Capítulo 9.

**CONTEXTO**

¿Recuerdan a Sonia, la directora del programa de seguros de salud del Capítulo 6? Bueno, ella también ha vuelto para más ayuda. Su proyecto de agrupamiento de k-means fue tan útil para encontrar grupos de personas que podrían beneficiarse de sus programas, que quiere hacer más. Esta vez, su preocupación es ayudar a quienes han sufrido ataques cardíacos. Quiere ayudarlos a mejorar sus opciones de estilo de vida, incluido el control del peso y el estrés, para mejorar sus posibilidades de no sufrir un segundo ataque cardíaco. Sonia se pregunta si, con los datos de capacitación adecuados, podemos predecir las probabilidades de que los asegurados de su empresa sufran segundos ataques cardíacos. Siente que realmente podría ayudar a algunos de sus asegurados que han sufrido ataques cardíacos ofreciendo clases o grupos de apoyo para el control del peso, el colesterol y el estrés. Al reducir estos factores de riesgo clave de ataque cardíaco, los clientes de su empleador vivirán vidas más saludables y el riesgo de su empleador de tener que pagar los costos asociados con el tratamiento de segundos ataques cardíacos también disminuirá. Sonia cree que incluso podría educar a los asegurados sobre formas de ahorrar dinero en otros aspectos de sus vidas, como por ejemplo en las primas de su seguro de vida, al poder demostrar que ahora son titulares de pólizas con menor riesgo.

**OBJETIVOS DE APRENDIZAJE**

Después de completar la lectura y los ejercicios de este capítulo, podrá:

* Explicar qué es la regresión logística, cómo se utiliza y los beneficios de su uso.
* Reconocer el formato necesario para los datos a fin de realizar una regresión logística predictiva.
* Desarrollar un modelo de minería de datos de regresión logística en RapidMiner utilizando un conjunto de datos de entrenamiento.
* Interpretar los resultados del modelo y aplicarlos a un conjunto de datos de puntuación a fin de implementar el modelo.

**COMPRENSIÓN ORGANIZACIONAL**

El deseo de Sonia es ampliar sus actividades de minería de datos para determinar qué tipos de programas debería desarrollar para ayudar a las víctimas de ataques cardíacos a evitar que vuelvan a sufrirlos. Sabe que varios factores de riesgo como el peso, el colesterol alto y el estrés contribuyen a los ataques cardíacos, particularmente en aquellos que ya han sufrido uno. También sabe que el costo de proporcionar programas desarrollados para ayudar a mitigar estos riesgos es una fracción del costo de brindar atención médica a un paciente que ha sufrido múltiples ataques cardíacos. Conseguir que su empleador se comprometa a financiar los programas es la parte fácil. Averiguar qué pacientes se beneficiarán de qué programas es más complicado. Ella espera que le proporcionemos alguna orientación, basada en la minería de datos, para determinar qué pacientes son buenos candidatos para qué programas. La conclusión de Sonia es que quiere saber si es probable que algo (un segundo ataque cardíaco) suceda o no, y si es así, qué probabilidad hay de que suceda o no. La regresión logística es una herramienta excelente para predecir la probabilidad de que algo suceda o no.

