UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escola Tècnica Superior d'Enginyeria Agronòmica i del Medi Natural

Escuela Técnica Superior de ingeniería agronómica y del medio natural (ETSIAMN)

Grado en biotecnología

Clasificación de Solanum lycopersicum y parientes silvestres mediante técnicas de aprendizaje automático partiendo de datos genéticos.

Autor: Fabián Robledo Yagüe

Tutor: José Miguel Blanca Postigo

Universitat Politècnica de València

Curso académico 2018-2019

València, junio de 2019

Resumen

Clasificación de *Solanum lycopersicum* y parientes silvestres mediante técnicas de aprendizaje automático partiendo de datos genéticos.

Desde la domesticación del tomate (*Solanum lycopersicum*), y especialmente tras su introducción en Europa en el siglo XVI, esta planta ha evolucionado en función de las necesidades de las poblaciones humanas que lo cultivaban, dando lugar a multitud de variedades. Estas variedades se diferencian tradicionalmente en base a caracteres morfológicos, pero ahora también disponemos de herramientas genómicas que nos permiten comparar sus genomas. Con esta información se pueden crear grupos genéticos que podemos comparar con las variedades basadas en los caracteres morfológicos. El objetivo de este trabajo es analizar la capacidad de distintos sistemas de clasificación automático para realizar estas clasificaciones. A partir de un conjunto de datos, que contiene las diferencias genéticas entre las muestras, se ha realizado una serie de agrupaciones automáticas de las distintas muestras en grupos en función de sus variaciones genéticas. Dichas clasificaciones se han comparado con otras hechas manualmente para comprobar la fiabilidad del modelo y la capacidad del sistema de clasificar las muestras en la categoría anteriormente asignada. Esta comparación ha permitido concluir la utilidad de un modelo automático para la clasificación de distintas especies salvajes emparentadas con el tomate, pero también ha demostrado la dificultad de clasificar la diversidad del tomate moderno.

Palabras clave: Solanum lycopersicum, Agrupación, clasificación, aprendizaje automático.

Abstract

Clasification of *Solanum lycopersicum* and wild relatives with machine learning techniques and genetic data

Since tomato (*Solanum lycopersicum*) was domesticated, and particularly after it was introduced in Europe in 16th century, this plant has evolved according to the needs of the human population that grew it, giving as a result many different varieties. The differences between them are traditionally morphologic traits, but nowadays genomic tools enable comparison of their genomes. With this data, we can now create genetic groups we that can be compared with morphology based varieties. The main goal of this study is to analyse the ability of different automatic machine learning systems to make these classifications. Starting from a dataset that contains genetic differences between samples, several automatic clusterings were performed, placing samples into groups according to their genetic variations. These classifications were compared with other manually made to check the reliability of the model and the system's ability to assign the sample to the manually assigned category. This comparison concluded that the model was useful to classify different tomato-related wild species, but also showed the existent hardship to classify the intraspecies diversity of modern tomato.

Keywords: Solanum lycopersicum, clustering, classification, machine learning.

Autor: Fabián Robledo Yagüe
Tutor: José Miguel Blanca Postigo
Valencia, junio de 2019

Agradecimientos

Este trabajo no es solo el resultado de la última fase de realización del grado, sino la culminación de una época que me ha marcado especialmente que ha sido muy importante para mí, en la que he aprendido no solo aquello que me permitirá ganarme la vida en el futuro si no también a ser mejor persona y a comprender a aquellos a mi alrededor.

En primer lugar, quiero agradecer a todos aquellos que han estado en el laboratorio donde he realizado este trabajo: tanto a Jose por ser mi tutor, como a Ximo, Peio, Estefanía, Victor, David y Marta por los momentos de trabajo, los momentos de ayuda, los momentos de charla y los momentos de risas. También a Javier Palanca por la ayuda que me ha dado y las correcciones que me ha hecho.

Agradecer también todos los momentos que me han aguantado mis padres cuando estaba en época de exámenes aun con todo el estrés encima, y me han apoyado y motivado a continuar por el camino que he elegido. También agradecer mi hermano que siempre ha estado ahí para ayudarme en cualquier problema que haya tenido y que gracias a el he conseguido llegar hasta donde he llegado y que siempre ha intentado sacar lo mejor de mi aunque no siempre

Parte de mis conocimientos también debo agradecérselos a todos aquellos divulgadores científicos, algunos de los cuales tuve el honor de conocer en persona, gracias a los cuales no solo ha mejorado mi comprensión del mundo científico y me han motivado a continuar por éste camino, todo lo que he aprendido en innumerables horas de charlas y vídeos que de alguna forma han afectado a éste trabajo, en particular a Dani García (@SoyBiotec). Añadir también el aprecio, el compañerismo y todos los momentos buenos que viví gracias a mis compañeros de iGEM y todo lo que he aprendido de multitud de campos a los que era ajeno.

Gracias también a esos pocos profesores que, en contra de lo esperado, hacen las clases teóricas amenas y motivan a los alumnos como yo a acudir cada día y aprender cosas nuevas, en lugar de dedicarse a leer diapositivas, y que realmente hacen que los alumnos entendamos la materia en lugar de memorizarla para aprobar un examen.

Por último, pero no menos importantes por ello, quiero agradecer su apoyo, su amistad, las risas y también el estrés de clase y los trabajos. Las horas interminables en la biblioteca o por internet, pero también el cine y las aventuras que hemos vivido juntos. Las horas de juego tanto juntos como online, a todas aquellas personas maravillosas a las que he conocido estos cuatro años y puedo con orgullo llamar amigos: Irene, Mario, Claudia, Sara, Josep, Samuel, Adrià, Arturo, Paula, Juanjo, Paloma, Almudena y Belén. Nada hubiese sido lo mismo sin ellos.

Gracias a todos vosotros por formar un pedacito de mi.

Índice de contenido

Índice de Figuras y Tablas

1 Introducción

1.1 Historia del tomate y su clasificación

El tomate (*Solanum lycopersicum*) es una especie cultivada que, al igual que otros miembros del género *Solanum*, como la patata o el pepino dulce, es originaria de la zona andina. Fue en esta región en la que el tomate se domesticó a partir de la especie silvestre *Solanum pimpinelifollium*. El proceso de domesticación consiste en la modificación de la especie silvestre mediante el uso de la selección artificial. Actualmente, se considera que el proceso de domesticación del tomate transcurrió en dos etapas: en la primera la variedad *Solanum lycopersicum* var. Cerasiforme fue domesticada en la región andina de Perú y Ecuador, posteriormente estos tomates se exportaron a la región Mesoamericana en dónde fueron modificados de nuevo creado un tomate muy similar a nuestras variedades tradicionales actuales. Por otro lado, el tomate sufrió un proceso de mejora durante el siglo XX dando lugar a los cultivares modernos con nuevas características por compañías privadas *(Bergougnoux, 2014)*.

Las distintas especies y variedades relacionadas con el el tomate cultivados se caracterizan por tener rasgos morfológicos y agronómicos diferenciados. Estos rasgos han sido utilizados tradicionalmente para clasificar estos materiales en distintas especies, poblaciones y variedades. Sin embargo, al utilizarse rasgos morfológicos y agronómicos sometidos a selección, es posible que esta clasificación no refleje el desarrollo histórico ni el proceso evolutivo subyacente que las ha creado.

Complementariamente se puede realizar clasificaciones genéticas basadas en las distancias genéticas. Esto puede hacerse, por ejemplo, utilizando análisis multivariantes, como Análisis de Componentes Principales (ACP) o Análisis de Coordenadas Principales (ACoP). Éstas técnicas estadísticas permiten simplificar datos multidimensionales a un reducido número de dimensiones, a cambio de perder una parte de la información original (*Bridges et al., 2011*). La limitación de esta aproximación es que dado que la clasificación la crea un investigador siempre tendrá un cierto carácter subjetivo. Alternativamente, se pueden utilizar métodos de clasificación automáticos como el implementado en el software Structure, que dividen las muestras genotipadas en poblaciones genéticas. La limitación de estos métodos es que hacen asunciones sobre cómo ha transcurrido la historia evolutiva que puede que no se correspondan con la realidad.

