Clasificación sobre dataset

Y. Sarahi García González[*]

Centro de Investigación en Matemáticas CIMAT[**]

07 de Junio de 2024

En este proyecto, se exploró el potencial de varios métodos de clasificación multiclase sobre el conjunto de datos Dry Bean, el cual consta de 13,611 instancias de siete tipos diferentes de frijoles secos, caracterizados por 16 extraídas de imágenes de alta resolución. Las características incluyen 12 dimensiones y 4 descriptores de forma (área, perímetro, longitud del eje mayor y redondez). Se utilizaron los clasificadores Random Forest, Support Vector Machine (SVM) y Regresión Logística. Además, a través de Grid Search con Cross-Validation se realizó un ajuste de hiperparámetros y se hizo Análisis de Componentes Principales (PCA) para la reducción de dimensionalidad. Se incluyen métricas de rendimiento y herramientas de visualización que evaluan evaluar la efectividad de cada modelo. En los resultados se destacan las fortalezas y limitaciones de cada enfoque.

I. INTRODUCCIÓN

El conjunto de datos Dry Bean es una colección de imágenes de alta resolución de frijoles secos, diseñada para facilitar el desarrollo de modelos avanzados de clasificación. El conjunto de datos incluye imágenes de 13,611 frijoles de siete variedades registradas: Seker, Barbunya, Bombay, Cali, Dermason, Horoz y Sira. Cada imagen de frijol fue procesada para extraer 16 características distintivas.

```
DataFrame Info:

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 13611 entries, 0 to 13610
Data columns (total 17 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Area             13611 non-null  int64
 1   Perimeter        13611 non-null  float64
 2   MajorAxisLength  13611 non-null  float64
 3   MinorAxisLength  13611 non-null  float64
 4   AspectRation     13611 non-null  float64
 5   Eccentricity     13611 non-null  float64
 6   ConvexArea       13611 non-null  int64
 7   EquivDiameter    13611 non-null  float64
 8   Extent           13611 non-null  float64
 9   Solidity         13611 non-null  float64
 10  roundness        13611 non-null  float64
 11  Compactness      13611 non-null  float64
 12  ShapeFactor1     13611 non-null  float64
 13  ShapeFactor2     13611 non-null  float64
 14  ShapeFactor3     13611 non-null  float64
 15  ShapeFactor4     13611 non-null  float64
 16  Class            13611 non-null  int64
dtypes: float64(14), int64(3)
memory usage: 1.8 MB
```

Figura 1. DryBean consta de 13311 muestras etiquetadas. Hay siete clases distintas. Todas las caracteríticas son números enteros o flotantes.

Estas características capturan aspectos críticos de la forma del frijol por lo que es recurso valioso para tareas de clasificación.

En este contexto, en el presente proyecto se emplean varias técnicas, procurando optimizar el rendimiento y precisión mediante el ajuste de hiperparámetros y selección de características.

La evaluación del rendimiento se realizó a través de métricas como precisión, recall y F1-score, junto con matrices de confusión para visualizar los errores de clasificación.

II. ANÁLISIS EXPLORATORIO

En la Figura 2 se muestra la distribución de lasdistintas características del conjunto de datos Dry Bean.

Algunas características, como Área, Perímetro, Longitud del Eje Mayor, Longitud del Eje Menor y Área Convexa, presentan distribuciones sesgadas a la derecha, indicando que la mayoría de los frijoles tienen valores menores para estas características. El Diámetro Equivalente y la Redondez también muestran una ligera asimetría a la derecha.

En contraste, la Relación de Aspecto y la Excentricidad presentan distribuciones más equilibradas, sugiriendo una variabilidad más uniforme en estas características entre los diferentes tipos de frijoles. De manera similar, los Factores de Forma 1 y 2 tienen distribuciones relativamente simétricas. Por otro lado, la Solidez y el Factor de Forma 4 tienen valores altamente concentrados en un rango estrecho, indicando poca variabilidad entre los frijoles para estas características.

Las distribuciones sesgadas sugieren la posible necesidad de normalizar el conjunto para mejorar el rendimiento del modelo.

