UNIDAD TEMÁTICA 3: Algoritmos Lineales

Trabajo de Aplicación 6 - Aplicación de Análisis Discriminante Lineal utilizando Rapid Miner, con predicción y análisis

ESCENARIO

El "Maestro", convencido de su capacidad para vislumbrar estrellas deportivas, ha puesto una academia para ayudar a jóvenes deportistas a lograr su mayor desempeño. En esta academia, el Maestro se enfoca particularmente en cuatro deportes: Fútbol, Basketball, Voleibol y Rugby.

Si bien ha visto que la mayoría de atletas jóvenes disfrutan practicando varios deportes, más adelante podrían preferir especializarse en uno en particular.

Datos

Al haber trabajado con atletas por muchos años, el Maestro ha ido recolectando un extenso conjunto de datos, y se pregunta ahora si sería posible sacar provecho a toda esa información para predecir el deporte más apropiado para los nuevos atletas. Él desearía poder recomendar a estos atletas el deporte en que tendrían mayor éxito si se especializaran en él.

Todos los atletas que han pasado por la academia fueron sometidos a una batería de tests, y también conoce cuál ha sido el deporte escogido por ellos.

Entre los datos disponibles, el dataset cuenta con:

- **Edad**: edad del atleta en años (con un decimal), al momento de efectuar las pruebas
- **Fuerza**: fortaleza del participante, medida en base a una serie de ejercicios de levantamiento de pesos, y expresada en una escala de 0 (fortaleza limitada) a 10 (realizó todos los levantamientos sin dificultad). Ningún participante obtuvo 8, 9 0 10 en esto, pero algunos obtuvieron 0.
- **Velocidad:** rendimiento del participante en un test de velocidad de respuesta. Se les contabilizó el tiempo en que podían presionar botones cuando se iluminaban o cuánto tardan en saltar al sonar una bocina. Los tiempos de respuesta fueron tabulados en una escala de 0 a 6, en donde 6 indica una velocidad de respuesta extremadamente alta y 0 una muy baja. Las valoraciones de los diferentes participantes se distribuyeron en todo el rango.
- **Lesiones**: es una columna simple con valores Si / No (1 / 0) que indica si el atleta ha sufrido previamente alguna lesión que fuera suficientemente severa como para requerir cirugía u otra intervención médica importante. Las lesiones leves tratadas con hielo, estiramiento, descanso, etc., fueron registradas como 0. Las lesiones que tomaron más de tres semanas en curar, o que requirieron terapia o cirugía fueron indicadas como 1.
- **Visión**: se evaluó de dos formas: la primera fue el test habitual de escala de visión 20/20, y la segunda utilizando tecnología de seguimiento de los movimientos oculares para determinar qué tan bien los atletas podían seguir visualmente objetos. Esta prueba desafía al participante a identificar elementos que se mueven rápidamente a través de su

campo de visión, y a estimar la velocidad y dirección de esos objetos. Las evaluaciones se registraron en una escala de 0 a 4, en que el valor 4 corresponde a una visión e identificación de objetos móviles perfecta. Ningún participante obtuvo una calificación de 4; los valores variaron entre 0 y 3.

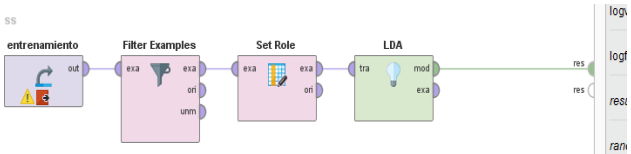
- **Resistencia**: Se aplicó a los participantes una batería de pruebas de aptitud física incluyendo carrera, ejercicio aeróbico y cardiovascular y natación. El rendimiento fue evaluado con una escala de 0 a 10, en donde el valor 10 representa la capacidad de realizar todas las tareas sin fatiga de ningún tipo. Los valores oscilaron entre 0 y 6 para los participantes. El Maestro nos ha reconocido que ni siquiera los atletas profesionales mejor entrenados serían capaces de obtener 10 en estas evaluaciones, ya que se han diseñado específicamente para probar los límites de la resistencia humana.
- **Agilidad**: calificación del participante en una serie de pruebas de su habilidad para moverse, girar, saltar, cambiar dirección, etc. Los participantes fueron evaluados con una escala de 0 a 100 en este atributo, y los valores variaron entre 13 y 80.
- **Capacidad de Decisión**: esta parte de la batería evalúa el proceso del atleta para decidir qué hacer en diferentes situaciones atléticas. Los atletas participaron en simulaciones que analizaron sus elecciones relativas o si pasar una pelota, moverse a una posición potencialmente ventajosa del campo de juego, etc. Las evaluaciones habrían de ser registradas en una escala de 0 a 100, pero el Maestro nos ha indicado que nadie que haya completado el test podría haber tenido una calificación menor que 3, ya que 3 puntos se otorgaban simplemente por participar. También nos ha indicado que tiene certeza que todos los atletas considerados participaron en este test, y sin embargo en los datos hay algunos valores menores que 3 y también algunos mayores que 100, así que sabemos que habrá que hacer un poco de preparación de los datos.
- **Deporte Primario:** este atributo indica en qué deporte se especializó cada uno de los atletas del dataset, luego de dejar la academia. Este es el atributo que el Maestro espera poder predecir para sus clientes actuales. Para los jóvenes en este estudio, este atributo será uno de 4 deportes: **Fútbol**, **Basketball**, **Voleibol** y **Rugby**.

