Reconocimiento y clasificación de marcas de Tequila a través de espectroscopia UV-Visible y análisis multivariante

Victor Ulises Lev Contreras Loera

Contenido

- Objetivo
- Motivación
- Introducción
- Antecedentes
- Adquisición de datos
- Análisis
- Validación
- Conclusiones

Motivación

- Gran cantidad de marcas
- Identificar adulteraciones
- Control de calidad

Objetivo :

- Elaborar una técnica rápida, no destructiva basada en métodos ópticos y estadísticos capaz de reconocer y clasificar Tequilas.

Introducción

Categorías

Tipos

Tequilas

Mixtos

100 % Agave

Blancos

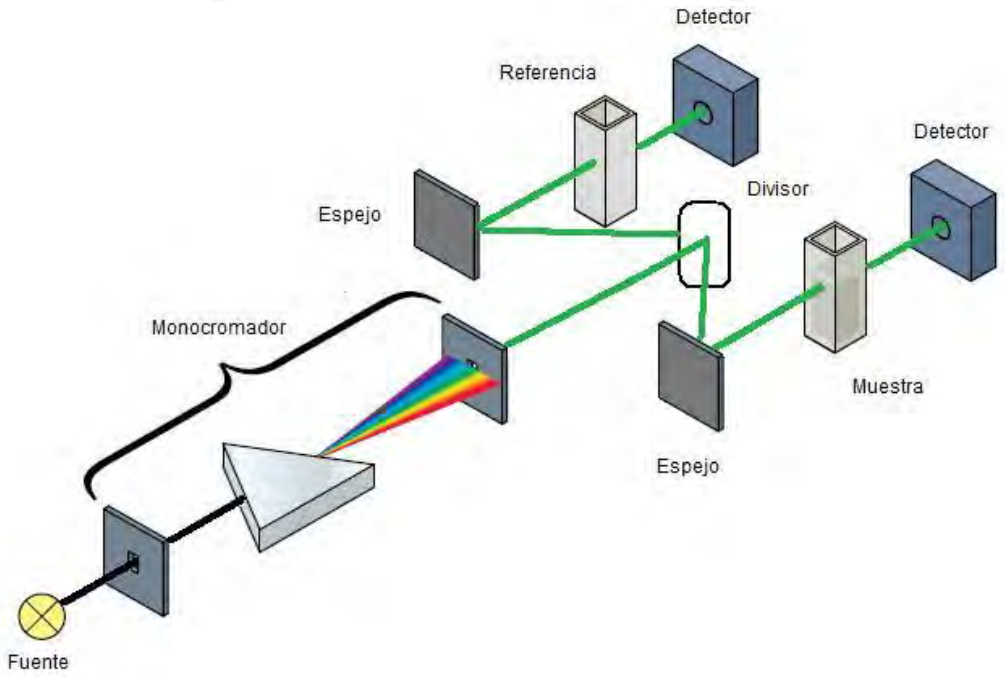
Reposados

Añejos

Metodología

- Métodos Ópticos: Adquisición de datos
- Métodos Estadísticos (PCA, SVM): Reducción de dimensionalidad y clasificación
- Otros (LDA, Elipses de confiabilidad)

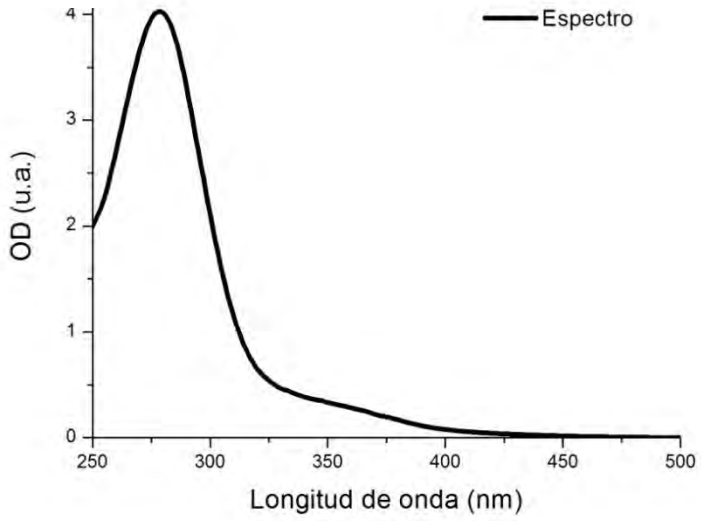
Espectroscopia de absorción



Espectrofotómetro

Ley de Beer

$$A = \log\left(\frac{I_0}{I}\right) = \varepsilon l c$$



Componentes principales:

Combinación lineal de las variables originales con máxima varianza

$$z_i = Xa_i$$

$$z'_i z_i = a'_i X' X a_i = a'_i S a_i$$

Restricción:

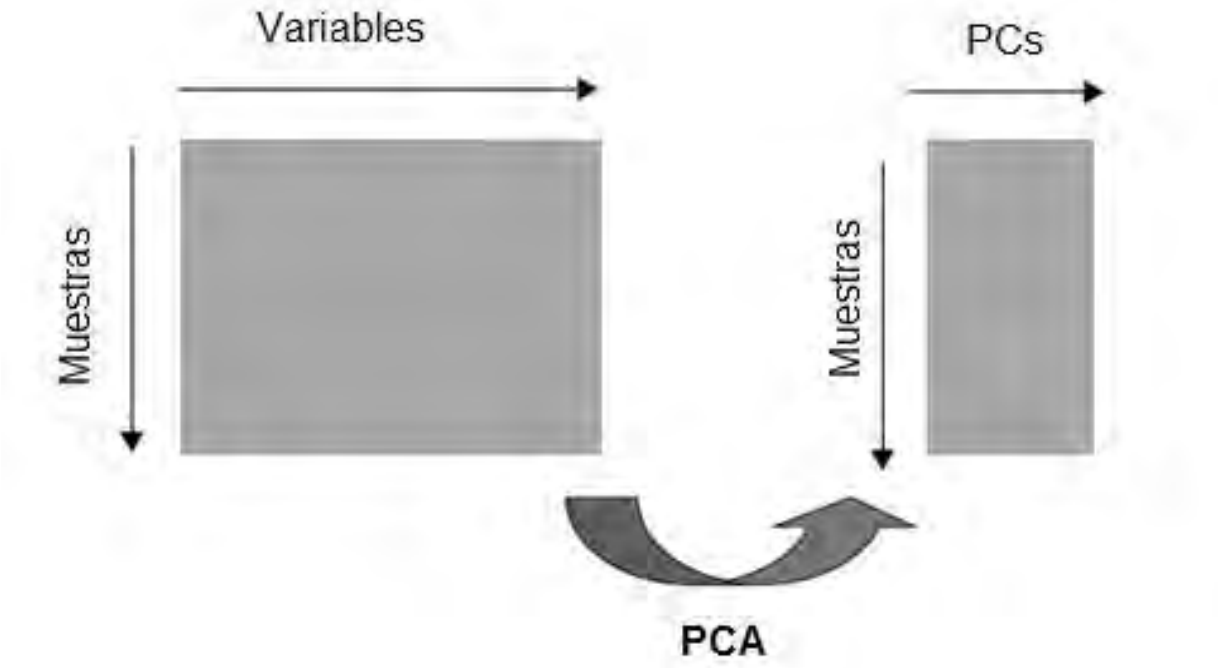
$$a'_i a_i = 1$$

Multiplicadores de Lagrange:

$$a'_i S a_i = \lambda$$

Análisis de Componentes Principales (PCA)

