



# Proyecto Machine Learning

Equipo NN



NN

BD CONSULTING

## INDICE

INTRODUCCION .....	2
DATOS/ATRIBUTOS DE DATASET .....	2
DATASET ORIGINAL.....	2
LISTADO DE ATRIBUTOS .....	2
ATRIBUTOS/VALORES DEPRICADOS Y/O NO DESEADOS .....	3
TRANSFORMACION .....	4
ACCIONES REALIZADAS .....	4
FUNCIONES APLICADAS .....	5
DATOS FINALES DATASET .....	6
DESCRIPCION COLUMNAS .....	6
REFERENCIA/TERMINOS PROPIOS DEL TEMA .....	7
MODELADO DE MACHINE LEARNING .....	8
AUTOMODELADO.....	8
SCREENSHOTS INICIALES .....	8
COMPARACION POR MÉTRICAS .....	9
GRÁFICOS DE PREDICCIÓN.....	10
MODELADO MANUAL.....	12
A GRAN ESCALA .....	12



BD CONSULTING

## INTRODUCCION

Para el proyecto de la capacitación de Machine Learning hemos seleccionado como tema de estudio el juego multiplataforma PUBG (Player's Unknown Battleground), en específico las puntuaciones y datos de los jugadores generados a raíz de las partidas registradas en el dataset utilizado.

El dataset en cuestión ha sido descargado de la página web Kaggle (enlace <https://www.kaggle.com/c/pubg-finish-placement-prediction/data>).

Enlace a carpeta de GoogleDrive de proyecto ([https://drive.google.com/open?id=1oV87FfRfa-4edbG8XOVLxRO8ntJkW\\_Mu](https://drive.google.com/open?id=1oV87FfRfa-4edbG8XOVLxRO8ntJkW_Mu)).

## DATOS/ATRIBUTOS DE DATASET

### DATASET ORIGINAL

#### LISTADO DE ATRIBUTOS

A continuación se lista los atributos del dataset sustraído del website Kaggle.

- Id
- groupId
- matchId
- assists
- boosts
- damageDealt
- DBNOs
- headshotKills
- heals
- killPlace
- killPoints
- kills
- killStreaks
- longestKill
- matchDuration
- matchType
- maxPlace
- numGroups
- rankPoints
- revives
- rideDistance



BD CONSULTING

- roadKills
- swimDistance
- teamKills
- vehicleDestroys
- walkDistance
- weaponsAcquired
- winPoints

#### ATRIBUTOS/VALORES DEPRICADOS Y/O NO DESEADOS

En el caso de la revisión de los atributos se pudo identificar atributos a no utilizar desde el inicio del contacto con el dataset. A continuación se listan estos atributos:

Nombre atributo	Tipo	Razon para ser removido
<b>longestKill</b>	Atributo (columna)	En la descripción del atributo se aclara que el mismo no es una métrica totalmente exacta.
<b>matchDuration</b>	Atributo (columna)	Se genera una columna posteriormente de duración de tiempo expresada en minutos, ya que originalmente este atributo esta definido en segundos.
<b>flarefpp</b>	Valor de atributo (matchType)	Es un tipo de partida no standard, únicamente eventual en base a eventos y/o temporadas. De tal manera no es de interés para nuestro estudio.
<b>crashfpp</b>	Valor de atributo (matchType)	Es un tipo de partida no standard, únicamente eventual en base a eventos y/o temporadas. De tal manera no es de interés para nuestro estudio.
<b>flaretpp</b>	Valor de atributo (matchType)	Es un tipo de partida no standard, únicamente eventual en base a eventos y/o temporadas. De tal manera no es de interés para nuestro estudio.
<b>crashtpp</b>	Valor de atributo (matchType)	Es un tipo de partida no standard, únicamente eventual en base a eventos y/o temporadas. De tal manera no es de interés para nuestro estudio.





BD CONSULTING

rankPoint		El proveedor indica que este atributo en un futuro cercano no será relevado.
-----------	--	--

## TRANSFORMACION

### ACCIONES REALIZADAS

En pro de contar con la información coherente, verificada y de interés se ha realizado una revisión de esta.

Cabe mencionar que el archive test\_V2.csv es manipulado y accedido a través del programa Microsoft Excel.