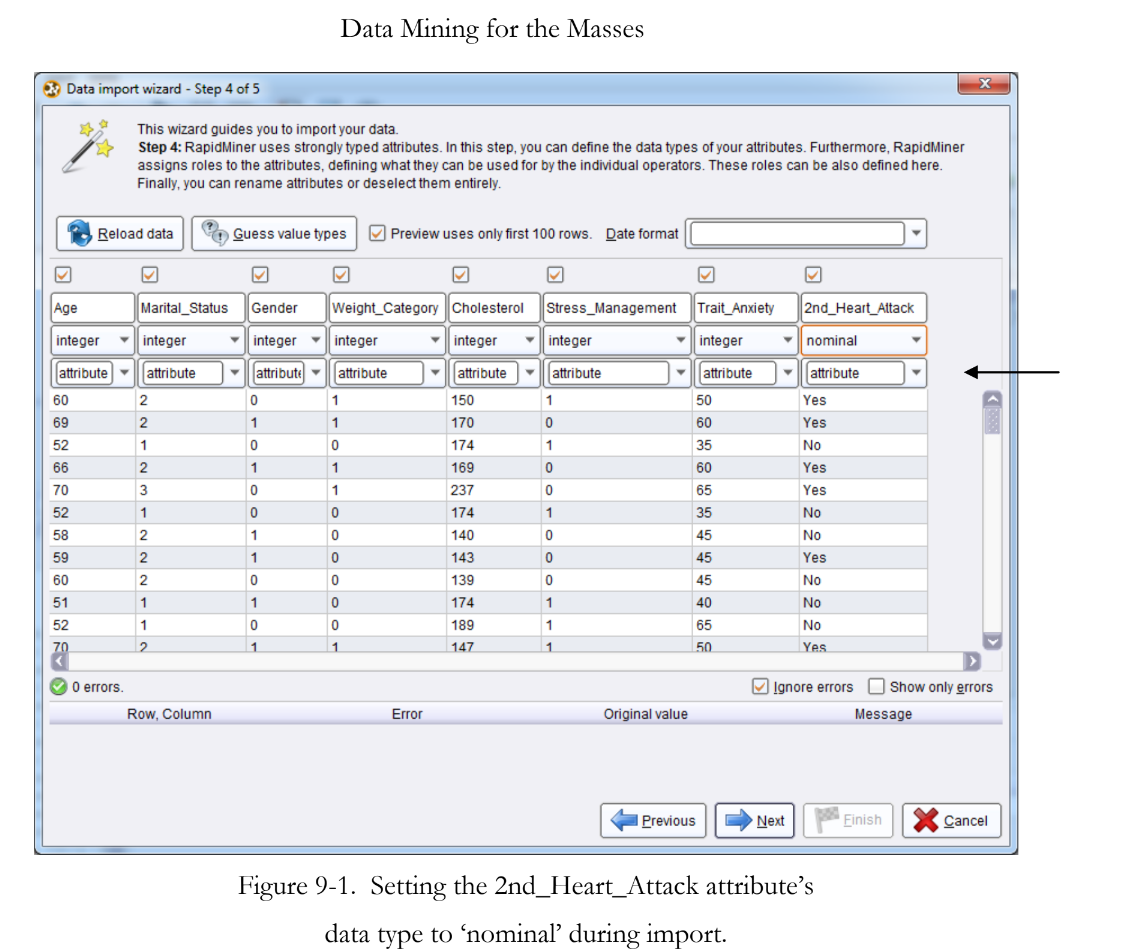
**COMPRENSIÓN DE DATOS**

* **Age:** La edad en años de la persona, redondeada al año entero más cercano.
* **Marital\_Status:** El estado civil actual de la persona, indicado por un número codificado: 0– Soltero, nunca casado; 1– Casado; 2– Divorciado; 3– Viudo.
* **Gender:** El sexo de la persona: 0 para mujer; 1 para hombre.
* **Weight\_Category:** El peso de la persona categorizado en uno de tres niveles: 0 para rango de peso normal; 1 para sobrepeso; y 2 ​​para obeso.
* **Cholesterol:** El nivel de colesterol de la persona, según lo registrado en el momento de su tratamiento para su infarto más reciente (su único infarto, en el caso de aquellos individuos en el conjunto de datos de puntuación).
* **Stress\_Management:** Atributo binario que indica si la persona ha asistido previamente o no a un curso de manejo del estrés: 0 para no; 1 para sí.
* **Trait\_Anxiety:** Puntuación en una escala de 0 a 100 que mide el nivel de estrés natural de cada persona y sus capacidades para afrontarlo. Poco tiempo después de que cada persona de cada uno de los dos conjuntos de datos se hubiera recuperado de su primer ataque cardíaco, se les administró una prueba estándar de ansiedad natural. Sus puntuaciones se tabulan y registran en este atributo en incrementos de cinco puntos. Una puntuación de 0 indicaría que la persona nunca siente ansiedad, presión o estrés en ninguna situación, mientras que una puntuación de 100 indicaría que la persona vive en un estado constante de agobio e incapacidad para afrontar sus circunstancias.
* **2nd\_Heart\_Attack:** Este atributo solo existe en el conjunto de datos de entrenamiento. Será nuestra etiqueta, el atributo de predicción o de destino. En el conjunto de datos de entrenamiento, el atributo se establece en "sí" para las personas que han sufrido un segundo ataque cardíaco y en "no" para las que no lo han sufrido.

**PREPARACIÓN DE DATOS**

Se han preparado dos conjuntos de datos que están disponibles para descargar desde el sitio web complementario. Se denominan Chapter09DataSet\_Training.csv y Chapter09DataSet\_Scoring.csv. Si desea seguir el ejemplo de este capítulo, descargue estos dos conjuntos de datos ahora y siga los pasos siguientes:

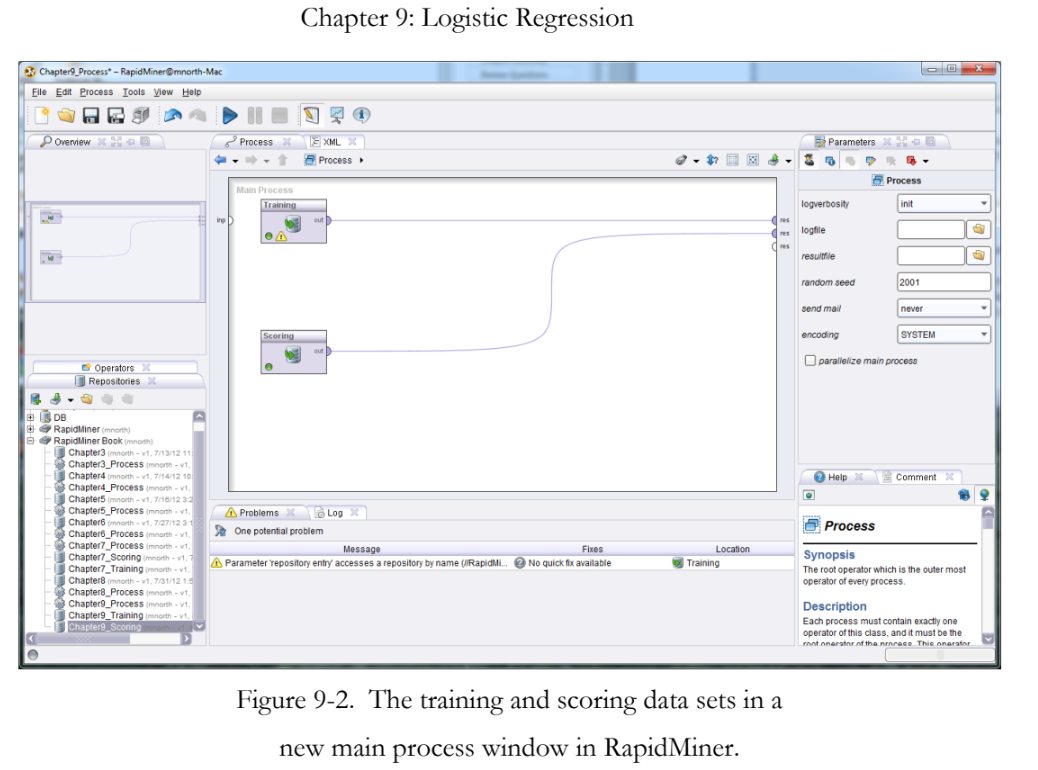
Comience el proceso de importación del conjunto de datos de entrenamiento primero. En su mayor parte, el proceso será el mismo que el que ha realizado en capítulos anteriores, pero para la regresión logística, hay algunas diferencias sutiles. Asegúrese de establecer la primera fila como los nombres de los atributos. En el cuarto paso, al establecer los tipos de datos y los roles de los atributos, deberá realizar al menos un cambio. Asegúrese de establecer el tipo de datos 2nd\_Heart\_Attack como "nominal", en lugar de binominal. Aunque es un campo sí/no, y RapidMiner lo establecerá como binominal por defecto debido a eso, el operador de regresión logística que usaremos en nuestra fase de modelado espera que la etiqueta sea nominal. RapidMiner no ofrece operadores binomiales a nominales o enteros a nominales, por lo que debemos asegurarnos de establecer este atributo de destino en el tipo de datos necesario de "nominal" cuando lo importamos. Esto se muestra en la Figura 9-1:



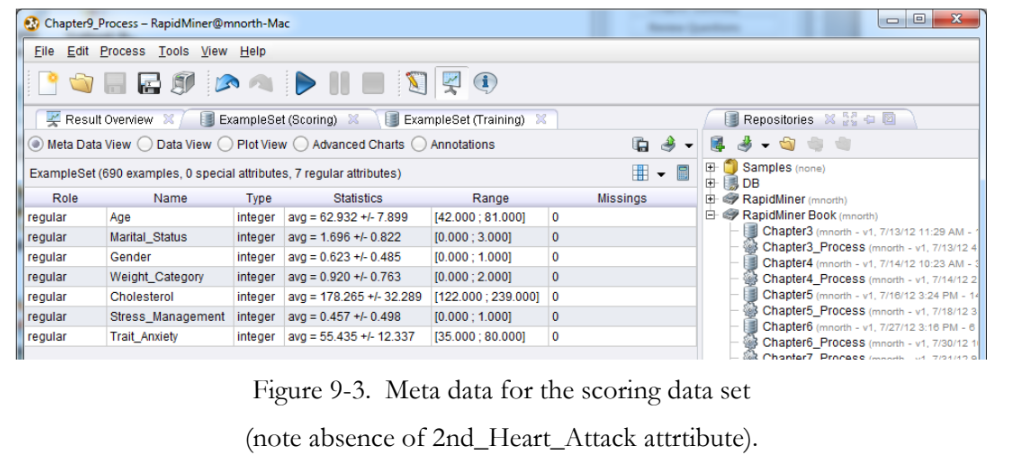
2) En este momento, también puede cambiar el rol del atributo 2nd\_Heart\_Attack a ‘etiqueta’, si lo desea. No lo hemos hecho en la Figura 9-1 y, posteriormente, agregaremos un operador de rol de conjunto a nuestra secuencia a medida que continuamos con la preparación de datos.

3) Complete el proceso de importación de datos para los datos de entrenamiento, luego arrastre y suelte el conjunto de datos en un nuevo proceso principal en blanco. Cambie el nombre del operador de recuperación del conjunto de datos a Entrenamiento.

4) Importe el conjunto de datos de puntuación ahora. Asegúrese de que el tipo de datos para todos los atributos sea ‘entero’. Este debería ser el valor predeterminado, pero puede que no lo sea, así que vuelva a verificarlo. Dado que el atributo 2nd\_Heart\_Attack no está incluido en el conjunto de datos de puntuación, no necesita preocuparse por cambiarlo como lo hizo en el paso 1. Complete el proceso de importación, arrastre y suelte el conjunto de datos de puntuación en su proceso principal y cambie el nombre del operador de recuperación de este conjunto de datos a Puntuación. Su modelo ahora debería verse similar a la Figura 9-2.