$$X \cdot a = z$$

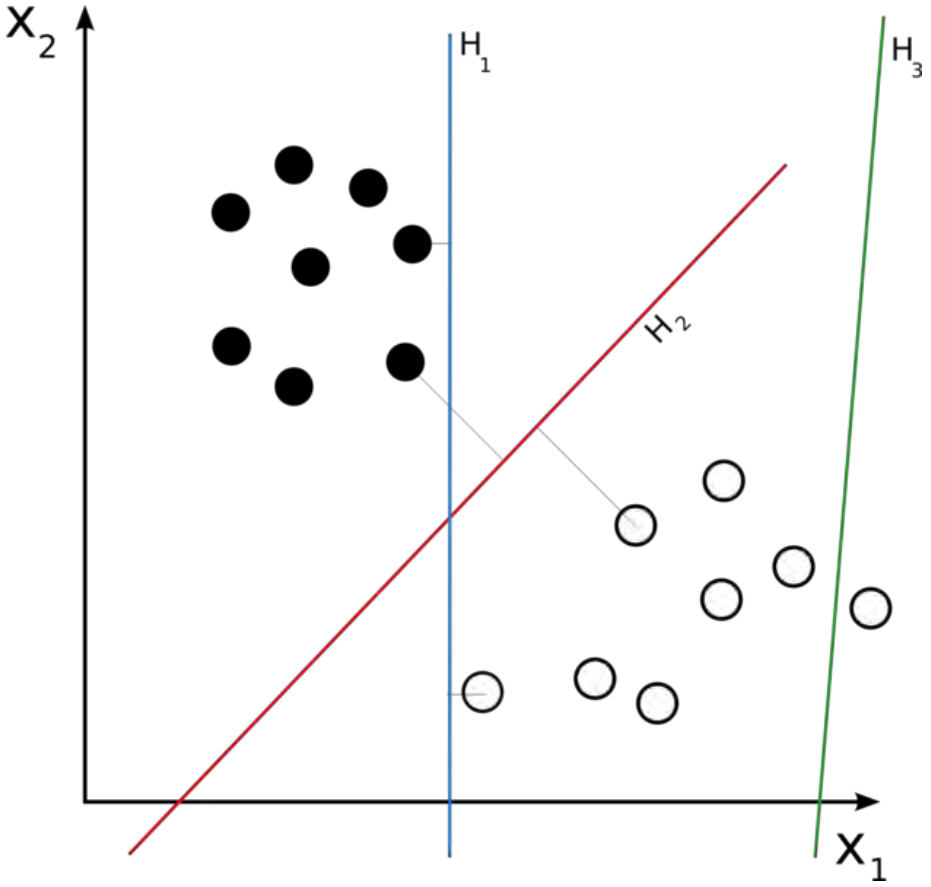


Transformación lineal

Support Vector Machines

$$D = \left(x_i, c_i \right) \middle| x_i \in R^p, c_i \{-1,1\}$$

$$w \cdot x - b = 0$$



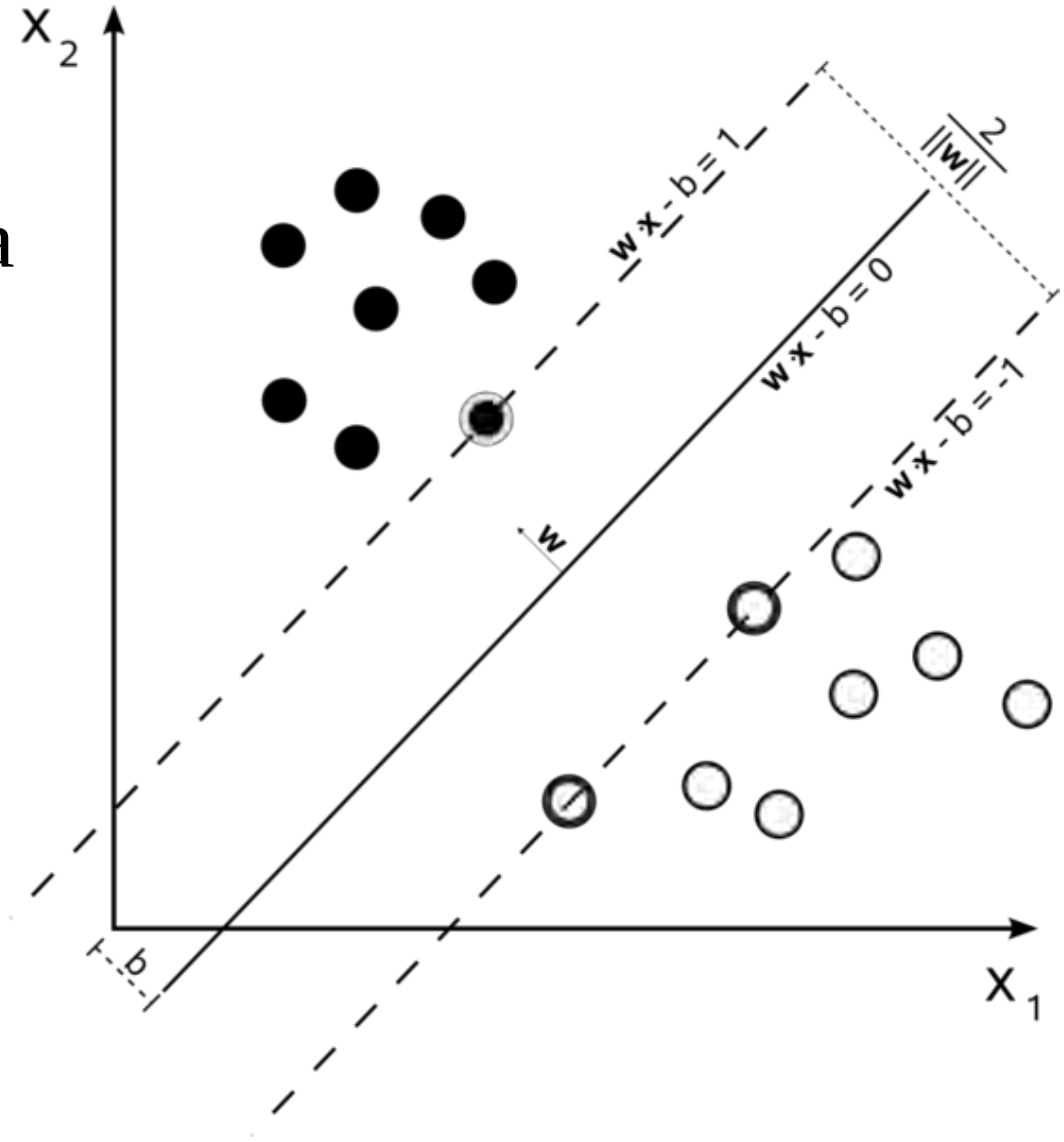
Support Vector Machines

encontrar w, b para

$\min imizar \|w\|$

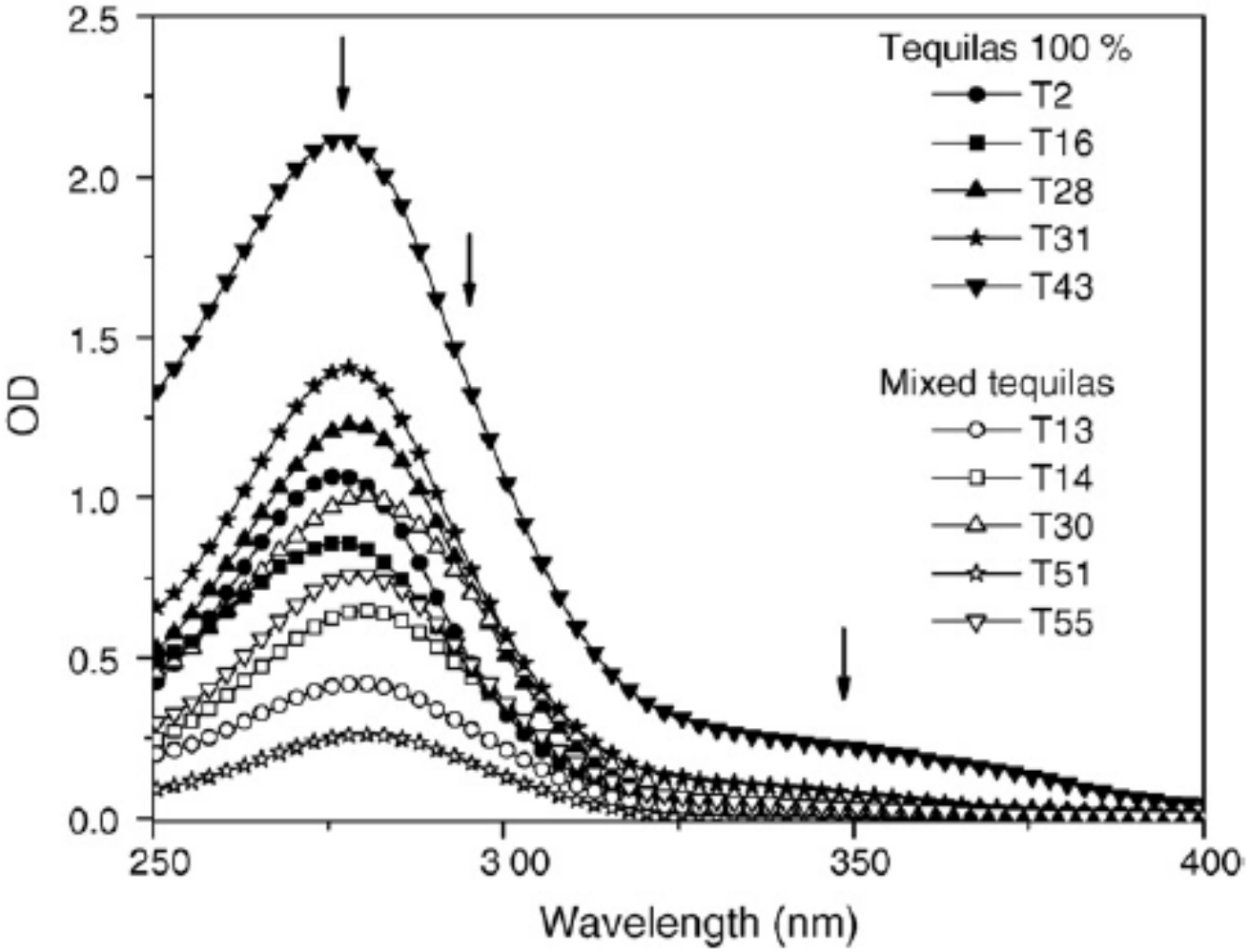
sujeto a

$c_i\left(w \cdot x_i - b\right) \geq 1$



Antecedentes

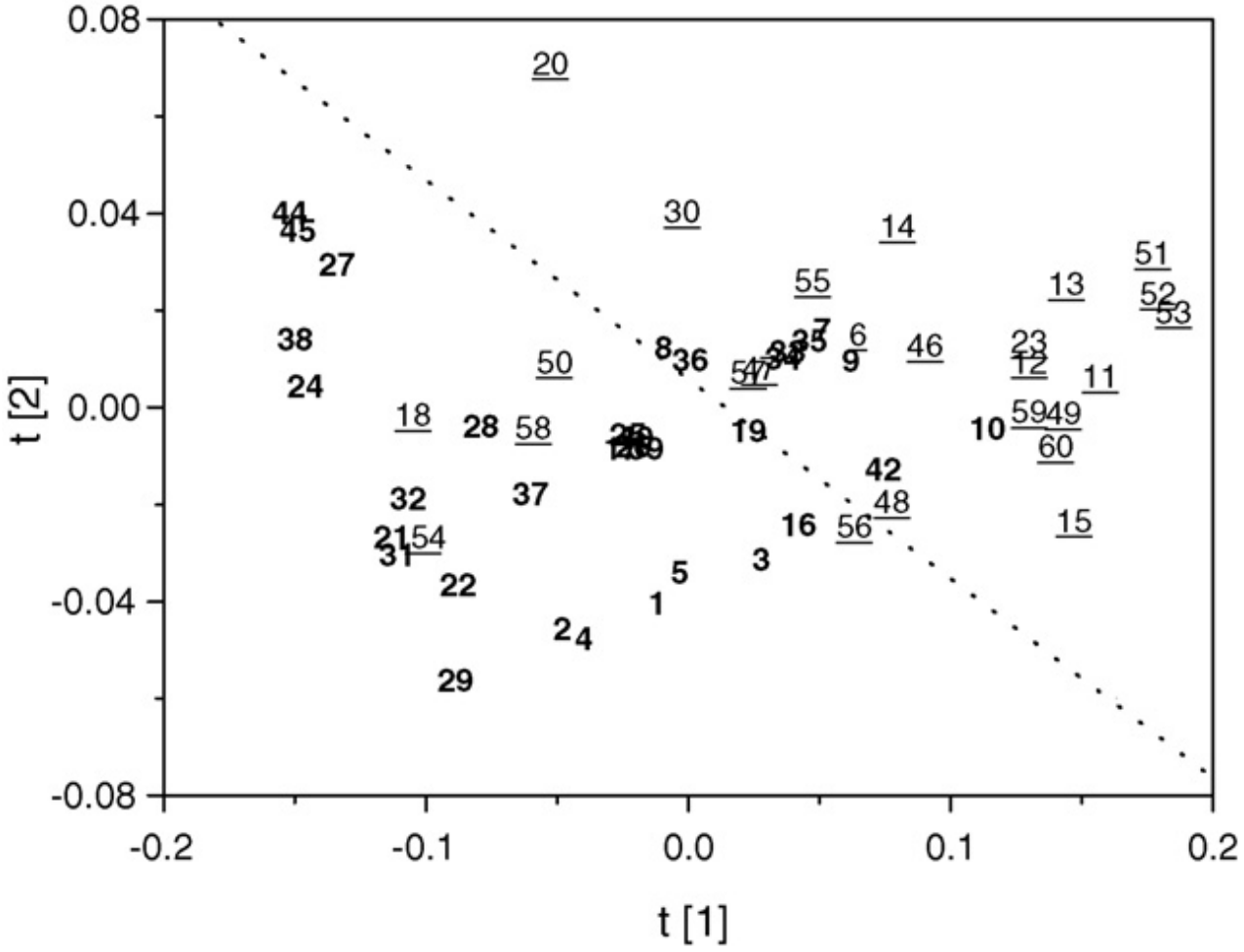
Tequilas Blancos



Absorción en tequilas blancos, 100% agave y mixtos

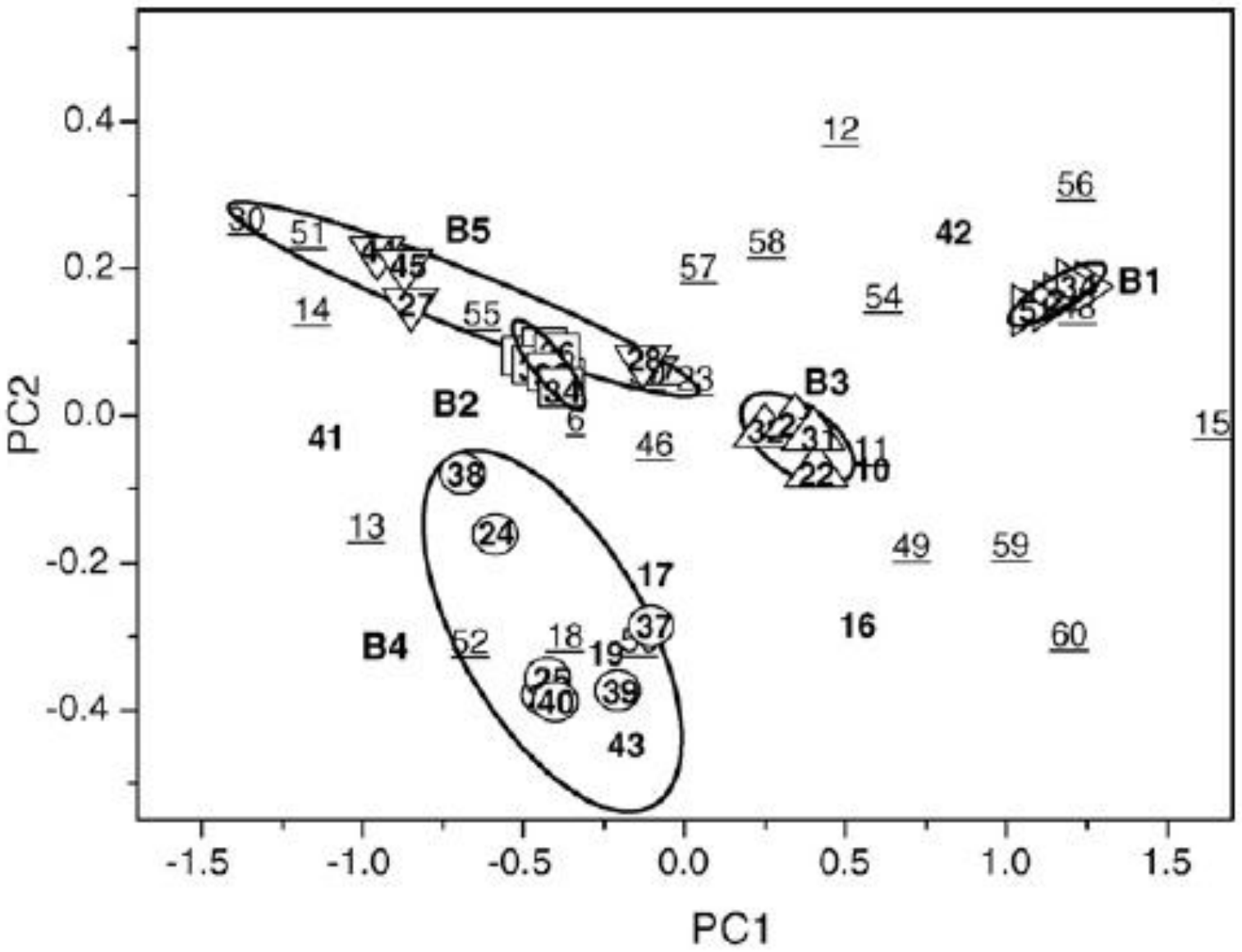
Referencias del articulo "UV-Vis absorption spectroscopy and multivariate analysis as a method to discriminate tequila" O. Barbosa et al.

Métodos Ópticos + Métodos Multivariantes



Referencias del articulo "UV-Vis absorption spectroscopy and multivariate analysis as a method to discriminate tequila" O. Barbosa et al

Componentes Principales

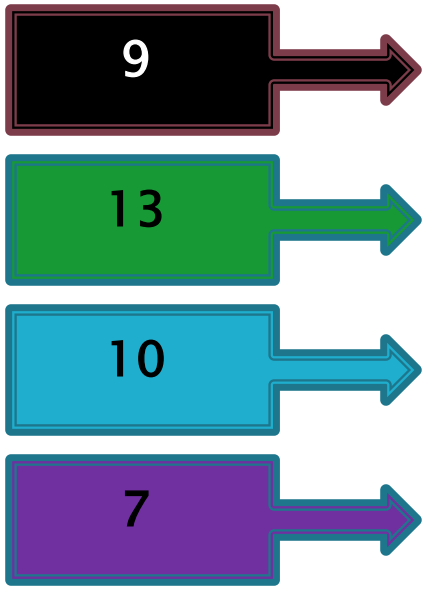


Referencias del articulo "UV-Vis absorption spectroscopy and multivariate analysis as a method to discriminate tequila" O. Barbosa et al