1. Se ejecuta "Text to Columns" de manera de eliminar las comas (",") de la data y además asignar cada atributo, así como sus valores a una columna única.
2. Se coloca un filtro en la fila que tiene las etiquetas de los atributos, es decir en la fila 1.
3. Se crea una columna llamada ViewType que corresponde a la perspectiva de la partida logrando discriminar si son del tipo fpp (first person perspective) o tpp (third person perspective).
  - a. Esto es seleccionado por cada jugador antes de solicitar la búsqueda de una partida de juego.
  - b. Originalmente la perspectiva venia mezclada con el modo de partida (solo, duo o squad)
4. Se crea una columna totalDistance para obtener sumatoria de las distintas columnas que miden la distancia recorrida.
  - a. Se encuentran recorridos por vehículo (lancha, moto, auto, etc), recorrido a pie y recorrido a nado
5. Se agrega una columna para el cálculo de bots (jugadores automáticos)
  - a. Ya que cada partida tiene una capacidad de 100 jugadores, la cual siempre deberá ser alcanzada. De lo contrario el servidor ingresara bots a la partida.
  - b. En este caso detectamos la anomalía de que 724 registros contaban con valores negativos de bot. Pudiendo identificar que en algún caso el servidor asigna bots a equipo con jugadores reales. Estos registros fueron eliminados del dataset.
6. En todos aquellos valores de ID, groupID y matchID donde se encontraba únicamente número se decidió eliminar los registros ya que lo regular es que sea un registro alfanumérico. Se eliminaron un total de 4998 filas (0.5% del total de dataset).
  - a. Esto se logra haciendo una columna auxiliar temporal (denominada SOY\_ALFANUMERICO?) la cual da como TRUE si ID, groupID y/o matchID en su valor es únicamente numérico. Luego de eso se usa el ordenamiento del filtro y se filtra únicamente por true, se seleccionan estas filas y se eliminan las filas.



7. Se identifico que en las columnas con valores numéricos en algunos valores estaban formateados como texto, denotando que estos registros estaban alineados sobre la izquierda y no al centro como el normal. Por lo cual se parsean todos los valores a valor numérico.
  - a. Esto se identifico inicialmente a nivel visual y en pruebas de RapidMiner las cuales daban un margen de error en automodel de mas de 20%, luego se realizo con algunos valores de muestra que cumplían esta condición la ejecución de la función ISNUMBER, dando false como resultado. Para finalmente decidir el procedimiento anteriormente mencionado.
8. Se agrega una columna denominada classPlace para clasificar la fila según la posición final del jugador en la partida. Los valores ingresados a esta columna son: CAT1(TOP1-10), CAT2(TOP11-30), CAT3(TOP31-100)

## FUNCIONES APLICADAS

En base a las anteriores acciones realizadas se describirá la función utilizada. Nótese que al referir columnas se hace el supuesto de manejo en base a tabla definida, en este caso no se realizo debido a que el archivo Excel contenía mas de 1 millón de registros.

- matchDurationmins
  - =[@matchDuration]/60
- viewType
  - IFERROR(IF(SEARCH("fpp";[@matchType];1)>0;"First Person";"Third Person")
    - La función SEARCH busca un carácter o cadena de caracteres dentro de un texto indicado por el usuario, en caso de que se encuentra dará un numero mayor a cero. De lo contrario la formula lanzara una excepción de error, por lo cual se encapsula con un IFERROR.
    - Se tienen dos clases de perspectivas por cada modo de juego, primera y tercera persona. Aun así, en el matchType solo se indica la de primera persona con fpp (first person perspective), el resto es de tercera persona.
- classPlace
  - IF([@maxPlace]>30;"CAT3";IF([@maxPlace]<=10;"CAT1";"CAT2"))
    - Se utiliza condicionales IF anidados buscando de peor a mejor.
- Bots
  - IFERROR(IF(SEARCH("duo";[@matchType];1)>0;100-([@numGroups]\*2);0;0)+IFERROR(IF(SEARCH("solo";[@matchType];1)>0;100-[@numGroups];0;0)+IFERROR(IF(SEARCH("squad";[@matchType];1)>0;100-([@numGroups]\*3);0;0))
    - Se utiliza un distintos IF sin anidar, básicamente cada IF analiza el tipo de partida y la cantidad de equipos, según eso también se decide cuantos jugadores por equipo deben ser. Dando como resultado por cada IF un número final, también se maneja con IFERROR, de manera que si un



BD CONSULTING

resultado da error se asigna cero asegurando no se modifique incorrectamente esta medida.

- Como referencia hemos tomado matchType duo -> 2 jugadores por equipo, solo -> 1 jugador por equipo, squad -> 3 jugadores por equipo. En cuanto a la modalidad de squad los equipos pueden ser de 3 jugadores o 4 jugadores, esto aplica para todos los equipos, aun así, como no se especifica tomamos el valor mínimo para squad.