5) Ejecute el modelo y compare los rangos de todos los atributos entre las pestañas de resultados de entrenamiento y puntuación (Figuras 9-3 y 9-4, respectivamente). Debería encontrar que los rangos son los mismos. Como fue el caso con la regresión lineal, los valores de puntuación deben estar todos dentro de los límites inferior y superior establecidos por los valores correspondientes en el conjunto de datos de entrenamiento. Podemos ver en las Figuras 9-3 y 9-4 que este es el caso, por lo que nuestros datos están muy limpios, fueron preparados durante la extracción de la base de datos de origen de Sonia y no necesitaremos hacer más preparación de datos para filtrar observaciones con valores inconsistentes o modificar valores faltantes.



A screenshot of a data mining

Description automatically generated

6) Vuelva a la perspectiva de diseño y agregue un operador Establecer rol a su flujo de entrenamiento. Recuerde que si designó a 2nd\_Heart\_Attack para que tenga un rol de "etiqueta" durante el proceso de importación, no necesitará agregar un operador Establecer rol en este momento. No hicimos esto en el ejemplo del libro, por lo que necesitamos el operador para designar a 2nd\_Heart\_Attack como nuestra etiqueta, nuestro atributo de destino:

A screenshot of a computer

Description automatically generated

Con el atributo de etiqueta establecido, ahora estamos preparados para comenzar...

7) Utilizando el campo de búsqueda en la pestaña Operadores, localice el operador Regresión logística. Verá que si simplemente busca la palabra "logístico" (como se hizo en la Figura 9-6), hay varios operadores logísticos y de regresión logística disponibles en RapidMiner. Usaremos el primero en este ejemplo, sin embargo, lo alentamos a experimentar con los otros como desee. Arrastre el operador Regresión logística a su flujo de entrenamiento.

A computer screen shot of a computer program

Description automatically generated

8) El operador de regresión logística generará coeficientes para cada uno de nuestros atributos predictores, de manera muy similar a como lo hizo el operador de regresión lineal. Si desea verlos, puede ejecutar su modelo ahora. La fórmula algebraica para la regresión logística es diferente y un poco más complicada que la de la regresión lineal. Ya no estamos calculando la pendiente de una línea recta, sino que estamos tratando de determinar la probabilidad de que una observación caiga en un punto dado a lo largo de una línea imaginaria curva y menos definida a través de un conjunto de datos. Los coeficientes de la regresión logística se utilizan en esa fórmula.

9) Si ejecutó su modelo para ver sus coeficientes, regrese ahora a la perspectiva de diseño. Como lo hizo en los ejemplos de los capítulos más recientes, agregue un operador Aplicar modelo a su flujo para reunir los conjuntos de datos de entrenamiento y puntuación. Recuerde que es posible que deba desconectar y volver a conectar algunos puertos, como hicimos en el Capítulo 7 (paso 13), para fusionar sus dos flujos. Asegúrese de que los puertos de laboratorio y mod estén conectados a los informes.

A screenshot of a computer

Description automatically generated

Hemos terminado de construir el modelo. Ejecútalo ahora y procederemos a...

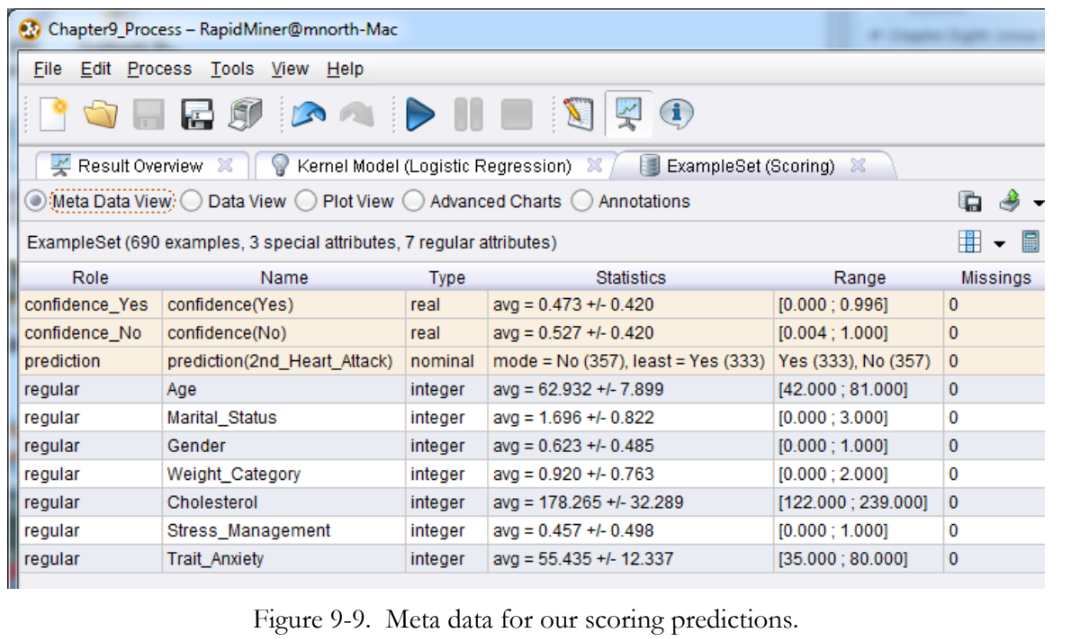
**EVALUACIÓN**

A screenshot of a computer

Description automatically generated

La pestaña inicial que se muestra en la perspectiva de resultados es una lista de nuestros coeficientes. Estos coeficientes se utilizan en el algoritmo de regresión logística para predecir si cada persona en nuestro conjunto de datos de puntuación sufrirá o no un segundo ataque cardíaco y, de ser así, qué tan seguros estamos de que la predicción se hará realidad.