Adquisición de datos

Tequilas blancos

4
marcas



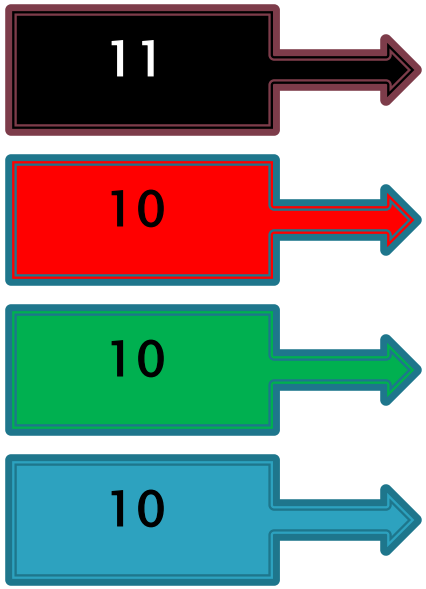
39
muestras

Marcas
Certificadas

Diferentes
lotes

Tequilas reposados

4
marcas



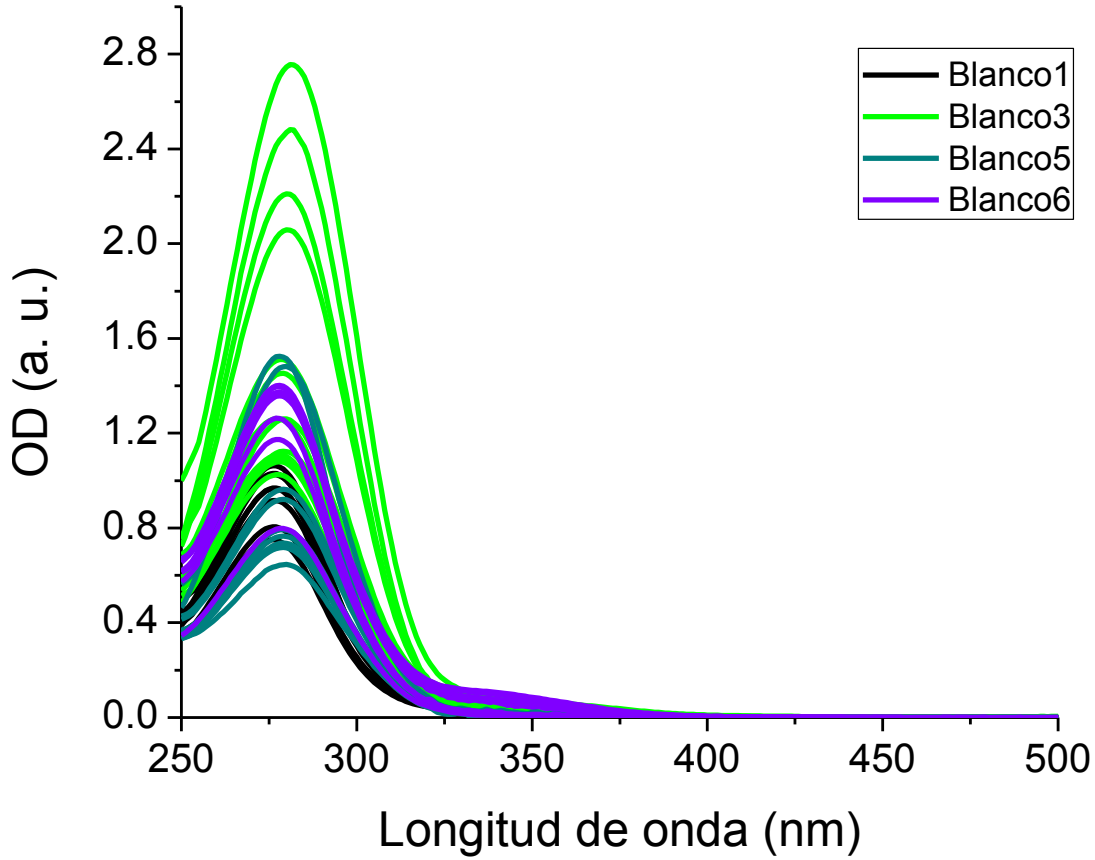
41
muestras

Marcas
Certificadas

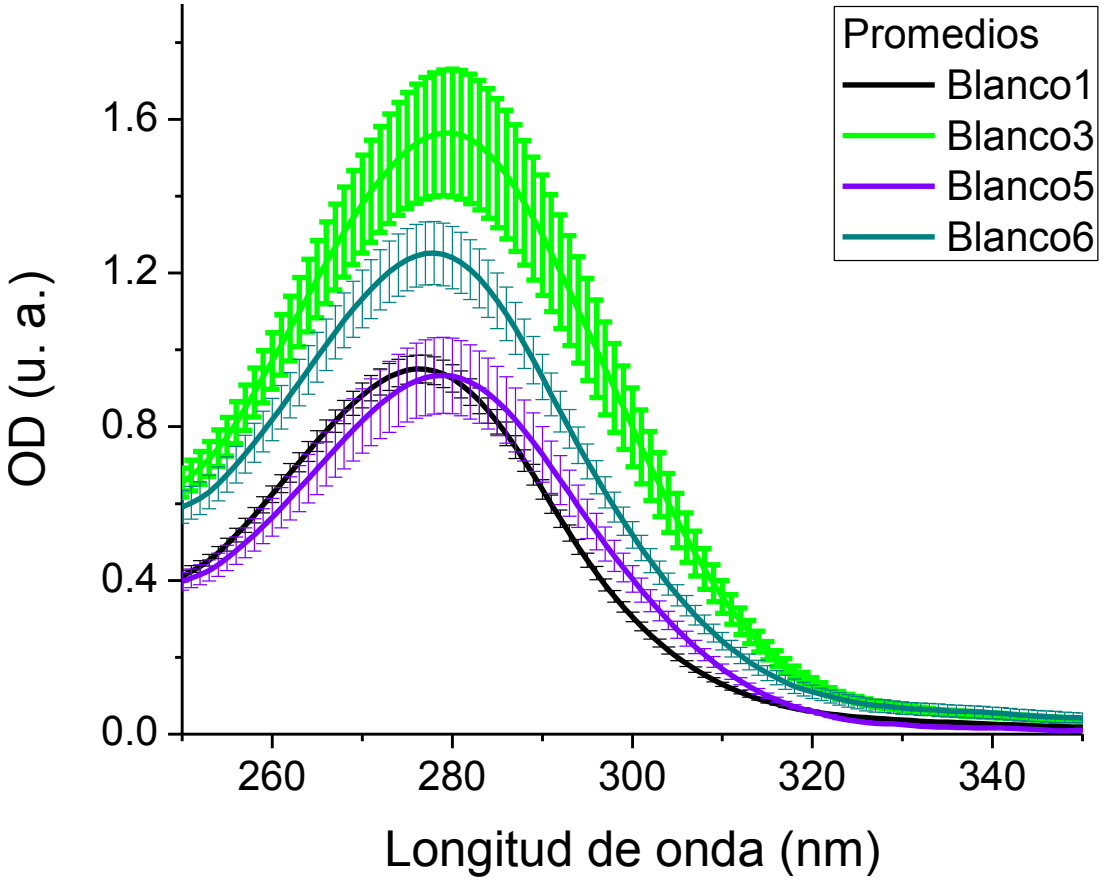
Diferentes
lotes

Resultados

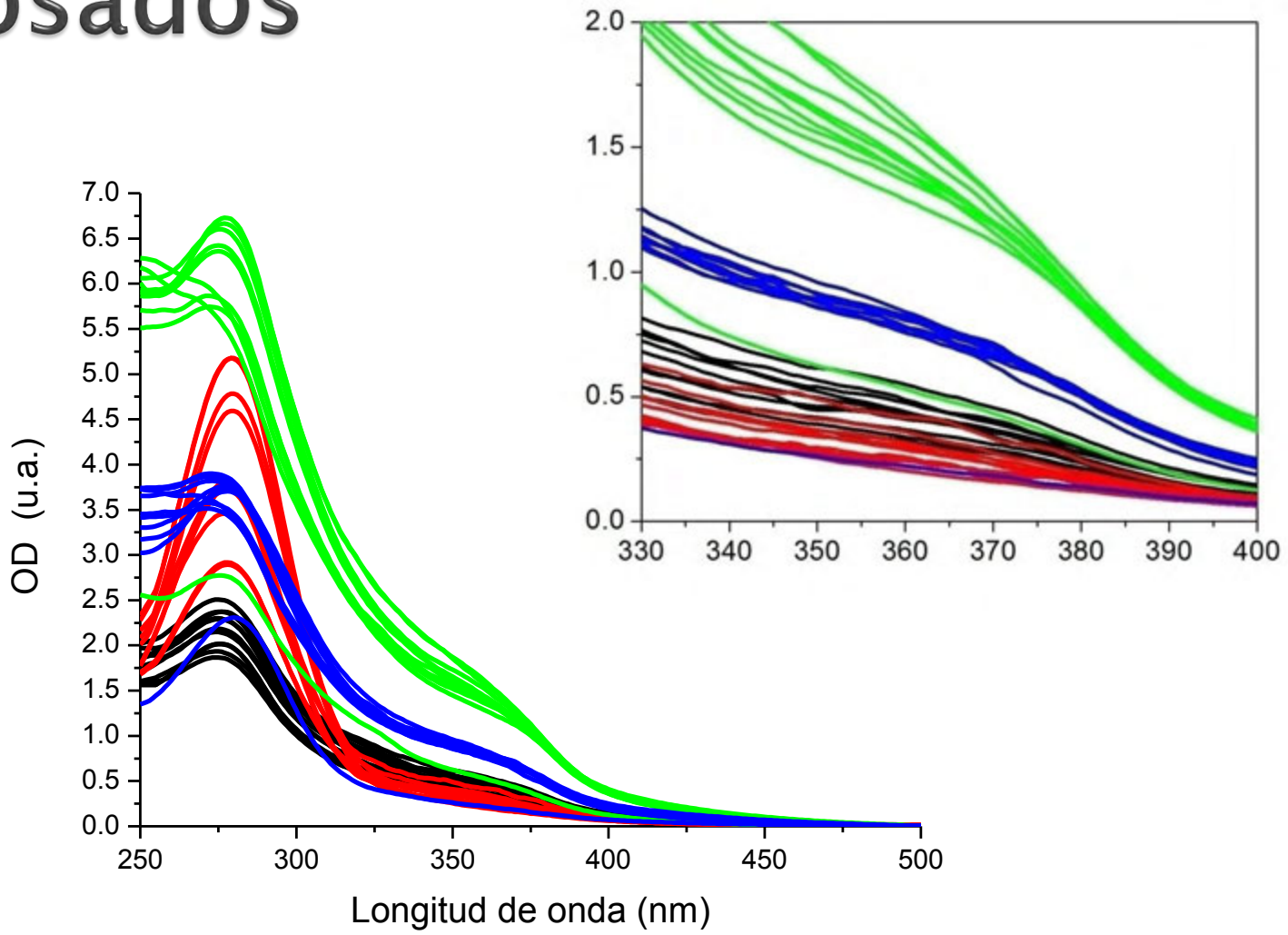
Espectros de absorción

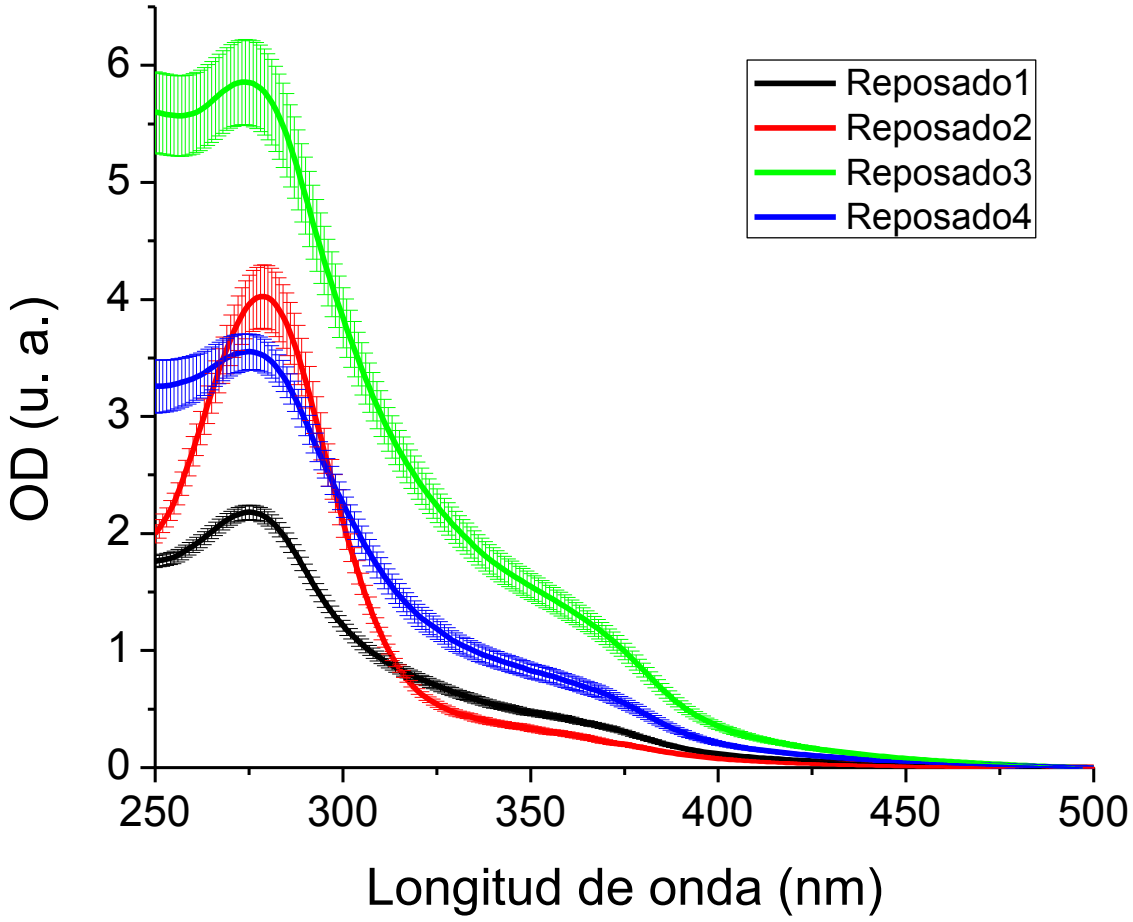


Promedio


Promedios
Blanco1
