# Clasificación y Clustering de Granos Secos

1ero Victor Herrera
*Estudiante de Ingeniería en Sistemas*
*Universidad de Cuenca*
Cuenca, Ecuador
victor.herrera@ucuenca.edu.ec

2do Pablo Solano
*Estudiante de Ingeniería en Sistemas*
*Universidad de Cuenca*
Cuenca, Ecuador
pablo.solanoc98@ucuenca.edu.ec

***Resumen*—La clasificación de granos es una tarea importante para el cultivo selectivo y la comercialización de los mismos. Esta tarea puede ser automatizada mediante técnicas de Visión por Computador y Machine Learning. En este trabajo se aplican y se comparan varios modelos de clasificación como SVM, MLP, XGB y KNN para determinar que tipo de clasificador es el mejor para realizar esta tarea. Por otro lado, se aplican y se comparan modelos de clustering como AC, K-means y HDBSCAN para determinar que método permite encontrar agrupamientos lo más cercanos posibles a la clasificación real de los granos. Como resultados principales se obtienen que SVM es el clasificador más adecuado para este problema y que de los métodos de clustering analizados ninguno se acerca lo suficiente a la clasificación real de los granos.**

***Index Terms*—Clasificación de granos secos, Clusterización de granos secos, Procesamiento de Imágenes, Técnicas de Machine Learning**

## I. Introducción

En la última década, el Ecuador ha tenido un resurgimiento en la producción agrícola, lo que ha hecho que la producción de granos en el Ecuador se eleve a un nivel cada vez más importante dentro la producción agrícola de la Sierra [1]. Esto se debe y se evidencia en las mejoras en tecnología de cultivo y pos-cosecha, y en el creciente interés del lado nacional e internacional por este tipo de alimentos saludables, lo que a su vez, ha llevado al continuo aumento del precio de estos productos [2].

En base a lo anterior, y debido a que la comercialización de granos secos basados en sus características incrementa el valor de estos en el mercado, considerando que las cualidades de las semillas son esenciales en el cultivo, pues el uso de semillas poco adecuadas provoca una menor cantidad de cosecha incluso si se proporcionan todas las condiciones óptimas de cultivo, se cree que la clasificación de las variedades de granos integrados en el proceso de producción del Ecuador puede ser de gran ayuda para productores agrícolas con necesidades de formar un estándar base para plantaciones y mercadeo [3].

Con el objetivo de conseguir un método automático de identificación/clasificación, se propone usar las cualidades de varios tipos de granos, obtenidas de imágenes por medio de técnicas de Visión por Computador (CV), consiguiendo medidas o cualidades esenciales de cada tipo de grano a identificar como: área, perímetro, solidez, redondez, etc. Estas características constituirán las entradas al problema.

Con estas medidas obtenidas se busca aplicar y comparar varios modelos de Aprendizaje de Máquina (ML) para la clasificación de cada uno de los granos que fueron procesados. La clasificación de los granos (es decir las clases indicadas) corresponden a la salida del problema. Adicionalmente, se propone utilizar técnicas de clusterización como una alternativa a los métodos de clasificación, para observar como estos se ajustan a las clases reales y poder concluir si estos se podrían usar en problemas similares en donde no existan etiquetas.

Por lo tanto, el objetivo de este trabajo es determinar el mejor modelo de clasificación posible, para la identificación de tipos granos en base a características geométricas y de forma. Además, se pretende usar clusterización con la intención de conseguir un modelo no supervisado capaz de agrupar dichos tipos de granos de una forma lo más cercana posible a la verdad fundamental.

## II. Trabajos Relacionados

El problema de la clasificación de granos de cualquier especie usando procedimientos de CV y ML requiere del uso de diferentes técnicas según la clase de problema que se requiera resolver. Distintos enfoques pueden ser utilizados dependiendo de, por ejemplo, la variabilidad que exista entre granos de una misma o distinta clase [4].

De una revisión de literatura comprensiva del problema de la clasificación de granos [4], se puede extraer que los distintos enfoques para la resolución de este problema varían a través de distintas aristas como el tipo de imágenes utilizado (HSV, RGB, etc.), los procesos de segmentación (detección de bordes, clustering, etc.), los modelos de clasificación (KNN, SVM, etc.) y la aplicación (identificación, análisis de calidad, etc.).

A nivel internacional, se han abordado enfoques para la clasificación de distintos granos de maíz procesando imágenes digitales para la obtención de características híbridas (histograma, análisis espectral, etc.) para posteriormente alimentar modelos como Random Forest (RF) o Multi Layer Perceptron (MLP) [5]. Por otro lado, también se han abordado enfoques con tecnología Near-Infrared Hyperspectral Imaging y Deep Learning (DL) para la clasificación de granos de arroz híbridos de calidad [6].

