**Explorar la clasificación con Azure Machine Learning Designer**

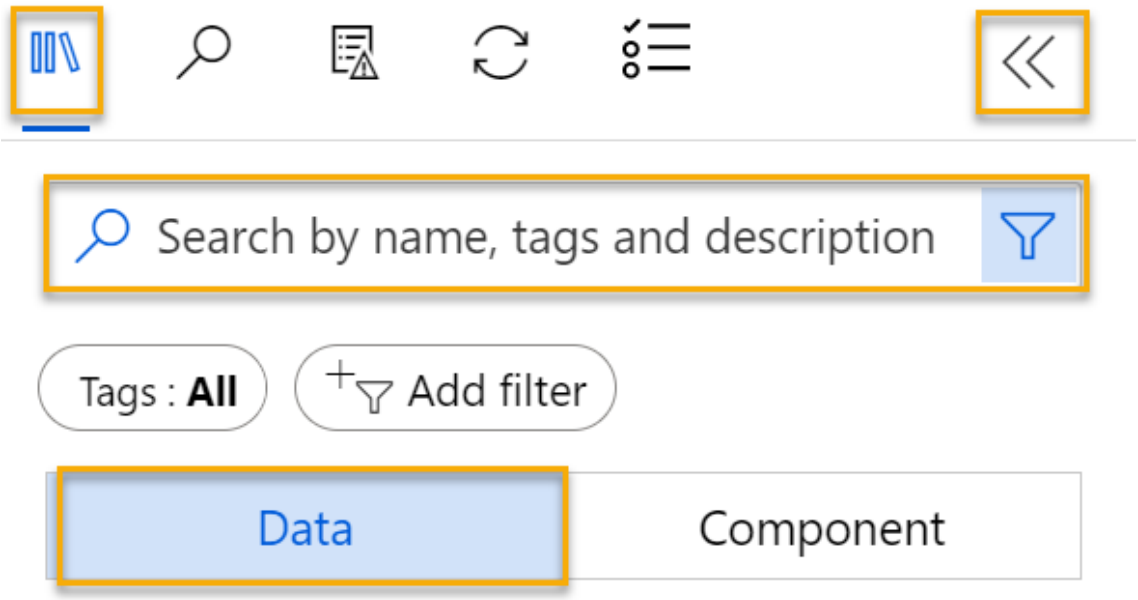
**Crear un conjunto de datos**

1. En Azure Machine Learning Studio , expanda el panel izquierdo seleccionando el ícono de menú en la parte superior izquierda de la pantalla. Seleccione la página **Datos** (en **Activos** ). La página Datos contiene archivos de datos o tablas específicos con los que planea trabajar en Azure ML. También puede crear conjuntos de datos desde esta página.
2. En la página **Datos** , en la pestaña **Activos de datos , seleccione + Crear** . A continuación, configure un activo de datos con los siguientes ajustes:
    - **Tipo de datos** :
        - **Nombre** : diabetes-data
        - **Descripción** : Datos sobre la diabetes
        - **Tipo de conjunto de datos** : tabular
    - **Fuente de datos** : De archivos web
    - **URL web** :
        - **URL web** : https://aka.ms/diabetes-data
        - **Omitir validación de datos** : *no seleccionar*
    - **Ajustes** :
        - **Formato de archivo** : Delimitado
        - **Delimitador** : Coma
        - **Codificación** : UTF-8
        - **Encabezados de columna** : solo el primer archivo tiene encabezados
        - **Saltar filas** : Ninguna
        - **El conjunto de datos contiene datos de varias líneas** : *no seleccionar*
    - **Esquema** :
        - Incluir todas las columnas excepto **Ruta**
        - Revisar los tipos detectados automáticamente
    - **Revisar**
        - Seleccionar **Crear**
3. Una vez creado el conjunto de datos, ábralo y acceda a la página **Explorar** para ver una muestra de los datos. Estos datos representan detalles de pacientes a los que se les realizó la prueba de diabetes.

**Cree una canalización en Designer y cargue datos en el lienzo**

Para comenzar a utilizar el diseñador de Azure Machine Learning, primero debe crear una canalización y agregar el conjunto de datos con el que desea trabajar.