1.2 Proyecto TRADITOM

En la actualidad, existen cientos de variedades de tomate tradicional europeas que derivan de las plantas traídas a Europa a partir del siglo XVI. El programa Horizon 2020 ha financiado un proyecto, denominado TRADITOM (traditom.eu), que busca conocer y aprovechar la diversidad genética existente de las distintas variedades de tomate tradicional existentes. Las variedades tradicionales suelen están bien valoradas por los consumidores,

aunque suelen tener un comportamiento agronómico muy inferior, por ejemplo suelen tener una menor productividad y uniformidad y resisten peor a algunas enfermedades.

El objetivo principal del proyecto TRADITOM es impulsar la conservación de las variedades tradicionales y aumentar su competitividad, dando valor a todo el acerbo genético existente.

Las muestras del proyecto TRADITOM incluyen tanto variedades tradicionales como otras muestras de referencia que incluyen variedades modernas y plantas silvestres.

1.3 Inteligencia artificial

En la actualidad, ramas como la genómica están generando grandes cantidades de datos que requieren nuevas formas de tratar datos complejos (*Libbrecht y Noble, 2015*). Una posible opción para crear clasificaciones automáticas consiste en utilizar herramientas estadísticas de clasificación automática (también conocidas como aprendizaje automático o *Machine Learning*). La librería Sci-kit learn (*Pedregosa et al, 2013*), para el lenguaje de programación Python, implementa varios de estos algoritmos y es ampliamente utilizada en distintos campos de la inteligencia artificial (*Buitinck et al, 2019*)

Las técnicas de aprendizaje automático pueden clasificarse en dos grandes familias: las que hacen un aprendizaje supervisado y las no supervisadas. En el primer caso, disponemos una serie de muestras etiquetadas. El algoritmo, analizando un conjunto de estas muestras etiquetadas, busca patrones que posteriormente podrá utilizar para etiquetar muestras nuevas. A este tipo de algoritmos se les denomina supervisados porque se les proporcionan las muestras ya etiquetadas. En la actualidad estos algoritmos son muy utilizados. Por ejemplo es común usarlos para clasificar imágenes en función de su contenido. Un problema típico consiste en darle imágenes con distintas letras manuscritas y pedir al algoritmo que aprenda a distinguir los caracteres manuscritos en nuevas imágenes. De manera similar, puede utilizarse la misma metodología para detectar la posibilidad de supervivencia al cáncer (*Gupta et al, 2014*), entre otras muchas aplicaciones en el ámbito biomédico y biotecnológico.

En la segunda familia de algoritmos, los no supervisados, las muestras no están etiquetadas previamente. En este caso lo que se persigue es crear un modelo estadístico de los datos de partida. Uno de las tareas más comunes consiste en realizar agrupaciones (clustering) de distintas muestras en función de sus similitudes. Idealmente el algoritmo debe agrupar las muestras en grupos separados atendiendo a su similitud.

Existen una gran variedad de algoritmos distintos que pueden realizar agrupaciones. Cada uno de ellos se basa en diferentes criterios matemáticos, utilizados para introducir una muestra en un grupo concreto y para determinar el total de grupos que se van a formar. Por ejemplo, el algoritmo K-means crea grupos minimizando la varianza dentro de cada grupo, mientras que el algoritmo DBSCAN coloca las muestras que se encuentran a una distancia determinada en el mismo grupo, Al tener distintos enfoques, un algoritmo puede ser apropiado para un conjunto de datos en particular pero ser inadecuado para otro conjunto de datos distinto,

Disponiendo de las categorías a las que pertenece cada muestra, podemos realizar tanto una clasificación supervisada o una agrupación no supervisada. Sin embargo, la clasificación disponible está basada criterios morfológicos y se ha realizado atendiendo a los nombres que tradicionalmente reciben las distintas variedades de tomate. Esta clasificación podría no corresponderse, en muchos casos, con una clasificación genética, que tenderá a estar más relacionada con la historia de esas variedades que con su morfología. Por lo tanto nuestro objetivo será crear clasificaciones moleculares que posteriormente compararemos con las clasificaciones tradicionales disponibles.

Esta es una aproximación que puede ser utilizada para clasificar las muestras de tomate en grupos genéticos.

1.4 Evaluación del modelo

Para evaluar la diferentes clasificaciones obtenidas por los distintos algoritmos se puede utilizar un conjunto de coeficientes estadísticos. Cada uno de ellos tiene unas características que los hacen adecuados para distintos casos, en función del algoritmo de agrupación utilizado y los datos previos disponibles. Además, si se dispone de una clasificación previa se puede evaluar en qué medida las clasificaciones generadas se corresponden con esta clasificación.

1.4.1 Coeficiente de Shilouette

El coeficiente de Shilouette (*Rousseeuw, 1986*) compara la distancia de una muestra respecto al centro de su grupo con la distancia al grupo más cercano, otorgándole un valor entre -1 y 1. Un valor negativo indica que la muestra debería encontrarse en otro grupo, mientras que un valor de 1 indicaría que está junto a otras muestras del grupo el en que se encuentra.

Si consideramos todas las muestras en lugar de una individual, el coeficiente de silhouette puede utilizarse para conocer si los grupos se encuentran dispersos o compactos, o incluso si solapan entre sí. Para ello, se calcula la media de todos las muestras. Valores positivos elevados indican que los grupos son compactos y están separados, mientras que valores cercanos a 0 indicarían grupos solapantes. Un valor negativo, indicaría que existen muestras que deberían clasificarse en otros grupos más cercanos.

$$s_i = \frac{b - a}{max(a, b)}$$

Figura 1: Cálculo del coeficiente de Silhouette para cada muestra, siendo a la distancia media a otro punto del grupo y b la distancia media a puntos que pertenecen a un grupo distinto.

$$CS = \frac{1}{n} \sum_{i=0}^{n} s_i$$

Figura 2: Coeficiente de silhouette para un dataset, que consiste en la media de los coeficientes de silhouete para cada uno de las muestras. S_i indica el valor de silhouette para la muestra i, y n el total de muestras

En general, cuanto mayor sea el coeficiente de Shilouette, menos dispersos y más separados se encuentran los grupos.

1.4.2 Índice Calinski-Harabasz

Otro de los coefiencientes utilizados en agrupamientos no supervisados, es el índice de Calisnki-Harabasz (*Calisnki y Harabasz, 1974*), que permite comparar dos agrupamientos entre sí para conocer cual de ellos tiene sus grupos mejor definidos; es decir, cuando mayor sea éste parámetro, sus muestras se encuentran menos dispersadas y, a su vez, alejadas de otros grupos. Este algoritmo no está acotado superiormente por lo que es difícil evaluarlo de forma aislada, siempre debe ser utilizado para comparar distintas clasificaciones.

$$s(k) = \frac{\mathrm{Tr}(B_k)}{\mathrm{Tr}(W_k)} \times \frac{N-k}{k-1}$$

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T$$

$$B_k = \sum_{q} n_q (c_q - c)(c_q - c)^T$$

Figura 3: Indice de Calinski-Harabasz, donde N es el numero total de muestras, k el de clusters, Wk la dispersión intra-cluster y Bk la dispersión inter-cluster. Nq el numero de muestras del cluster Q, Cq es el punto central del cluster Q y x es la posición de cada muestra. Obtenido de https://scikit-learn.org/stable/modules/clustering.html#calinski-harabasz-index. Editado por mi.

Este es un parámetro rápido de calcular y suele utilizarse para determinar el número de grupos que debe incluir una clasificación.

1.4.3 Índice Davies-Bouldin

Por último, el índice de Davies-Bouldin (*Davies y Bouldin, 1979*) busca minimizar una función objetivo: la distancia de las muestras al centro de su grupo. Cuanto más cercano sea el valor a 0, mejor se considera la agrupación, pues implica que las muestras están juntas en sus grupos y a su vez están muy separados del resto.