A nivel nacional, existen trabajos que se enfocan en clasificar granos de cacao frescos, usando clustering K-means para remover el fondo de las imágenes y Support Vector Machine (SVM) como modelo de clasificación [7]. Además, otros trabajos se centran en comparar varias técnicas de CV

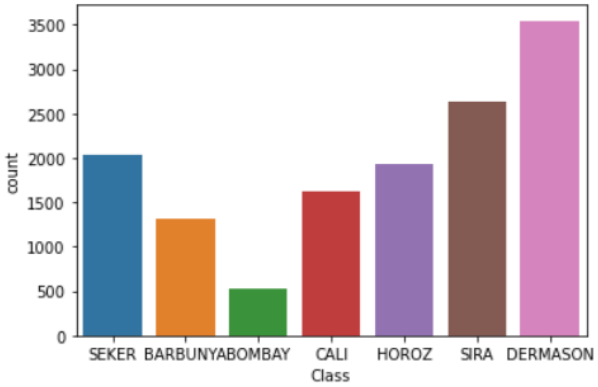


Figura 1. Distribución de Clases.

para determinar el grado de fermentación de granos de cacao a partir de la información de color en formato RGB [8].

Por último, es importante destacar el trabajo realizado para lograr una clasificación mutliclase de frijoles secos usando modelos de clasificación como SVM, MLP, Árboles de Decisión (DT) y K-nearest neighbors (KNN) [3]. De este último trabajo se obtiene el dataset que se utiliza para la realización del trabajo descrito en este documento.

## III. Descripción del Dataset

El dataset utilizado proviene de un estudio realizado en Turquía para la clasificación de granos [3]. El dataset consiste en las características geométricas y de forma que se obtuvieron de un conjunto de imágenes usando técnicas de CV. El dataset, que contiene 13,611 instancias, se encuentra limpio y procesado de manera que sus 16 atributos contienen valores numéricos únicamente, exceptuando las etiquetas de clase que corresponden a 7 tipos de granos secos de la familia de las fabáceas (Burbunya, Bombay, Cali, Dermason, Horoz, Seker, Sira), cuya distribución se puede observar en la Figura 1.

La Figura 1 permite observar que se trata de un dataset que presenta un problema de desbalanciamiento de clases. Este hecho se considerará más adelante para la definición de métricas de rendimiento y para las técnicas de validación de modelos.

Respecto de los 16 atributos numéricos, una descripción de los mismos en términos de mínimo, máximo, media y desviación estándar se puede observar en el Cuadro I. Aquí se puede observar que los valores numéricos de los 16 atributos tienen distintas escalas. Para evitar que existan atributos que dominen el problema debido a la diferencia de escalas, se realiza un escalamiento de normalización entre mínimo y máximo de 0 a 1.

Por otro lado, es útil realizar un análisis de correlación entre atributos para poder visualizar la influencia que tiene un atributo sobre otro, y en especial sobre la variable de etiqueta de clase. En la Figura 2 se puede observar un diagrama de correlación para estos atributos.

En este diagrama se pueden observar la intensidad de la correlación entre atributos, donde un azul intenso indica una

Cuadro I
Descripción estadística de atributos

| | *media* | *std* | *min* | *max* |
|---|---|---|---|---|
| ***Area*** | 53048.284549 | 29324.095717 | 20420 | 254616 |
| ***Perimeter*** | 855.283459 | 214.289696 | 524.736 | 1985.37 |
| ***MajorAxisLength*** | 320.141867 | 85.694186 | 183.601165 | 738.860153 |
| ***MinorAxisLength*** | 202.270714 | 44.970091 | 122.512653 | 460.198497 |
| ***AspectRation*** | 1.583242 | 0.246678 | 1.024868 | 2.430306 |
| ***Eccentricity*** | 0.750895 | 0.092002 | 0.218951 | 0.911423 |
| ***ConvexArea*** | 53768.200206 | 29774.915817 | 20684 | 263261 |
| ***EquivDiameter*** | 253.06422 | 59.17712 | 161.243764 | 569.374358 |
| ***Extent*** | 0.749733 | 0.049086 | 0.555315 | 0.866195 |
| ***Solidity*** | 0.987143 | 0.00466 | 0.919246 | 0.994677 |
| ***roundness*** | 0.873282 | 0.05952 | 0.489618 | 0.990685 |
| ***Compactness*** | 0.799864 | 0.061713 | 0.640577 | 0.987303 |
| ***ShapeFactor1*** | 0.006564 | 0.001128 | 0.002778 | 0.010451 |
| ***ShapeFactor2*** | 0.001716 | 0.000596 | 0.000564 | 0.003665 |
| ***ShapeFactor3*** | 0.64359 | 0.098996 | 0.410339 | 0.974767 |
| ***ShapeFactor4*** | 0.995063 | 0.004366 | 0.947687 | 0.999733 |