[*] yesenia.garcia@cimat.mx

[**] REP, a cargo de Dr. Johan Van horebeek

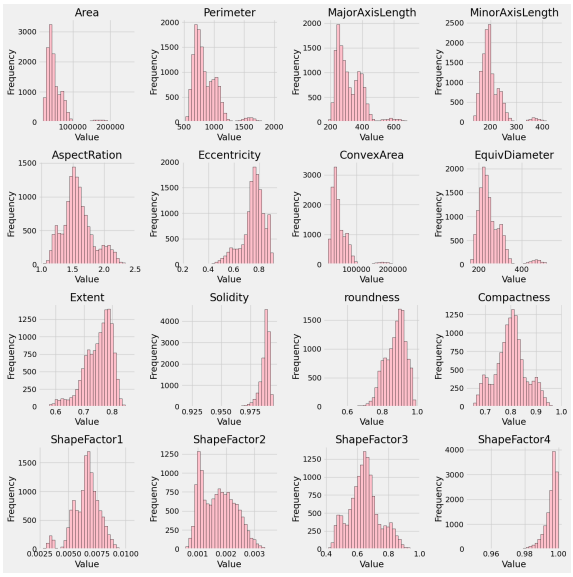

Figura 2. Histogramas de cada una de las 16 características del conjunto de datos Dry Bean

El rendimiento de los modelos podría verse afectado dado que hay varías características con distribuciones sesgadas, lo que sugiere la posible necesidad de estandarizar los datos para así reducir el sesgo. Por otro lado, las características con poca variabilidad (Solidez y Factor de Forma 4) podrían ser menos útiles para la clasificación.

A continuación en la Figura 3 se encuentra la gráfica de correlación de todas las variables.

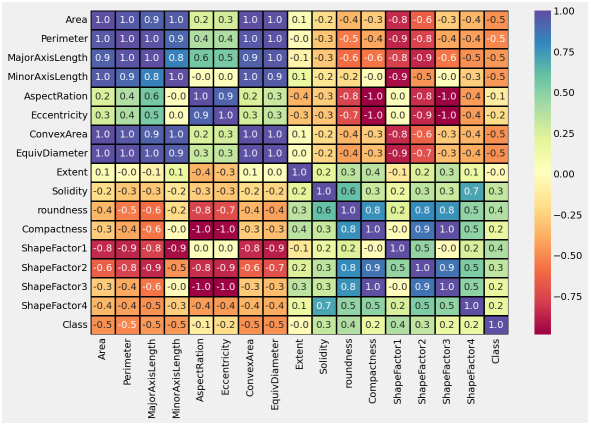

Figura 3. Correlación de pearson entre las variables del dataset DryBean

Las fuertes correlaciones positivas entre características (como el Área, el Perímetro y la Longitud del Eje Mayor) sugieren redundancia en la información aportada por ellas. Por lo que técnicas de reducción de dimensionalidad (como PCA) pueden ser útiles para simplificar el modelo y eliminar esa redundancia.

En cuanto a las etiquetas, se muestra Figura 4, una la gráfica de barras de las siete clases del dataset, en orden decreciente: DERMASON, SIRA , SEKER , HOROZ , CALI , BARBUNYA y BOMBAY.

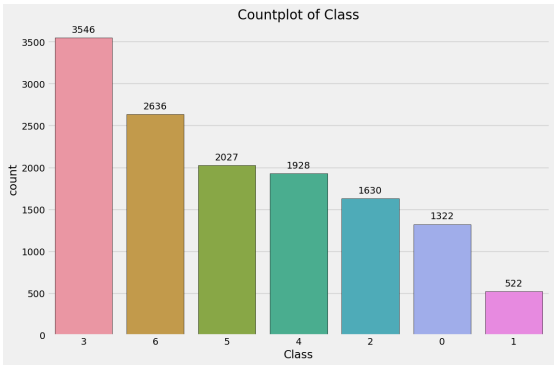

Figura 4. Gráfica de barras de las 6 clases

La distribución revela un claorr desbalance entre las clases. La clase Seker es la más prevalente (3546 instancias). Las variedades Barbunya, Cali y Horoz presentan frecuencias similares (entre 1928 y 2027). La clase Bombay es la menos representada con solo 522 instancias.

Debido al desbalance será importante considerar métricas de evaluación sensibles al desbalance, como el recall, la precisión y el F1-score, además de la precisión general.

III. RESULTADOS

Debido a la fuerte correlación observada entre varias variables en 3, se aplicó PCA a los datos. Los resultados de la varianza explicada acumulada se pueden observar a continuación:

Número de Componente	Varianza Explicada Acumulada
1	0.554664
2	0.818974
3	0.899040
4	0.950181
5	0.977573
6	0.989071
7	0.996048
8	0.999298
9	0.999815
10	0.999906
11	0.999971
12	0.999990
13	0.999999
14	1.000000
15	1.000000
16	1.000000

Figura 5. Tabla de la varianza explicada acumulada de cada una de las 16 compoentes.

