Clasificación para leucocitos blastos, linfocitos reactivos y linfocitos normales*
Análisis avanzado de datos

Nicolás Rojas
MACC
Universidad del Rosario
@urosario.edu.co

Samuel Pérez
MACC
Universidad del Rosario
victor.perez@urosario.edu.co

Andrés Zapata
MACC
Universidad del Rosario
@urosario.edu.co

Sara Palacios
MACC
Universidad del Rosario
@sara.palaciosc@urosario.edu.co

Resumen

Para el `Abnormal blood cells`, hemos realizado un breve análisis estadístico. En primer lugar se muestran las características de los datos, sus distribuciones y relaciones entre si. Se ha tomado la mayor cantidad de descriptores para el desarrollo del análisis. A lo largo de todo el proyecto se toma como variable objetivo `tipoCelula` pues el objetivo es clasificar leucocitos (glóbulos blancos) Blastos (BL), linfocitos reactivos (RL), linfocitos normales (N) en función de sus características geométricas. Se han desarrollado modelos de regresiones logísticas, análisis a discriminante lineal, modelos de clasificación dentro del marco generalizado de GLM y GAM y finalmente un modelo de clustering usando modelos mixtos Gaussianos.
una regresión múltiple regularizada con `Lasso` y `Ridge` para obtener mejores resultados. Además propusimos modelos con ajustes polinómicos y splines naturales y finalmente usamos algunos métodos kernel para el suavizado de las regresiones.

Introducción

El `Abnormal blood cells` se trata de un conjunto de datos creado por el profesor Santiago Alférez, fue realizado mediante técnicas de visión por computador, sobre imágenes de células de sangre periférica. Este recopila información acerca de algunos leucocitos (glóbulos blancos) Blastos (BL), linfocitos reactivos (RL), linfocitos normales (N). Así como lo menciona el docente, sobre cada imagen de célula, se desarrolló un algoritmo de segmentación para extraer las regiones del núcleo, célula (completa) y una región circundarte que rodea a la célula. Con estas regiones se calcularon descriptores referentes a la geometría y forma de las regiones y, a su color y textura.

El dataset contiene 214 descriptores cada uno con 8830 observaciones. Con esta información, se realizaremos modelos como la clasificación logística, LDA, GLM, GAM y GMM. Tomaremos todos los descriptores para lograr tener una mejor clasificación para la variable de salida que en este caso sería `tipoCelula` es decir, clasificaremos leucocitos (glóbulos blancos) Blastos (BL), linfocitos reactivos (RL) y linfocitos normales (N) en función de sus características geométricas.

Para lo anterior usamos el lenguaje de programación `Python`, junto con las librerias expuestas en el Jupyter Notebook anexo.

Análisis exploratorio

Teniendo en cuenta que tenemos muchas variables para detallar, realizaremos el análisis exploratorio sobre dos grupos: célula y núcleo. Para cada grupo veremos las correlaciones y distribuciones de las variables `Area EquivDiameter, Eccentricity, Perimeter, Solidity, Extent, circularity, Elongation,roundnessCH, convexity, circleVariance, ellipVariance`.

Podemos, entonces, a través de las librerías importadas, visualizar algunas características de los datos. A través de `heatmap` podemos identificar las correlaciones de las variables ya sean positivas o negativas.

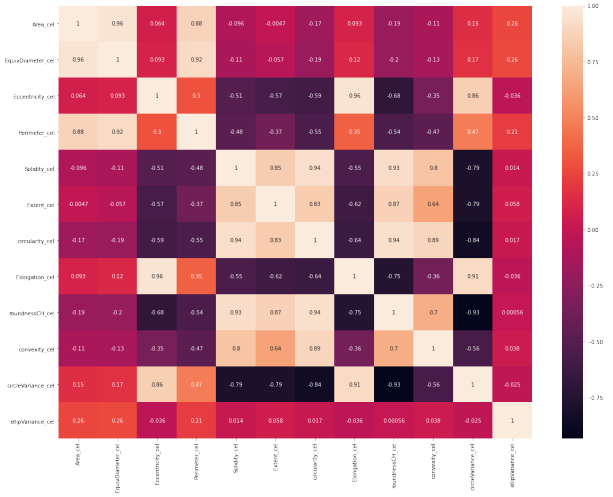

Figure 1: Heatmap para células

*Segundo proyecto del curso Análisis Avanzado de Datos de Matemáticas Aplicadas y Ciencias de la Computación de la Universidad del Rosario, 2020-2.