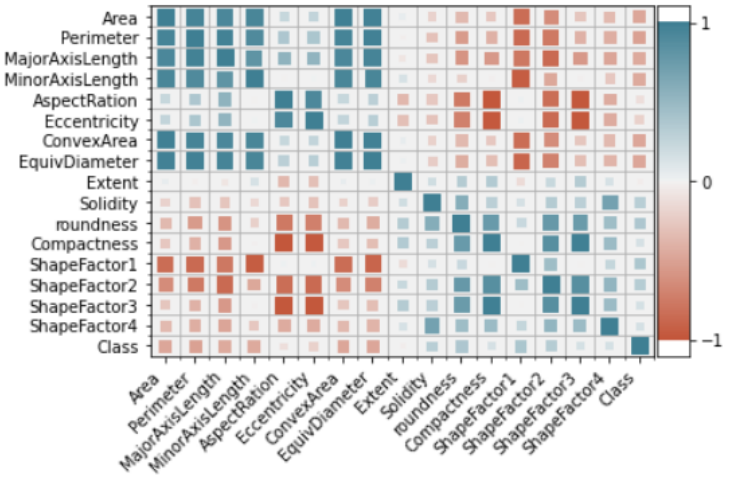


Figura 2. Diagrama de Correlación entre Atributos.

correlación directamente proporcional (positiva) entre atributos, y un naranja intenso indica una correlación inversamente proporcional (negativa) entre atributos. Los colores menos intensos indican relaciones de correlación débiles en cualquiera de ambos sentidos.

Se puede observar, por ejemplo, que los atributos geométricos como área, perímetro, longitud de ejes mayor y menor están altamente correlacionadas entre sí de manera positiva. Por otro lado, los atributos como la compacidad y excentricidad están altamente correlacionadas entre sí de manera negativa.

No obstante, se puede observar que ningún atributo está altamente correlacionado ya sea de manera positiva o negativa, con la etiqueta de clase. Es por esto, y a pesar de las correlaciones entre atributos anteriormente descritas, que se decide no descartar ningún atributo para la aplicación de tareas de ML. Adicionalmente, según el estudio de donde sale el dataset [3], todos estos atributos son parte de las características más importantes a considerar para la clasificación de granos secos de este tipo.

## IV. METODOLOGÍA

Todo el trabajo aquí descrito se realiza utilizando Python 3 y sus librerías para ML como sklearn. La metodología a aplicar en este trabajo se divide en dos flujos principales. El primero, se centra en aplicar y comparar modelos de clasificación, para escoger el mejor modelo en base a evaluaciones de hipótesis para comparar el rendimiento de los modelos. El segundo, busca aplicar y comparar modelos de clusterización para encontrar un agrupamiento lo más cercano posible a la verdad fundamental. En las siguientes subsecciones se explican a detalle cada uno de los flujos anteriomente descritos.

### *IV-A. Clasificación*

Para la aplicación y comparación de modelos de clasificación primero se debe considerar la división del dataset para efectos de entrenamiento, validación y prueba. Para esto, lo primero que se hace es dividir el dataset en un set de entrenamiento (70 %) y un set de prueba (30 %). Esta división se hace de manera estratificada para mantener la distribución de clases original en ambos datasets.

Luego, se debe considerar que para la validación durante el entrenamiento de los modelos, se aplica cross-validation al dataset de entrenamiento (70 %). Esto se hace usando un enfoque Stratified 10-fold cross-validation para tratar con el desbalanciamiento de clases.

*IV-A1. Stratified K-fold cross-validation:* Con Stratified K-cross validation el dataset se divide en K folds de forma pseudoaleatoria asegurándose de mantener la proporción o distribución de clases en cada fold [9]. En este caso se utiliza Stratified 10-fold cross-validation para usar el 10 % del dataset para validación en cada iteración.

Esta técnica de validación se puede usar con una técnica de tuneo de hiper parámetros como Grid Search o Random Search [10], para encontrar los mejores valores posibles de los hiper parámetros de cada modelo. Para esto, se debe considerar que en este trabajo se utilizan 4 modelos de clasificación de distintos tipos.

*IV-A2. Modelos Utilizados:* A continuación se detallan los modelos utilizados:

- Support Vector Machines (SVM)
- Multi Layer Perceptron (SVM)
- eXtreme Gradient Boosting (XGB)
- K-nearest neighbors (KNN)

Se escogieron métodos de distinta naturaleza ya que se desea comprobar que tipo de clasificador funciona mejor para este problema de clasificación de granos secos.

*IV-A3. Tuneo de hiper parámetros:* Para el tuneo de hiper parámetros se realizan procesos de Grid Search o Random Search (dependiendo la complejidad computacional del modelo) con el mismo Stratified 10-fold cross validation para cada modelo. En el Cuadro II se puede observar el tuneo de hiper parámetros realizado para cada modelo.

Cabe destacar que la medida de rendimiento utilizada para la determinación de los mejores hiper parámetros en el proceso de tuneo de los modelos es Balanced Accuracy. Se usa esta medida ya que la misma permite realizar una buena evaluación al lidiar con el problema de clases desbalanceadas [11].