Una agrupación que indique grupos compactos y separados tendrá valores inferiores a 0,7 en este índice, mientras que un valor superior puede indicar grupos solapantes o dispersos

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} R_{ij}$$

Figura 4: Índice de Davies-Bouldin. S_i es la distancia media de una muestra del cluster i a su centroide. S_j la distancia media de una muestra del cluster j al centroide del cluster j. d_{ij} la distancia entre los clusteres i y j.

Junto con el coeficiente de silhouette y el índice de Calinski-Harabasz, este parámetro nos da una idea de como de buena es la agrupación y si podemos utilizarla para comparar las variedades de tomate. Comparando los tres parámetros, podemos seleccionar el modelo que mejor optimice estos coeficientes para establecer cual es el mejor modelo.

2 Objetivos

El objetivo principal de presente trabajo es evaluar el desempeño de los distintos algoritmos de clasificación automática no supervisada basada en las muestras el genotipadas por el proyecto TRADITOM.

Para ello se evaluarán seis algoritmos distintos de clasificación automática.

Los algoritmos serán utilizados en dos problemas distintos: 1) la clasificación de las muestras en grupos genéticos silvestres y cultivados y 2) la clasificación de las muestras cultivadas en distintas variedades de tomate

Para todas las clasificaciones se calcularán los parámetros Silhouette, Calinski-Harabasz y Davies-Bouldin.

3 Material y métodos

3.1 Información genética

Para poder realizar este trabajo se han utilizado los datos del proyecto TRADITOM. Una parte de éste consistió en la secuenciación y genotipado de cientos de variedades de tomate tradicional europeo. En particular, los datos genéticos de partida son los resultantes del genotipado mediante GBS (Genotyping By Sequencing) de hasta 70000 SNP (Single Nucleotide Polymorphism) en cada una de éstas variedades, tanto silvestres como domésticas. La información está almacenada en archivos VCF (Variant Call Format). En estos archivos se indica, para cada SNP, la posición en el genoma, los alelos que han sido detectados y el genotipo para cada una de las muestras. Esta información se utilizará para generar una matriz de coordenadas, indicando las diferencias entre las muestras, que a su vez permitirá realizar un Análisis de Coordenadas principales (PCoA) que reducirá toda la información de los archivos a 5 coordenadas,

Este conjunto de datos consta de 1628 muestras de decenas de distintas variedades de tomate, así como de varias especies silvestres relacionadas: *Solanum pimpinellifolium Solanum galapagense* y *Solanum chesmaniae*. Hay un 9% de las muestras de las que se desconoce la variedad a la que pertenecen.

De ese total de 1628 muestras, 435 pertenecen a la categoría de muestras de tomates silvestres y cultivados, mientras que las 1193 restantes pertenecen a la categoría de tomates domesticados. A su vez, cada una de estas dos categorías se subdivide en las distintas variedades a las que pertenecen las muestras: Las variedades de tomate silvestre consta de un total de 33 variedades, mientras que las variedades domesticadas tiene 30 variedades distintas

Cada una de las variedades tiene asignadas las muestras que corresponden a esa variedad. Hay mucha variación en la cantidad de muestras de cada categoría, desde una única muestra hasta 37 en las variedades silvestres, o 200 en las variedades domesticadas.

3.2 Herramientas informáticas

Tabla 1: Versiones del software utilizado

Librería	Versión
Python	3.6.7
Jupyter notebook	4.4.0
Variation	N/D
Sci-kit learn	0.20.2
Matplotlib	2.2.2
Curly Whirly	1.19.03.26

3.2.1 Python

Python es el lenguaje de programación interpretado que se ha utilizado en el presente trabajo. Python fue creado en 1991 por Guido Van Rossum, y ha crecido en popularidad hasta ser uno de los lenguajes de programación más utilizados actualmente La versión utilizada en este trabajo corresponde a la 3.6.7. Python es uno de los lenguajes más utilizados en bioinformática, data science e inteligencia artificial debido a su simplicidad, facilidad de lectura y a la existencia de multitud de librerías especializadas para estos campos.

3.2.2 Jupyter notebook

La mayor parte del código fue desarrollado y ejecutado utilizando Jupyter notebook, una herramienta diseñada para facilitar el análisis de datos de forma interactiva.

3.2.3 Variation

La librería *Variation* permite trabajar con grandes cantidades de datos genotípicos. Esta librería dispone de herramientas para eliminar SNPs que no satisfagan diversos criterios de calidad, como, por ejemplo, porcentaje de datos faltantes. Además, también dispone de funciones para realizar Análisis de Coordenadas Principales (ACP) a partir de datos genotípicos.

3.2.4 Sci-kit learn

Sci-kit learn es una librería de código abierto que implementa distintos algoritmos de inteligencia artificial. Permite la producción de modelos tanto para clasificación supervisada como no supervisada y, además, dispone de herramientas que permiten analizar los resultados obtenidos.

3.2.4.1 Agrupación no supervisada

Sci-kit learn dispone de 7 algoritmos de agrupación no supervisada: K-means, Affinity, Mean-shift, Spectral, Jerárquico, DBSCAN, Gausiano y Birch. No todos estos algoritmos son igual de adecuados para nuestro problema. En nuestro caso las muestras de las que disponemos no están distribuidas uniformemente por variedad de tomate. De algunas variedades se dispone de hasta 30 muestras mientras que de otras solo tenemos 1. Esto hace que el uso de algoritmos que asumen que los grupos obtenidos van a tener tamaños similares, como K-means, Spectral, Birch y gausiano, no parezcan recomendables.

Los 3 algoritmos restantes, (Affinity, DBSCAN y Jerárquico) parecen ser apropiados para nuestros datos, ya que todos permiten obtener grupos de distintos tamaños y son adecuados para datos con más de 3 dimensiones con diferente densidad de puntos.

Cada uno de los algoritmos requiere distintos parámetros, que se le han de proporcionar. Por ejemplo, el algoritmo jerárquico necesita dos parámetros para funcionar: el criterio de enlace, que varia el comportamiento del algoritmo y puede tomar cuatro valores distintos (*single, complete, average y ward*), y el número de grupos que debe formar el algoritmo. Variando cualquiera de los dos cambian los distintos grupos generados.

En el caso de DBSCAN y Affinity, no es necesario indicarle el número de grupos, ya que son capaces de determinar cuantos grupos deben producir. Por ejemplo, DBSCAN, introduce en el mismo grupo aquellas muestras que están a una distancia menor a un valor deseado, que se le proporciona como parámetro. El numero de grupos variará en función de la distancia entre las muestras y el valor límite proporcionado: Las muestras que están lo suficientemente alejadas de todos los grupos se consideran no clasificadas, y se introduce en un grupo especial, con el resto de muestras no clasificadas

El algoritmo Jerárquico dispone de una serie de parámetros que permiten adaptarlo a distintos problemas. El más importante y uno de los que más afectan al resultado final es el criterio de enlace (*linkage*) que tiene 4 opciones distintas (*single, complete, average y ward*). En estos casos, es necesario indicar el número de grupos distintos que se quieren generar. En las variedades silvestres se han utilizado un total de 10 y en las variedades domésticas, 16. Utilizar menos acaba formando pocos grupos muy grandes, mientras que utilizar más provocaba grupos muy dispersos. DBSCAN y Affinity, a diferencia de éste, reciben distintos parámetros específicos, indicando como se generaran los grupos, y el algoritmo decidirá cuantos grupos utilizar.

3.2.5 Visualización de datos:

Tras la realización de las distintas clasificaciones el resultado se almacena en un fichero en el que también se incluye la variedad de tomate a la que pertenece cada muestra. Se han utilizado diversas herramientas para analizar el contenido de estos ficheros de resultados.

3.2.5.1 CurlyWhirly

CurlyWhirly es un software que nos permite visualizar interactivamente puntos en un espacio tridimensional. Además, tiene la capacidad de variar el color de dichos puntos en función de las clasificaciones que le demos.

3.2.5.2 Matplotlib

Matplotlib es una librería que permite crear distintos tipos de gráficas y figuras para facilitar la comprensión visual del lector.