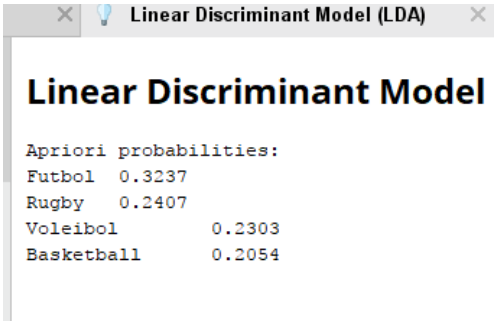
Ejercicio 1 - Preparación de Datos

1. Importa en RM el dataset de entrenamiento ("***sport-training.csv***").
   a) Agrega el dataset a un nuevo proceso, y renombra el operador "retrieve" del dataset a ""entrenamiento".
   b) Sabemos, de la descripción, que habrá que hacer algo de preparación, en particular en el atributo "**CapacidadDecision**". Analizar los valores incorrectos (< 3 o >100). En principio, se sugiere eliminar los ejemplos si son pocos…. (utilizar el operador "filter examples")
   c) Setea este atributo como "label" utilizando un operador "set role"
   d) Analiza las distribuciones, outliers y otras características de los demás predictores. Para cada uno, describe los hallazgos

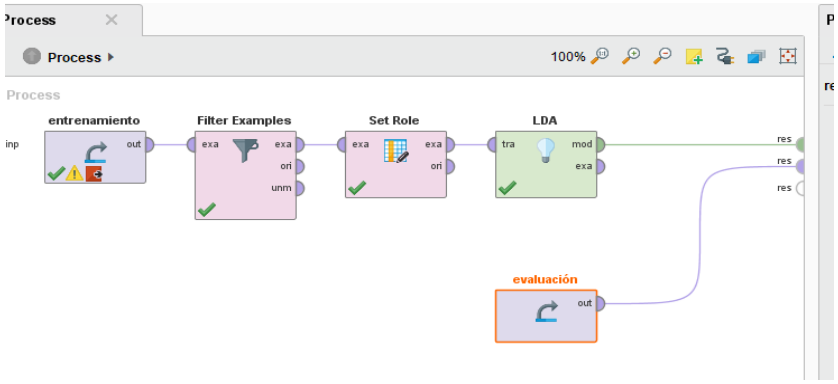
Ejercicio 2 - Modelado

1. Agregar un operador “LDA”, conectar la salida del “set role” a su entrada “tra” y su salida “mod” a un puerto de salida
2. Ejecuta el modelo: observa ahora las probabilidades “a priori” calculadas

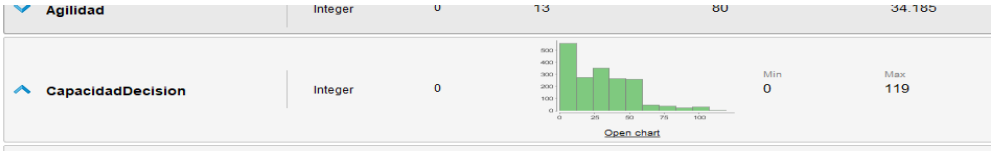


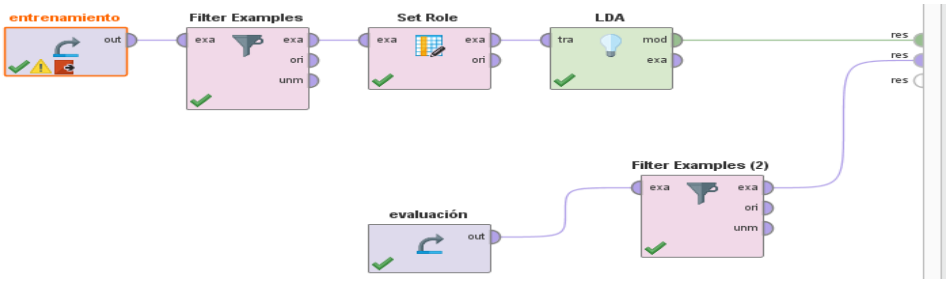


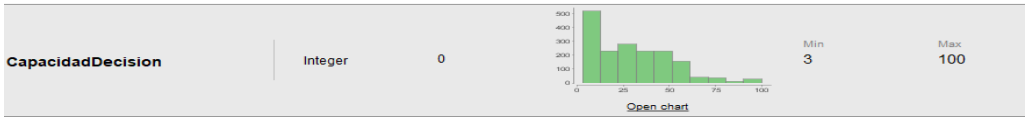
3. Estas probabilidades suman 1.
4. ¿Cómo se calculan? Demostrarlo para las 4 clases.
5. Estas probabilidades, junto con los valores para cada atributo, se utilizarán para predecir la clasificación de “DeportePrimario” para cada uno de los clientes actuales de la academia del Maestro, que están representados en el dataset ***“sport-scoring.csv”.***
6. Agrega el dataset ***“sport-scoring.csv”*** al modelo, y conéctalo directo a la salida.