- totalDistance
  - SUM([@rideDistance];[ @swimDistance];[ @walkDistance])
    - Se hace la sumatoria de las 3 distancias
- Parseo de columnas a numérico
  - NUMBERVALUE(columna)
    - Se realizo a las columnas assists, boosts, damageDealt, DBNOs, headshotKills, heals, killPlace, killPoints, kills, killStreaks, matchDurationMins, maxPlace, numGroups, Bots, revives, rideDistance, roadKills, swimDistance, teamKills, vehicleDestroys, walkDistance, TotalDistance, weaponsAcquired, winPoints
- soyAlfanumerico?
  - IF(OR(ISNUMBER([@id])=TRUE, ISNUMBER([@groupID])=TRUE, ISNUMBER([@matchID])=TRUE), "TRUE", "FALSE")

## DATOS FINALES DATASET

### DESCRIPCION COLUMNAS

Se adjunta la table de descripción de columnas.

Nombre atributo	Descripción
<b>Id</b>	ID de jugador
<b>groupId</b>	ID de equipo
<b>matchId</b>	ID de partida
<b>Assists</b>	Cantidad de asistencias
<b>Boosts</b>	Cantidad de boosts
<b>damageDealt</b>	Daño recibido (excluyendo el daño autoinfligido)
<b>DBNOs</b>	Cantidad de knock outs
<b>headshotKills</b>	Cantidad de asesinatos por disparo a la cabeza
<b>Heals</b>	Cantidad de curaciones



<b>killPlace</b>	Ranking interno de partida en base a cantidad de enemigos asesinados
<b>killPoints</b>	Cálculo de puntos basado en ranking externo de asesinatos
<b>Kills</b>	Cantidad de asesinatos
<b>killStreaks</b>	Mayor número de enemigos asesinados en un periodo de corto de tiempo
<b>matchDurationMins</b>	Duración de partida en minutos
<b>matchType</b>	Tipo de partida por modalidad de juego
<b>ViewType</b>	Perspectiva de juego
<b>maxPlace</b>	Ranking final de la partida por jugador
<b>ClassPlace</b>	Categoría de TOP según el maxPlace
<b>numGroups</b>	Cantidad de equipos en la partida
<b>Bots</b>	Cantidad de bots
<b>Revives</b>	Cantidad de reanimaciones
<b>rideDistance</b>	Distancia recorrida en vehículo
<b>roadKills</b>	Asesinatos generados al conducir
<b>swimDistance</b>	Distancia recorrida a nado
<b>teamKills</b>	Cantidad de asesinatos por fuego amigo (compañeros de equipo)
<b>vehicleDestroys</b>	Cantidad de vehículos destruidos
<b>walkDistance</b>	Cantidad de distancia recorrida a pie
<b>TotalDistance</b>	Total de distancia recorrida
<b>weaponsAcquired</b>	Cantidad de armas adquirida en partida
<b>winPoints</b>	Porcentaje de probabilidad de ganar del jugador (no específica como se genera)

#### REFERENCIA/TERMINOS PROPIOS DEL TEMA

- Asistencia: Cuando dos o más jugadores de un mismo equipo atacan y matan a otro jugador solo el último en dispararle cuenta como el que lo mato, pero los otros jugadores del equipo que le dispararon reciben como valor de Asistencia el daño que causaron.
- Boosts: Aplicación de objetos para obtención de stamina como inyección de adrenalina, bebida energizante, entre otros. Permitirá al jugador obtener un extra de velocidad por un instante corto en la partida, así como también en caso de estar gravemente herido recuperar la velocidad normal del jugador en el campo.



- Knock out: Cada vez que un jugador derriba a otro, esto se logra a través de la inflación de daño, produciendo que el jugador enemigo no pueda caminar mas y únicamente pueda arrastrarse en el piso.
- Curaciones: Cada vez que el jugador consume un ítem para restablecer su nivel de vida, entre ellos podemos mencionar píldoras.
- Reanimación: Cada vez que un jugador reanima a un compañero de equipo, es decir le asiste con “curaciones” para sacarlo del estado del knock out. Esto únicamente permitirá al compañero recobrar la movilidad a pie y un monto reducido de vida, su stamina se ve reducida de tal manera aún tiene efecto de “lentitud”.

## MODELADO DE MACHINE LEARNING

### AUTOMODELADO

Se inicia el uso de la aplicación RapidMiner a través del automodelado debido a que esta manera se podrá elegir el modelo mas optimo para aplicar al dataset.

En particular se han seleccionado 3 modelos a comparar Generalized Linear Model, Deep Learning y Gradient Boosted Tress. En referencia a otros modelos posibles el tiempo de ejecución era de gran magnitud lo cual no permitía obtener un resultado final. Se seleccionó la operación de predicción para el atributo maxPlace.