3.3 Métricas de agrupamiento

Existen parámetros que reflejan el grado de dispersión de los grupos obtenidos en las distintas clasificaciones. En este trabajo hemos utilizado: el coeficiente de Shilouette, el índice Calinski-Harabasz y el índice Davies-Bouldin.

4 Resultados y discusión

Se han utilizado 6 algoritmos distintos de clasificación automática. Cada uno de estos algoritmos ha sido empleado para realizar dos clasificaciones: una para separar los grupos genéticos en silvestres y tradicionales y otra para tratar de clasificar las muestras cultivadas en variedades.

Una posible aproximación consiste en realizar PCAs sobre los que podemos crear una clasificación manualmente. La principal limitación de esta aproximación es que sólo podemos ver 3 proyecciones a la vez y esto hace que siempre estemos ignorando una parte de la variación, la que no es recogida por estas tres primeras proyecciones. Además, estas clasificaciones manuales siempre incorporarán un cierto grado de subjetividad. Es por ello que el desarrollo de métodos automáticos de clasificación puede ser muy interesante.

Cada clasificación automática generada fue comparada con clasificación previa disponible. Para ello, se representaron figuras en las que el eje X representó los distintos grupos generados automáticamente (numerados de 0 al numero de grupos existentes; en caso de DBSCAN, numerados de -1, que representa a las muestras que no pertenecen a ningún grupo al total de grupos) mientras que el eje Y representa la clase de la clasificación previa.

4.1 DBSCAN

En las muestras silvestres DBSCAN produjo una clasificación que incluyó un solo grupo con la mayoría de las muestras (76% del total). El resto de muestras o bien fueron clasificadas en otros 6 grupos minoritarios o quedaron sin clasificar (Figura 5). Las muestras que no fueron asignadas a ningún grupo se encuentran en la figura como clasificadas en el grupo -1. Si se ignora el problema del grupo mayoritario, el resto de grupos propuestos por DBSCAN se correspondieron en gran medida con la clasificación previa disponible para estas muestras silvestres (Figuras 5a).

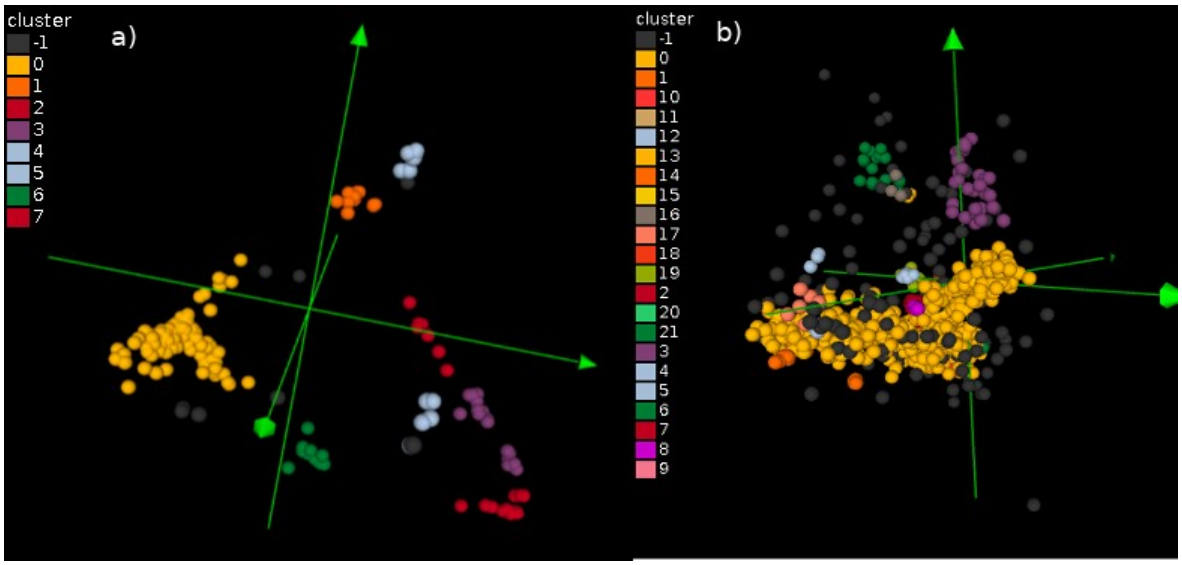

Figura 5: PCA de las muestras utilizadas y su clasificación. a) Muestras silvestres y cultivadas, b) muestras domésticas. Cada color indica un grupo distinto generado por DBSCAN. El color gris oscuro representa a los grupos que no fueron clasificados en ningún grupo

Aparentemente, los índices calculados para evaluar la calidad de la clasificación (Silhouette, Calinski-Harabasz y Davies-Bouldin) (Tabla 2) son bastante buenos. Si atendemos a los grupos que ha formado, el grupo 0, que contiene una mayoría de muestras, que corresponden a las variedades cultivadas, mientras que el resto de grupos corresponden a variedades silvestres

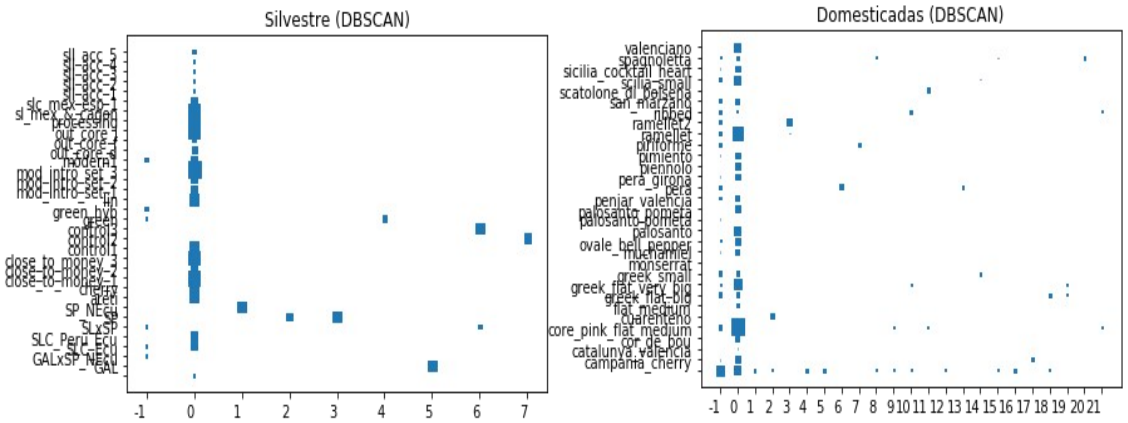

Figura 6: Comparación de las variedades existentes con los grupos generados con el algoritmo DBSCAN para a) las variedades silvestres y b) las variedades domesticadas. Los puntos indican muestras en común para cada par categoría y variedad. Cuanto más grande es el punto, más muestras tienen en común.

El resultado en el caso de las muestras cultivadas fue muy similar, un grupo mayoritario con la mayor parte de las muestras. En este caso los parámetros de calidad no son tan buenos. Esto probablemente se deba a que hay bastantes muestras cultivadas que el DBSCAN no ha clasificado en ningún grupo (Figura 6).

Tabla 2: Parámetros de agrupamiento para la clasificación de variedades silvestres y domésticas con el algoritmo DBSCAN

DBSCAN			
Clasificación	**Silhouette**	**Calinski-Harabasz**	**Davies-Bouldin**
Silvestres	0.797	656.345	0.883
Domesticadas	-0.133	32.811	1.638

4.2 Affinity

Affinity, a diferencia de DBSCAN, produjo, tanto para las muestras silvestres como para las domestidadas, clasificaciones con un elevado número de grupos y con las muestras repartidas equitativamente entre estos grupos (Figura 7).