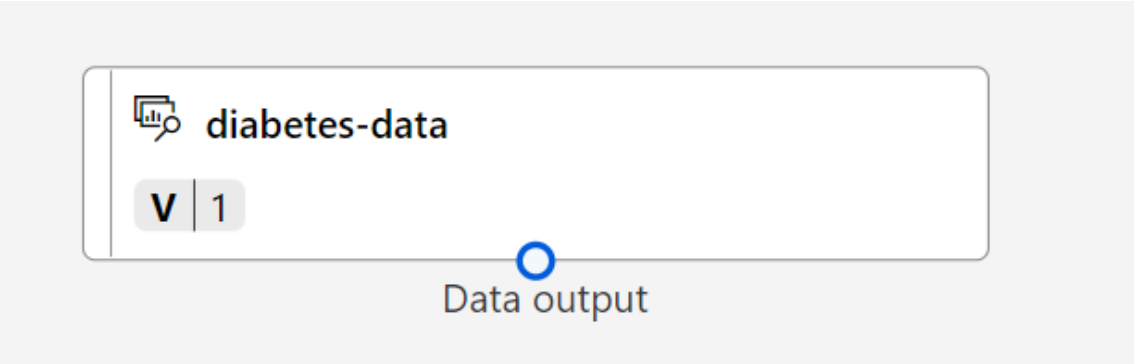
1. En el estudio de Azure Machine Learning, en el panel izquierdo, seleccione el elemento **Diseñador (en Creación** ) y, luego, seleccione **+** para crear una nueva canalización.
2. Cambie el nombre del borrador de **Pipeline-Created-on-** ***date*** a **Capacitación en diabetes** .
3. Luego, en el proyecto, junto al nombre de la tubería a la izquierda, seleccione el ícono de flechas para expandir el panel si aún no está expandido. El panel debería abrirse de manera predeterminada en el panel **de la biblioteca de activos** , indicado por el ícono de libros en la parte superior del panel. Tenga en cuenta que hay una barra de búsqueda para ubicar activos. Observe dos botones, **Datos** y **Componente** .




4. Seleccionar **datos** . Busque y coloque el conjunto de **datos de diabetes** en el lienzo.
5. Haga clic con el botón derecho (Ctrl+clic en una Mac) en el conjunto de **datos de diabetes** en el lienzo y seleccione **Vista previa de datos** .
6. Revise el esquema de los datos en la pestaña *Perfil* , observando que puede ver las distribuciones de las distintas columnas como histogramas.
7. Desplácese hacia abajo y seleccione el encabezado de la columna **Diabético** y observe que contiene dos valores, **0** y **1.** Estos valores representan las dos clases posibles para la *etiqueta* que predecirá su modelo: un valor de **0** significa que el paciente no tiene diabetes y un valor de **1** significa que el paciente es diabético.
8. Desplácese hacia arriba y revise las otras columnas, que representan las *características* que se utilizarán para predecir la etiqueta. Tenga en cuenta que

la mayoría de estas columnas son numéricas, pero cada característica está en su propia escala. Por ejemplo, los valores **de Edad** varían de 21 a 77, mientras que los valores **de DiabetesPedigree** varían de 0,078 a 2,3016. Al entrenar un modelo de aprendizaje automático, a veces es posible que los valores más grandes dominen la función predictiva resultante, lo que reduce la influencia de las características en una escala más pequeña. Por lo general, los científicos de datos mitigan este posible sesgo normalizando *las* columnas numéricas para que estén en escalas similares.

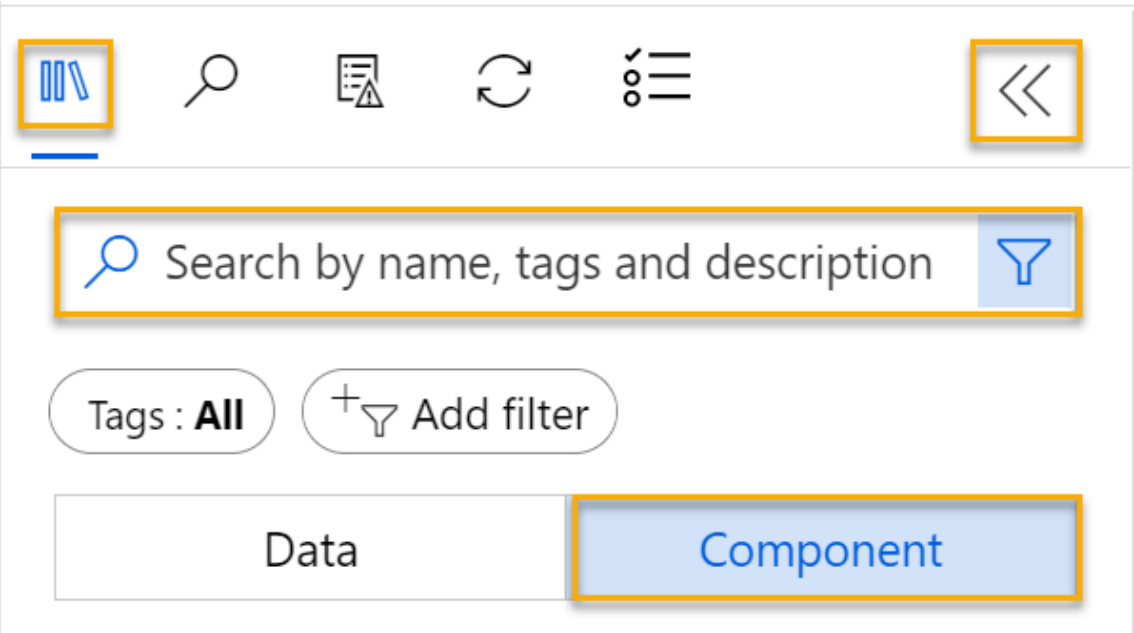
9. Cierre la pestaña **DataOutput** para poder ver el conjunto de datos en el lienzo de esta manera:




**Agregar transformaciones**

Antes de poder entrenar un modelo, normalmente es necesario aplicar algunas transformaciones de preprocesamiento a los datos.

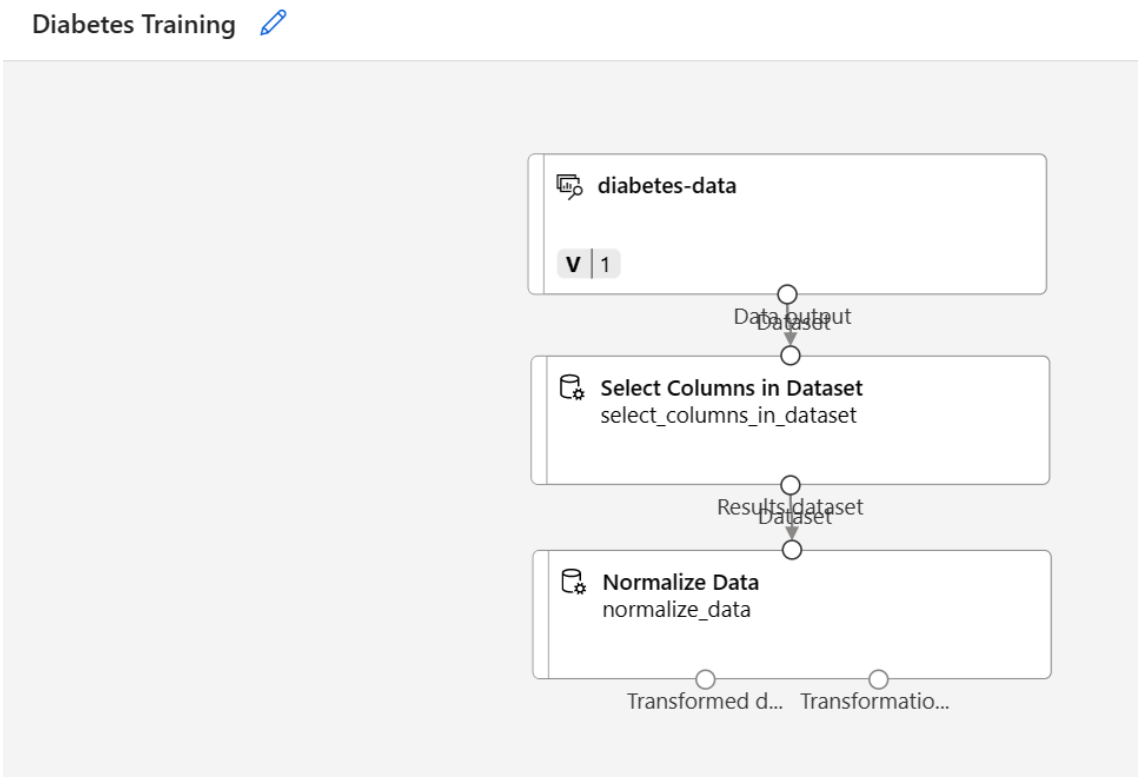
1. En el panel **de la biblioteca de activos** de la izquierda, seleccione **Componente** , que contiene una amplia gama de módulos que puede utilizar para la transformación de datos y el entrenamiento de modelos. También puede utilizar la barra de búsqueda para localizar módulos rápidamente.



2. Busque el módulo **Seleccionar columnas en el conjunto de datos** y colóquelo en el lienzo debajo del conjunto **de datos de diabetes** . Luego, conecte la salida de la parte inferior del conjunto **de datos de diabetes** con la entrada en la parte superior del módulo **Seleccionar columnas en el conjunto de datos** .