Se tomó un umbral del 98 %, de manera que los siguientes modelos se ajustaron tomando en cuenta sólo 6 caracte

Para realizar el ajuste de hiperparámetros de cada modelo se dividió en dataset en tres conjuntos de Train, Val y Test. El procedimiento general fue:

1. **Definición de la Grid**: Posibles combinaciones de valores para cada modelo basado en las características de éste y en las del dataset.
2. **Configuración de GridSearch**: Se configuró *GridSearchCV* de sklearn python y Cross-Validation de 5 pliegues para evaluar cada combinación de hiperparámetros.
3. **Entrenamiento y Evaluación**: Se entrenó cada modelo con las distintas las combinaciones de hiperparámetros en el conjunto de Validación.
4. **Selección de Mejores Hiperparámetros**: La combinación de hiperparámetros que proporcionó el mejor rendimiento, según la métrica de exactitud, se seleccionó para el modelo final que se entreno en el conjunto Train.

Random Forest

El clasificador Random Forest, con una exactitud global del 92 %, demuestra ser particularmente efectivo en la mayoría de las clases. Las clases 1 y 4 alcanzan una precisión y recall perfectos. Sin embargo, la clase 6 presenta un desempeño ligeramente inferior con un F1-score de 0.87.

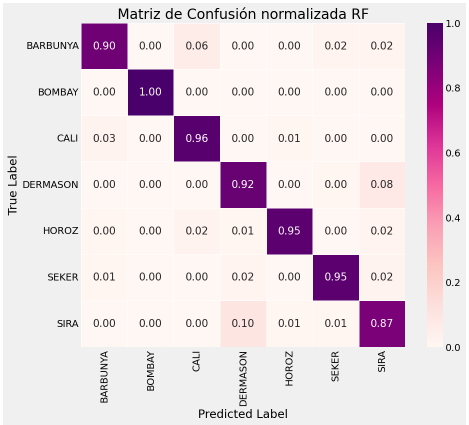

Figura 6. Matriz de Confusión para el método Random Forest

Clase	Precisión	Recall	F1-score	Soporte
0	0.94	0.90	0.92	261
1	1.00	1.00	1.00	117
2	0.92	0.96	0.94	317
3	0.91	0.92	0.91	671
4	0.97	0.95	0.96	408
5	0.97	0.95	0.96	413
6	0.86	0.87	0.87	536
Accuracy			0.92	

Cuadro I. Resultados del Random Forest con los mejores parámetros

SVM

El clasificador SVM, con una exactitud global del 93 %, también muestra un alto rendimiento excelente. Similar al Random Forest, las clases 1 y 4 alcanzan valores perfectos de precisión y recall. La muestra un F1-score de 0.89, lo que es una ligera mejora sobre el rendimiento del Random Forest.

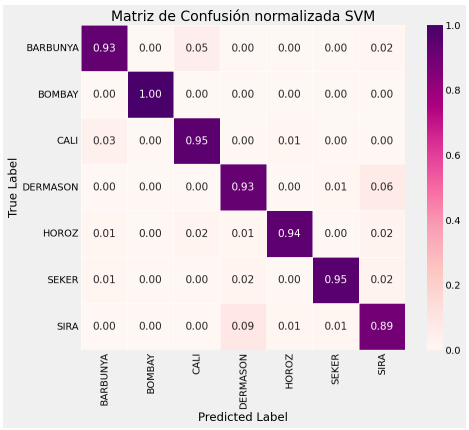

Figura 7. Matriz de Confusión para el método SVM

Clase	Precisión	Recall	F1-score	Soporte
0	0.93	0.93	0.93	261
1	1.00	1.00	1.00	117
2	0.94	0.95	0.94	317
3	0.91	0.93	0.92	671
4	0.97	0.94	0.96	408
5	0.97	0.95	0.96	413
6	0.88	0.89	0.89	536
Accuracy			0.93	

Cuadro II. Resultados del SVM con los mejores parámetros

Regresión Logística

El clasificador de Regresión Logística muestra un desempeño comparable con una exactitud global del 92 %. Las

clases 1 y 4 nuevamente alcanzan altos valores de precisión y recall. La clase 0 y la clase 2 también presentan altos F1-scores de 0.91 y 0.94. Sin embargo, la clase 6 vuelve a tener un F1-score bajo de 0.87.