Debemos mencionar que nos vimos forzados a utilizar el dataset de Test, no así el de Train, debido a limitaciones de software y hardware relacionadas al tamaño del Train Dataset. Aun así el Test Dataset cuenta con mas 1 millón de registros.

### SCREENSHOTS INICIALES

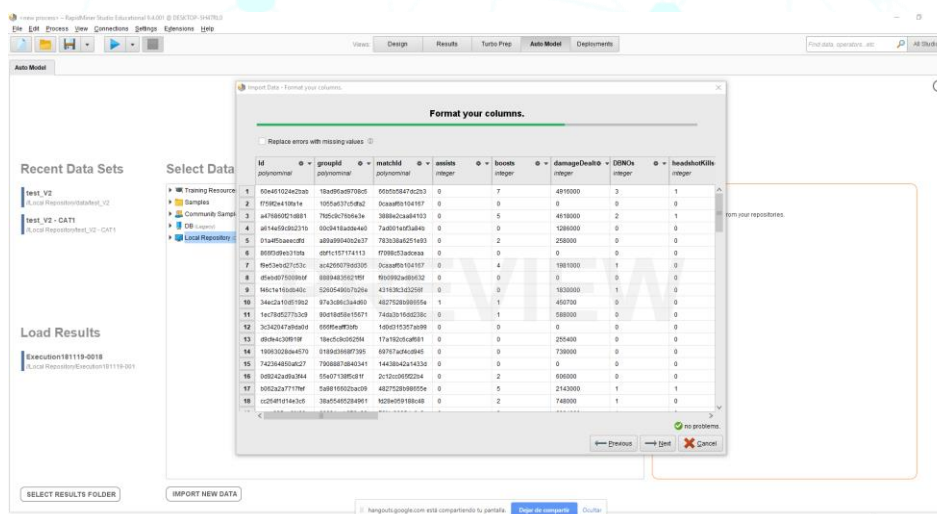


Ilustración 1 Importación de dataset al RapidMiner



NN  
BD CONSULTING

Results

Generalized Linear Model

Model

Simulator

Performance

Predictions

Predictions Chart

Production Model

Deep Learning

Gradient Boosted Trees

Model

Simulator

Performance

Optimal Parameters

Predictions

Predictions Chart

Production Model

General

Data

Statistics

Weights by Correlation

Correlations

SAVE RESULTS

Data

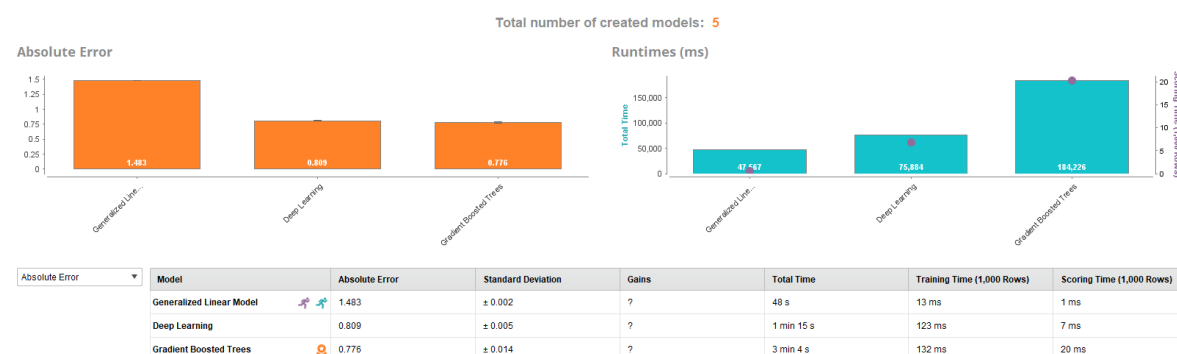
maxPlace Number	assists Number	boosts Number	Deaths Number	DamageDealt Number	DONs Number	headshotKills Number	heals Number	kills Assists	matchDuration...	matchType Category	revives Number	TotalDistance Number	weaponsAcq...
50	0	7	0	4916000	3	1	4	5	30.880	duo	1	131050000	3
50	0	0	0	0	0	0	0	0	31.480	duo	0	24711000	5
50	0	5	0	4618000	2	1	9	4	25.350	duo	1	11450000	4
50	0	0	0	1286000	0	0	1	1	31.280	duo	0	31308000	3
50	0	2	0	258000	0	0	0	0	24.080	duo	0	28768400	5
50	0	0	0	0	0	0	0	0	33.350	duo	0	29270000	5
50	0	4	0	1981000	1	0	3	1	31.480	duo	2	10040000	5
50	0	0	0	0	0	0	0	0	23.500	duo	0	124000	1
50	0	0	0	1830000	1	0	0	0	23.850	duo	0	1506000	3
50	1	1	0	450700	0	0	0	0	32.130	duo	0	76400000	3
50	0	1	0	588000	0	0	0	0	24.270	duo	0	13290000	3
50	0	0	0	0	0	0	0	0	24.420	duo	0	716800	1
50	0	0	0	255400	0	0	0	0	23.200	duo	0	2818000	1
50	0	0	0	738000	0	0	0	0	22.620	duo	0	12280000	3
50	0	0	0	0	0	0	0	0	32.900	duo	0	5680000	2
50	0	2	0	606000	0	0	4	1	31.870	duo	0	48730000	5
50	0	5	0	2143000	1	1	0	2	32.130	duo	0	70163900	4