3. Haga doble clic en el módulo **Seleccionar columnas en el conjunto de datos** para acceder a un panel de configuración a la derecha. Seleccione **Editar columna** . Luego, en la ventana **Seleccionar columnas , seleccione Por nombre** y **Agregar todas** las columnas. Luego, elimine **PatientID** y haga clic en **Guardar** .

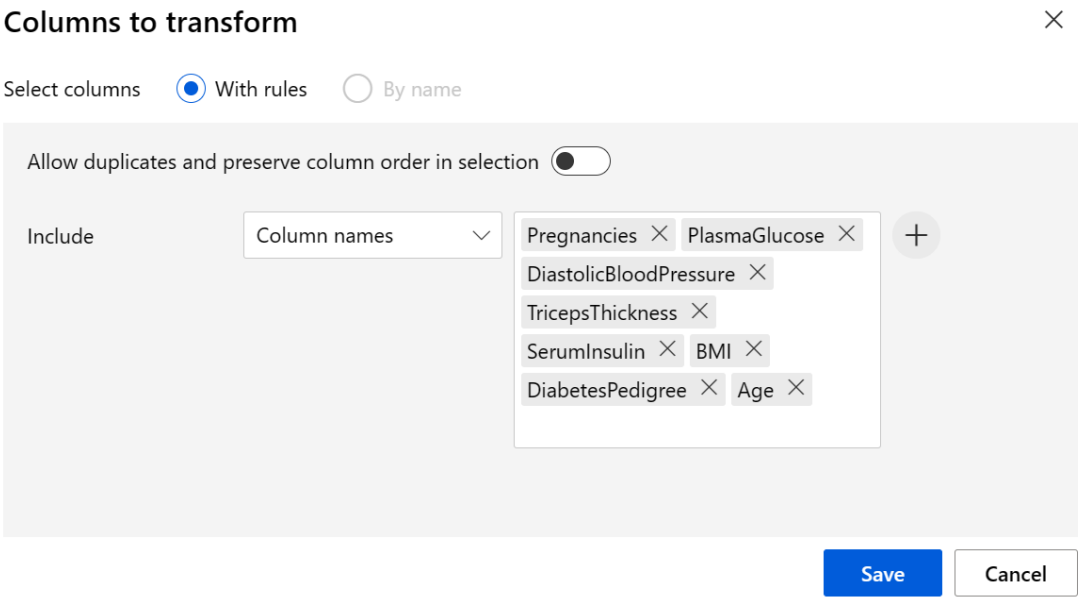
4. Busque el módulo **Normalizar datos** y colóquelo en el lienzo debajo del módulo **Seleccionar columnas en el conjunto de datos** . Luego, conecte la salida de la parte inferior del módulo **Seleccionar columnas en el conjunto de datos** con la entrada de la parte superior del módulo **Normalizar datos** , de la siguiente manera:




5. Haga doble clic en el módulo **Normalizar datos** para ver su configuración; tenga en cuenta que requiere que especifique el método de transformación y las columnas que se transformarán.

6. Establezca el *método de transformación* en **MinMax** y la *opción Usar 0 para columnas constantes cuando esté marcada* en **Verdadero** . Edite las columnas que desea transformar con **Editar columnas** . Seleccione Columnas **con reglas** y copie y pegue la siguiente lista en los nombres de las columnas incluidas:

Pregnancies, PlasmaGlucose, DiastolicBloodPressure, TricepsThickness, SerumInsulin, BMI, DiabetesPedigree, Age




Haga clic en **Guardar** y cierre el cuadro de selección.

La transformación de datos normaliza las columnas numéricas para ponerlas en la misma escala, lo que debería ayudar a evitar que las columnas con valores grandes dominen el entrenamiento del modelo. Normalmente, aplicarías un montón de transformaciones de preprocesamiento como esta para preparar tus datos para el entrenamiento, pero en este ejercicio simplificaremos las cosas.

**Ejecutar**

Para aplicar sus transformaciones de datos, debe ejecutar la canalización como un experimento.