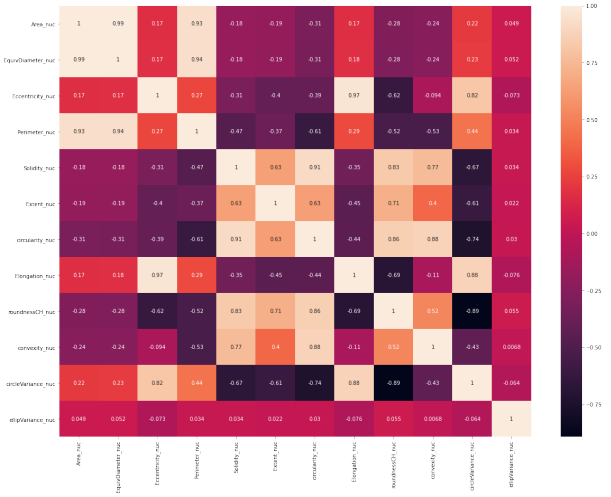

Figure 2: Heatmap para núcleos

Asi, podemos ver algunas correlaciones interesantes, en ambos casos la variable `circleVariance` esta relacionada (negativamente) con `roundness`, `circularity`, `extent` y `solidity`. Por otro lado, la variable `circularity` tiene una alta correlación (positiva) con las variables `convexity`, `roundnessCH`, `extent`, `solidity`. Por esto, hemos reducido un poco más estos datos para visualizar las distribuciones.

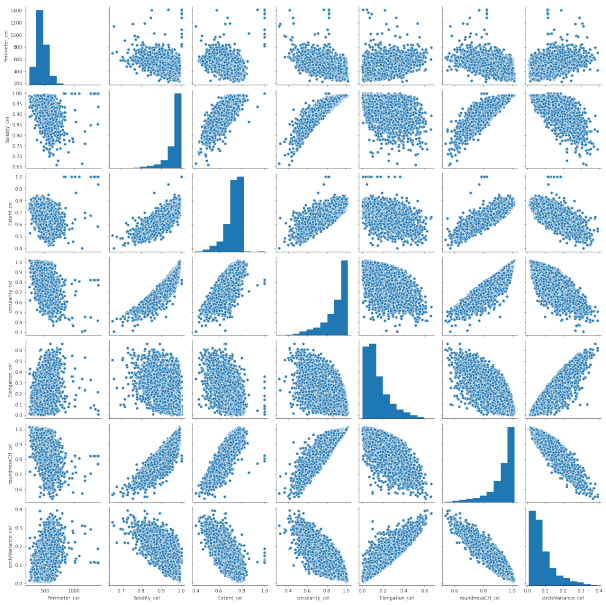

Figure 3: Pairplot para células

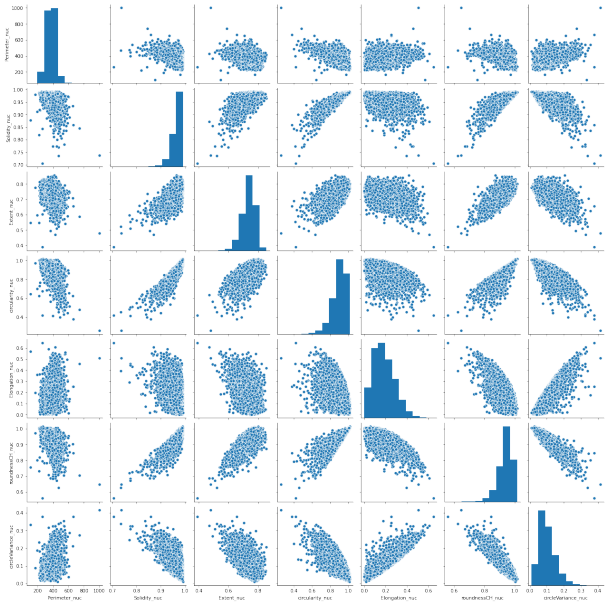

Figure 4: Pairplot para núcleos

Podemos ver cómo se distribuyen de manera más clara los valores en las diferentes variables. También es posible ver que existen unos puntos, para cada variable, que están muy lejanos del grupo principal, lo cual puede generar un poco de sesgo al realizar el análisis. Sin embargo, la gran mayoría de estas variables están muy correlacionadas, lo que nos lleva a tener en cuenta una gran cantidad de variables para plantear los modelos.