Blanco3
Blanco5
Blanco6
OD (u. a.)
0.0
0.4
0.8
1.2
1.6
260
280
300
320
340
Longitud de onda (nm)

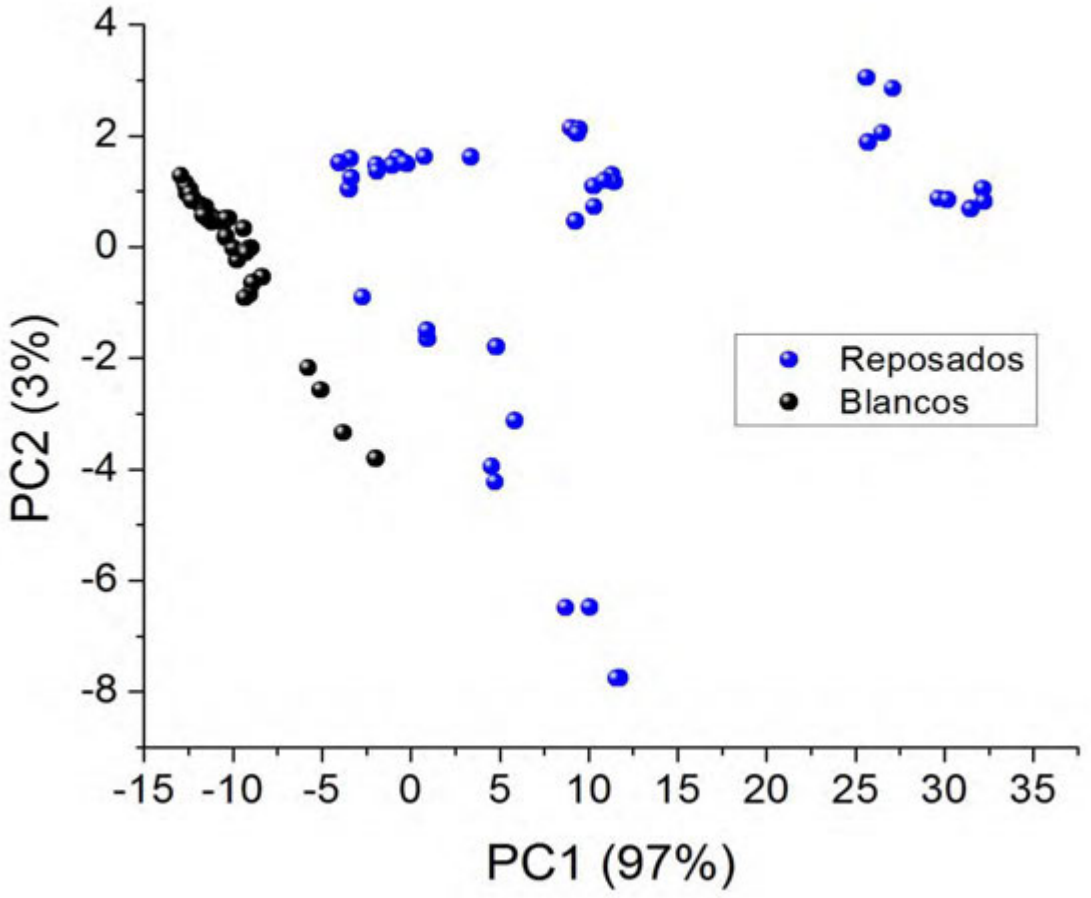
Reposados


OD (u.a.)
Longitud de onda (nm)
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
250
300
350
400
450
500
2.0
1.5
1.0
0.5
0.0
330
340
350
360
370
380
390
400

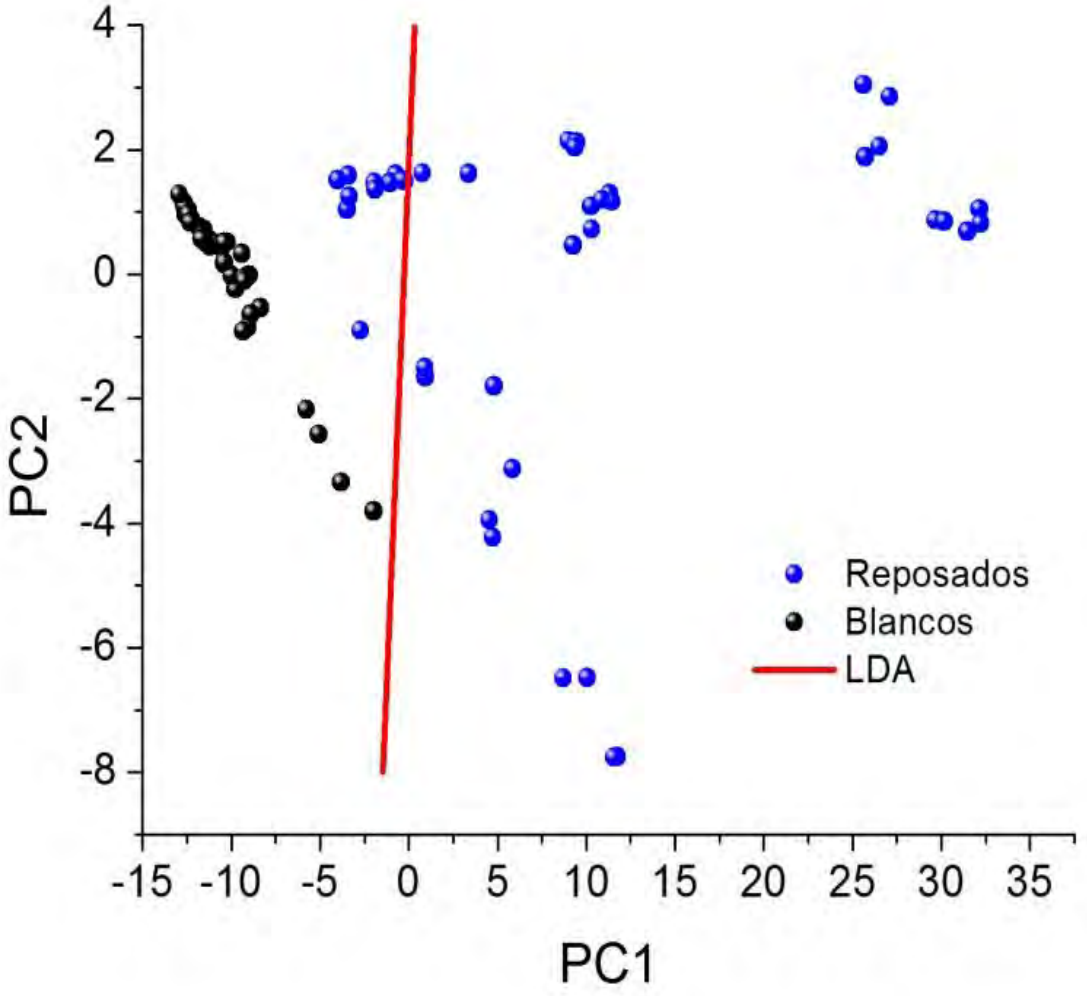

Reposado1
Reposado2
Reposado3
Reposado4
OD (u. a.)
0
1
2
3
4
5
6
250
300
350
400
450
500
Longitud de onda (nm)

