TALLER 5 ANÁLISIS DISCRIMINANTE UTILIZANDO SPSS

El análisis discriminante es un análisis estadístico multivariado cuya finalidad es analizar si existen diferencias significativas entre grupos de objetos respecto a un conjunto de variables medidas sobre los mismos. El análisis tiene dos objetivos:

- Pretende encontrar relaciones lineales entre las variables continuas que mejor discriminen en los grupos dados a los objetos. Estas combinaciones lineales de las variables deben maximizar la varianza entre los grupos y minimizar la varianza dentro de los grupos.

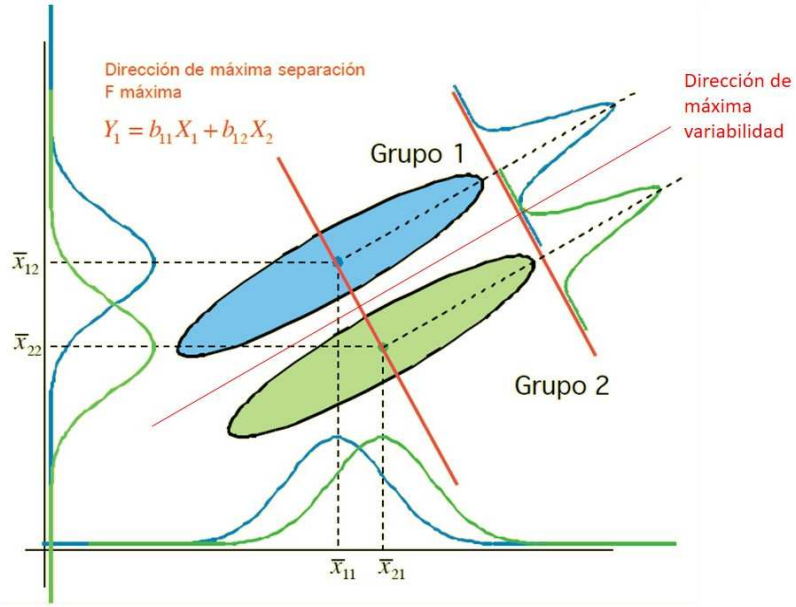

Figura 1. El análisis discriminante tiene como fin encontrar una combinación lineal de variables que maximicen la varianza entre grupos y minimice la varianza entre los grupos

- Construir una regla de decisión que asigne un objeto nuevo, que no sabemos clasificar previamente, a uno de los grupos prefijados con un cierto grado de riesgo.

En este taller usted llevará a cabo un análisis de estabilidad para entender su relación con las variables morfológicas e hidrológicas de la cuenca de La Arenosa ubicada en el municipio de San Carlos, Antioquia.

Para esto, llevará a cabo el análisis discriminante para diferenciar las celdas de la zona de estudio entre estables e inestables. Las variables que incluirá entre el análisis son:

- **Altitud:** Altura sobre el nivel del mar de la celda
- **Pendiente:** Pendiente, en grados, de cada celda
- **Aspecto:** Azimut de la dirección de buzamiento de la celda
- **Profundidad del suelo:** Profundidad, en metros de la unidad de suelos residuales
- **Distancia a drenajes:** Distancia en metros al drenaje más cercano a cada celda, medidos en dirección perpendicular al flujo
- **Acumulación de flujo:** Número de celdas que drenan a la celda determinada

La idea es entender cuáles variables inciden más sobre la estabilidad de la zona, y al final, hallar la ecuación discriminante con la que evaluará la estabilidad de toda la cuenca, y probará su desempeño comparándolo con la cartografía de los deslizamientos que se presentaron allí después del episodio de lluvias intensas del 21 de septiembre de 1990.

1. PREPARACIÓN DE LA TABLA

Para realizar el análisis discriminante usted debe ingresar a SPSS una tabla, la cual debe incluir en las filas los objetos, y en las columnas los atributos de dichos objetos. Para este caso, debemos ingresar una tabla donde cada fila corresponda a una celda, y cada columna corresponda al valor de cada atributo de dicha celda. A cada una de estas variables las llamaremos de aquí en adelante variables independientes. La última columna de la tabla debe ser la variable dependiente de tipo categórica binaria que distinga la celda entre estable (0) o inestable (1).

Usted podría ingresar en la tabla todos los pixeles de la zona de estudio (cerca de 200.000), pero no es recomendable dado el gran tiempo de procesamiento y que más adelante deseamos comprobar el desempeño del modelo en el resto de la zona de estudio. Por ello, armará una tabla con una muestra representativa que incluya celdas inestables y estables.