1. Seleccione **Configurar y enviar** en la parte superior de la página para abrir el cuadro de diálogo **Configurar trabajo de canalización** .
2. En la página **Básico** , seleccione **Crear nuevo** y establezca el nombre del experimento como **mslearn-diabetes-training,** luego seleccione **Siguiente** .
3. En la página **Entradas y salidas** , seleccione **Siguiente** sin realizar ningún cambio.
4. En la página **Configuración de tiempo de ejecución** , aparece un error porque no tienes un cálculo predeterminado para ejecutar la canalización. En el menú **desplegable Seleccionar tipo de cálculo** , selecciona *Clúster de cálculo* y, en el menú **desplegable Seleccionar clúster de cálculo de Azure ML** , selecciona el clúster de cálculo que creaste recientemente.
5. Seleccione **Revisar + Enviar** para revisar el trabajo del pipeline y luego seleccione **Enviar** para ejecutar el pipeline de entrenamiento.
6. Espere unos minutos hasta que finalice la ejecución. Puede comprobar el estado del trabajo seleccionando **Trabajos** en **Activos** . Desde allí, seleccione el experimento **mslearn-diabetes-training** y, luego, el trabajo **Entrenamiento en diabetes .**

**Ver los datos transformados**

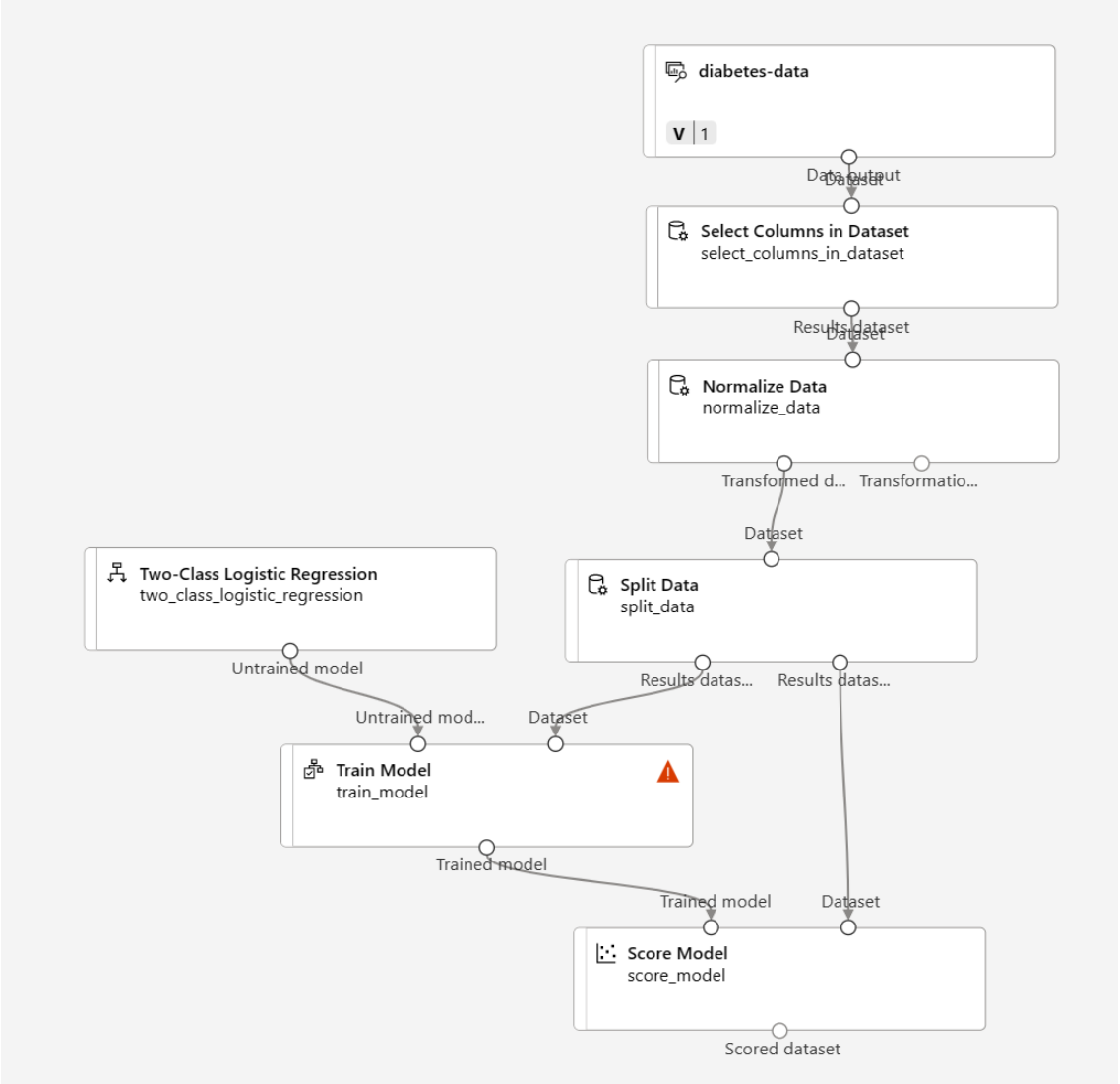
Una vez finalizada la ejecución, el conjunto de datos estará preparado para el entrenamiento del modelo.