Finalmente podemos observar más detalladamente cada variable con sus histogramas:

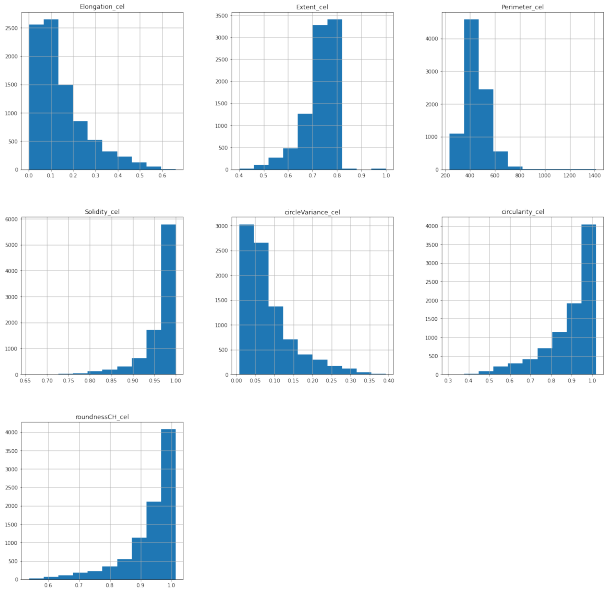

Figure 5: Histograma para células

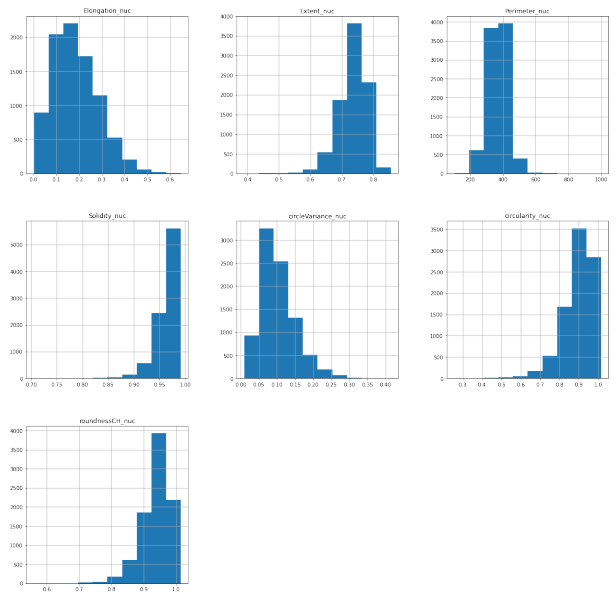

Figure 6: Histograma para núcleos

De aquí es posible observar que las variables `Enlongation y circleVariance` intentan tener una distribución normal. Además, los valores de `Solidity, roundness, circylarity y extent` tienen a ser muy altos(mayores a 0.5), por otro lado las variables `Elongation, circleVariance` toman valores muy bajos. Por último, la mayoría de valores de `Perimeter` se centran en valores entre 200 y 600.

Como se ha mencionado anteriormente, las variables tienen una alta relación entre si. Es por esto que para los modelos planteados en las siguientes secciones se utilizó la mayor parte del dataset. Se excluyeron las variables `entidad, historia, frotis, archivos` pues no describen los **tipos de células** (BL, RL, N) los cuales van a ser el foco de análisis.

Con esta contextualización de los datos, se busca tener un mejor entendimiento de nuestras manipulaciones en las siguientes secciones, en las cuales ya entraremos a tratar con modelos más sofisticados.

Clasificación con regresión logística y LDA

Se dividió el dataset en conjuntos de entrenamiento y de prueba, como se mencionó anteriormente, la variable objetivo es `tipoCelula` y se toman todos los descriptores del conjunto. Para realizar un mejor modelo se normalizaron las variables de entrada con *StandardScaler* del paquete *sklearn*. Luego, se hizo una comparación de p-values para encontrar los mejores descriptores. Si el p-value era menor a 0.01 entonces se consideraba importante para la clasificación de cada tipo de célula, así, encontramos que son 211 los que nos ayudan a hacer el modelo.