Para preparar la tabla, siga los siguientes pasos:

- En ArcMap, abra el raster Fuentes_Arenosa.tif. En él, se muestran las celdas que fallaron en el evento del 21 de septiembre de 1990 con el valor 1, mientras que el resto de las celdas con valor 0 fueron celdas estables.
- Abra el raster P_Muestra_Arenosa.tif. Este raster tiene únicamente 2000 celdas, de las cuales 1000 corresponden a celdas estables y 1000 a celdas inestables. Este raster se obtuvo tomando puntos aleatorios en la zona de estudio de tal forma que hayan igual número de muestras estables e inestables, y es el que utilizará para construir la ecuación discriminante.
- Abra los raster DEM_Arenosa.tif, Pendiente_Arenosa.tif, Aspectos_Arenosa.tif, Prof_suelo_Arenosa.tif, Dist_drenajes_Arenosa.tif, y Acumulacion_Flujo_Arenosa. Observe las características de la zona de estudio. Estos raster contienen la información de las variables independientes que utilizará para construir la tabla.
- En el ArcToolbox, seleccione Spatial Analyst > Extraction > Sample. Esta herramienta permite crear una tabla que muestra el valor de las celdas de varios raster con una ubicación definida. En el campo Input Rasters, agregue cada uno de los raster que contienen las variables independientes (DEM, Pendiente, Aspecto, Profundidad del suelo, Distancia a Drenajes y Acumulación de flujo) y el raster con la variable categórica (Fuentes). En el campo Input location raster or point features, agregue el raster con la ubicación de las celdas de muestra (P_Muestra_Arenosa). Seleccione la locación donde desea guardar la tabla y asígnele un nombre. Deje el resto de las opciones por defecto. Haga clic en OK.
- La tabla que acaba de crear aparecerá en la Tabla de Contenidos. Haga clic derecho sobre ella y seleccione Open. Observe que para cada una de las celdas de muestra se extrajeron los valores de cada una de las variables independientes, y finalmente si la celda es inestable o estable. Despliegue las Opciones de Tabla y seleccione la opción Export. Exporte todos los registros en formato dBase.

2. ANÁLISIS DISCRIMINANTE EN SPSS

Una vez tenga la información para ingresar al programa, lleve a cabo el análisis discriminante en SPSS:

- Abra IBM SPSS. Seleccione Archivo > Abrir > Datos. En el menú desplegable de tipo de archivo, seleccione dBase (*.dbf) y abra la tabla de datos que creó.
- Una vez se carguen los datos, seleccione Analizar > Clasificar > Discriminante.
- En la nueva ventana, ingrese la variable de agrupación, es decir la variable categórica (Fuentes Arenosa), y defina su rango (0 y 1). Ingrese en las variables independientes las variables que utilizará en el análisis (DEM, Pendiente, Aspecto, Profundidad del suelo y Acumulación de Flujo).
- Debajo de la ventana de variables independientes, debe seleccionar el método con el cual SPSS llevará a cabo el análisis. El primer método es incluir en el análisis y en la ecuación discriminante todas las variables que ingresó. Las variables que tengan poca incidencia en la estabilidad de una celda tendrán poco peso en la ecuación. El segundo método es el de inclusión por pasos, en el cual se van incluyendo las variables más significativas una por una, y eliminando las variables que contribuyen poco a la diferenciación de celdas estables e inestables. En este taller, usted llevará a cabo el análisis discriminante incluyendo todas las variables primero para entender cuáles juegan un papel más significativo en diferenciar la inestabilidad, y después llevará a cabo el análisis por pasos para construir la ecuación discriminante sólo con las variables más representativas.
- Haga Clic sobre Estadísticos. En esta parte usted debe especificar qué pruebas estadísticas de sus datos desea realizar antes del análisis discriminante. El fin de estas pruebas es conocer a priori qué tan adecuado es realizar el análisis discriminante con sus datos. En algunos casos, estas pruebas pueden dar resultados que indiquen que ninguna variable discrimina correctamente los grupos, por lo que el análisis discriminante no tendrá un buen resultado. Algunas de las pruebas que es importante realizar son:
    - **Medias:** Proporciona el vector de medias (los centroides) y desviaciones típicas de cada variable para cada grupo
    - **ANOVAs univariados:** Análisis de la varianza. Contrasta igualdad de medias entre los grupos para cada variable
    - **M de Box:** Comprueba la igualdad entre las matrices de covarianzas de los grupos. Dicha prueba tiene como hipótesis nula que las matrices de covarianzas son iguales.
    - **Matriz de Correlación intra-grupos:** Permite saber qué variables están altamente correlacionadas y deben ser eliminadas del análisis para evitar imprecisiones.
    - **Coeficientes de Fisher:** Coeficientes de la función de clasificación
    - **Coeficientes no estandarizados:** Coeficientes de la función discriminante canónica de Fisher centrados

  Una vez haya seleccionado los estadísticos, haga clic sobre Continuar.