Lo anterior lleva al hecho de qué métricas de rendimiento utilizar para evaluar el desempeño de cada uno de los modelos con los mejores hiper parámetros encontrados. La medida común para evaluar el rendimiento de un modelo es el Accuracy, sin embargo, en un problema con clases desbalanceadas esta medida podría no ser suficiente.

*IV-A4. Métricas de Rendimiento:* Para hacer una mejor evaluación de los modelos se consideran las métricas que pueden ser calculadas a partir de una matriz de confusión como el Balanced Accuracy, Precision, Recall y F1-Score. Estas tres últimas métricas pueden ser calculadas para cada clase, pero también se puede calcular un promedio de las mismas. Para este estudio, se considera el promedio ponderado (weighted) de estas tres métricas ya que este considera el desbalanciamiento de clases [11].

El uso de estas métricas ayudan a evaluar el rendimiento de los modelos y podrían ayudar a compararlos de una manera básica. No obstante, escoger un modelo como el mejor en base a una comparación de estas métricas solamente podría no ser lo más adecuado. Para solventar este problema, en este estudio se hace uso de pruebas estadísticas.

*IV-A5. Pruebas Estadísticas:* Para la comparación de los modelos se realiza un test de McNemar para los resultados obtenidos en el data set de prueba por cada par de modelos [12]. Con esta prueba se espera obtener al menos un par de modelos de los que pueda decirse que son mejores al menos que otro modelo. Con estos modelos se procedería a realizar un desempate utilizando una prueba estadística 5x2cv paired t test [12]. Para este prueba se usa la métrica Balanced Accuracy para la evaluación de hipótesis.

*IV-A6. Intento de Mejora con Reducción de Dimensiones:* Una vez se ha determinado cuál es el mejor modelo de clasificación para el problema (al menos en base a las pruebas estadísticas mencionadas anteriormente), se busca comprobar si los resultados del mejor modelo pueden ser mejorados aplicando transformaciones de reducción de dimensiones al dataset.

Para este estudio, se consideran tres tipos de transformaciones de reducción de dimesiones para probar: PCA (al menor número de dimensiones que permita explicar la mayor varianza posible), uno de los métodos más antiguos para la reducción de dimensiones [10], t-SNE (a dos dimensiones) una técnica moderna diseñada en especial para la visualización de datos [10]; y una transformación mixta PCA + t-SNE para comprobar su funcionamiento en conjunto.

### *IV-B. Clusterización*

Para la aplicación de modelos de clusterización con el objetivo de encontrar un agrupamiento lo más cercano posible a la verdad fundamental de la clasificación de estos granos secos, se debe considerar que se ha tratar con modelos de clustering disjunto y exhaustivo [13]. Ya que la verdad fundamental consiste en clases de granos excluyentes y además cada grano debe pertenecer a una clase.

Cuadro II
ESPACIO DE PARÁMETROS DE CADA MODELO

| | Tipo de Búsqueda | Parámetros Empleados | | |
|---|---|---|---|---|
| **SVM** | GridSearchCV | C = (0.1, 1, 10, 100) | Gamma = (1, 0.1, 0.01, 0.001) | Kernel = rbf |
| **MLP** | GridSearchCV | Capas Ocultas = (2, 4, 10, 20) | Activación = (logistic, relu) | |
| **XGB** | RandomizedSearchCV | # Estimadores = (50, 200, 350) | Profundidad Máx. = (2, 9, 14) | Tasa Aprendizaje = (0.1, 0.5, 0.9) |
| **KNN** | GridSearchCV | # Vecinos = (3, 5, 7, 9, 11, 13, 15, 17, 19) | Pesos = (uniform, distance) | |

*IV-B1. Modelos Utilizados:* Considerando lo anterior, y que se requiere probar modelos de distinta naturaleza (basados en centroides, densidad, jerarquía, etc.) a continuación se detallan los modelos utilizados:

- K-means
- Agglomerative Clustering (AC)
- Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)

*IV-B2. Versiones del Dataset:* La aplicación de los modelos de clusterización se realiza sobre cuatro versiones diferentes del dataset, estas versiones corresponden al dataset normalizado con todas las variables, una versión reducida aplicando PCA (al menor número de dimensiones que permita explicar la mayor varianza posible), una versión aplicando t-SNE (a dos dimensiones) y otra versión aplicando una reducción PCA + t-SNE. Estas distintas versiones tienen el propósito de probar varias configuraciones del dataset y ver si aportan para mejorar los modelos de clusterización.

*IV-B3. Visualización:* Para visualizar los resultados de la clusterización de cada modelo se utilizan gráficas de tipo Scatter Plot. Para cada versión del dataset se muestran cuatro gráficas comparativas (Ground Truth, K-means, AG, HDBSCAN) vistas desde varias perspectivas. Estas perspectivas corresponden a una en la que se muestran las instancias vistas a través de dos atributos del dataset (estos atributos son dos de los que más correlación tienen con la etiqueta de clase y menos correlación entre ellos mismos), otra en la que se aplica PCA (a dos dimensiones) y otra aplicando t-SNE (a dos dimensiones) para propósitos de visualización únicamente.

