación por media: Se debe calcular la media o promedio de los valores existentes de la columna y aplicar a los valores faltantes dentro de la misma.

Imputación

Como primer paso del proceso de prepocesamiento de nuestro conjunto de datos se obtienen los datos generales del mismo, como visualización de las columnas, promedio, desviasión estandar y más métricas de las columnas que tienen valores continuos.

1. **Columna Age** Para esta columna se observa uns distribución de datos concentrada entre 20 y 70 años, el índice inferior en términos de atletas podría sonar lógico, sin embargo el índice superior podría ser un poco dudosa, hay muchos deportes olimpico que no implican movimientos físicos. El atleta olimpico más viejo de la historia fue Oscar Swahn en la disciplina de tiro con arco, todo arriba de esta edad, se considerará como outlier.

   En la gráfica de frecuencia se puede ver que las edades más frecuentes para atletas van entre 15 y 25. Esta columna cuenta con 9,474 datos faltantes.

Para esta columna se ha seleccionado una técnica determinsta, como es el método de imputación usando la media de los valores, es decir calcular la media de la columna y aplicarla a los espacios con datos faltantes.

2. **Columna Height** La frecuencia de datos se encuentran entre 160 y 190. Esta columma cuenta con 60,171 datos faltantes. Para este conjunto de datos la primera técnica que se aplicará es una técnica estocástica, la imputación aleatoria.

   En esta columna será muy complicado encontrar outliers debido a que el peso de un atleta depende mucho de la disciplina a la que fue inscrito.

3. **Columna Weight**

   La distribución de los datos se encuentra entre 40 y 125, la frecuencia tiene un rango mayormente entre 50 y 90. Esta columna cuenta con 62875 valores faltantes, es la cantidad más grande dentro de este ejercicios. A esta columna se aplicará una técnica determinista que será por vecinos cercanos (KNN).

   El objetivo de esto es importar y utilizar la librería sklearn y sus funciones de imputación. Para esto se requiere utilizar algunas técnicas extra para mandar los valores de la columna a la librería. Ya que nuestra data se encuentra en formato dataframe de pandas, se requiere pasar a formato de numpy y cambiar su shape para mandarlo a la librería KNNImputer. Una vez que se aplica este método, se regresará a su shape original y se va a sustituir los nuevos valores de la columna en el dataset original. Este método tardó aproximadamente 6 minutos en aplicarse a la columna exitosamente.

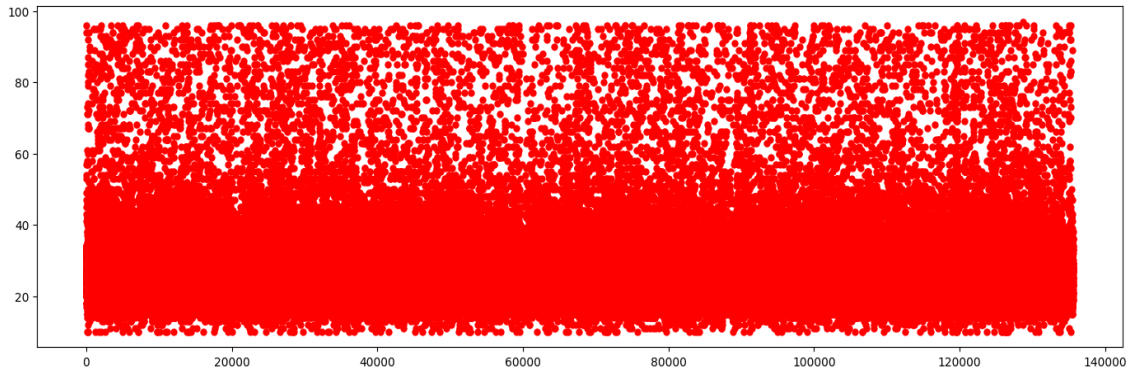
Normalización

En este documento se aplciarán 2 técnicas de normalización, las cuales son Min-Max y Z-score. Para la columna Height se aplicará el método Min-Max y para la columna Weight se usará el método Z-score sin el uso de librerías. Para la Min-Max se planea escalar los datos dento un rango de 10 a 1.