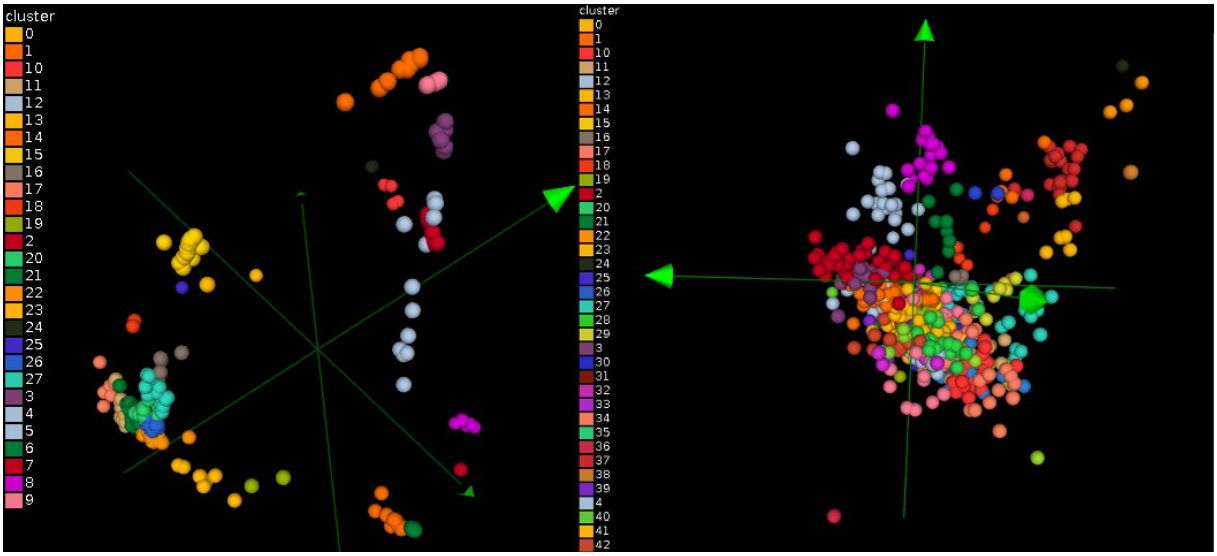

Figura 7: PCA de las muestras utilizadas y su clasificación. a) Muestras silvestres y cultivadas, b) muestras domésticas. Cada color indica un grupo distinto generado por Affinity.

En el caso de Affinity la clasificación no se correspondió con la clasificación previa disponible para la variedades silvestres ni para las domesticadas. La mayor parte de los grupos previos fue dividida por Affinity en varios grupos pequeños con muy pocas muestras (Figura 8).

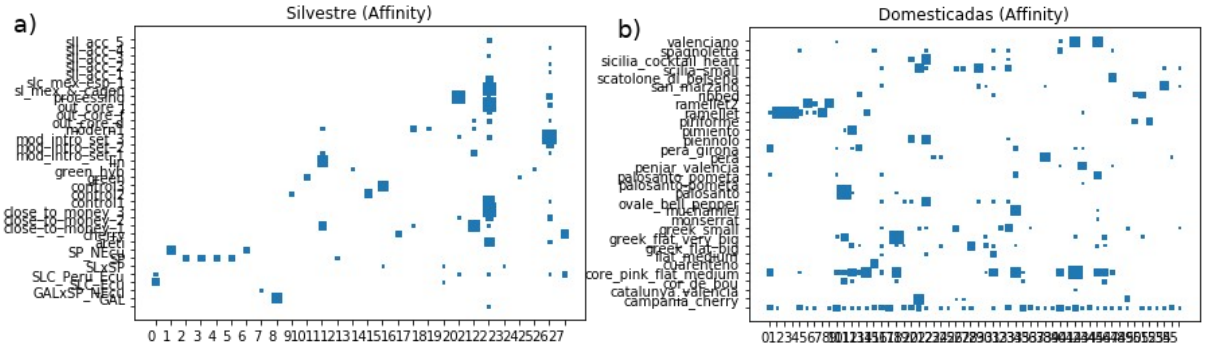

Figura 8: Comparación de las variedades existentes con los grupos generados con el algoritmo Affinity para a) las variedades silvestres y b) las variedades domésticadas. Los puntos indican muestras en común para cada par categoría y variedad. Cuanto mayor sea el punto, más muestras tienen en común

Los parámetros de calidad para Affinity no fueron buenos (Índice de silhouette de 0,48 y Davies-Bouldin de 0.808) (Tabla 3). Probablemente los grupos creados fueron demasiado numerosos. Además, en el caso de las variedades domésticas fueron grupos parcialmente solapantes.

Tabla 3: Parámetros de agrupamiento para la clasificación de variedades silvestres y domésticas con el algoritmo Affinity

Affinity			
Clasificación	**Silhouette**	**Calinski-Harabasz**	**Davies-Bouldin**
Silvestres	0.488	2280.429	0.808
Domesticadas	0.248	237.616	1.029

4.3 Jerárquico

4.3.1 Single

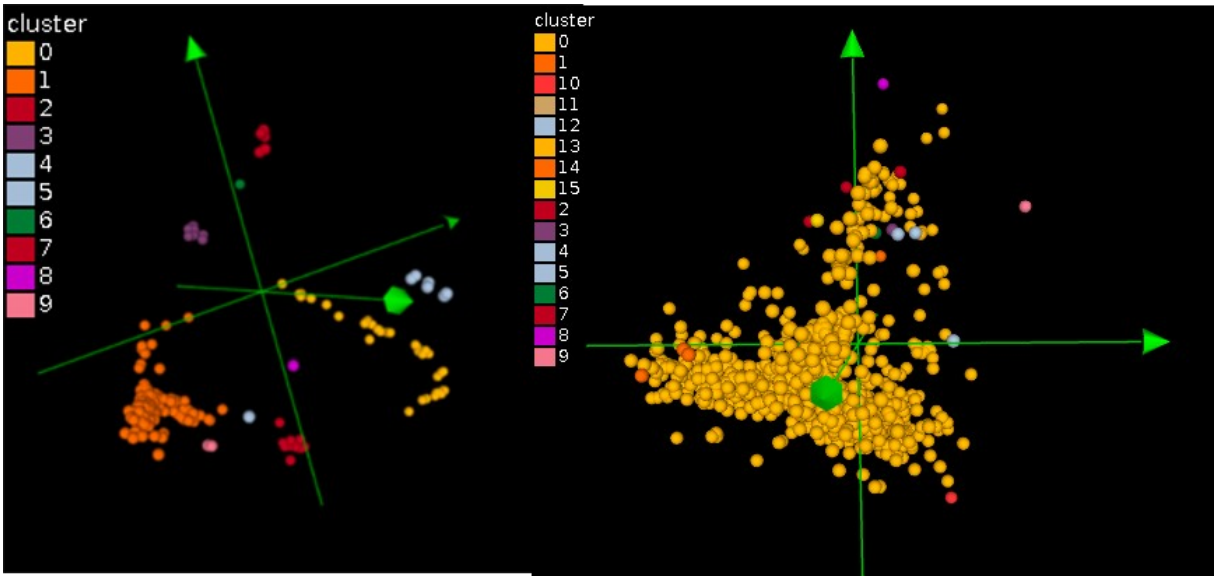

Figura 9: Figura 9: PCA de las muestras utilizadas y su clasificación. a) Muestras silvestres, b) muestras domésticas. Cada color indica un grupo distinto generado por el algoritmo jerárquico Single.

