## Proceso:

## Objetivo: El objetivo de este trabajo grupal es realizar una Regresión y Aprendizaje no Supervisado con un dataset a elección

## Dataset:

## Link: <https://www.kaggle.com/datasets/zadafiyabhrami/athlete-performance-prediction-dataset>

## Descripción: Este dataset está diseñado para proporcionar un análisis profundo de los numerosos aspectos que afectan al rendimiento deportivo. Incluye información sobre la condición física del atleta, su programa de entrenamiento, su nutrición y su rendimiento previo. Científicos deportivos, entrenadores y analistas de datos pueden utilizar este conjunto de datos para pronosticar resultados de rendimiento, optimizar las estrategias de entrenamiento y mejorar el rendimiento deportivo general.

## Cantidad de ejemplos y de atributos:

El dataset cuenta con 21 columnas y 12601 ejemplos

|  |  |
| --- | --- |
| **ATRIBUTO** | **TIPO** |

|  |  |
| --- | --- |
| Athlete\_ID | Categorico, String |
| Athlete\_Name | Categorico, String |
| Sport\_Type | Categorico, String |
| Event | Categorico, String |
| Training\_Hours\_per\_Week | Numerical, float |
| Average\_Heart\_Rate | Numerical, float |
| BMI | Numerical, float |
| Sleep\_Hours\_per\_Night | Numerical, float |
| Daily\_Caloric\_Intake | Numerical, float |
| Hydration\_Level | Numerical, float |
| Injury\_History | Categorical, string |
| Previous\_Competition\_Performance | Numerical, float |
| Training\_Intensity | Categorical, string |
| Resting\_Heart\_Rate | Numerical, float |
| Body\_Fat\_Percentage | Numerical, float |
| VO2\_Max | Numerical, float |
| Event\_Distance | Numerical, float |
| Altitude\_Training | Categorical, string |
| Mental\_Focus\_Level | Numerical, float |
| Competition\_Date | Datetime, string |

|  |  |
| --- | --- |
| Performance\_Metric | Target, Numerical, float |

## Tareas:

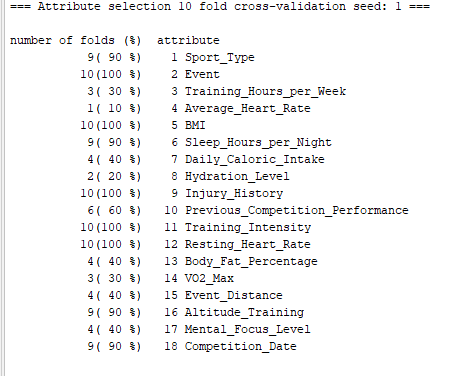
## Limpieza: hay alta presencia de datos faltantes o nulos en el dataset, para reducirlos se utilizó un script de python para hacer esta tarea más sencilla, lo que se realizó fue limpiar aquellas filas las cuales tenían al menos 3 valores faltantes, las columnas eliminadas fueron: ['Athlete\_ID', 'Athlete\_Name', 'Average\_Heart\_Rate', 'BMI', 'Resting\_Heart\_Rate', 'Body\_Fat\_Percentage', 'VO2\_Max'].

Para la ultima corrida del modelo se eliminarion las siguientes columnas [training hours per week, avera hearth rate, daily calopric intake, hydration level, body fat percentage, vo2 max y event distance]

## dejando al dataset con un total de 14 columnas restantes y 11199 datos

## Selección: El dato a seleccionar es Performance\_Metric, en la tarea de grecia se utilizara para encontrar algo <agregar info>

Se utiliza “select attribute” de weka para conocer cuales atributos tienes mas pesos que otros



Se eliminaron las columnas que tengan un 40% o menor

## Integración: No se integraron datos ya que el dataset proporcionó todo lo que se necesitaba para este proceso de minería de datos

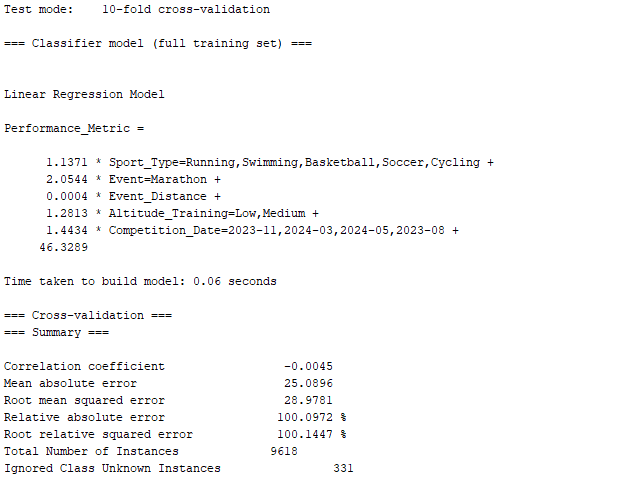
## Transformación: Competition\_Date se modificó para que en vez de tener una fecha completa solo tenga año y mes, luego es una ultima corrida del modelo se volvio a modificar competition date para que esta solo tenga año de competencia

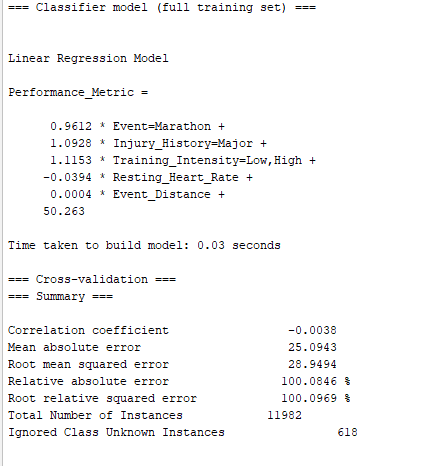
## Minería:

Se aplicaron los siguientes algoritmos:

Se utilizaron LinearRegression y RegressionByDiscretization utilizando la visualización en weka para comprender los resultados

El primero se realizó con linar regression





Se volvió a pasos anteriores para corroborar si el problema de no encontrar una relación lineal fuera de la poca cantidad de atributos, sin embargo, no se encontraron mejorías.

## Evaluación:

Se utilizo cross validation k folds de 10, tomando en relación el error absoluto y los gráficos visualizados en weka, analizando los resultados se comprueba que no hay una relacion lineal entre las horas entrenadas por semana y trainin intensity y la performance metric la cual es el atributo a elegir

## Conclusiones:

No se logró encontrar una relación lineal entre las variables, por ende este dataset no es adecuado para la tarea de regresión planteada