- Si está realizando el análisis con el método de introducir todas las variables independientes juntas, la opción Método estará deshabilitada. Si seleccionó la opción de inclusión de variables por pasos, debe especificar en esta sección cuál desea que sea el criterio para incluir o eliminar las variables del análisis. En cada paso, el programa incluirá o sacará una

variable del análisis de acuerdo a un criterio que usted escoja. Puede consultar todos los métodos, pero el más común es la Lambda de Wilks. Este valor equivale a las desviaciones de la media dentro de cada grupo, entre las desviaciones de la media total sin distinguir grupos. Si una variable tiene un valor pequeño de Lambda de Wilks, la variable discrimina mucho. Así mismo, debe elegir el valor de corte ya sea para valor o probabilidad de F para incluir o excluir una variable del análisis. En este caso, utilice los valores pre-establecidos de F. El valor de F, llamado también F de Snedecor, corresponde a las desviaciones de las medias de cada uno de los grupos a la media total, entre las desviaciones a la media dentro de cada grupo para cada variable. Si F es grande para cada variable, entonces las medias de cada grupo están muy separadas y la variable discrimina bien, y si F es pequeña para cada variable, la variable discrimina poco, ya que habrá poca homogeneidad en los grupos y éstos estarán muy próximos. Deje activada la opción de resumen de los pasos. Haga clic en Continuar.

- Haga clic sobre Guardar. Active los campos de guardar la pertenencia a grupos pronosticada y las puntuaciones discriminantes. Esto almacenará los resultados de la ecuación discriminante y el grupo pronosticado a cada celda como una nueva variable en la tabla de datos.
- Haga clic sobre Clasificar. En esta ventana debe especificar algunas opciones de procesamiento y visualización de los resultados. Las probabilidades previas son las probabilidades a priori de un individuo de pertenecer a cada grupo. Si usted fuera a trabajar con una muestra que incluya más celdas estables que inestables, podría determinar que las probabilidades previas se determinen según el tamaño de los grupos, y por ende la probabilidad de una celda de pertenecer al grupo de estables será mayor que la probabilidad de ser inestable. Dado que en este taller trabajará con una muestra de igual número de individuos estables e inestables, puede seleccionar cualquiera de las opciones, ya que la probabilidad de pertenecer a cualquier grupo es del 50%. En las opciones de visualización, asegúrese de activar la tabla de resumen, la clasificación dejando uno fuera, y la matriz de covarianza intra-grupos.
- Una vez haya determinado las condiciones para el análisis haga clic en Aceptar.

3. ANÁLISIS DE RESULTADOS

Cuando SPSS termina de realizar el análisis discriminante, le mostrará la hoja de resultados. Estos incluyen:

- **Resumen de proceso del caso de análisis:** Muestra los individuos tenidos en cuenta en el análisis y cuales fueron excluidos.
- **Estadísticas de grupo:** Se muestran los estadísticos descriptivos de media y desviación estándar de cada variable para los dos grupos por separado y para toda la población.
- **Prueba de igualdad de medias de grupos:** Tiene como objetivo conocer si las variables introducidas en el análisis tienen poder discriminatorio. Para ello se contrasta la prueba de igualdad de las medias de los grupos para cada variable, llamada Lambda de Wilks. Esta prueba nos indica si las medidas de cada variable son distintas en cada grupo. Si p-valor (Sig.) < 0.05, entonces las variables son significativas, por lo que las varianzas son distintas. Por el contrario, si p-valor (Sig.) > 0.05 las variables no son significativas, lo que se traduce

a que las varianzas de los grupos estables e inestables son iguales, en cuyo caso la variable no tiene mucho poder discriminatorio a nivel univariante. La información de esta tabla suele utilizarse como prueba preliminar para detectar si los grupos difieren en las variables de clasificación seleccionadas; sin embargo, hay que considerar que una variable no significativa a nivel univariante podría aportar información discriminativa a nivel multivariante (Figura 2).