Una vez tomados los descriptores generamos los modelos de regresión logística y el análisis de discriminante lineal, ambos evaluados con validación cruzada de 10 iteraciones.

Así, comparamos los modelos según el accuracy de cada uno:

```
Accuracy para Regresión Logística: 0.9715260023735954
Accuracy para LDA: 0.9710405654803915
```

Figure 7: Accuracy

Como se logra ver, el accuracy para la regresión logística es un poco más alto que el de LDA, por lo tanto es mejor. Entonces, evaluamos la regresión logística en el conjunto de prueba y graficamos la matriz de confusión para nuestra variable objetivo.

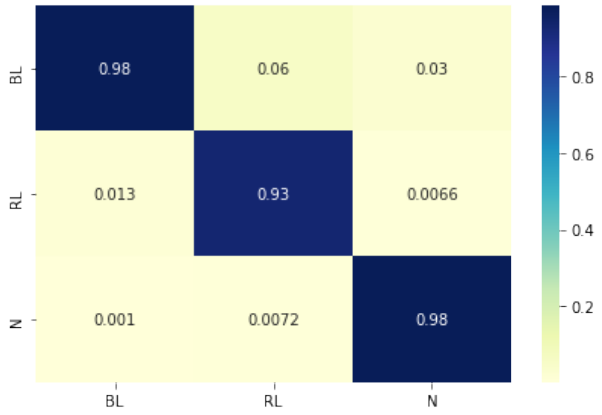

Figure 8: Matriz de confusión

Clustering con GMM

Para el desarrollo del modelo mixto Gaussiano, escogimos todos los descriptores así como lo hicimos en el anterior modelo. Teniendo en cuenta que queremos realizar la clasificación de cada tipo de célula, tomamos 3 componentes para el análisis, uno para cada tipo.

Una vez realizado el modelo hicimos un análisis con las métricas BIC, AIC y silhouette para encontrar la mejor `covariance_type` para el modelo mixto.

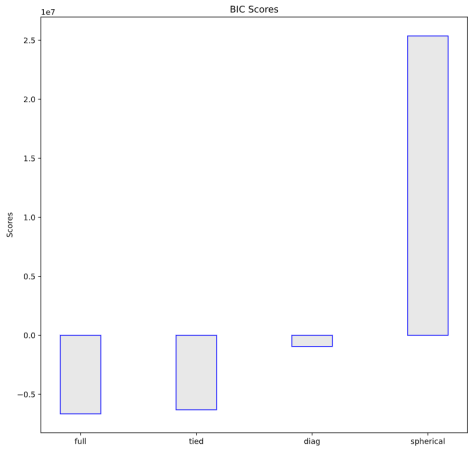

Figure 9: Métrica BIC

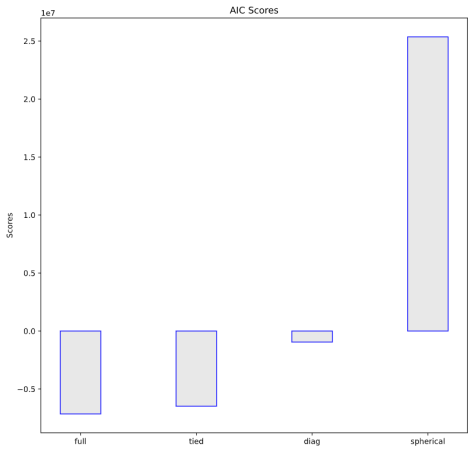

Figure 10: Métrica AIC

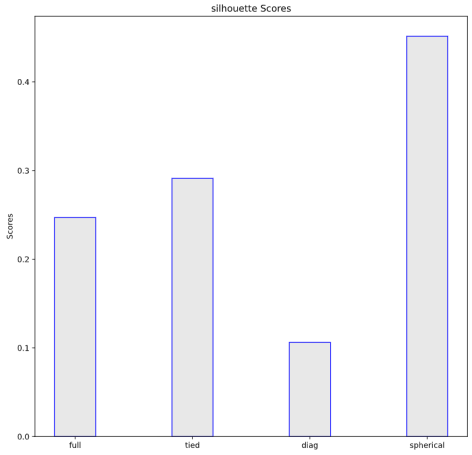

Figure 11: Métrica Silhouette

GLM y GAM

Finalmente se han creado modelos de clasificación dentro del marco generalizado de GLM y GAM. Se tuvieron el cuenta tres modelos, cada uno con variables específicas que serán listadas más adelante. En primer lugar se hizo un GAM, es decir, hicimos splines para cada variable para luego usar la función logit como función de enlace y así obtener un GLM clasificatorio.