1. Haga clic con el botón derecho (Ctrl+clic en una Mac) en el módulo **Normalizar datos** en el lienzo y seleccione **Vista previa de datos** . Seleccione **Conjunto de datos transformado** .
2. Vea los datos y observe que las columnas numéricas que seleccionó se han normalizado a una escala común.
3. Cerrar la visualización de resultados de datos normalizados. Volver a la pestaña anterior.

Después de utilizar transformaciones de datos para preparar los datos, puede usarlos para entrenar un modelo de aprendizaje automático.

Es una práctica habitual entrenar el modelo utilizando un subconjunto de los datos, mientras se reservan algunos datos para probar el modelo entrenado. Esto le permite comparar las etiquetas que predice el modelo con las etiquetas conocidas reales en el conjunto de datos original.

En este ejercicio, trabajará en los pasos necesarios para ampliar el proceso **de capacitación en diabetes,** como se muestra aquí:




Siga los pasos a continuación, utilizando la imagen de arriba como referencia mientras agrega y configura los módulos necesarios.

1. Regrese a la página **Diseñador** y seleccione el canal **de Capacitación en diabetes** .
2. En el panel **de la biblioteca de activos** de la izquierda, en **Componente** , busque y coloque un módulo **Datos divididos** en el lienzo debajo del módulo **Normalizar datos** . Luego, conecte la salida *Conjunto de datos transformado* (izquierda) del módulo **Normalizar datos** a la entrada del módulo **Datos divididos** .

**Consejo:** Utilice la barra de búsqueda para localizar módulos rápidamente.

3. Seleccione el módulo **Datos divididos** y configure sus ajustes de la siguiente manera:
    - **Modo de división** : Dividir filas
    - **Fracción de filas en el primer conjunto de datos de salida** : 0,7
    - **División aleatoria** : Verdadero

- **Semilla aleatoria** : 123
- **División estratificada** : Falso

4. En la **biblioteca de activos** , busque y coloque un módulo **Entrenar modelo** en el lienzo, debajo del módulo **Dividir datos** . Luego, conecte la salida del *conjunto de datos Resultados 1* (izquierda) del módulo **Dividir datos** a la entrada *Conjunto de datos* (derecha) del módulo **Entrenar modelo** .
5. El modelo que estamos entrenando predecirá el valor **Diabético , así que seleccione el módulo Entrenar modelo** y modifique su configuración para establecer la **columna Etiqueta** en **Diabético** .

La etiqueta **de diabético** que predecirá el modelo es una clase (0 o 1), por lo que necesitamos entrenar el modelo utilizando un algoritmo de *clasificación* . En concreto, hay dos clases posibles, por lo que necesitamos un algoritmo *de clasificación binario* .

6. En la **biblioteca de activos** , busque y coloque un módulo **de regresión logística de dos clases** en el lienzo, a la izquierda del módulo **Datos divididos** y sobre el módulo **Entrenar modelo** . Luego, conecte su salida a la entrada *Modelo no entrenado* (izquierda) del módulo **Entrenar modelo** .

Para probar el modelo entrenado, necesitamos usarlo para *puntuar* el conjunto de datos de validación que conservamos cuando dividimos los datos originales; en otras palabras, predecir etiquetas para las características en el conjunto de datos de validación.

7. En la **biblioteca de activos** , busque y coloque un módulo **Score Model** en el lienzo, debajo del módulo **Train Model** . Luego, conecte la salida del módulo **Train Model** a la entrada *Trained model* (izquierda) del módulo **Score Model** ; y conecte la salida *Results dataset2* (derecha) del módulo **Split Data** a la entrada *Dataset* (derecha) del módulo **Score Model** .

**Ejecutar el proceso**

Ahora está listo para ejecutar el proceso de entrenamiento y entrenar el modelo.