Si bien estas técnicas de visualización contribuyen de forma gráfica a la tarea de evaluación y comparación de los modelos de clusterización en su objetivo de aproximar la verdad fundamental del dataset, no deja de ser una tarea subjetiva y difícil de realizar. Para solventar esto se propone el uso de la métrica de similaridad entre clusterings Adjusted Rand Index (ARI).

*IV-B4. Adjusted Rand Index:* Rand Index (RI) calcula una medida de similitud entre dos clusterings (en este caso el clustering de los modelos y el ground truth) considerando todos los pares de instancias y contando los pares que se asignan en el mismo cluster o en clusters diferentes en los clusterings predichos y reales. Adjusted Rand Index (ARI) es una medida ajustada de la anterior que tiene un valor cercano a 0 para el etiquetado aleatorio independientemente del número de clusters e instancias y es exactamente 1 cuando los clusters son idénticos (incluyendo permutaciones) [14].

Cuadro III
MEJORES PARÁRAMETROS DE CADA MODELO

| | Mejores Parámetros | | |
|---|---|---|---|
| **SVM** | C = 100 | Gamma = 1 | |
| **MLP** | Capas Ocultas = 20 | Activación = relu | |
| **XGB** | # Estimadores = 200 | Prof. Máx. = 9 | T. Aprendizaje = 0.1 |
| **KNN** | # Vecinos = 13 | Pesos = distance | |

## V. RESULTADOS Y DISCUSIÓN

La obtención de un proceso automático de clasificación de granos secos puede ayudar en gran medida a la comercialización y cultivo de estos productos. Para conseguir esto, en este estudio se ha propuesto usar modelos de ML y compararlos para encontrar el mejor posible, lo que resultara en un modelo confiable capaz de clasificar los diferentes tipos de granos.

Para resolver esta tarea se hizo uso de información recolectada por medio de técnicas de visión de computador donde modelos trataron de clasificar siete tipos de granos diferentes. Esta información consistió de 13,611 instancias donde se tenia el valor numérico de 16 cualidades las cuales describían un tipo de grano especifico. Los resultados independientes de los dos tipos de aprendizaje usados fueron los siguientes.

### *V-A. Clasificación*

Sobre el dataset a clasificar, primero se realizó un escalamiento de todas las variables que este posee. Posteriormente se realizó un análisis de la distribución de las etiquetas de cada instancia, y se descubrió que esta era una distribución desbalanceada. Por lo que cualquier entrenamiento con este requirió una estratificación y posterior validación con Stratified 10-fold cross-validation.

Por otro lado, se realizó un tuneo de parámetros con los cuales se obtuvo la mejor combinación modelo-parámetros posible de acuerdo a el espacio de parámetros indicado en el Cuadro II. Los resultados de este tuneo se presenta en el Cuadro III.

Una vez obtenidos los mejores hiper parámetros para cada modelo se paso a probar cada uno de los modelos entrenados con el dataset de prueba. Las métricas de rendimiento de todos los modelos se pueden observar en el Cuadro IV.

De acuerdo con el Cuadro IV, podemos ver que las métricas importantes de rendimiento de MLP y KNN son inferiores a las de SVM y XGB. Para MLP y KNN Precision, Recall y F1 son inferiores a el 93 %. También podemos ver que todas estas medidas por parte de SVM y XGB llegan a ser mayores a 93 %. La única diferencia entre estos dos modelos son los dígitos decimales, donde si las comparamos, SVM supera en todas a XGB.

Cuadro IV
MÉTRICAS DE RENDIMIENTO

| | Métricas Ponderadas | | | |
|---|---|---|---|---|
| | **Balanced Accuracy** | **Precision** | **Recall** | **F1** |
| **SVM** | 93.96 % | 93.15 % | 93.14 % | 93.13 % |
| **MLP** | 92.97 % | 92.19 % | 92.16 % | 92.15 % |
| **XGB** | 93.96 % | 93.03 % | 93.02 % | 93.01 % |
| **KNN** | 93.52 % | 92.7 % | 92.67 % | 92.68 % |

,

Cuadro V
RESULTADOS DE TEST DE MCNEMAR

| Modelo 1 | Modelo 2 | p-value | Mejor Rendimiento |
|---|---|---|---|
| **SVM** | **MLP** | 0.000625000978493 | SVM |
| **SVM** | **XGB** | 0.711531417761122 | Equivalentes |
| **MLP** | **XGB** | 0.011044764603214 | XGB |
| **SVM** | **KNN** | 0.110211836965136 | Equivalentes |
| **MLP** | **KNN** | 0.08519169643588 | Equivalentes |
| **XGB** | **KNN** | 0.297953061608165 | Equivalentes |

Dado este análisis, podríamos concluir que SVM es el mejor modelo para la clasificación de granos, sin embargo el pequeño rango de diferencia ente todos estos modelos nos impulsa a realizar otro tipo de test con el objetivo de tener una segunda referencia al determinar el modelo más apropiado.