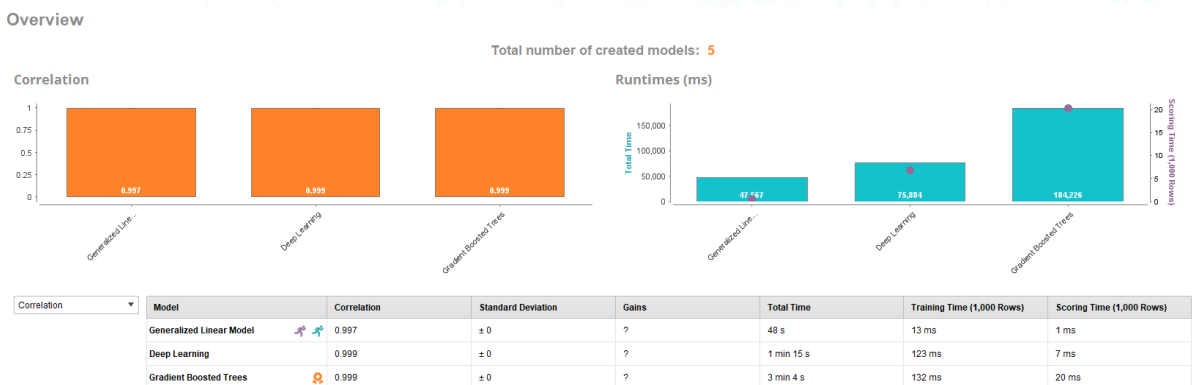
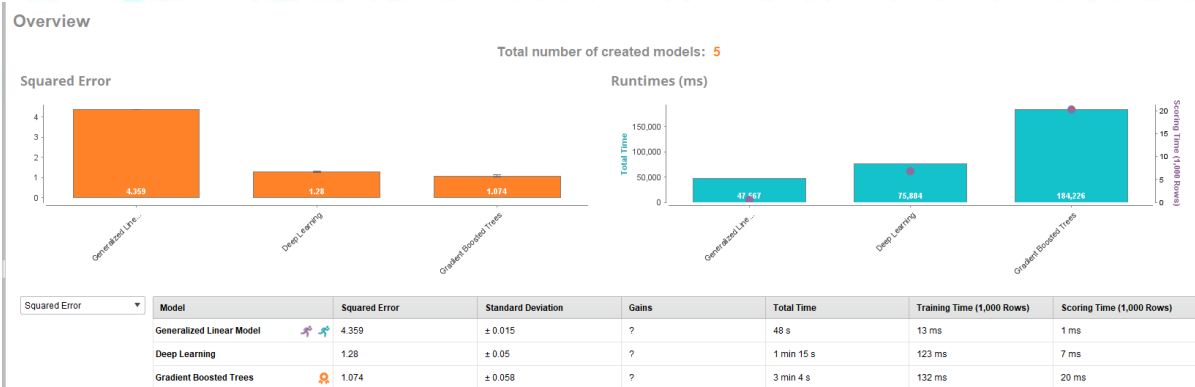
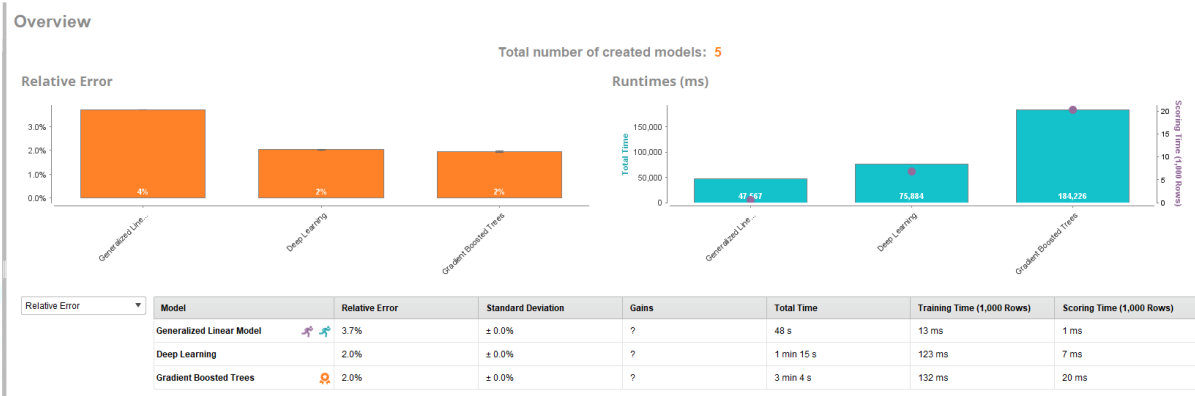
Ilustración 2 Atributos seleccionados