1. Seleccione **Configurar y enviar** y ejecute el proceso utilizando el experimento existente llamado **mslearn-diabetes-training** .
2. Espere a que finalice la ejecución del experimento. Esto puede tardar 5 minutos o más.
3. Para comprobar el estado del trabajo, seleccione **Trabajos** en **Activos** . Desde allí, seleccione el experimento **mslearn-diabetes-training** y, a continuación, seleccione el trabajo más reciente **de Capacitación en diabetes .**
4. En la nueva pestaña, haga clic derecho (Ctrl+clic en una Mac) en el módulo **Modelo de puntuación** en el lienzo, seleccione **Vista previa de datos** y luego seleccione **Conjunto de datos puntuados** para ver los resultados.
5. Desplácese hacia la derecha y observe que junto a la columna **Diabético** (que contiene los valores verdaderos conocidos de la etiqueta) hay una nueva columna llamada **Etiquetas puntuadas** , que contiene los valores de etiqueta previstos, y una columna **Probabilidades puntuadas** que contiene un valor de probabilidad

entre 0 y 1. Esto indica la probabilidad de una predicción *positiva* , por lo que las probabilidades mayores de 0,5 dan como resultado una etiqueta prevista de ***1*** (diabético), mientras que las probabilidades entre 0 y 0,5 dan como resultado una etiqueta prevista de ***0*** (no diabético).

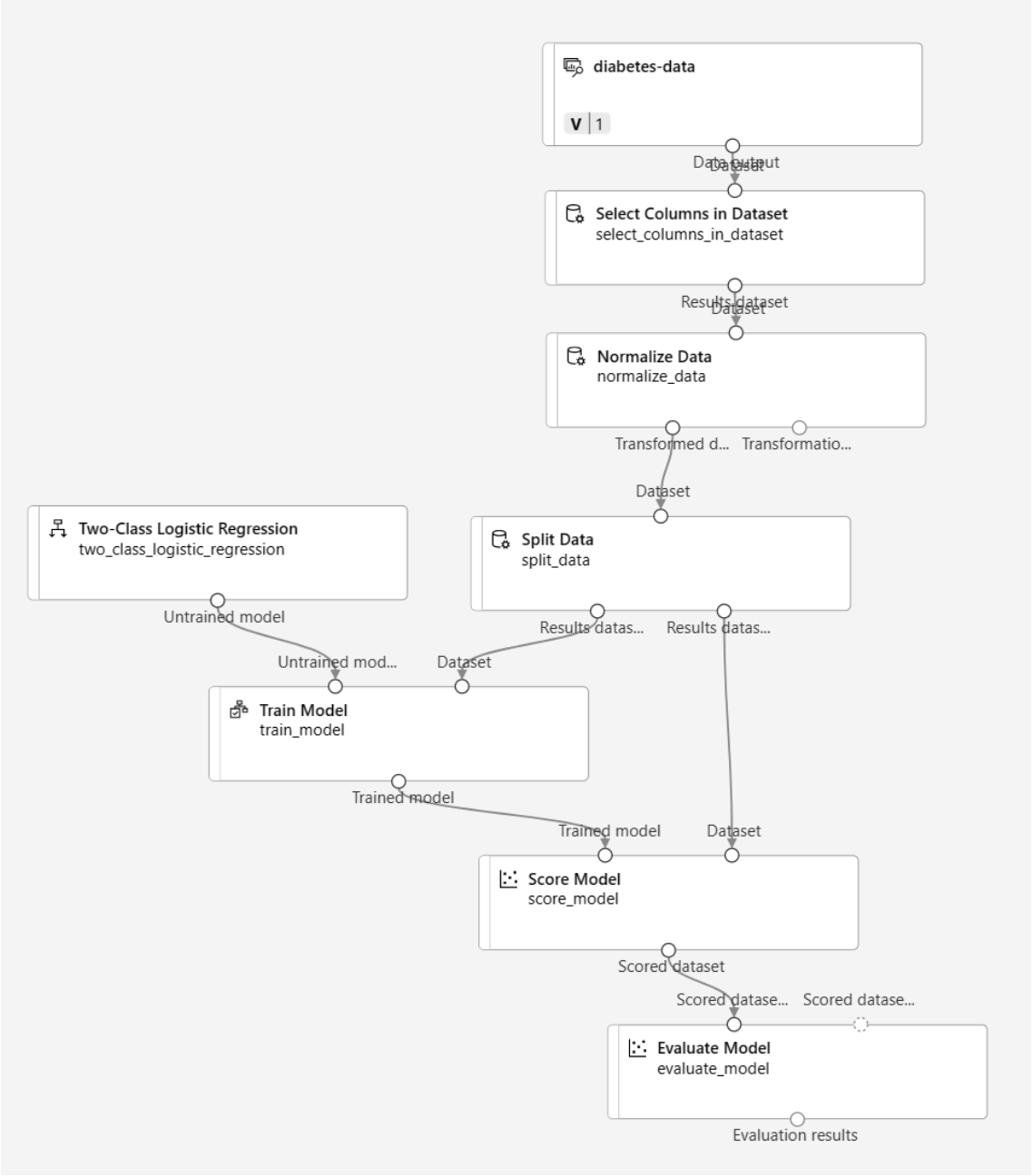
6. Cierre la pestaña **Scored_dataset** .