Así, se ha realizado el test estadístico de McNemar entre cada par de modelos utilizados. Los resultados de estos tests se los puede ver en el Cuadro V.

Se debe recalcar, que para todo test de McNemar realizado en este estudio, la suma de b + c siempre fue mayor a 25 por lo que la distribución chi-cuadrado es apropiada para obtener los p-values que se expresan en el Cuadro V [3].

En el Cuadro V se puede observar como la mayoría de los tests de McNemar dan como resultado a p-values mayores a $\alpha = 0{,}05$, por lo que en estos casos no podemos rechazar la hipótesis nula y debemos asumir que no hay una diferencia significativa entre los dos modelos.

En el caso que el p-value del test resulte en ser menor que $\alpha = 0{,}05$, podemos rechazar la hipótesis nula de que ambos modelos tienen un rendimiento similar, por lo que se deben observar aquellos modelos tienen mayor aciertos en sus clasificaciones para determinar el mejor. Esto resulta en que en el caso de SVM vs MLP se declare a SVM como un mejor modelo, y en el caso de MLP vs XGB se declare a XGB como el modelo con un mejor rendimiento.

Con este análisis, se puede decir que SVM y XGB son el par de modelos con un mejor rendimiento, pero esto no es una respuesta definitiva para nuestra tarea. Es decir, no hemos encontrado aún un único y distintivo modelo que sea el mejor para el caso de clasificación de granos, es por esto que realizamos un último test, llamado 5x2cv paired t-test para comparar el rendimiento de estos dos modelos que sobresalieron. Al realizar este test y al analizar el p-value que nos devolvió, resultó que p-value $> \alpha$ con un p-value = 0.289, por lo que este último y definitivo test no ha podido encontrar una diferencia entre el rendimiento de SVM y XGB.

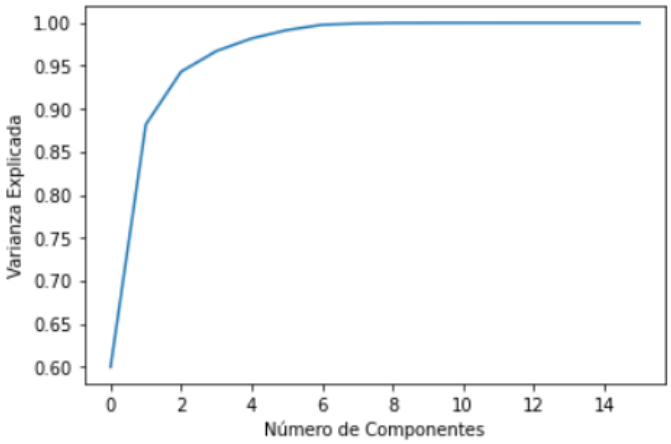


Figura 3. PCA - Varianza Explicada.

Cuadro VI
RESULTADOS DE TEST DE MCNEMAR DADO TIPO DE REDUCCIÓN DE DIMENSIONES

| Modelo | Tipo de Reducción de Dimensiones | p-value | Mejor Rendimiento |
|---|---|---|---|
| SVM | SVM(PCA) | 0.7547764 | Equivalentes |
| SVM | SVM(t-SNE) | 1.9874126e-25 | SVM |
| SVM | SVM(PCA + t-SNE) | 7.3135693e-26 | SVM |

Debido a que no se pudo encontrar un modelo definitivo por medio de métricas de rendimiento o pruebas de hipótesis, tomaremos como mejor modelo al que tiene un menor tiempo de ejecución. De entre SVM y XGB, este resulta ser SVM, el cual se logró entrenar en un tiempo de 49.8 segundos, que es una gran diferencia a comparación de XGB que entrenó en un largo tiempo de 6.2 minutos.

*V-A1. Posibles mejoras aplicando reducción de dimensiones:* Una vez obtenido el mejor modelo de clasificación se intentó mejorar aún más el rendimiento de este aplicando técnicas de reducción de dimensiones al dataset. Entre estas técnicas consta PCA, para lo cual es clave obtener la cantidad de dimensiones a las cuales reducir el dataset. Para obtener esta cantidad, se uso la técnica de la varianza explicada que se puede visualizar en la Figura 3.

Se puede ver en la figura 3 como se consigue explicar casi el 100 % de la varianza con seis componentes. Por lo que cualquier reducción realizada por medio de PCA se la hizo a 6 dimensiones.

Continuando, una vez aplicada cada una de las transformaciones (PCA, t-SNE, PCA+t-SNE) se tomó su resultado y se aplicó SVM a cada uno con los parámetros encontrados anteriormente. Una vez entrenado los modelos se pasa nuevamente a evaluar por el test de McNemar a cada uno de estos por pares. Los p-values de todos estos test se los puede observar en el Cuadro VI.