Análisis

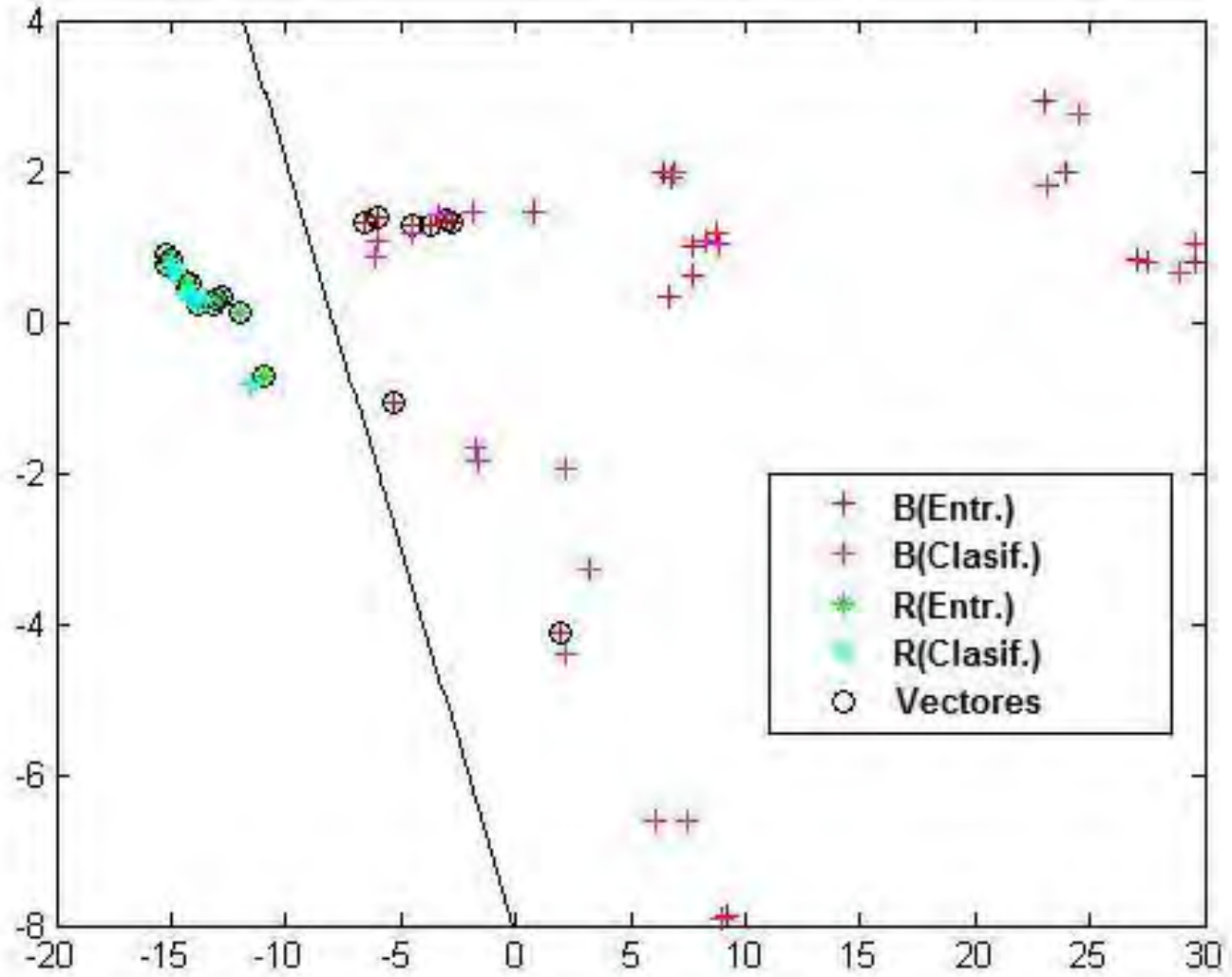
Tipos: blanco y reposado



LDA

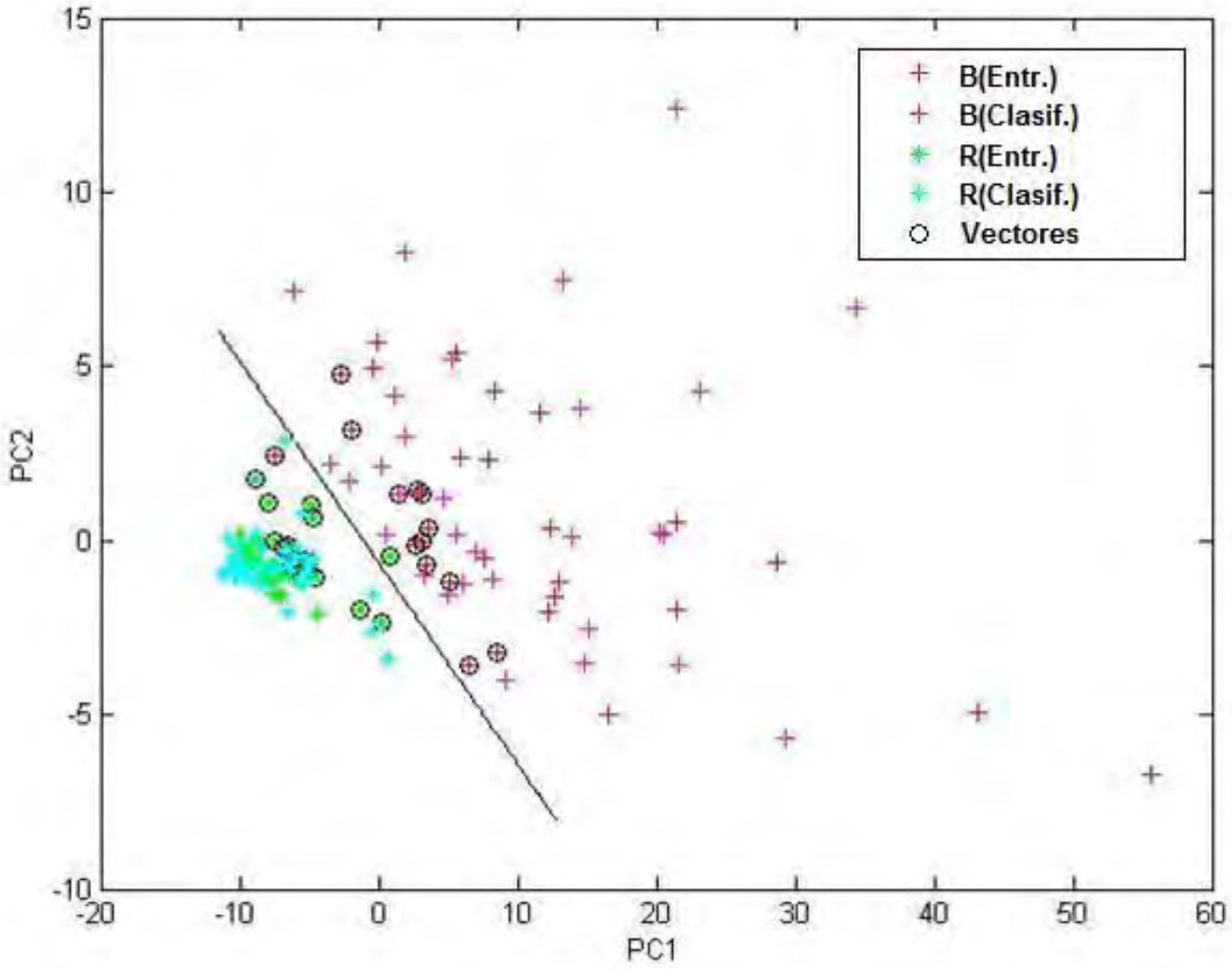

4
2
0
-2
-4
-6
-8
PC2
-15
-10
-5
0
5
10
15
20
25
30
35
PC1
Reposados
Blancos
LDA

PCA
SVM

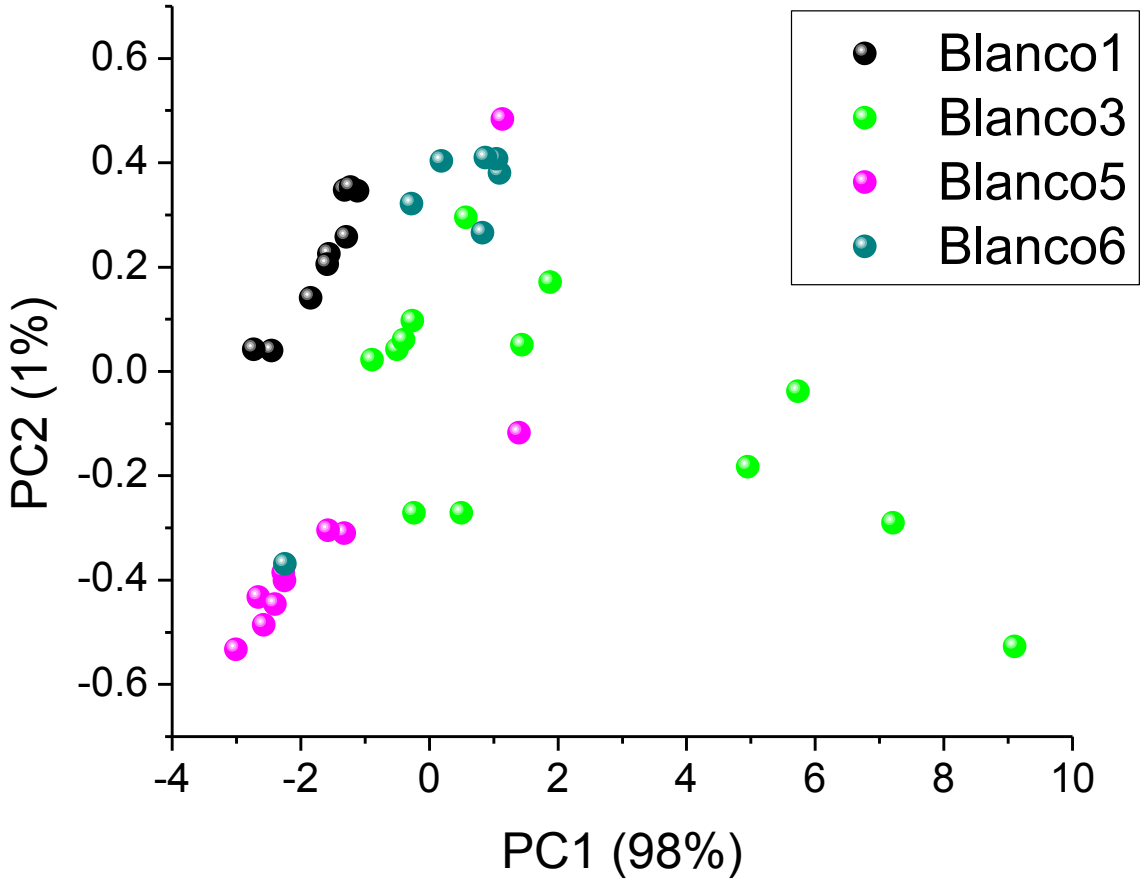

B(Entr.)
B(Clasif.)
R(Entr.)
R(Clasif.)
Vectores
4
2
0
-2
-4
-6
-8
-20
-15
-10
-5
0
5
10
15
20
25
30