El modelo predice valores para la etiqueta **de diabetes** , pero ¿qué tan confiables son sus predicciones? Para evaluar eso, es necesario evaluar el modelo.

Los datos de validación que retuvo y utilizó para puntuar el modelo incluyen los valores conocidos de la etiqueta. Por lo tanto, para validar el modelo, puede comparar los valores reales de la etiqueta con los valores de la etiqueta que se predijeron cuando puntuó el conjunto de datos de validación. En función de esta comparación, puede calcular varias métricas que describen el rendimiento del modelo.

**Agregar un módulo Evaluar modelo**

1. Regrese al **Diseñador** y abra el proceso **de capacitación en diabetes** que creó.
2. En la **biblioteca de activos** , busque y coloque un módulo **Evaluar modelo** en el lienzo, debajo del módulo **Puntuar modelo** , y conecte la salida del módulo **Puntuar modelo** a la entrada Conjunto *de datos puntuado* (izquierda) del módulo **Evaluar modelo** .
3. Asegúrese de que su canalización se vea así:




4. Seleccione **Configurar y enviar** y ejecute el proceso utilizando el experimento existente llamado **mslearn-diabetes-training** .

5. Espere a que finalice la ejecución del experimento.

6. Para comprobar el estado del trabajo, seleccione **Trabajos** en **Activos** . Desde allí, seleccione el experimento **mslearn-diabetes-training** y, a continuación, seleccione el trabajo más reciente **de Capacitación en diabetes .**

7. En la nueva pestaña, haga clic con el botón derecho (Ctrl+clic en una Mac) en el módulo **Evaluar modelo en el lienzo, seleccione Vista previa de datos** y luego seleccione **Resultados de evaluación** para ver las métricas de rendimiento. Estas métricas pueden ayudar a los científicos de datos a evaluar qué tan bien predice el modelo en función de los datos de validación.

8. Desplácese hacia abajo para ver la *matriz de confusión* del modelo. Observe los valores previstos y reales para cada clase posible.
9. Revise las métricas a la izquierda de la matriz de confusión, que incluyen:
    - **Precisión** : En otras palabras, ¿qué proporción de predicciones sobre diabetes acertó el modelo?
    - **Precisión** : En otras palabras, de todos los pacientes que *el modelo predijo* que tendrían diabetes, el porcentaje de veces que el modelo es correcto.
    - **Recordatorio** : En otras palabras, de todos los pacientes *que realmente tienen* diabetes, ¿cuántos casos de diabetes identificó correctamente el modelo?
    - **Puntuación F1**
10. Utilice el control deslizante **Umbral** ubicado sobre la lista de métricas. Intente mover el control deslizante Umbral y observe el efecto en la matriz de confusión. Si lo mueve completamente hacia la izquierda (0), la métrica de Recordación se convierte en 1, y si lo mueve completamente hacia la derecha (1), la métrica de Recordación se convierte en 0.
11. Observe por encima del control deslizante Umbral la **curva ROC** y la métrica **AUC** que se enumeran junto con las otras métricas que se muestran a continuación. Para tener una idea de cómo esta área representa el rendimiento del modelo, imagine una línea diagonal recta desde la parte inferior izquierda hasta la parte superior derecha del gráfico ROC. Esto representa el rendimiento esperado si simplemente adivinara o lanzara una moneda al aire para cada paciente: podría esperar obtener aproximadamente la mitad de aciertos y la otra mitad de errores, por lo que el área debajo de la línea diagonal representa un AUC de 0,5. Si el AUC de su modelo es mayor que esto para un modelo de clasificación binaria, entonces el modelo funciona mejor que una suposición aleatoria.
12. Cierre la pestaña **Evaluation_results** .

El rendimiento de este modelo no es tan bueno, en parte porque solo realizamos un preprocesamiento e ingeniería de características mínimos. Puede probar un algoritmo de clasificación diferente, como **Two-Class Decision Forest** , y comparar los resultados. Puede conectar las salidas del módulo **Split Data** a varios módulos **Train Model** y **Score Model** , y puede conectar un segundo módulo **Score Model** al módulo **Evaluate Model** para ver una comparación en paralelo. El objetivo del ejercicio es simplemente presentarle la clasificación y la interfaz del diseñador de Azure Machine Learning, ¡no entrenar un modelo perfecto!