Modelo 1

Este modelo se ajustó con las siguientes 9 variables para clasificar entre células que fueran linfocitos normales (N) y celulas que no lo fueran:`Hairiness`, `Area_cel`, `EquivDiameter_cel`, `Extent_cel`, `roundnessCH_cel`, `Eccentricity_nuc`, `Perimeter_nuc`, `Extent_nuc`, `roundnessCH_nuc`. Con ánimo de encontrar el mejor modelo para clasificar los linfocitos normales la variable objetivo tomará dos valores, uno para el caso en el que sea linfocito normal y otro en caso de que no lo sea. Luego realizaremos la clasificación tomando como descriptores las variables mencionadas anteriormente. Consiguiendo con este una precisión del 0.98 en el conjunto de entrenamiento y un 0.97 en el conjunto de prueba.

Los splines obtenidos para cada variable por este modelo fueron los siguientes.

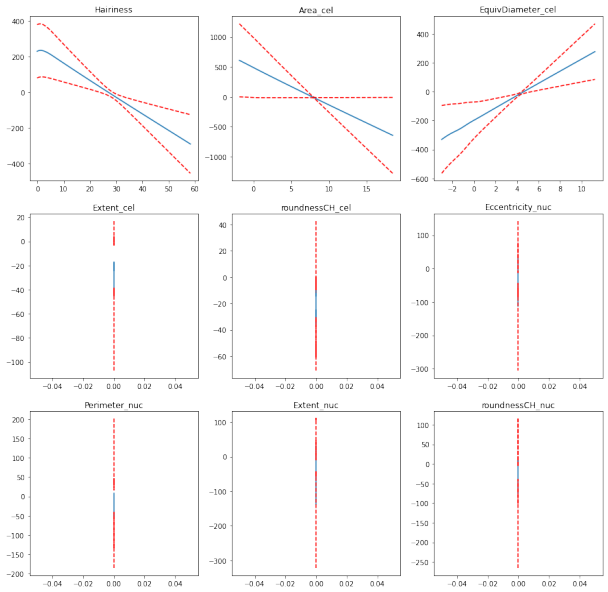

Figure 12: Modelo linfocitos normales

Modelo 2

Este modelo tomó las variables `Hairness`,`Area_cel`, `EquivDiameter_cel`, `Perimeter_cel`, `solidity_cel`, `circularity_cel`, `convexity_cel`, `elipVariance_cel`, `Area_nuc`, `EquivDiameter_nuc`, `Perimeter_nuc`, `circularity_nuc`,

`enlongation_nuc`, `convexiti_nuc`, `circleVariance_nuc`, `elipVariance_nuc`.
Con ánimo de encontrar el mejor modelo para clasificar los linfocitos reactivos la variable objetivo tomará dos valores, uno para el caso en el que sea linfocito reactivo y otro en caso de que no lo sea. Luego realizaremos la clasificación tomando como descriptores las variables mencionadas anteriormente. Consiguiendo una precisión para el conjunto de entrenamiento de 0.942 y un accuracy de 0.939 para el conjunto de prueba.

Los splines obtenidos para cada variable por este modelo fueron los siguientes.