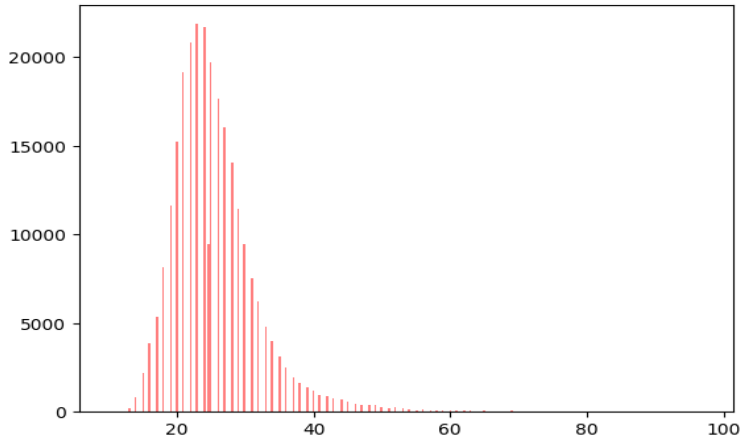
Conclusiones y Resultados

1. **Columna Age**

   Como primera opción se aplicó la técnica de la imputación aleatoria, el valor máximo es de 97 y minimo es del 10. Por lo que se colocaron 9474 valores aleatorios entre 10 y 97. Al finalizar se obtiene la gráfica que se muestra a continuación. Se puede apreciar que no es la técnica más adecuada ya que la distribución de los datos se muestra bastante afectada.

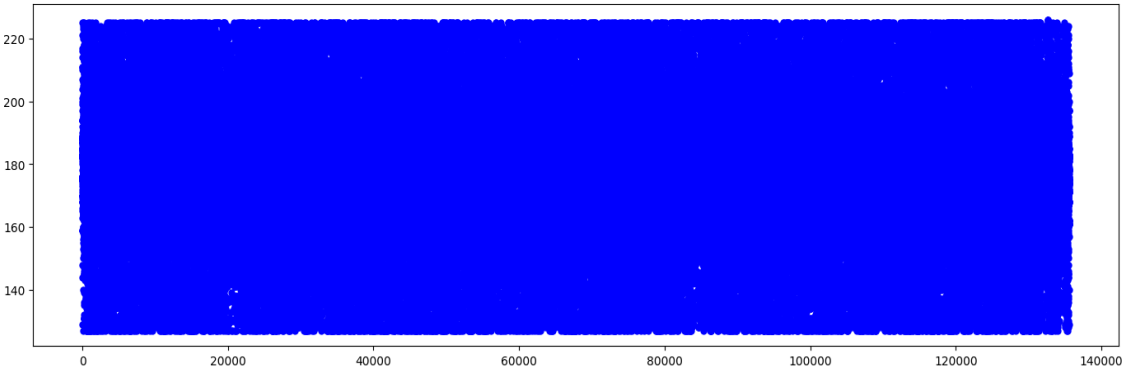


La técnica de imputación por la media aplicada a la columna Age se puede considerar adecuada debido a que la distribución de los datos conservó su forma, el cambió en la desviasión estandar es minimo, el valor inicial fue de 6.39, después de hacer la imputación cambió a 6.28. Si hay un cambio pero no es considerable. Sin embargo, si se puede aprecia una pequeña afección en la gráfica de frecuencia de datos, ahora el valor 25.66 aparece en la gráfica con un conteo de 9474 instancias, que era el número de datos faltantes, es un total del 3 por ciento de la columna. El resultado se puede ver en la gráfica siguiente.

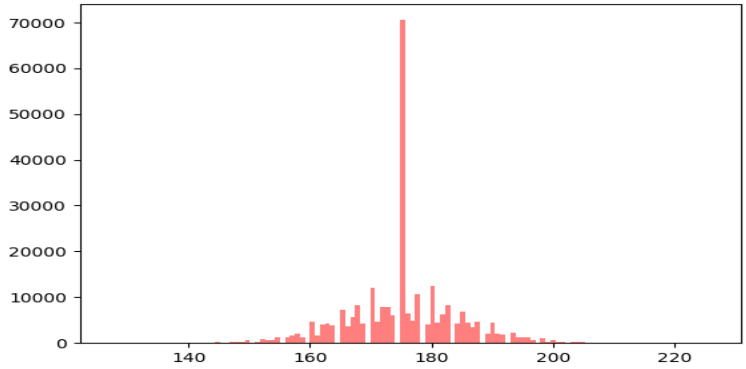


2. **Columna Height**

Igual que la primera columna, se aplicó una imputación estocástica de tipo aleatoria como primera opción, sin obtener los mejores resultados, esto se puede apreciar en la imagen siguiente. La distribución de datos se pierde por completo.