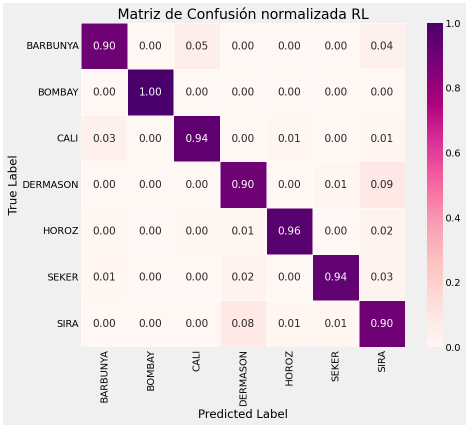

Figura 8. Matriz de Confusión para el método de Regresión Logística

Clase	Precisión	Recall	F1-score	Soporte
0	0.93	0.90	0.91	261
1	1.00	1.00	1.00	117
2	0.95	0.94	0.94	317
3	0.92	0.90	0.91	671
4	0.97	0.96	0.97	408
5	0.96	0.94	0.95	413
6	0.84	0.90	0.87	536
Exactitud		0.92		

Cuadro III. Resultados de la Regresión Logística con los mejores parámetros

Finalmente, la figura 9 muestra una idea de como es la frontera de desición de cada método. Para generar las gráficas se ajustaron los modelos (con los mejores hiperparámetros) sobre el conjunto train restringido a dos features.

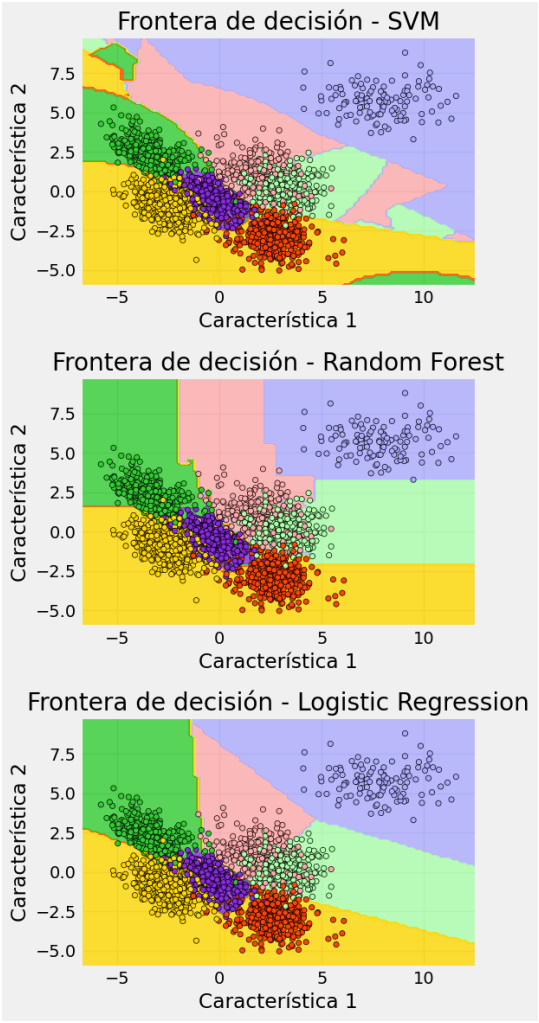

Figura 9. Frontera de desición en 2D

Podemos apreciar que a pesar de toda la información que no se muestra y que no se utilizó para ajustar los modelos, se capturan bien casi todas las clases.

IV. CONCLUSIONES

En resumen, los clasificadores evaluados:Random Forest, SVM y Logistic Regression: muestran un alto rendimiento en la tarea de clasificación multiclase del conjunto de datos Dry Bean. El SVM fue ligeramente superior en términos de precisión global.

Por otro lado, como era de esperar la menor precisión se alcanzo en el la clase 6 (BOMBAY) que es la menos reprsentada, por lo que deben considrarse mejoras al ajustar los modelos, por ejemplo, técnicas de balanceo tales como el sobremuestreo de las clases minoritarias o el submuestreo de las clases mayoritarias.

REFERENCIAS

1. Machine Learning Repository: Dry Bean Dataset https://archive.ics.uci.edu/dataset/602/dry+bean+dataset
2. Multinomial Logistic Regression.Recuperado de https://rpubs.com/fhernanb/mlr
3. Random Forest with python. Recuperado de https://cienciadedatos.net/documentos/py08_random_forest_python