Pase a la pestaña de resultados de la puntuación. Primero, veremos los metadatos (Figura 9-9).



En esta figura podemos ver que RapidMiner ha generado tres nuevos atributos para nosotros: confianza (Sí), confianza (No) y predicción(2do\_ataque\_cardíaco). En nuestra columna Estadísticas,

encontramos que, de las 690 personas representadas, estamos prediciendo que 357 no sufrirán segundos ataques cardíacos y que 333 sí. La esperanza de Sonia es poder involucrar a estas 333, y quizás a algunas de las 357 con niveles bajos de confianza en su predicción "No", en programas para mejorar su salud y, por lo tanto, sus posibilidades de evitar otro ataque cardíaco. Pasemos a la Vista de datos.

A screenshot of a computer

Description automatically generated

En la Figura 9-10, podemos ver que a cada persona se le ha dado una predicción de “No” (no sufrirán un segundo ataque cardíaco) o “Sí” (lo sufrirán). Es de vital importancia recordar en este punto de nuestra evaluación que, si esto fuera real, y no un ejemplo de libro de texto, estas serían personas reales, con nombres, familias y vidas. Sí, estamos usando datos para evaluar su salud, pero no deberíamos tratar a estas personas como números. Esperamos que nuestro trabajo y análisis ayuden a nuestra cliente imaginaria Sonia en sus esfuerzos por servir mejor a estas personas. Al extraer datos, siempre debemos tener en cuenta el elemento humano, y hablaremos más sobre esto en el Capítulo 14.

Entonces, tenemos estas predicciones de que algunas personas en nuestro conjunto de datos de puntuación están en camino de sufrir un segundo ataque cardíaco y otras no, pero ¿cuán seguros estamos de estas predicciones? Los atributos de confianza (Sí) y confianza (No) pueden ayudarnos a responder esa pregunta. Para comenzar, consideremos a la persona representada en la Fila 1. Se trata de un hombre soltero (nunca casado) de 61 años. Se lo ha clasificado como con sobrepeso, pero tiene un colesterol más bajo que el promedio (la media que se muestra en nuestros metadatos en la Figura 9-9 es un poco más de 178). Obtuvo una puntuación justo en el medio en nuestra prueba de ansiedad a los 50 años y ha asistido a clases de manejo del estrés. Con estos atributos personales, comparados con los de nuestros datos de entrenamiento, nuestro modelo nos ofrece un nivel de confianza del 86,1 % de que la predicción "No" es correcta. Esto nos deja con un 13,9 % de duda en nuestra predicción. Los valores “No” y “Sí” siempre sumarán 1, o, en otras palabras, 100 %. Para cada persona en el conjunto de datos, sus atributos se incorporan al modelo de regresión logística y se calcula una predicción con porcentajes de confianza.