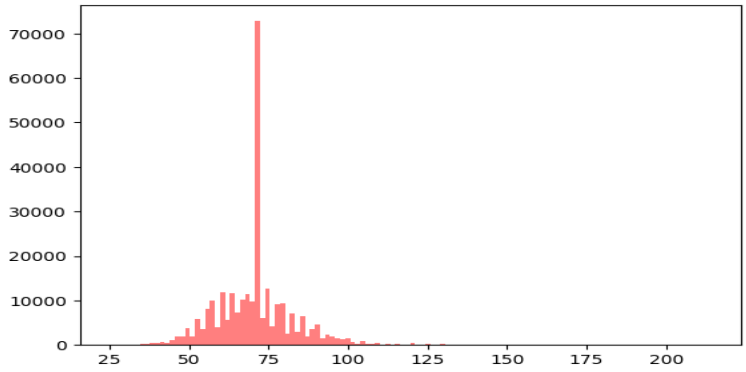


Por lo que se tomó la opción de aplicar otra técnica, imputación por la media. Al ser un gran número de datos faltantes (60171), es evidente que se tendrán modificaciones en distribución o frencuencia. Esta técnica afecta a la frecuencia, sin embargo se conserva mejor la distribución de los datos.



3. **Columna Weight**

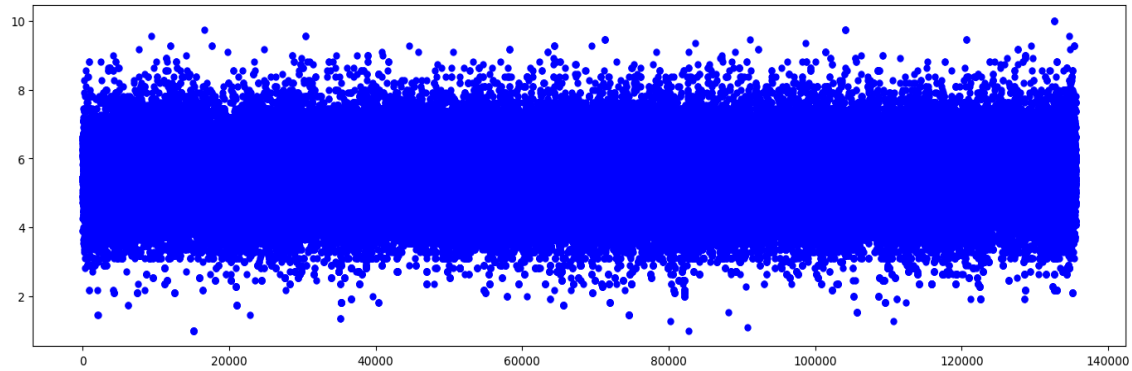
   Se mostrará en la gráfica de frecuencia que la librería calcula un número para aplicar a todos los datos faltantes, el método fit transform que ofrece el módulo de imputer en sklearn permité entrenar y aplicar el valor obtenido del entrenamiento en el mismo proceso. Como se puede ver en la gráfica siguiente, esto hace que se aplique el mismo valor a todos los datos faltantes dentro de la columna y la frecuencia cambie considereblemente debido a la cantidad de faltantes (62875)



   Con este método se cierra la etapa de imputación en este ejercicio y se puede ver que se tienen que probar varios métodos a las columnas para apreciar que método afecta menos a los datos originales, para esto se deben comparar la informacion estadística del dataset antes y después de aplicar cualquier técnica. Evidentemente se tendrán resultados diferentes a los originales pero el objetivo es reducir el margen de error en el preprocesamiento de los datos.

4. **Normalización**

   En la columna que se aplicó el método Min-Max (Height) se puede observar que si se aplica la fórmula correctamente, se van a escalar todos los datos dentro del rango elegido, en este caso de 10 a 1. La gráfica muestra la misma distribución de datos dentro del rango mencionado. Por lo que la normalización Min-Max fue aplicada exitosamente.



Para la columma Weight se aplicó el método de Z-score para escalar los datos. Se debe obtener desviación estandar y media de la columna para aplicar la fórmula. Una vez que se tienen estos datos se aplica dicha fórmula a cada dato faltante y en la imagen a continuación se puede ver el resultado. Los datos se escalaron de -4 a 12, la distribución de datos se puede apreciar de la misma manera que la original por lo que la normalización se considera correcta.