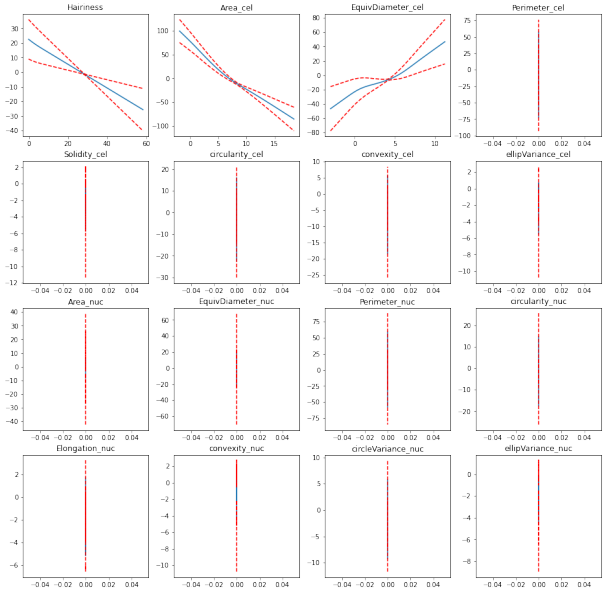

Figure 13: Modelo linfocitos reactivos

Modelo 3

Este modelo tomó las variables `Hairiness`,`Area_cel`, `EquivDiameter_cel`, `Perimeter_cel`, `circularity_cel`, `roundnessCH_cel`, `convexity_cel`, `ellipVariance_cel`, `Area_nuc`, `EquivDiameter_nuc`, `Perimeter_nuc`, `Solidity_nuc`, `Elongation_nuc`, `convexity_nuc`, `circleVariance_nuc`, `ellipVariance_nuc`.
Con ánimo de encontrar el mejor modelo para clasificar los leucocitos blastos la variable objetivo tomará dos valores, uno para el caso en el que sea leucocito blasto y otro en caso de que no lo sea. Luego realizaremos la clasificación tomando como descriptores las variables mencionadas anteriormente.Consiguiendo una precisión para el conjunto de entrenamiento de 0.923 y un accuracy de 0.9173 para el conjunto de prueba.
Los splines obtenidos para cada variable por este modelo se observan en la siguiente figura.

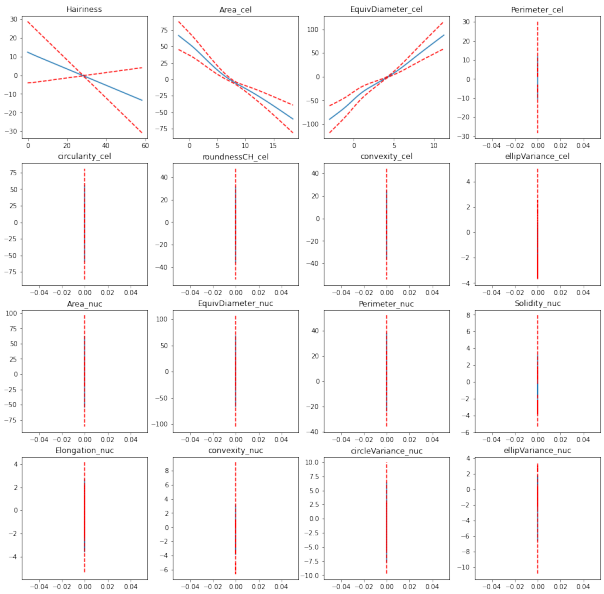

Figure 14: Modelo blastos

Conclusiones

En primer lugar, es posible concluir que para todos los casos, los descriptores elegidos(la gran mayoría) logran describir muy bien la variable objetivo. Es decir, los leucocitos (glóbulos blancos) Blastos (BL), linfocitos reactivos (RL), linfocitos normales (N) tienen características geométricas muy específicas que permiten clasificarlos muy bien.

Para el caso del análisis de discriminante lineal y regresión logística es posible concluir que el mejor modelo es el LDA pues obtiene un mejor accuracy y, por ser tan alto, se ajusta mejor a los datos.

Además, con las métricas tomadas para el modelo GMM podemos concluir que las métricas BIC y AIC nos sugieren que el mejor modelo es con el `covariance_type='full'`. Sin embargo, es interesante ver que en el caso de la métrica silhouette, se infiere que el modelo más consistente es con `covariance_type='spherical'`. Esto se puede ser causado debido a un sobrelapamiento en los clusters, haciendo que las clasificaciones se hagan de una "buena manera", esta suposición tiene en cuenta las anteriores métricas donde esta configuración tiene un score excesivamente alto.

Dentro de los modelos utilizando GAM y GLM se obtuvieron altos valores de accuracy en ambos conjuntos (train y test) utilizando pocas variables. Por otro lado, los splines conseguidos por el modelo GAM no son muy congruentes con los datos, para variables distintas a `Hairiness`, `Area_cel`, `EquivDiameter_cel`, las cuales son las tres primeras variables en común de los tres modelos.