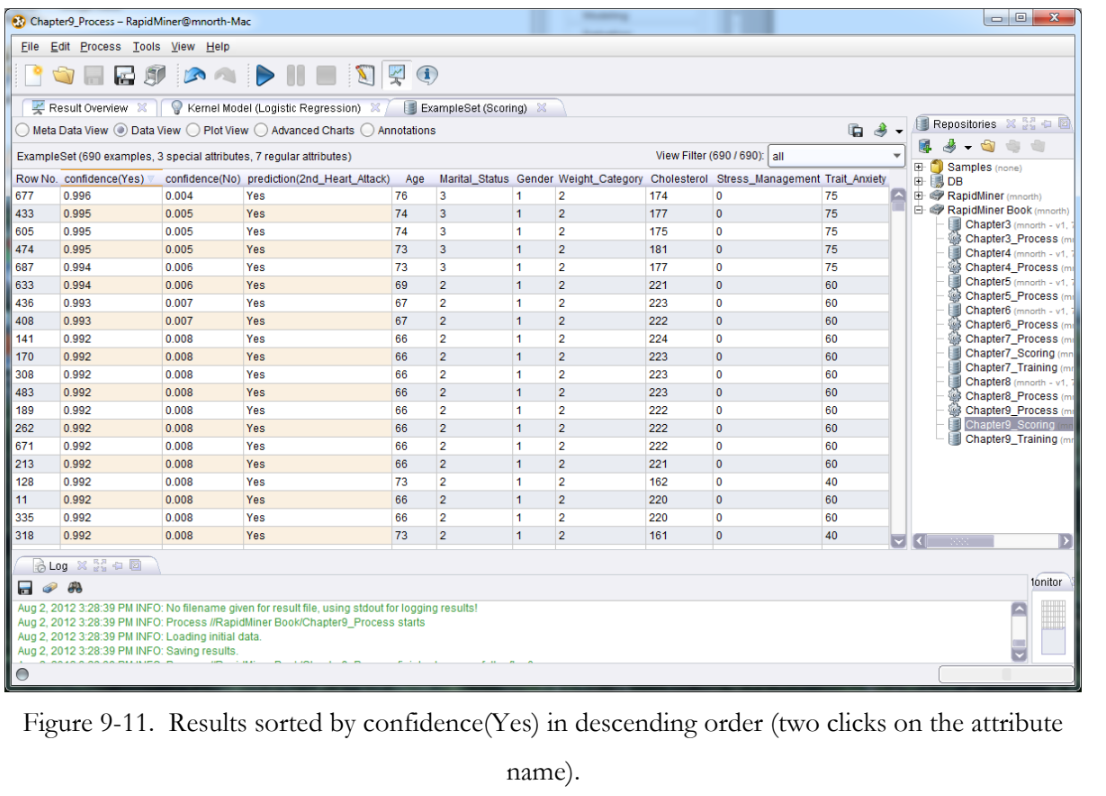
Consideremos a otra persona como ejemplo en la Figura 9-10. Observe la fila 11. Se trata de un hombre de 66 años que se ha divorciado. Está por encima de los valores promedio en todos los atributos. Si bien no es tan viejo como algunos de nuestros datos, está envejeciendo y es obeso. Su colesterol está entre los más altos de nuestro conjunto de datos, obtuvo una puntuación más alta que el promedio en la prueba de ansiedad y no ha asistido a una clase de manejo del estrés. Estamos prediciendo, con un 99,2 % de confianza, que este hombre sufrirá un segundo ataque cardíaco. Las señales de advertencia están todas ahí y Sonia ahora puede verlas con bastante facilidad.

Una vez que comprende cómo leer el resultado, Sonia puede proceder a...

**DESPLIEGUE**

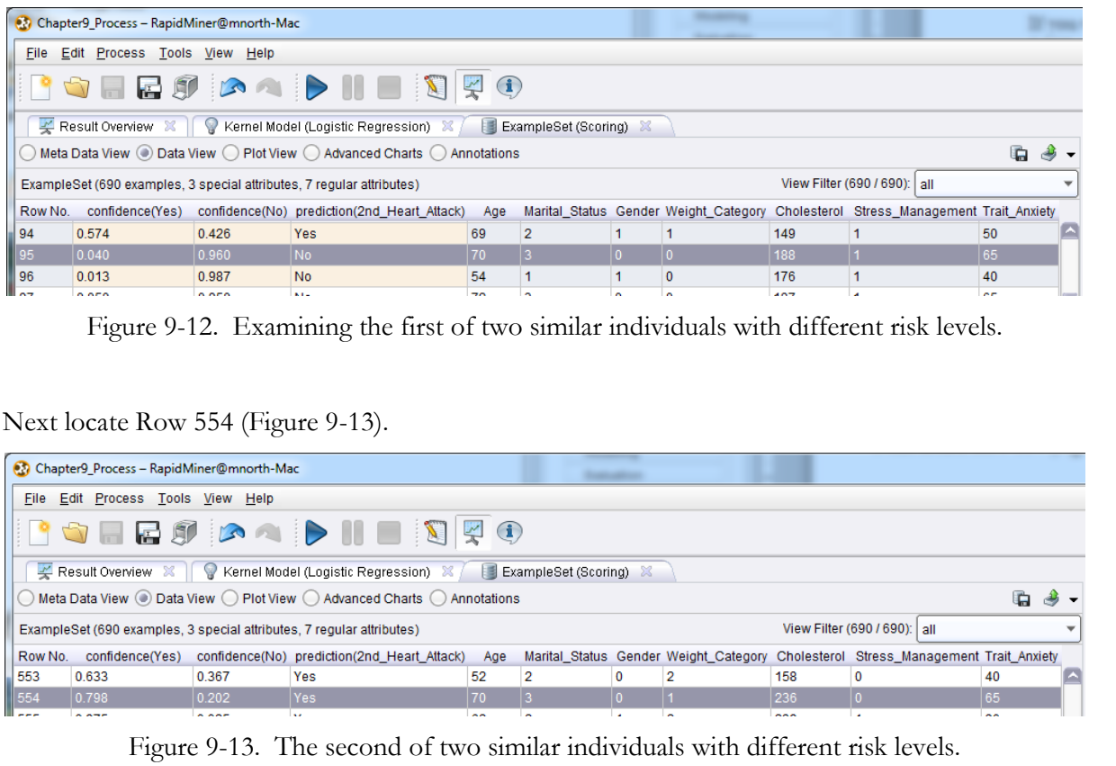
En el contexto de la persona representada en la fila 11, parece bastante obvio que Sonia debería intentar acercarse a este caballero de inmediato y ofrecerle ayuda en todos los aspectos. Es posible que desee ayudarlo a encontrar un grupo de apoyo para la pérdida de peso, como Comedores Compulsivos Anónimos, brindarle información sobre cómo lidiar con el divorcio y/o el estrés, y alentar a la persona a trabajar con su médico para regular mejor su colesterol a través de la dieta y quizás también con medicamentos. Puede haber una cantidad de las 690 personas que claramente necesitan ayuda específica. Haga clic dos veces en el nombre del atributo confianza(Sí). Al hacer clic en el encabezado de una columna (el nombre del atributo) en la perspectiva de resultados de RapidMiner, se ordenará el conjunto de datos por ese atributo. Haga clic una vez para ordenar en orden ascendente, dos veces para volver a ordenar en orden descendente y una tercera vez para devolver el conjunto de datos a su estado original.