Blancos - Madurados

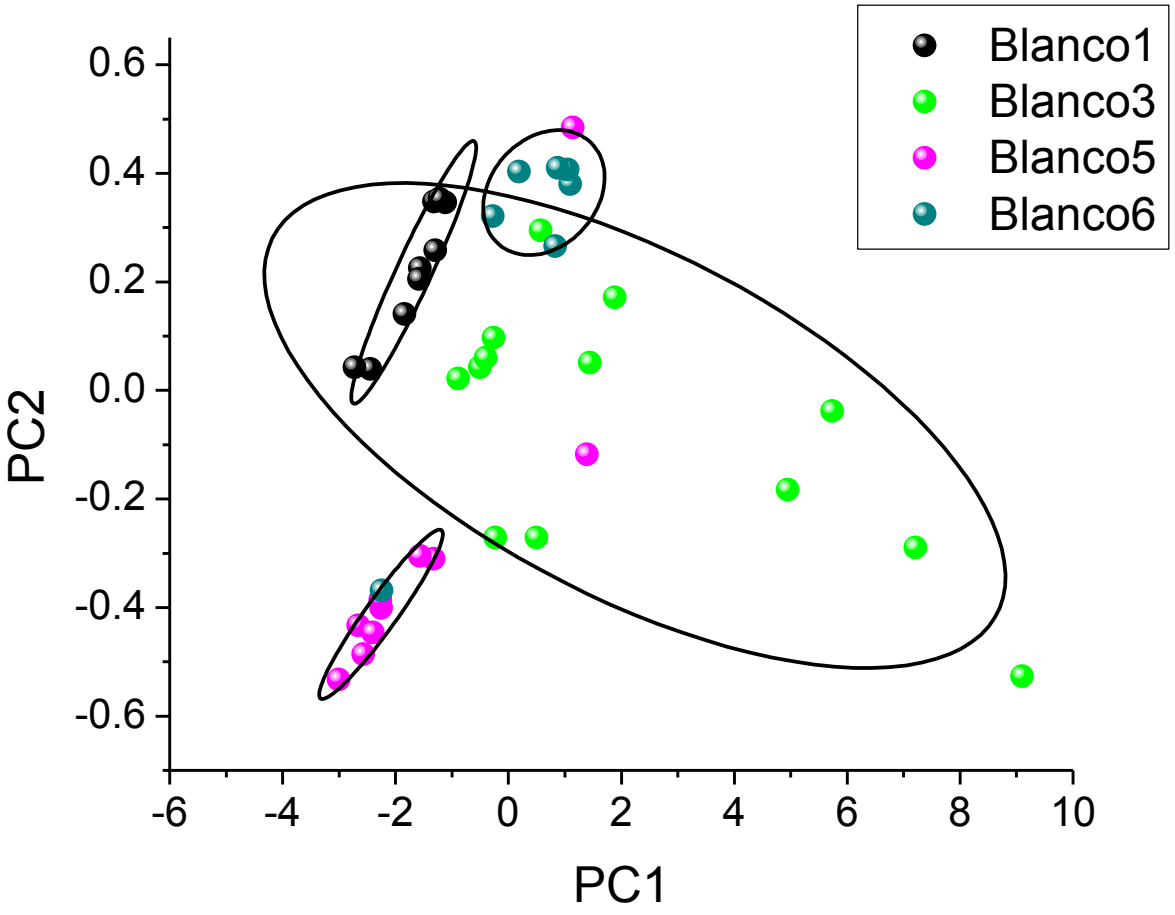
PCA
SVM



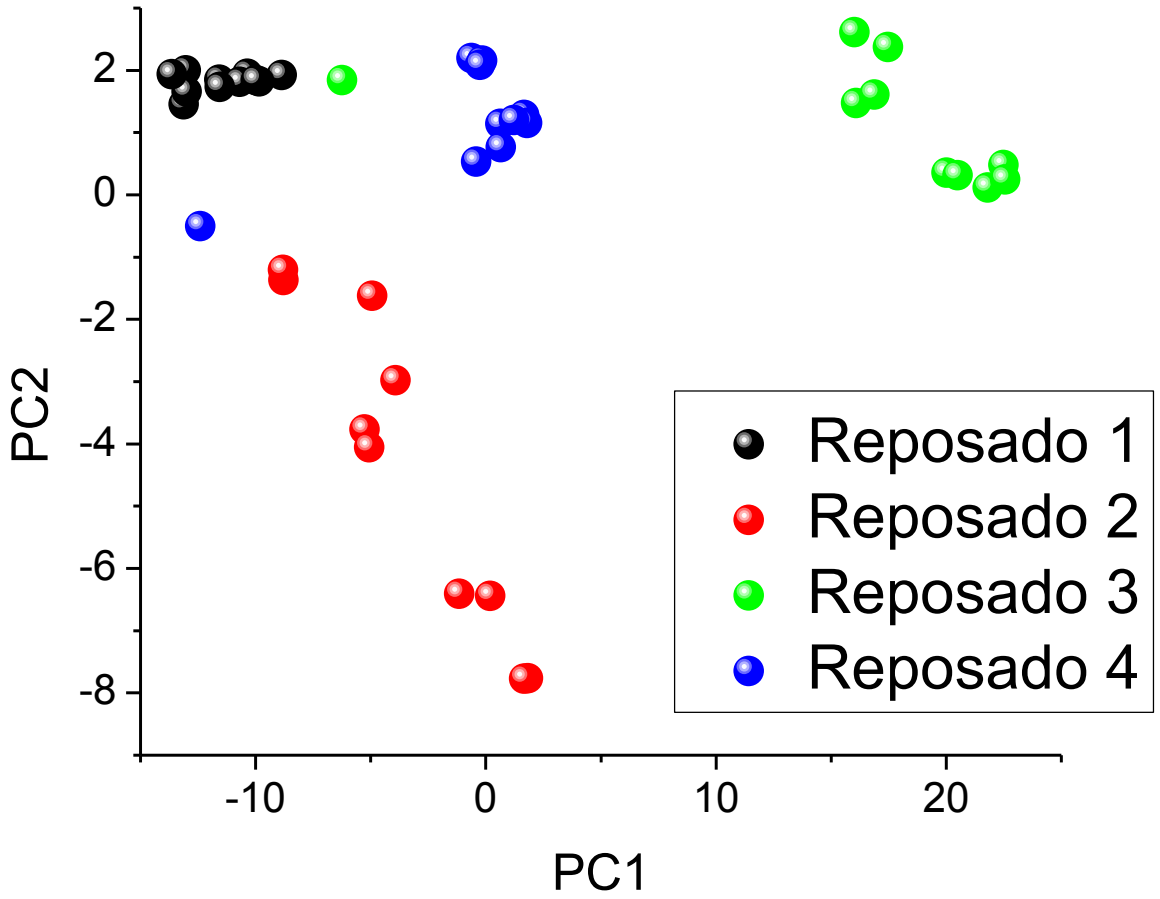
Blancos: Análisis por marca



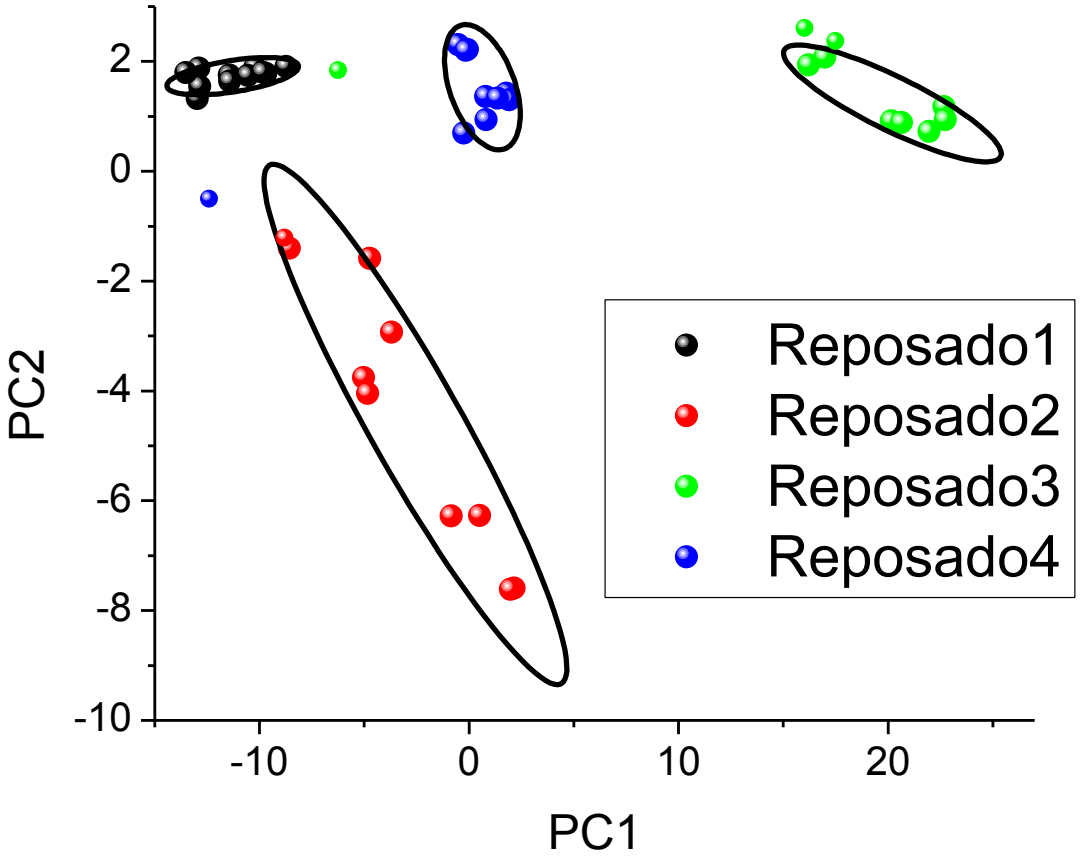
Agrupación


Blanco1
Blanco3
Blanco5
Blanco6
PC2
PC1
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-6
-4
-2
0
2
4
6
8
10