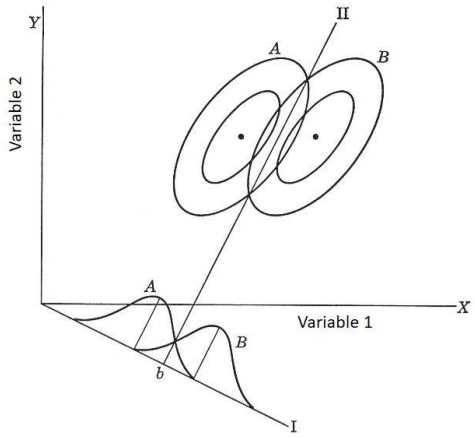

Figura 2. Una variable puede no discriminar los grupos a nivel univariante, como en este caso la Variable 2, pero cuando se combina linealmente con otra variable, puede discriminar muy bien los grupos.

- **Matrices dentro de grupos combinados:** Muestra la matriz de correlación entre las variables. Variables altamente relacionadas (r>0.8 o r<-0.8), agregan ruido al modelo y una de las variables debe ser eliminada del análisis (normalmente aquella que tenga menor poder discriminatorio).
- **Resultado de la prueba M de Box:** Prueba de igualdad de las medias de los grupos. Uno de los supuestos del análisis discriminante es que las matrices de varianzas-covarianzas poblacionales correspondientes a cada grupo son diferentes entre sí. Para interpretar su resultado se utiliza el estadístico F (Si Sig. (p-valor) < 0.05, se interpreta que las covarianzas son distintas y que el análisis discriminante es aplicable. Si Sig. (p-valor) > 0.05, las covarianzas de los dos grupos son iguales, y no es aplicable análisis discriminante). En la tabla del Log determinante muestra los logaritmos de los determinantes de todas las matrices utilizadas en el cálculo del estadístico M, los cuales permiten comprobar qué grupos (cuando hay más de dos) difieren más.
- **Estadísticas por pasos:** Sólo se muestran cuando se realiza el análisis con la metodología de inclusión de variables por pasos
  - **Variables entradas/eliminadas, Variables en y no en el análisis:** Estas tablas muestran, paso por paso, cuáles variables son eliminadas y agregadas al análisis de acuerdo a su valor de F: Desviación de la media de cada uno de los grupos a la media total entre las desviaciones a la media dentro de cada grupo. Finalmente se dejan las variables cuyo uso minimice el valor de Lambda de Wilks, o la relación entre las medias.

- **Resumen de funciones discriminantes canónicas:**

A partir de las siguientes dos pruebas se determina si es válido aplicar el análisis discriminante al conjunto de datos, ya que evalúan si hay diferencias entre cada grupo.

- **Autovalores:** Las correlaciones canónicas miden las desviaciones de las puntuaciones discriminantes entre grupos respecto a las desviaciones totales sin distinguir grupos. Si su valor es grande (próximo a 1) la dispersión será debida a las diferencias entre grupos, y en consecuencia, la función discriminará mucho. Cuanto más alto es el valor del autovalor, más eficaz será el análisis para clasificar a los sujetos. El valor mínimo es cero y no tiene un valor máximo.
- **Lambda de Wilks:** El estadístico Lambda de Wilks es el cociente entre la suma de cuadrados dentro de los grupos y la suma de cuadrados sin distinguir grupos. Esto equivale a las desviaciones a la media dentro de cada grupo entre las desviaciones a la media total sin distinguir grupos. Cuanto más cerca de 0 se encuentre mayor es el poder discriminante de las variables consideradas, y cuanto más cerca de 1 menor es el poder discriminante. Si el p valor < 0,05, conduce a rechazar la hipótesis nula de igualdad entre los dos vectores de medias. Es decir, las variables ejercen un efecto significativo en la separación de los grupos y el análisis discriminante es válido.

Los siguientes análisis tienen como finalidad conocer cuáles variables tienen mayor poder discriminatorio

- **Coeficiente de función discriminante canónica estandarizada:** Aparecen los coeficientes de la función discriminante canónica estandarizados. Estos coeficientes aparecen cuando se tipifican o estandarizan cada una de las variables clasificadoras para que tengan media 0 y desviación típica 1. De esta forma se evitan los problemas de escala que pudieran existir entre las variables y, consecuentemente, la magnitud de los coeficientes estandarizados son un indicador de la importancia que tiene cada variable en el cálculo de la función discriminante. Las comparaciones deben hacerse en porcentaje.
- **Matriz de estructuras:** Es conveniente conocer cuáles son las variables que tienen mayor poder discriminante en orden a clasificar a un individuo en uno de los grupos (estable, inestable). Una forma de medir ese poder discriminante es calculando el coeficiente de correlación entre cada una de las variables y la función discriminante. Esta es precisamente la información que se da en la tabla de matriz de estructuras. Las comparaciones deben hacerse siempre en valor absoluto. En el programa SPSS las variables aparecen ordenadas de acuerdo con el valor absoluto de los coeficientes de correlación.