La Figura 9-11 muestra nuestros resultados ordenados en orden descendente en el atributo confianza(Sí).



Si hiciera una cuenta regresiva desde el primer registro (fila 667) hasta el punto en el que nuestro valor de confianza (Sí) es 0,950, descubriría que hay 140 personas en el conjunto de datos para quienes tenemos un 95 % o más de confianza de que corren riesgo de sufrir un ataque cardíaco recurrente (y eso no es redondear a aquellos que tienen un 0,949 en la columna "Sí"). Por lo tanto, hay algunas personas que son bastante fáciles de detectar. Puede notar que muchas están divorciadas, pero varias también son viudas. La pérdida de un cónyuge por cualquier motivo es difícil, por lo que tal vez Sonia pueda comenzar por ofrecer más programas para apoyar a quienes se ajustan a esta descripción. La mayoría de estas personas son obesas y tienen niveles de colesterol superiores a 200, y ninguna ha participado en clases de manejo del estrés. Sonia tiene varias oportunidades para ayudar a estas personas y probablemente les ofrecería la posibilidad de participar en varios programas o crear un programa que ofrezca un enfoque holístico del bienestar físico y mental. Dado que hay una buena cantidad de estas personas que comparten tantos rasgos de alto riesgo, esta puede ser una excelente manera de crear grupos de apoyo para ellas.

Pero también hay personas en el conjunto de datos que tal vez necesiten ayuda, pero no son tan evidentes y tal vez solo necesiten ayuda en una o dos áreas. Haga clic en confianza (sí) una tercera vez para que los datos de los resultados vuelvan a su estado original (ordenados por número de fila). Ahora, desplácese hacia abajo hasta encontrar la fila 95 (resaltada en la Figura 9-12). Tome nota de los atributos de esta persona.



Las dos personas representadas en las filas 95 y 554 tienen mucho en común. En primer lugar, ambas están en este conjunto de datos porque han sufrido ataques cardíacos. Ambas son mujeres de 70 años cuyos maridos han muerto. Ambas tienen un rasgo de ansiedad de 65 puntos. Y, sin embargo, estamos prediciendo con un 96% de certeza que la primera no sufrirá otro ataque cardíaco, mientras que predecimos con casi un 80% que la otra sí. Incluso sus categorías de peso son similares, aunque el sobrepeso sin duda influye en el riesgo de la segunda mujer. Pero lo que es realmente evidente al comparar a estas dos mujeres es que la segunda mujer tiene un nivel de colesterol que casi toca el límite superior de nuestro rango en este conjunto de datos (el límite superior que se muestra en la Figura 9-9 es 239), y no ha asistido a clases de manejo del estrés. Tal vez Sonia pueda usar estas comparaciones para ayudar a esta mujer a entender cuán drásticamente puede mejorar sus posibilidades de evitar otro ataque cardíaco. En esencia, Sonia podría decir: “Hay mujeres que son muy parecidas a usted y que tienen casi cero posibilidades de sufrir otro ataque cardíaco. Si reduce su colesterol, aprende a controlar su estrés y quizás reduce su peso hasta un nivel más cercano al normal, puede eliminar prácticamente el riesgo de sufrir otro ataque cardíaco”. Sonia podría continuar ofreciendo programas específicos para esta mujer, dirigidos específicamente al control del colesterol, el peso o el estrés.

**RESUMEN DEL CAPÍTULO**

La regresión logística es una excelente manera de predecir si algo sucederá o no, y cuán confiados estamos en tales predicciones. Toma en cuenta una serie de atributos numéricos y luego los utiliza a través de un conjunto de datos de entrenamiento para predecir los resultados probables en un conjunto de datos de puntuación comparable. La regresión logística utiliza un atributo objetivo nominal (o etiqueta, en RapidMiner) para categorizar las observaciones en un conjunto de datos de puntuación en sus resultados probables.

Al igual que con la regresión lineal, los datos de puntuación deben tener rangos que se encuentren dentro de sus rangos de datos de entrenamiento correspondientes. Sin esos límites, es inseguro e imprudente hacer suposiciones sobre las observaciones en el conjunto de datos de puntuación, ya que no hay observaciones comparables en los datos de entrenamiento sobre los cuales basar sus suposiciones de puntuación. Sin embargo, cuando se utiliza dentro de estos límites, la regresión logística puede ayudarnos a predecir de manera rápida y sencilla el resultado de algún fenómeno en un conjunto de datos, y a determinar cuán confiados podemos estar en la precisión de esa predicción.