## COMPARACION POR MÉTRICAS

En base a esta comparación concluimos que el modelo Generalized Linear es el mas rápido en finalizarse de estas 3 elecciones, el de mejor performance en general ha sido Gradient Boosted Trees. Aún así este ultimo denota un tiempo de ejecución total bastante mayor a los otros modelos.

### Overview





## GRÁFICOS DE PREDICCIÓN

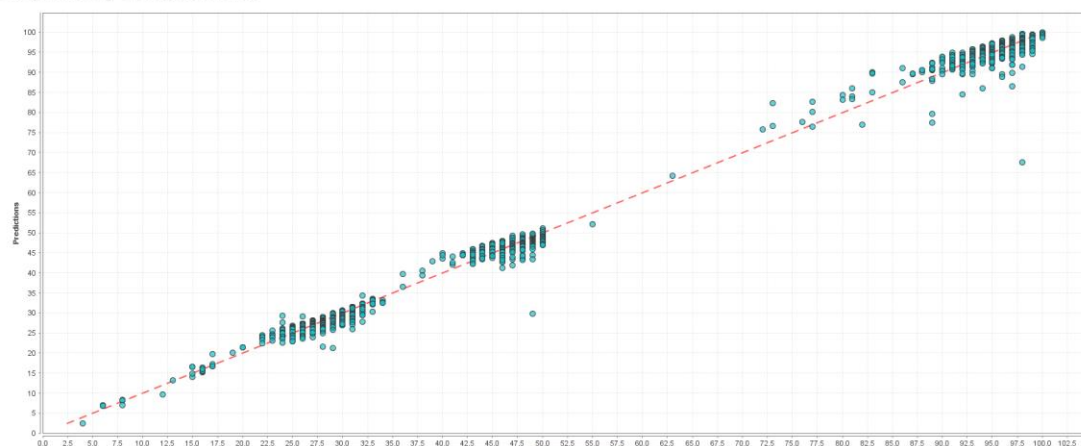
En base a los gráficos de predicción no resta duda que de los 3 modelos, cuales logran mejor predicción son Deep Learning y Gradient Boosted Tree. Aun así en Deep Learning parecen estar presenta anomalía o "outliers" sin contar un motivo específico de ello.