Construcción de la(s) ecuación(es) discriminante(s): La ecuación o ecuaciones discriminantes se pueden construir de dos formas. La escogencia de cualquier forma no afecta el resultado final.

La primera forma de construir la ecuación es usando los valores de la tabla de coeficientes de la función discriminante canónica (no estandarizada). En el caso del primer ejercicio donde se ingresan todas las variables al análisis, la ecuación tendría la forma:

$$d = (-0.002 * DEM) + (0.140 * Pendiente) - (0.006 * Aspectos) + (3.227 * Profundidad\ suelo) + (Acumulacion * 0) - 7.314$$

Para encontrar el punto de corte (valor límite para considerar un individuo como perteneciente a un grupo), se calcula con el valor de los centroides (Figura 3). Para el caso anterior, sería:

$$C = \frac{D_1 + D_2}{2} = \frac{-2.33 - 2.33}{2} = 0$$

Para este caso en concreto, el punto de corte es 0. Eso quiere decir que celdas cuya ecuación discriminante d tenga como resultado un valor negativo, pertenece al grupo 0 (estable) y si el valor es positivo, pertenece al grupo 1 (inestable).

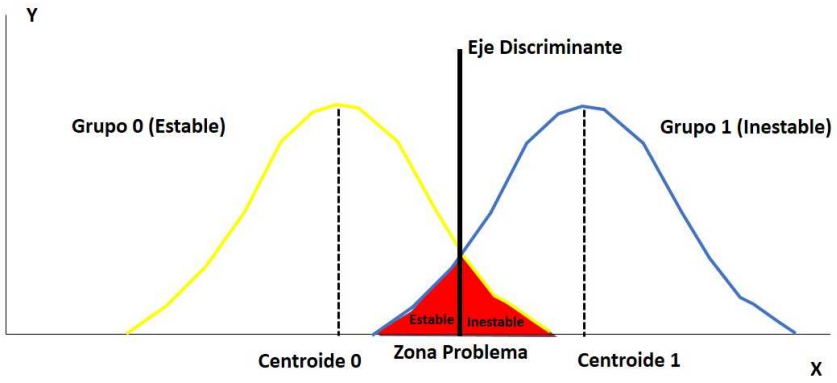

Figura 3. Representación del punto de corte definido usando los centroides de cada grupo.

La segunda forma de construir la regla discriminante es definir dos ecuaciones a partir de los valores de los coeficientes de función de clasificación (se encuentran en la sección de resultados de clasificación). Para el caso de ejemplo, estas ecuaciones serían:

$$F_0 = (0.036 * DEM) + (69.917 * Pendiente) - (0.024 * Aspectos) + (3249.052 * Profundidad\ suelo) + (0.010 * Acumulacion) - 4570.29$$

$$F_1 = (0.035 * DEM) + (69.982 * Pendiente) - (0.021 * Aspectos) + (3250.554 * Profundidad\ suelo) + (0.010 * Acumulacion) - 4574.233$$

Cada celda será asignada al grupo en el que obtenga un mayor valor de las funciones discriminantes.

- **Resultados de Clasificación:**

SPSS realiza una validación del desempeño de la ecuación discriminante. Para ello clasifica cada individuo según la ecuación discriminante y realiza la validación cruzada.

- **Clasificación con ecuación discriminante:** Cada individuo de la muestra es clasificado como estable o inestable usando la ecuación discriminante. Los resultados de la clasificación los puede ver en la tabla de datos como una nueva variable. El porcentaje de casos agrupados correctamente se muestran debajo de la tabla de resultados de clasificación.

- **Validación cruzada:** Se generan tantas funciones discriminantes como casos válidos tiene el análisis; cada una de esas funciones se obtiene eliminando un caso; después, cada caso es clasificado utilizando la función discriminante en la que no ha intervenido. El porcentaje de casos agrupados correctamente se muestra debajo de la tabla de resultados de clasificación.