**PREGUNTAS DE REPASO**

Para el ejercicio de este capítulo, utilizará la regresión logística para intentar predecir si los jóvenes que conoce terminarán o no la universidad. Complete los siguientes pasos:

1) Abra una nueva hoja de cálculo en blanco en OpenOffice Calc. En la parte inferior de la hoja de cálculo habrá tres pestañas predeterminadas denominadas Hoja1, Hoja2, Hoja3. Cambie el nombre de la primera a Formación y el de la segunda a Puntuación. Puede cambiar el nombre de las pestañas haciendo doble clic en sus etiquetas. Puede eliminar o ignorar la tercera hoja predeterminada.

2) En la hoja de formación, comenzando en la celda A1 y siguiendo a lo largo, cree etiquetas de tributo para cinco atributos: Parent\_Grad, Gender, Income\_Level, Num\_Siblings y Graduated.

3) Copie cada uno de estos nombres de atributo excepto Graduated en la hoja de Puntuación.

4) En la hoja de capacitación, ingrese valores para cada uno de estos atributos para varios adultos que conozca que estén en la edad en que podrían haberse graduado de la universidad. Estos pueden ser familiares, amigos y vecinos, compañeros de trabajo o estudiantes, etc. Intente hacer al menos 20 observaciones; 30 o más sería mejor. Ingrese las parejas de esposos como dos observaciones separadas. Use lo siguiente para guiar su ingreso de datos:

a. Para Parent\_Grad, ingrese un 0 si ninguno de los padres de la persona se graduó de la universidad, un 1 si uno de los padres lo hizo y un 2 si ambos padres lo hicieron. Si los padres de la persona obtuvieron un título de posgrado, podría experimentar para hacer que este atributo sea aún más interesante al usarlo para contener la cantidad total de títulos universitarios de los padres de la persona. Por ejemplo, si la persona representada en la observación tenía una madre que obtuvo una licenciatura, maestría y doctorado, y un padre que obtuvo una licenciatura y maestría, podría ingresar un 5 en este atributo para esa persona.

b. Para Sexo, ingrese 0 para mujer y 1 para hombre.

c. Para Nivel\_de\_ingreso, ingrese un 0 si la persona vive en un hogar con un nivel de ingresos por debajo de lo que consideraría por debajo del promedio, un 1 para promedio y un 2 para por encima del promedio. Puede estimar o generalizar. Sea sensible a los demás cuando recopile sus datos: no fisgonee demasiado o corra el riesgo de ofender a los sujetos de sus datos.

d. Para Número\_de\_hermanos, ingrese la cantidad de hermanos que tiene la persona.

e. Para Graduado, ingrese "Sí" si la persona se graduó de la universidad y "No" si no lo hizo.

5) Una vez que haya compilado su conjunto de datos de entrenamiento, cambie a la hoja de puntuación en OpenOffice Calc. Repita el proceso de ingreso de datos para al menos 20 (cuantos más, mejor) jóvenes de entre 0 y 18 años que conozca. Utilizará el conjunto de entrenamiento para intentar predecir si estos jóvenes se graduarán o no de la universidad y, de ser así, qué tan seguro está de su predicción. Recuerde que estos son sus datos de puntuación, por lo que no proporcionará el atributo Graduado, lo predecirá en breve.

6) Utilice la opción de menú Archivo > Guardar como en OpenOffice Calc para guardar sus hojas de entrenamiento y puntuación como archivos CSV.

7) Importa los dos archivos CSV a tu repositorio de RapidMiner. Asegúrate de darles nombres descriptivos.

8) Arrastre los dos conjuntos de datos a una nueva ventana de proceso. Si ha preparado bien los datos en OpenOffice Calc, no debería tener que lidiar con datos faltantes o inconsistentes, por lo que la preparación de datos debería ser mínima. Cambie el nombre de los dos operadores de recuperación para poder distinguir la diferencia entre los conjuntos de datos de entrenamiento y puntuación.

9) Un paso necesario en la preparación de datos es agregar un operador Establecer rol y definir el atributo Graduado como etiqueta en los datos de entrenamiento. Alternativamente, puede configurar el atributo Graduado como etiqueta durante la importación de datos.

10) Agregue un operador de regresión logística a su flujo de entrenamiento.

11) Aplique su modelo de regresión logística a sus datos de puntuación y ejecute su modelo. Evalúe e informe sus resultados. ¿Sus porcentajes de confianza son interesantes? ¿Sorprendentes? ¿Los valores de graduación previstos parecen razonables y coherentes con sus datos de entrenamiento? ¿Alguna variable independiente (atributo predictor) parece ser un predictor particularmente bueno de la variable dependiente (etiqueta o atributo de predicción)? Si es así, ¿por qué cree que es así?

Paso del reto.

12) Cambie el operador de regresión logística por un tipo diferente de operador logístico (por ejemplo, pruebe el operador logístico W de Weka). Vuelva a ejecutar el modelo. Considere realizar una investigación para conocer la diferencia entre los algoritmos que sustentan los distintos enfoques logísticos. Compare los nuevos resultados con los resultados originales de la regresión logística e informe cualquier hallazgo o diferencia interesante.