4 Marcas- Reposados

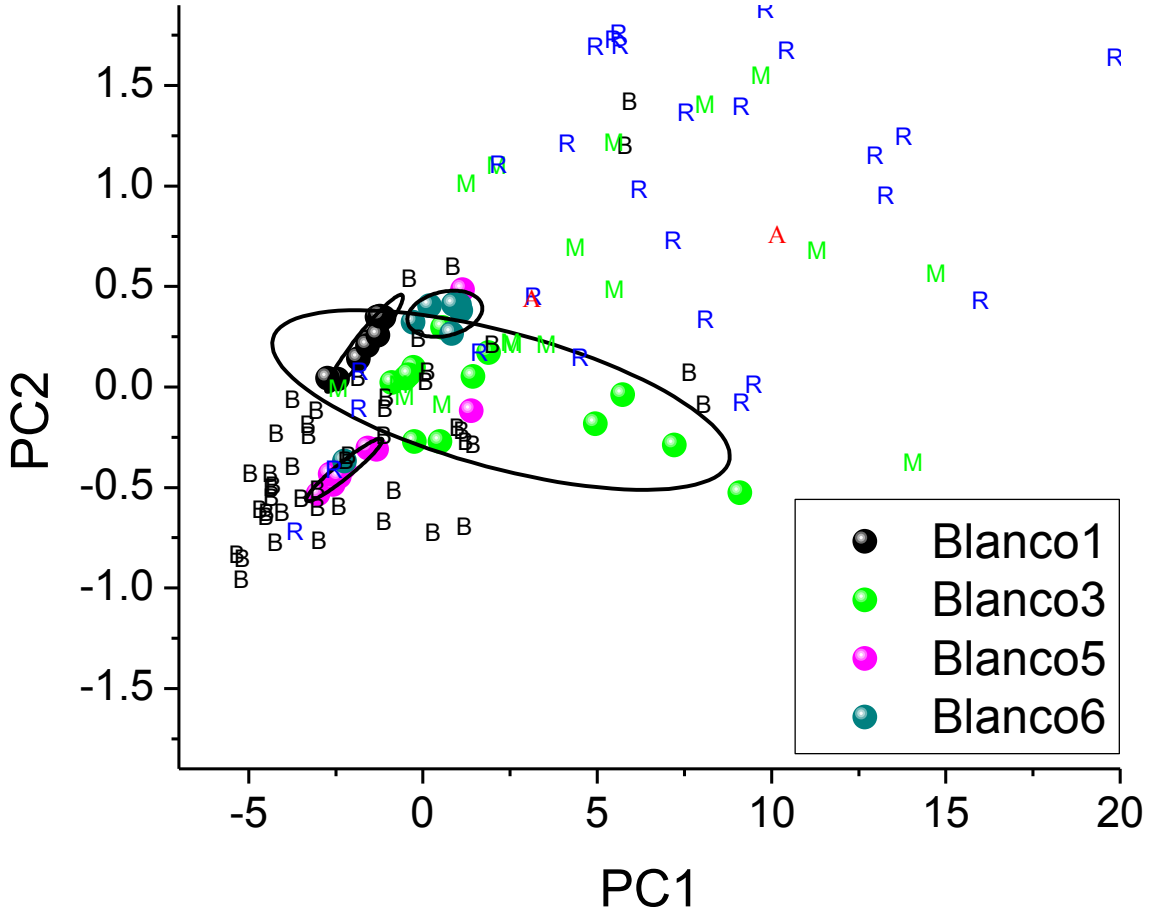


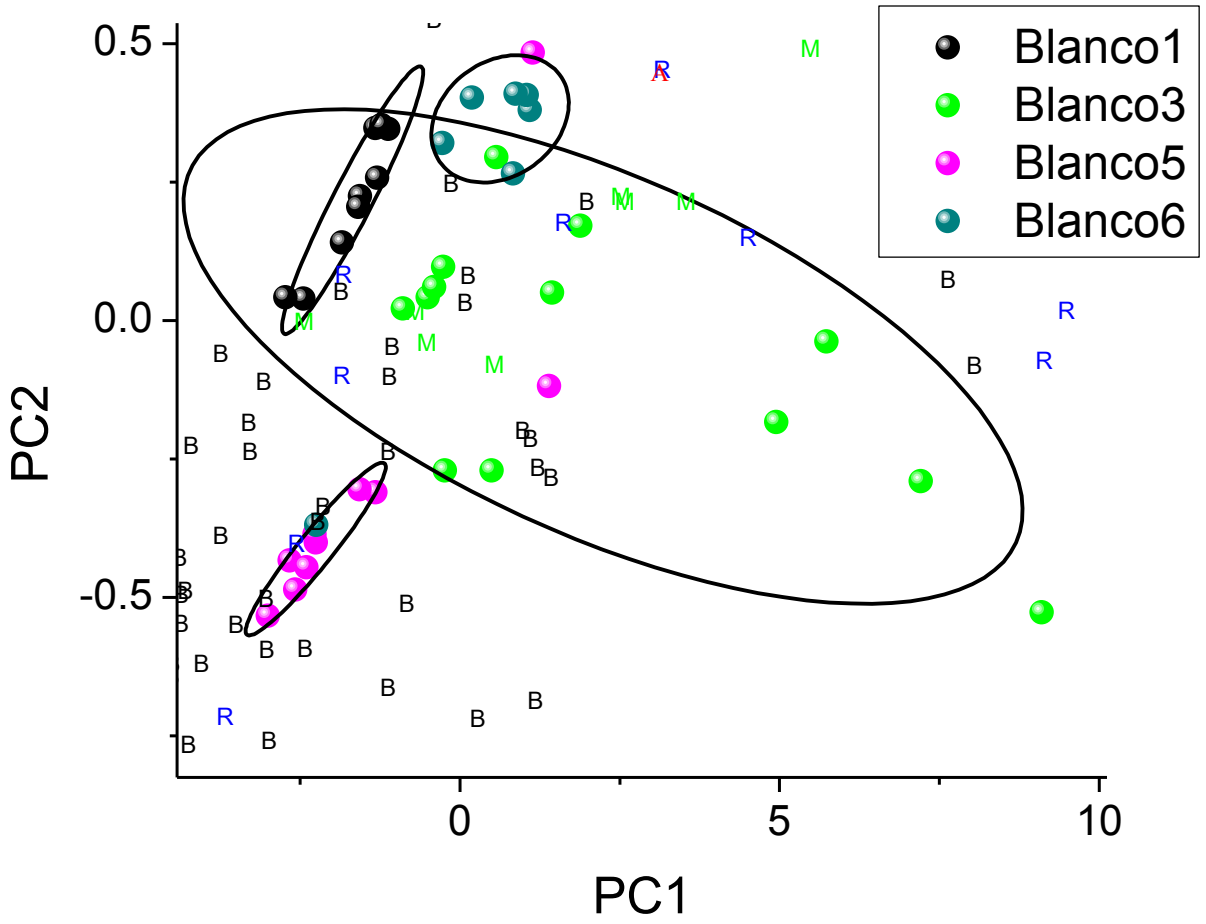
Agrupación


2
0
-2
-4
-6
-8
-10
PC2
-10
0
10
20
PC1
Reposado1
Reposado2
Reposado3
Reposado4

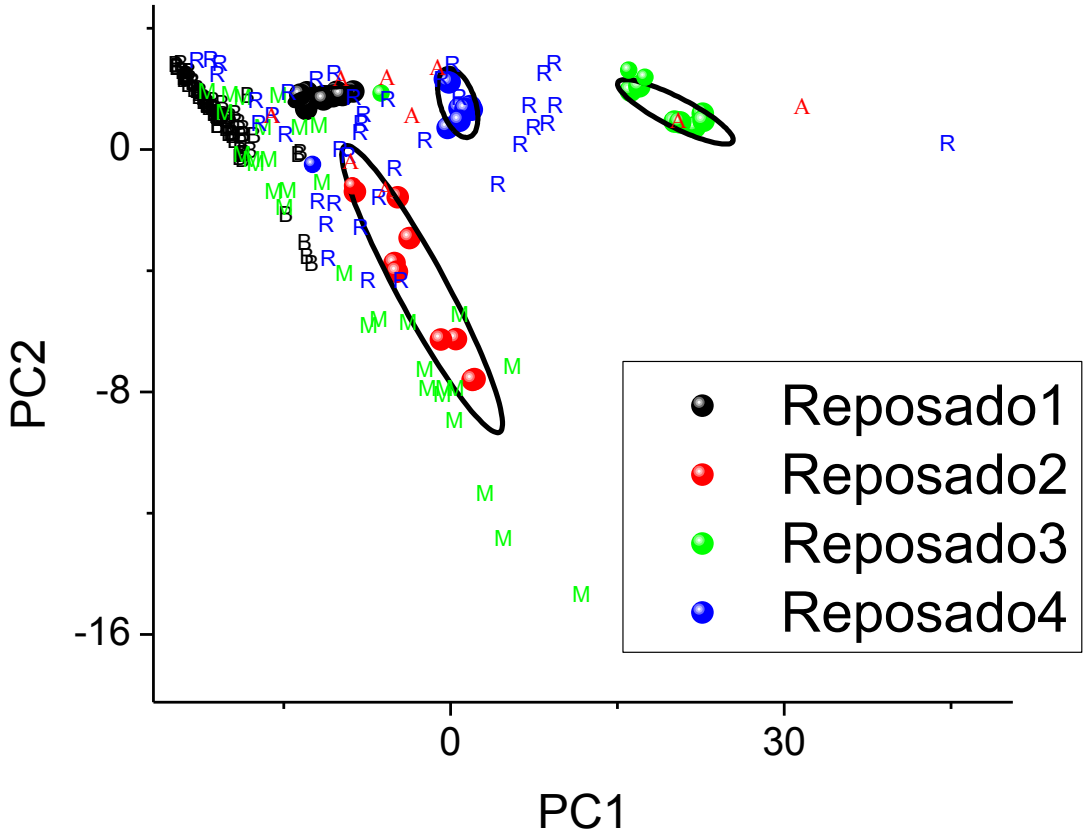
Validación

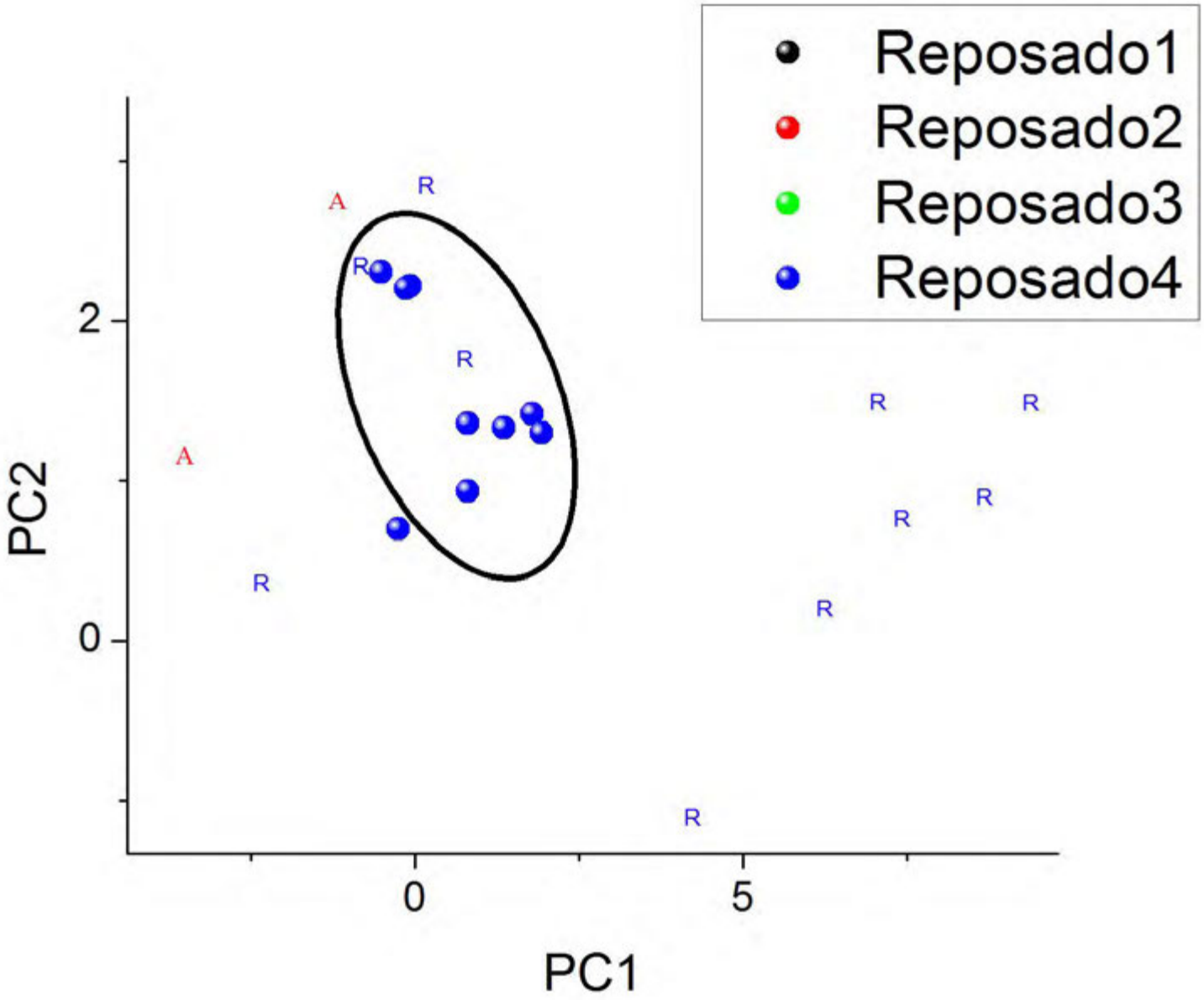
Blancos


PC2
PC1
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-5
0
5
10
15
20
Blanco1
Blanco3
Blanco5
Blanco6