7. Ejecuta el modelo nuevamente. Ahora RM mostrará un tab adicional con los datos del dataset “sport-scoring”.
    a. Verifica el tipo de datos de los atributos sea el correcto.
    b. Analiza los rangos de los mismos en comparación con los del dataset de entrenamiento
    c. Observa que también hay incongruencias en los valores del atributo “CapacidadDecision”. Resuelve este problema de la misma forma que se hizo para el dataset de entrenamiento.



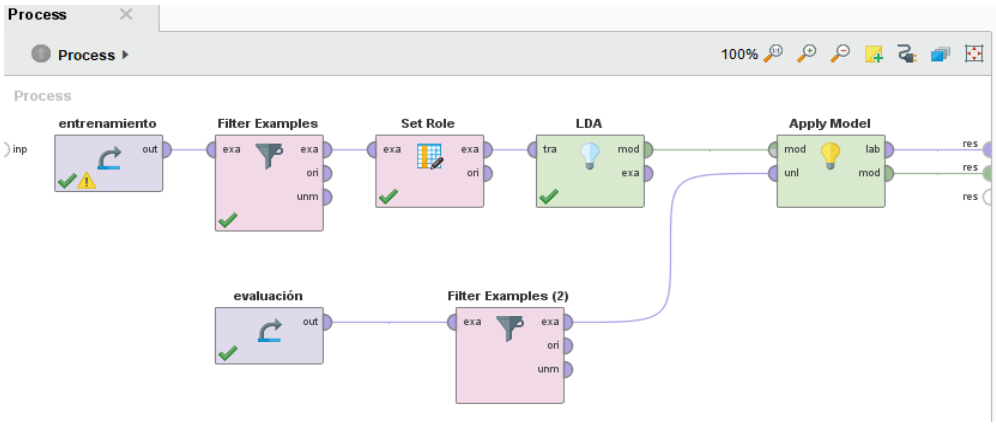




   d. De 1841 ejemplos, nos quedan 1767 después de los filtros.
   e. Completa el proceso de importación y renombrar el operador “retrieve” a “evaluación”

8. Corre el modelo y comparar los rangos de los atributos entre los datasets de entrenamiento y evaluación.
   a. ¿cómo son, comparativamente, estos rangos?
   b. ¿están todos los atributos de los ejemplos de evaluación / predicción en los rangos de los atributos del dataset de entrenamiento?
      i. ¿por qué tenemos que verificar esto?
   c. ¿hay más tareas de preparación previa de los datos para hacer?

9. **agrega** un operador “apply model”, y conectar a sus entradas:
   a. a “mod” conectar la salida del operador “LDA”
   b. a “unl” el dataset “scoring”
10. **verifica** que los puertos “lab” y “mod” estén conectados a las salidas “res” del proceso



Ejercicio 3 - Evaluación

11. **Ejecuta el proceso**, y en los resultados, observa que RM ha generado un nuevo atributo, “prediction(DeportePrimario”.
12. **Observa** las estadísticas correspondientes, en detalles y con gráficos
    a. ¿cómo son los valores de la predicción?
    b. ¿cómo son los valores de confianza en cada caso?

Ejercicio 4 – Deploy

El Maestro ahora cuenta con las predicciones para sus clientes actuales.

Podrá eventualmente tomar cada una de ellas y hablar con sus clientes.

Desde RM es posible extraer esta información de diferentes maneras. Si se trata de un conjunto relativamente pequeño, podríamos simplemente copiarlo en una planilla electrónica.

1. Crea una nueva planilla electrónica
2. En la vista de Data de los resultados de predicción, selecciona todos los ejemplos (“select all – ctrl+A”) y cópialos (“ctrl-C”)
3. Pégalos en la planilla. Puedes ahora manejar cómodamente los datos y trabajar con ellos.
    a. Ej: contar cuántas ocurrencias de cada deporte se han predicho
    b. Dado un identificador (en este caso sólo tenemos el número de tupla generado por RM), devolver todo el ejemplo, con la predicción y los valores de todos los atributos correspondientes, para así informar al cliente
4. Eventualmente sería conveniente generar un identificador único de ejemplo / cliente, ya en el dataset de entrada de scoring, para luego tener una referencia de identificación inmutable (recuerda que se han eliminado varios ejemplos del dataset)

Ejercicio 5 – Análisis de la performance del modelo

Utilizando las técnicas de validación que ya conoces (prueba con varias)

1. Separa el conjunto de entrenamiento en train / test
2. Ejecuta el modelo y registra los resultados
3. ¿qué consideraciones te merece?
4. Intenta mejorar los resultados, aplicando técnicas de preparación de los datos
5. ¿tienen los atributos las características más apropiadas para este algoritmo?