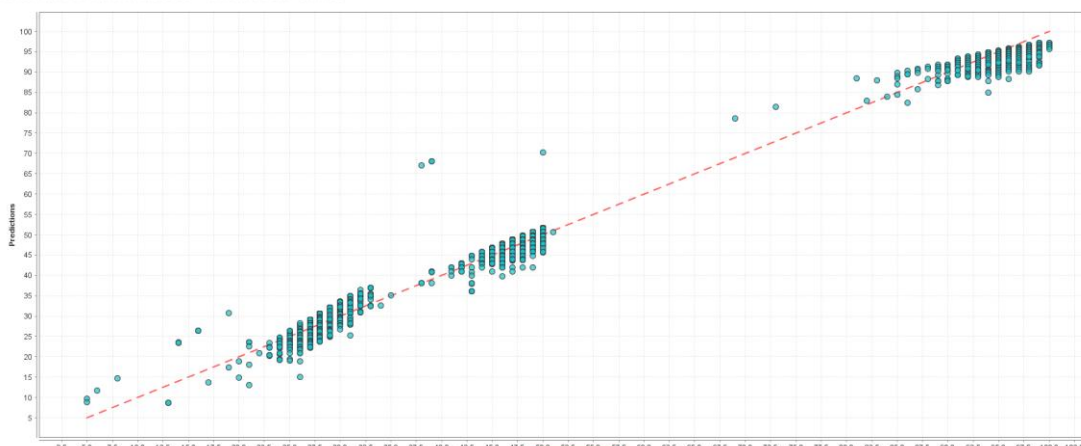
NN

BD CONSULTING

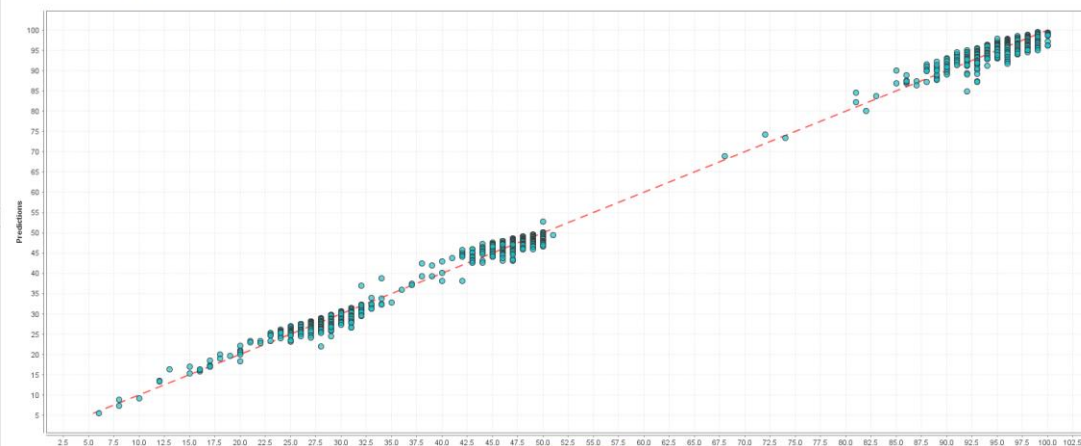
Deep Learning - Predictions Chart



Generalized Linear Model - Predictions Chart



Gradient Boosted Trees - Predictions Chart





BD CONSULTING

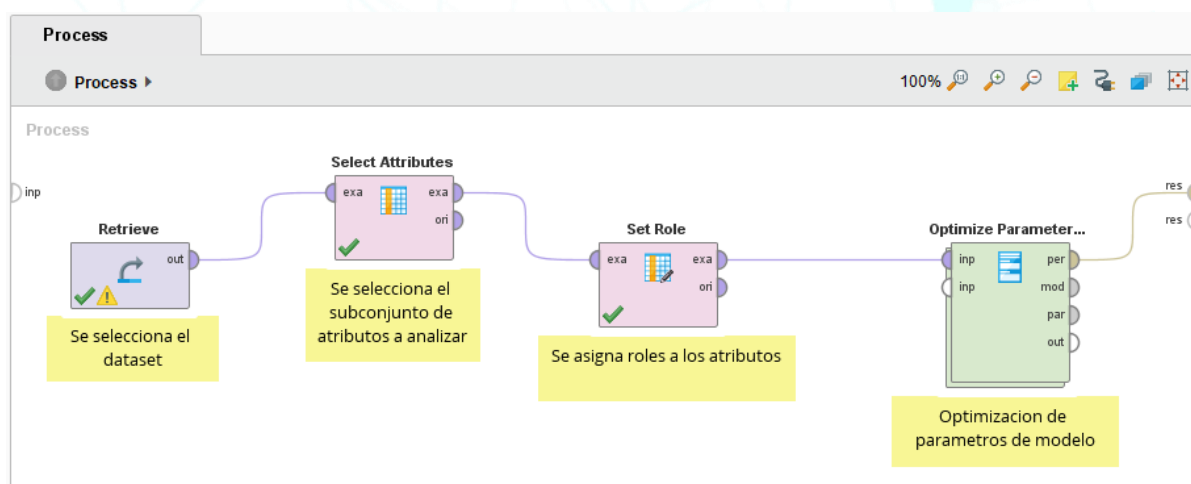
## MODELADO MANUAL

Se describe proceso manual para mejora del modelo más performante del automodelado, Gradient Boosted Trees, el mismo se realiza en rapidminer.

### A GRAN ESCALA

Se incorporan los elementos

- Retrieve
  - Se obtiene el dataset a trabajar
- Select attributes
  - Se obtiene un subconjunto de datos siguiendo los seleccionados durante el automodel
- Set Role
  - Se asignan roles a los atributos
- Optimize Parameters Grid
  - Se busca que el modelo Gradient Boosted Trees cuente con la mejor performance








BD CONSULTING

## SELECT ATTRIBUTES


**Parameters** ✕

 **Select Attributes**

attribute filter type subset ▼
















attributes Select Attributes...