EJERCICIO

Lleve a cabo el análisis discriminante en la zona de estudio usando las dos metodologías: Con la inclusión por pasos de las variables, y usando todas las variables en el análisis. Responda:

- ¿Cuáles variables tienen medias considerablemente diferentes en las celdas estables y en las inestables?
- ¿Cuáles variables tienen mayor poder discriminatorio a nivel unvariante? ¿Cuáles cumplen el supuesto de la prueba de igualdad de medias de grupos?
- ¿Cuáles variables del análisis se encuentran altamente relacionadas? ¿Cuál de ellas eliminaría del análisis?
- ¿Cuáles variables fueron eliminadas del análisis discriminante por pasos y cuáles fueron incluidas? ¿Tienen estos resultados relación con la prueba de igualdad de medias de grupo de cada variable? ¿Por qué una variable puede cumplir la prueba de igualdad de medias de grupo y aun así estar incluido dentro del análisis discriminante?
- ¿Qué valores obtuvo en la correlación canónica y Lambda de Wilks de la ecuación discriminante? ¿Se espera que la ecuación discriminante discrimine mucho o poco? ¿Es válido realizar el análisis discriminante con los datos que se le dieron?
- Organice las variables de mayor a menor incidencia en la estabilidad de la zona de estudio
- Para cada método, escriba las ecuaciones discriminantes. ¿Cuál método tuvo mejor desempeño?
- En ArcMap, cree un raster de estabilidad de la zona de estudio donde clasifique cada una de las celdas entre estable e inestable utilizando la ecuación discriminante del método con el que haya tenido mejor desempeño. Para esto, siga los siguientes pasos:
  - En el ArcToolbox abra Spatial Analyst Tools > Map Algebra > Raster Calculator. Agregue en la expresión una de las ecuaciones discriminantes y guarde el raster resultante. Haga lo mismo para la segunda ecuación discriminante. Llame a los raster F0 y F1 (Figura 4).
  - Usando el Raster Calculator, cree el raster final de estabilidad modelada, donde las celdas estimadas como estables tomen el valor de 0 y las inestables tomen el valor de 1. Para eso, cree una regla de decisión donde a cada celda se le asigne el valor de 1 o 0 dependiendo de cuál raster (F0 o F1) tomó un mayor valor en cada celda. Utilice el condicional Con, el cual tiene la siguiente estructura:

    ```
    Con(Condición, Valor si condición se cumple, Valor si condición no se cumple)
    ```

    Por ejemplo, para los raster de valores de las ecuaciones F0 y F1, el condicional es de la forma:

```
Con(F0>F1, 0, 1)
```

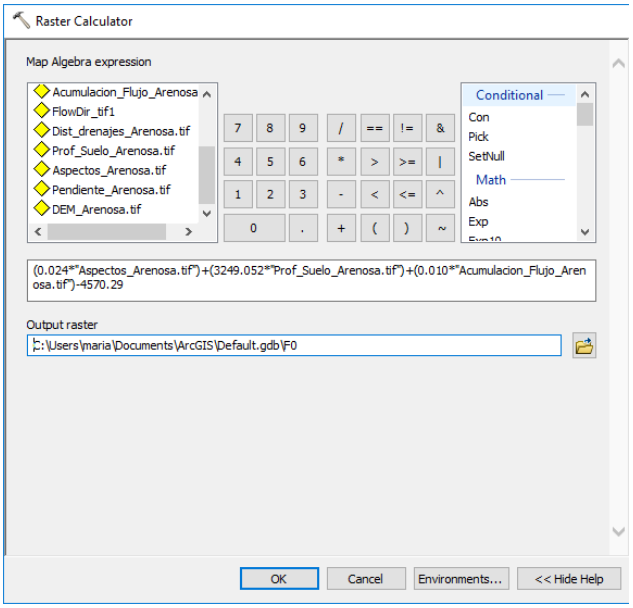

Figura 4. Cálculo en ArcGis de los puntajes discriminantes

- ¿Subestima o sobreestima el análisis la estabilidad de la zona de estudio? Calcule, usando el Raster Calculator el número de celdas estables clasificadas como estables (verdaderos positivos), celdas estables clasificadas como inestables (falsos negativos), celdas inestables clasificadas como estables (falsos positivos) y celdas inestables clasificadas como inestables (verdaderos negativos).

		Realidad	
		0	**1**
Analisis Discriminante	**0**		
	1		

Elaborado por: María Isabel Arango & Edier Aristizábal