En el Cuadro VI podemos ver que al probar SVM con un dataset al cual se le aplicó reducción de variables por medio de t-SNE o PCA + t-SNE, el p-value es menor que $\alpha = 0{,}05$, lo que significa que podemos rechazar la hipótesis nula de

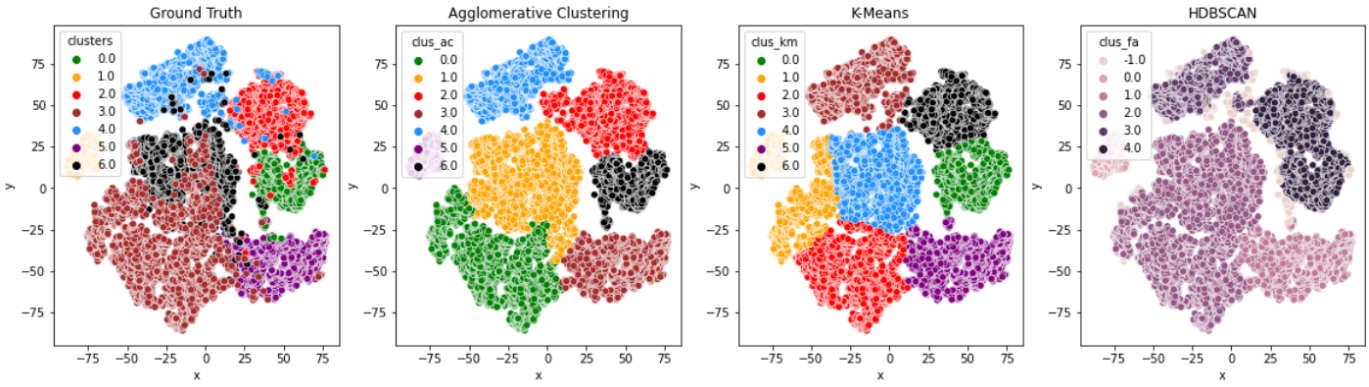


Figura 4. Clustering de dataset reducido con t-SNE.

Cuadro VII
ARI DE TÉCNICAS DE CLUSTERIZACIÓN APLICADAS SOBRE VARIOS TIPOS DE REDUCCIÓN DE DIMENSIONES

| **Método de Clusterización** | **Tipo de Reducción de Dimensiones** | | | |
|---|---|---|---|---|
| | **Ninguno** | **PCA** | **t-SNE** | **PCA + t-SNE** |
| AC | 0.6446 | 0.62424 | 0.7220 | 0.6969 |
| K-Means | 0.6253 | 0.6245 | 0.6430 | 0.6188 |
| HDBSCAN | 0.0342 | 0.0341 | 0.5774 | 0.0327 |

que ambos modelos tienen un rendimiento similar y podemos concluir que en estos casos SVM entrenado con todas las variables del dataset tiene un mejor rendimiento. Por último, al observar el p-value del test que se realizo con el SVM entrenado con el dataset transformado por PCA, podemos decir que el rendimiento es equivalente.

### *V-B. Clusterización*

La clusterización se la realizó sobre todas las variables del dataset normalizado, sobre el resultado de aplicación de PCA, el resultado de la aplicación de t-SNE y finalmente sobre la aplicación de PCA + t-SNE. Para todos estos tipos de input se aplicó los modelos: Aglomerative Clustering, K-Means y HDBSCAN. A continuación, en el Cuadro VII, se muestra los resultados de la aplicación de clustering de acuerdo al ARI.

En el Cuadro VII se puede ver una matriz donde en cada celda esta el valor de ARI calculado según la clusterización sobre cada tipo de reducción. Mediante este índice podemos notar que la mayoría de las técnicas de clusterización se comportan de la misma manera, teniendo un ARI en el rango de 0.61 - 0.68, a excepción de HDBSCAN que tiene índices muy bajos, llegando a ser 0.034.

A pesar de todo, lo que se debe destacar es que el mejor ARI que se consigue es al realizar AC con la reducción de dimensiones t-SNE, llegando a conseguir un ARI de 0.7220. Lo que se considera como una buena clusterización pues este buen desempeño puede ser confirmado al momento de visualizar (Figura 4) la clusterización realizada por AC con respecto la verdad fundamental.

## VI. CONCLUSIONES

Respecto de la clasificación de granos secos a partir de sus características geométricas y de forma (obtenidas mediante procedimientos de CV) utilizando modelos de ML como SVM, MLP, XGB y KNN, se puede concluir que en general estos modelos proveen buenos resultados en base a métricas como Balanced Accuracy, Precision, Recall o F1-Score. Se comparó el rendimiento de estos modelos usando pruebas estadísticas como el test de McNemar o 5x2cv paired t test, considerando además el tiempo de entrenamiento se obtuvo que el modelo SVM es el que ofrece mejores resultados. Se intentó mejorar el rendimiento de este modelo aplicando transformaciones como PCA o t-SNE a los datos, pero estas transformaciones no tuvieron mayor impacto en los resultados.