En el caso de las variedades silvestres, single produjo diez grupos distintos; uno de ellos contiene la mayoría de las muestras, correspondiente con las variedades cultivadas, y el resto de muestras se encuentran repartidas en los otros 9 grupos. En el caso de las variedades domésticas, creó 16 grupos, pero casi todas las muestras se localizaron en un solo, siendo el resto de los 15 grupos muy poco numerosos. (Figura 9)

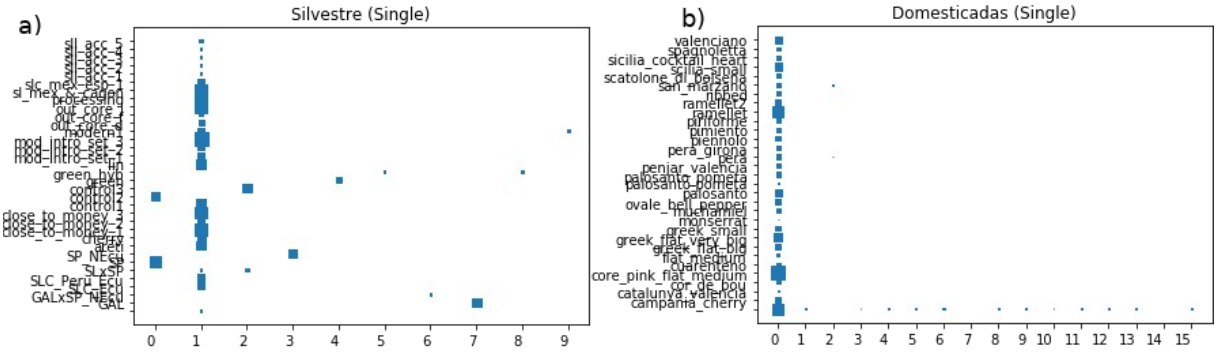

Figura 10: Comparación de las variedades existentes con los grupos generados con el algoritmo jerárquico Single para a) las variedades silvestres y cultivadas y b) las variedades domésticadas. Los puntos indican muestras en común para cada categoría y variedad. Cuanto mayor sea el punto, más muestras tienen en común

Los parámetros de calidad para las variedades silvestres fueron buenos, aunque, como en el caso del DBSCAN, la clasificación creada fue similar a la clasificación conocida, ya que contiene un grupo con las variedades cultivadas y en los 9 grupos resatantes contenian variedades silvestres En cambio, en el caso de las variedades domésticas, los parámetros fueron contradictorios; si bien dos de ellos fueron buenos (Silhouette y Davies-Bouldin), el tercero (Calinski-Harabasz) indicó que la clasificación era muy pobre. De las 1193 muestras distintas de variedades domésticas, 1171 (98,15%) se encontraron en el mismo grupo, mientras que los otros 15 grupos se repartieron solo las 22 muestras restantes (Figura 10b).

Tabla 4: Parámetros de agrupamiento para la clasificación de variedades silvestres y domésticas con el algoritmo Jerárquico Single

Jerárquico (Single)			
Clasificación	**Silhouette**	**Calinski-Harabasz**	**Davies-Bouldin**
Silvestres	0.687	496.031	0.341
Domesticadas	0.690	8.008	0.452

4.3.2 Complete

El resultado de complete fue similar al del single. Creó 10 grupos para las variedades silvestres y 16 para las variedades domésticas. En el caso de las variedades silvestres y cultivadas las variedades cultivadas no se encuentran repartidas en dos grupos El resto de grupos que formó contuvo muestras de varias variedades silvestres. (Figura 12)

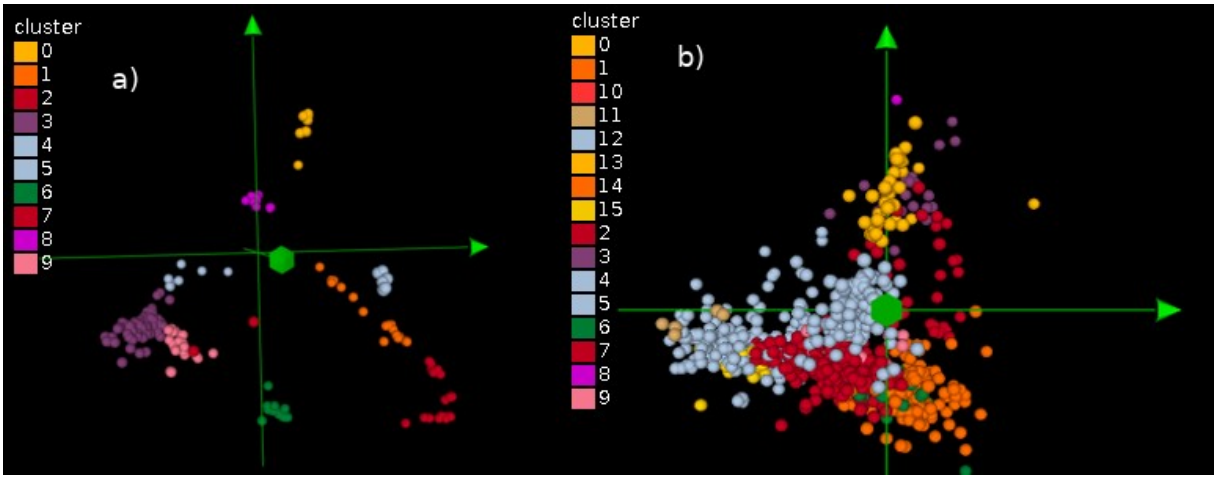

Figura 11: PCA de las muestras utilizadas y su clasificación. a) Muestras silvestres y cultivadas, b) muestras domésticas. Cada color indica un grupo distinto generado por el algoritmo jerárquico Complete.

En el caso de las variedades domésticas, las muestras se repartieron más uniformemente en varios grupos. Sin embargo, hay tres de ellos (grupos 1, 2 y 4, figura 12) que incluyeron una mayor proporción de muestras respecto a los demás.

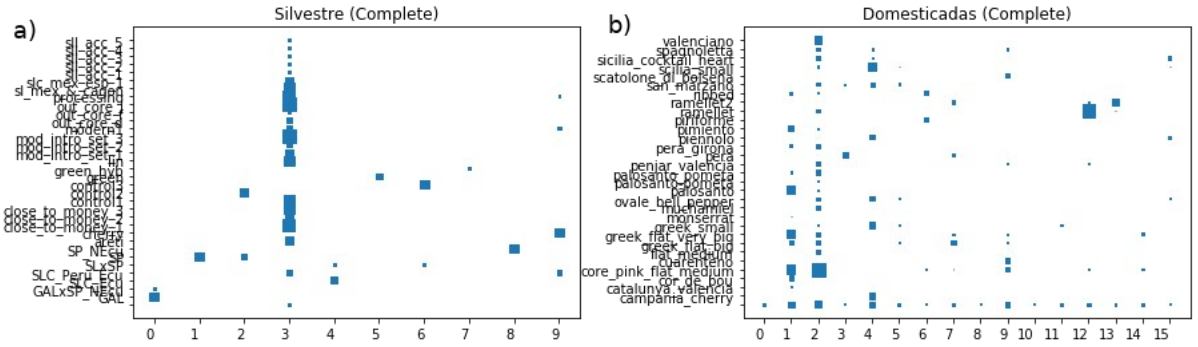

Figura 12: Comparación de las variedades existentes con los grupos generados con el algoritmo jerárquico Complete para a) las variedades silvestres y b) las variedades domesticadas. Los puntos indican muestras en común para cada par categoría y variedad. Cuanto mayor sea el punto, más muestras tienen en común

Si atendemos a los parámetros de calidad (Tabla 5), la clasificación de variedades silvestres fue buena, y varias variedades se correspondieron exactamente los grupos creados automáticamente, pero en las variedades domésticas, los resultados no fueron tan buenos. En éste caso, las variedades se encuentran totalmente repartidas en los distintos grupos generados. (Figura 12)

Tabla 5: Parámetros de agrupamiento para la clasificación de variedades silvestres y domésticas con el algoritmo Jerárquico Complete

Jerárquico (Complete)			
Clasificación	**Silhouette**	**Calinski-Harabasz**	**Davies-Bouldin**
Silvestres	0.642	955.075	0.505
Domesticadas	0.218	240.273	0.927