☐ invert selection

 **Select Attributes: attributes**  
The attribute which should be chosen.
















**Attributes**

Search ✕

-  ClassPlace
-  groupId
-  # killPlace
-  # killPoints
-  # killStreaks
-  matchId
-  # numGroups
-  # rideDistance
-  # roadKills
-  # swimDistance
-  # teamKills
-  # vehicleDestroys
-  ViewType
-  # walkDistance
-  # winPoints

**Selected Attributes**

Search + ✕

-  # assists
-  # boosts
-  # Bots
-  # damageDealt
-  # DBNOs
-  # headshotKills
-  # heals
-  Id
-  # kills
-  # matchDurationMins
-  matchType
-  # maxPlace
-  # revives
-  # TotalDistance
-  # weaponsAcquired

➔ ➜

✓ Apply ✕ Cancel



BD CONSULTING

## SET ROLES

Atributos seleccionados a asignar rol

- totalDistance
  - Se selecciona como label a la distancia total para entender como el recorrido en el mapa de un jugador se relaciona con su ranking general en la partida.
- Bots
  - Se entiende a priori que la cantidad de bots puede afectar positivamente el ranking de un jugador real, de tal manera se definió como peso. Dicho sea de paso es un factor externo al jugador, al menos de su control o voluntad ya que es realizado por el servidor.
- ID
  - El id de jugador se define como ID para identificar de manera única los jugadores.
- maxPlace
  - Es el valor que deseamos predecir del jugador en una partida, su ranking final.

**Parameters** ✕

**Set Role**

attribute name

TotalDistance

▼ ⓘ

target role

label

▼ ⓘ

set additional roles

Edit List (3)...

ⓘ

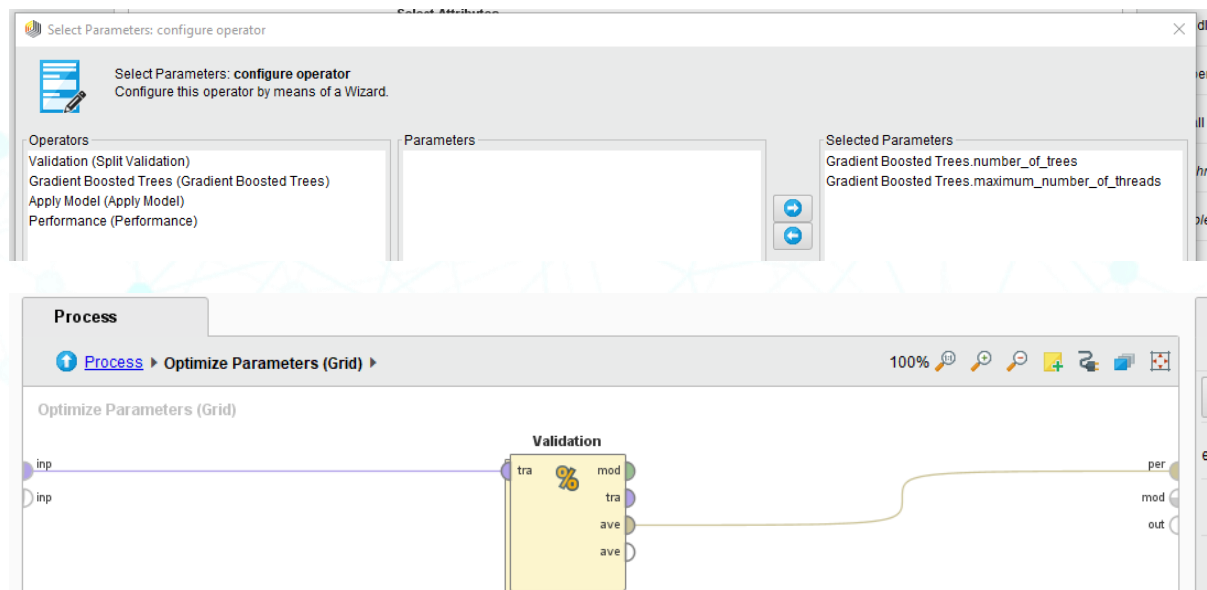
Edit Parameter List: set additional roles ✕

**Edit Parameter List: set additional roles**  
This parameter defines additional attribute role combinations.

attribute name	target role
maxPlace	prediction
Bots	weight
Id	id

## OPTIMIZER PERFORMANCE

Se solicita buscar el número de árboles e hilos más performante para el modelo conjunto a ello dentro del proceso se ejecuta un Split validation para dividir la data entre entrenamiento y test.



## SPLIT VALIDATION