Por otro lado, respecto de la clusterización de granos secos para lograr un agrupamiento lo más cercano posible a la clasificación real de los granos, se puede decir que los modelos como AC o K-means obtienen resultados regulares en la realización de esta tarea, los modelos como HDBSCAN proporcionan resultados nada útiles. Esto se cumple tanto para la versión original del dataset como para las versiones reducidas usando PCA, t-SNE y PCA + t-SNE. Sin embargo, se puede considerar que AC es el modelo que mejores resultados provee para esta tarea. Las conclusiones anteriores se pueden corroborar gráficamente y con la métrica ARI.

Por último, como trabajo futuro se plantea realizar un trabajo similar al aquí descrito utilizando granos de producción ecuatoriana. Dicho trabajo futuro implicaría la recolección de imágenes digitales de los granos y su posterior procesamiento con técnicas de CV para la extracción de características importantes. Con esto se podría conformar un dataset similar al utilizado en este trabajo y realizar un análisis de ML similar al aquí descrito. Los resultados de este trabajo podrían usarse para el desarrollo de un sistema comercial que ayude a productores agrícolas ecuatorianos en la clasificación automatizada y análisis de calidad de granos.

## CONTRIBUCIONES

**Víctor Herrera:** Conceptualización, Investigación, Recopilación de Referencias, Revisión de Trabajos Relacionados,

Análisis Exploratorio, Metodología, Programación, Validación y Evaluación, Análisis Final, Escritura. **Pablo Solano:** Conceptualización, Investigación, Recopilación de Referencias, Revisión de Trabajos Relacionados, Análisis Exploratorio, Metodología, Programación, Validación y Evaluación, Análisis Final, Escritura.

## REFERENCIAS


[1] D. Horton, "Investigación colaborativa de granos andinos en Ecuador", 2014 [Online]. Disponible en: http://repositorio.iniap.gob.ec/handle/41000/102.

[2] E. Peralta I., A. Murillo I., N. Mazón, E. Villacrés, y M. Rivera M., "Catálogo de variedades mejoradas de granos andinos: Chocho, quinua y amaranto, para la sierra de Ecuador", may 2013 [Online]. Disponible en: http://repositorio.iniap.gob.ec/handle/41000/2713.

[3] M. Koklu y I. A. Ozkan, "Multiclass classification of dry beans using computer vision and machine learning techniques", Computers and Electronics in Agriculture, vol. 174, p. 105507, jul. 2020, doi: 10.1016/j.compag.2020.105507.

[4] H. O. Velesaca, P. L. Suárez, R. Mira, y A. D. Sappa, "Computer vision based food grain classification: A comprehensive survey", Computers and Electronics in Agriculture, vol. 187, p. 106287, ago. 2021, doi: 10.1016/j.compag.2021.106287.

[5] A. Ali et al., "Machine learning approach for the classification of corn seed using hybrid features", International Journal of Food Properties, vol. 23, no 1, pp. 1110–1124, ene. 2020, doi: 10.1080/10942912.2020.1778724.

[6] P. Nie, J. Zhang, X. Feng, C. Yu, y Y. He, "Classification of hybrid seeds using near-infrared hyperspectral imaging technology combined with deep learning", Sensors and Actuators B: Chemical, vol. 296, p. 126630, oct. 2019, doi: 10.1016/j.snb.2019.126630.

[7] O. Oña y A. Jefferson, "Sistema de clasificación de granos de cacao frescos basados en visión computacional", ene. 2020 [Online]. Disponible en: http://bibdigital.epn.edu.ec/handle/15000/20676.

[8] T. J. Negrete Peña y J. G. Llaguno Vera, "Comparación de técnicas de visión artificial para determinar el grado de fermentación de varios tipos de granos de cacao en el proceso postcosecha en la región litoral del Ecuador", 2017 [Online]. Disponible en: http://dspace.ups.edu.ec/handle/123456789/15007.

[9] D. Krstajic, L. J. Buturovic, D. E. Leahy, y S. Thomas, "Cross-validation pitfalls when selecting and assessing regression and classification models", J Cheminform, vol. 6, no 1, p. 10, dic. 2014, doi: 10.1186/1758-2946-6-10.

[10] A. Burkov, The hundred-page machine learning book. Polen: Andriy Burkov, 2019.

[11] "API Reference — scikit-learn 0.24.2 documentation". [Online]. Disponible en: https://scikit-learn.org/stable/modules/classes.html#sklearn-metrics-metrics.

[12] T. G. Dietterich, "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms", Neural Computation, vol. 10, no 7, pp. 1895–1923, oct. 1998, doi: 10.1162/089976698300017197.

[13] H. Blockeel, "Machine Learning and Inductive Inference", p. 331.

[14] K. Y. Yeung y W. Ruzzo, "Details of the Adjusted Rand index and Clustering algorithms Supplement to the paper 'An empirical study on Principal Component Analysis for clustering gene expression data' (to appear in Bioinformatics)", Science, vol. 17, ene. 2001.