4.3.3 Ward

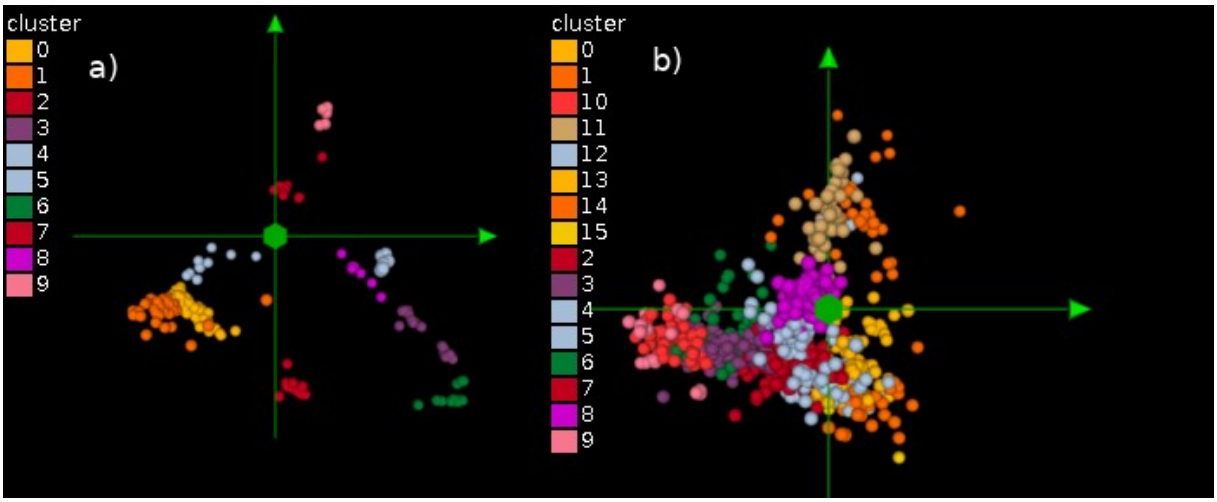

Figura 13: PCA de las muestras utilizadas y su clasificación. a) Muestras silvestres, b) muestras domésticas. Cada color indica un grupo distinto generado por el algoritmo jerárquica ward.

En las variedades silvestres ward produjo 10 grupos. tres de ellos (0, 1 y 4) corresponden a las variedades cultivadas y el resto a variedades silvestres(Figura 13a). En las variedades domésticas, los grupos se encuentran solapados (Figura 13b) y cada uno de ellos contiene muestras de varias de las variedades existentes (Figura 14b)

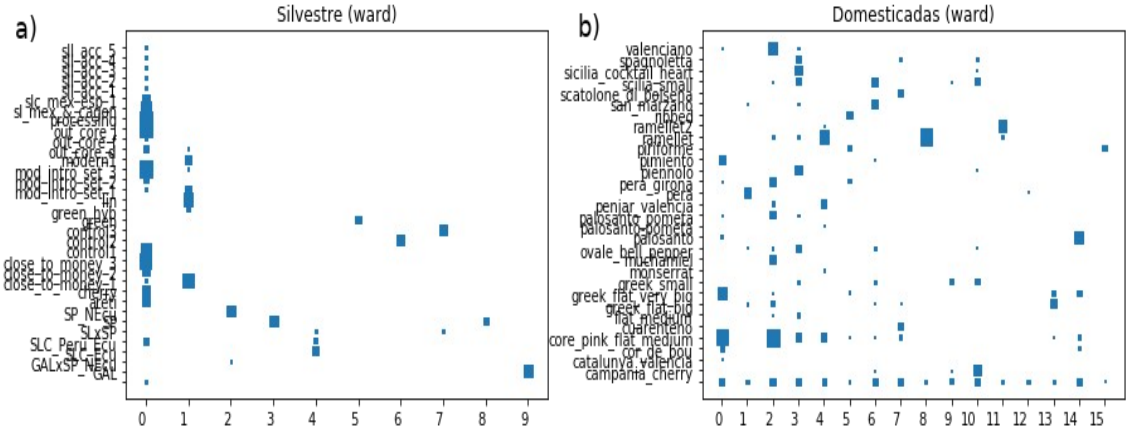

Figura 14:Comparación de las variedades existentes con los grupos generados con el algoritmo jerárquico ward para a) las variedades silvestres y b) las variedades domesticadas. Los puntos indican muestras en común para cada par categoría y variedad. Cuanto mayor sea el punto, más muestras tienen en común

La clasificación de las variedades silvestres, teniendo en cuenta los parámetros de calidad (Un índice de silhouette de 0.6 y un índice de Davies-Bouldin de 0.49) (Tabla 6) fue buena. En cambio, estos mismos parámetros para el caso de las variedades domésticas, no fueron tan buenos (Slihouete de 0.24 y Davies-Bouldin de 1.11).

Tabla 6: Parámetros de agrupamiento para la clasificación de variedades silvestres y domésticas con el algoritmo Jerárquico Average

Jerárquico (Ward)			
Clasificación	**Silhouette**	**Calinski-Harabasz**	**Davies-Bouldin**
Silvestres	0.602	1078.232	0.490
Domesticadas	0.240	323.947	1.116

4.3.4 Average

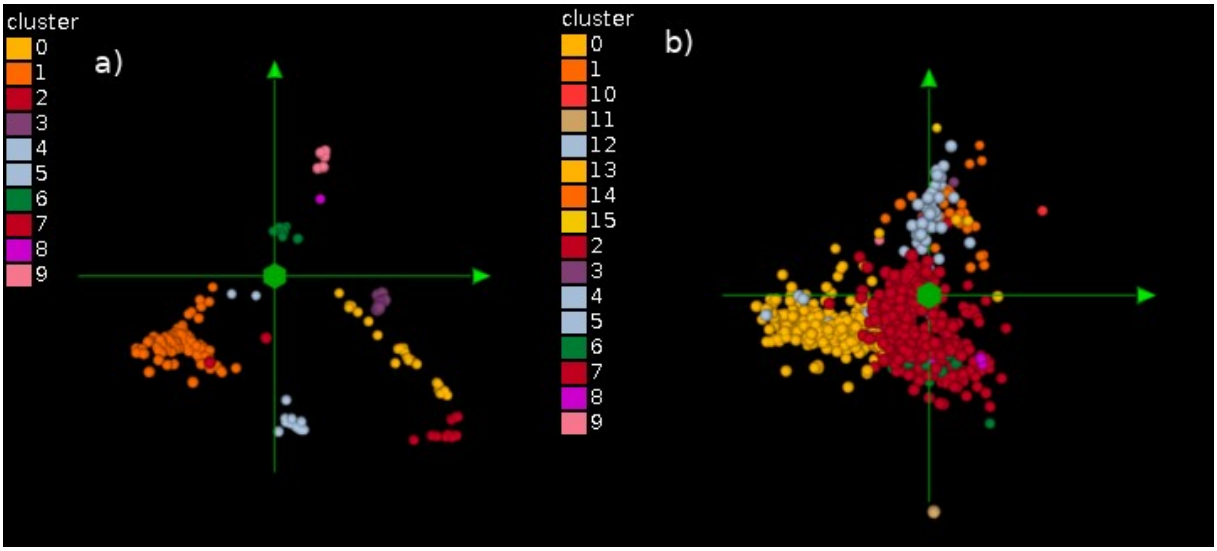

Figura 15: PCA de las muestras utilizadas y su clasificación. a) Muestras silvestres, b) muestras domésticas. Cada color indica un grupo distinto generado por el algoritmo jerárquico Average

En las muestras silvestres, average produjo 10 grupos. Un grupo fue mayoritario con 337 muestras (75,22%), que contiene las muestras cultivadas, mientras que en las variedades domésticas, 2 grupos incluyeron 1103 muestras (92,0% del total) (Figura 15).

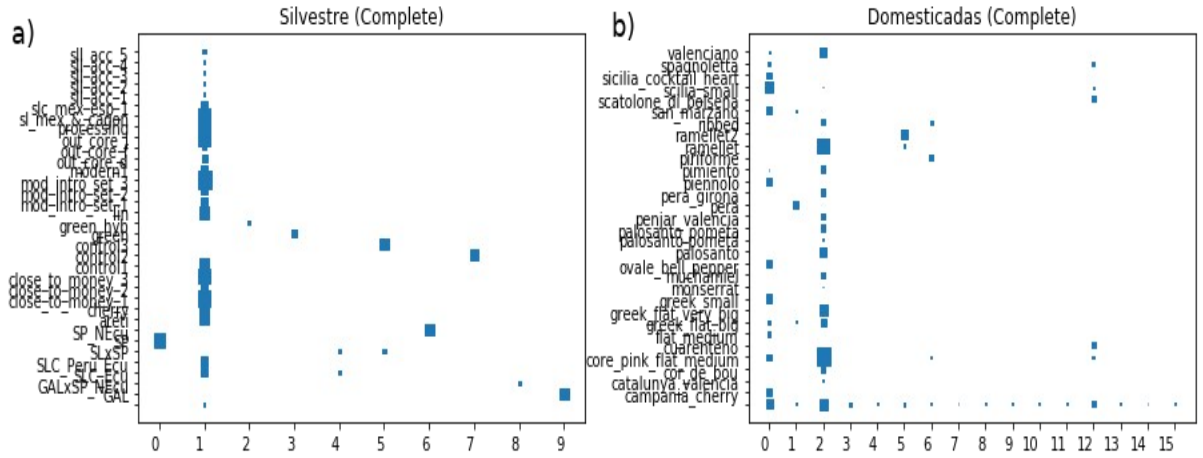

Figura 16:Comparación de las variedades existentes con los grupos generados con el algoritmo jerárquico average para a) las variedades silvestres y b) las variedades domesticadas. Los puntos indican muestras en común para cada par categoría y variedad. Cuanto mayor sea el punto, más muestras tienen en común

De acuerdo con los parámetros de calidad (Tabla 7), la clasificación de variedades silvestresfue buena, Con indice de silhouette de 0.756, y de davies-bouldin de 0.354, los grupos creados, correspondientes tanto a las variedades cultivadas (grupo 0, figura 15a) , mientras que las variedades silvestres se corresponden exactamente con los grupos creados automáticamente (Figura 16a). En cambio, las muestras variedades domésticas se concentran en dos grupos principalmente. Por ello, los índices en este caso son bastante bajos (Silhouette 0.30) (Tabla 7)

Tabla 7: Parámetros de agrupamiento para la clasificación de variedades silvestres y domésticas con el algoritmo Jerárquico Average.

Jerárquico (Average)			
Clasificación	**Silhouette**	**Calinski-Harabasz**	**Davies-Bouldin**
Silvestres	0.756	729.777	0.354
Domesticadas	0.308	118.849	0.627

4.4 Comparación de los distintos algoritmos

En el caso de las muestras silvestres los algoritmos pudieron crear clasificaciones que se correspondieron en buena medida con la clasificación previa disponible. Hasta 7 variedades distintas (dependiendo del algoritmo utilizado) pudieron separarse del resto de muestras. Sin embargo, todos los algoritmos generan un grupo con un porcentaje elevado de las muestras El algoritmo Average fue el que tuvo un mejor comportamiento para las variedades

silvestres, ya que clasificó correctamente, en las variedades existentes un mayor número de muestra.

De los 6 algoritmos distintos que se compararon, DBSCAN y las opciones Single y Average del algoritmo jerárquico formaron las clasificaciones más similares a las variedades ya existentes. Estas clasificaciones consiguen, por un lado, un grupo correspondiente a las variedades cultivadas, que es el que contiene la mayoría de las muestras, y por otro, pequeños grupos correspondientes a las variedades silvestres.

Affinity, en cambio, produjo muchos grupos pequeños que incluían pocas muestras. Cada grupo incluyó muestras que, de acuerdo con la información previa, pertenecían a distintas variedades que el algoritmo no fue capaz de diferenciar. Por último, las diferentes opciones del algoritmo jerárquico (single, complete, ward y average) produjeron resultados similares en el caso de las variedades silvestres, con grupos pequeños que correspondían con las clasificaciones previas, aunque ward y complete no clasificaron correctamente las muestras cultivadas, y las dividieron en varios grupos.

En general, ninguno de los seis algoritmos fue capaz de crear grupos que se correspondieses con las variedades domésticas conocidas. Éstas muestras se encuentran mezcladas entre sí y no se han podido separar en grupos homogéneos y poco dispersos. Las muestras de todas las variedades se reparten entre múltiples grupos distintos, dando lugar a grupos con muchas variedades, y variedades repartidas en muchos grupos.

5 Conclusiones

- Los algoritmos DBSCAN, Average y single fueron capaces de clasificar correctamente varias de las variedades silvestres y cultivadas.
- El algoritmo jerárquico Average produjo la clasificación más similar a la disponible previamente: un grupo que contiene las variedades cultivadas y otros 6 que cada uno contiene una variedad silvestre.
- Los algoritmos jerárquicos fueron capaces de formar grupos correspondientes a las variedades silvestres existentes, aunque no, seguido de DBSCAN, siendo el último Affinity, que no fue capaz de formar ninguno.
- Ninguno de los algoritmos utilizados fue capaz de clasificar correctamente las variedades domesticadas.
- Los algoritmos DBSCAN, Single y Complete produjeron clasificaciones en las que las variedades silvestres se encontraba en un grupo,
- El algoritmo Affinity produjo demasiados grupos, repartiendo las muestras correspondientes a las distintas variedades entre varios de ellos.
- El algoritmo jerárquico Single tiende a producir un grupo con elevada cantidad de muestras, mientras el resto de grupos tienen muy pocas.
- Los Índices de calidad de las clasificaciones caracterizadas que conseguían separar las variedades cultivadas de las variedades silvestres eran superiores a aquellos que no.

6 Anexos

Todo el código utilizado para generar los modelos de agrupamiento se incluye como anexo a éste documento. Éstos se estructuran de la siguiente forma:

- Un archivo python llamado “curly.py” que incluye el código desarrollado en esto proyecto y que contiene el código necesario
- Una serie de libretas Jupyter que permiten reproducir los algoritmos utilizados para obtener los resultados de cada uno de los agrupamientos. Éstas contienen el código y anotaciones varias.
    - “Clasificación de variedades silvestres y cultivadas.ipynb” que contiene la clasificación de todas las variedades silvestres, su distribución, la comparación con la clasificación tradicional de las muestras de esta categoría y los parámetros asociados a esta clasificación
    - “Clasificación de variedades domésticas.ipynb” que contiene la clasificación de todas las variedades domésticas, su distribución, la comparación con la clasificación tradicional de las muestras de esta categoría y los parámetros asociados a esta clasificación

7 Bibliografía

1. Bai, Y. and Lindhout, P. (2007). Domestication and Breeding of Tomatoes: What have We Gained and What Can We Gain in the Future?. *Annals of Botany*, 100(5), pp.1085-1094.

2. Bergougnoux, V. (2014). The history of tomato: From domestication to biopharming. *Biotechnology Advances*, 32(1), pp.170-189.

3. Bridges, M., Heron, E., O'Dushlaine, C., Segurado, R., Morris, D., Corvin, A., Gill, M. and Pinto, C. (2011). Genetic Classification of Populations Using Supervised Learning. *PLoS ONE*, 6(5), p.e14802.

4. Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., Vanderplas, J., Joly, A., Holt, B. and Varoquaux, G. (2013). *API design for machine learning software: experiences from the scikit-learn project*.

5. Calinski, T. and Harabasz, J. (1974). A Dendrite Method for Cluster Analysis. *Communications in Statistics - Simulation and Computation*, 3(1), pp.1-27.

6. Davies, D. and Bouldin, D. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2), pp.224-227.

7. Gupta, S., Tran, T., Luo, W., Phung, D., Kennedy, R., Broad, A., Campbell, D., Kipp, D., Singh, M., Khasraw, M., Matheson, L., Ashley, D. and Venkatesh, S. (2013). *Machine-learning prediction of cancer survival: a retrospective study using electronic administrative records and a cancer registry*. [online] BMjournal.

8. Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2001). *On cluster validation techniques. Journal of Intelligent Information Systems*, 17(2/3), pp.107-145.

9. Hunter, J. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3), pp.90-95.

10. Libbrecht, M. and Noble, W. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6), pp.321-332.

11. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, É. (2011). *Scikit-learn: Machine Learning in Python*.

12. Rousseeuw, P. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, pp.53-65.