Blanco1
Blanco3
Blanco5
Blanco6
PC1
PC2
0.5
0.0
-0.5
0
5
10

Reposados



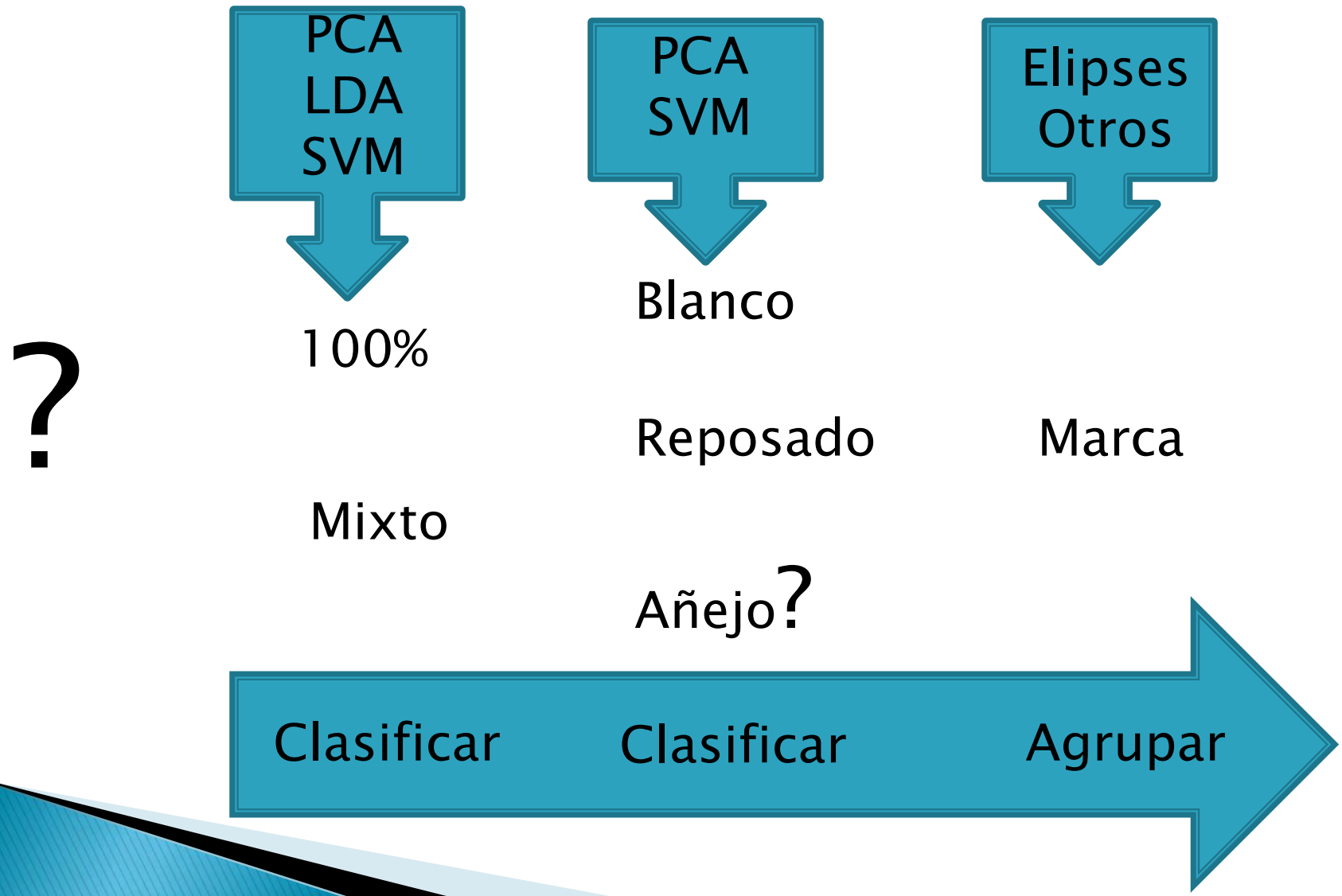

Reposado1
Reposado2
Reposado3
Reposado4
PC2
PC1
2
0
0
5
A
R

Conclusiones

- A través de métodos rápidos no invasivos:

- Se redujo la dimensionalidad del problema.
- Se discriminó el 100% de las muestras en sus respectivos tipos (blanco y reposado).
- Se agruparon las marcas de Tequila en el nuevo espacio generado por PCA.
- Se puede asociar una región característica a cada marca en el nuevo espacio.

Predicción de muestras



Gracias por su atención

Ecuaciones de elipses de confiabilidad

$$\left(\frac{X-\mu_x}{\sigma_x}\right)^2+\left(\frac{Y-\mu_y}{\sigma_y}\right)^2=k$$

Contenido

Lista de Figuras

3.2. Espectro de absorción y su primera derivada para una muestra de Tequila antes y después de un filtrado de partículas mayores a los 200nm.

3.3. Espectros de absorción en la región UV-Visible de 39 muestras de Tequila Blanco correspondientes a las 4 marcas analizadas.

3.4. Espectros de absorción de soluciones sintéticas de componentes individuales disueltos a ciertas concentraciones para una muestra de Tequila blanco. La línea punteada representa el espectro de absorción de un Tequila blanco. Imagen obtenida de la referencia [17].

3.5. a) Espectros de absorción promedio y su desviación estándar para el máximo de absorción para las cuatro marcas de Tequila blanco analizadas. b) Promedio de los espectros de absorción para las cuatro marcas de Tequila blanco analizadas y su error estándar a cada longitud de onda.

3.6. Espectros de absorción en la región UV-Visible del espectro electromagnético de las 41 muestras de Tequilas reposados.

3.7. a) Espectros de absorción promedio y su desviación estándar para el máximo de absorción para las cuatro marcas de Tequila reposado analizadas. b) Promedio de los espectros de absorción para las cuatro marcas de Tequila reposado analizadas y su error estándar a cada longitud de onda.

3.8. Espectros de absorción promedio de dos marcas de Tequila en los tipos blanco y reposado.

3.9. Diagrama ilustrativo de la reducción de dimensiones por PCA.

3.10. Gráfica de PC1 vs PC2 generada por el modelo de PCA para 80 muestras de Tequila, las esferas negras representan 39 muestras de Tequila blanco y las esferas azules representan 41 muestras de Tequila reposado, todas ellas proyectadas en el espacio PC1-PC2.

3.11. Gráfica de PC1 vs PC2 generada por el modelo de PCA para 80 muestras de Tequila, las esferas negras representan 39 muestras de Tequila blanco y las esferas azules representan 41 muestras de Tequila reposado, todas ellas proyectadas en el espacio PC1-PC2. La frontera de separación representada como una línea roja se calcula con el método Linear Discriminant analysis (LDA).

3.12. Grafica de PC1 vs PC2 generada a partir de un análisis de componentes principales para 80 muestras de Tequila, 39 blancos y 41 reposados. Se etiquetan las muestras de acuerdo a su tipo y clasifican con SVM. Se selecciona aleatoriamente el 50% de la información empleada para el entrenamiento y selección de vectores de soporte; el 50% restante valida dicha clasificación. Las muestras representadas por círculos corresponden a los vectores de soporte.

3.13. Gráfica de PC1 vs PC2 generada por el modelo de PCA para 39 muestras de Tequila blanco, los colores de las esferas indican las marcas de acuerdo a la etiqueta de la figura, todas las muestras están proyectadas en el espacio PC1-PC2.

3.14. Grafica de PC1 vs PC2 para 39 muestras de Tequila blanco de 4 marcas diferentes. Cada marca está representada por un color diferente como lo indica la etiqueta. Las elipses de confiabilidad se generaron con un 95% de probabilidad de encontrar una muestra dentro de la elipse correspondiente de acuerdo a su marca.

3.15. Gráfica de PC1 vs PC2 generada por el modelo de PCA para 41 muestras de Tequila reposado, los colores de las esferas indican las marcas de acuerdo a la etiqueta de la figura, todas las muestras están proyectadas en el espacio PC1-PC2.

3.16. Grafica de PC1 vs PC2 para 41 muestras de Tequila reposado de 4 marcas diferentes. Cada marca está representada por un color diferente como lo indica la etiqueta. Las elipses de confiabilidad se generaron con un 95% de probabilidad de encontrar una muestra dentro de la elipse correspondiente de acuerdo a su marca.

3.17. Validación del modelo PCA-4 marcas de Tequilas Blancos. Se utiliza un total de 145 bebidas alcohólicas, entre mezcales, Tequilas blancos, reposados y añejos. Los Tequilas blancos, reposados y añejos se representan por M, B, R y A respectivamente.

3.18. Validación del modelo PCA-Tequilas blancos por marca. La gráfica muestra el intervalo correspondiente a las elipses de confiabilidad de las cuatro marcas.

3.19. Validación del modelo PCA-4 marcas de Tequilas reposados. Se utiliza un total de 163 bebidas alcohólicas, entre mezcales y Tequilas blancos, reposados y añejos.

3.20. Validación del modelo PCA-4 marcas de Tequilas reposados. Se utiliza un total de 163 bebidas alcohólicas, entre mezcales y Tequilas blancos, reposados y añejos. Se grafica la sección correspondiente a las elipses de confiabilidad de las cuatro marcas.

A mis padres

Gracias por esta vida

y todo

lo que me han dado en ella.

Agradecimientos

Al Consejo Nacional de Ciencia y Tecnología CONACyT ya que sin su apoyo no habría sido posible la realización de mis estudios de posgrado.

Al Centro de Investigaciones en Óptica A. C. y a todas las personas que hicieron posible la culminación de este trabajo, en especial al Doctor J. Oracio C. Barbosa García, profesor y asesor de esta tesis, por su paciencia y tiempo dedicado en mi formación académica y humana.

Al Departamento de Formación Académica por su excelente trabajo y especialmente por brindarme su apoyo, amistad y confianza.

También agradezco de manera muy especial a los doctores Gabriel Ramos Ortiz, Juan Luis Pichardo Molina, Marco Antonio Meneses Nava y José Luis Maldonado Rivera por haber compartido sus conocimientos y experiencias pero sobre todo por el gran trato humano que me brindaron. Al Ing. Quim. Martín Olmos por su invaluable ayuda en el laboratorio.

Al departamento de Propiedades Ópticas de la materia, a mis profesores, compañeros y amigos de este Centro.

Resumen

A través de una técnica de espectroscopia basada en la absorción de luz en el rango UV visible y un análisis estadístico multivariable es posible identificar marcas de tequila para la protección de marca in situ. Se adquirieron muestras de Tequila tipo blanco y reposado de 4 marcas registradas para cada tipo en tiendas de licores y en algunos bares en las ciudades de León Gto. y de Zacatecas Zac. con el propósito de encontrar similitudes y diferencias que permitan una agrupación y una posible predicción de marcas para futuras muestras. Se describe el método para la discriminación de estas bebidas alcohólicas el cual es complementario a los métodos actuales basados en análisis químicos que son de mayor coste y complejidad. Los resultados muestran una buena agrupación de las marcas analizadas por lo que se puede identificar una marca de otra y a la vez discriminar entre tequilas blancos y reposados; estos resultados se obtienen en cuestión de minutos y pueden realizarse fuera de un laboratorio especializado.

Objetivo

En este trabajo se pretende generar un método de clasificación y reconocimiento de marcas de Tequila complementario a los métodos tradicionales basado en el análisis de absorción en la región UV-VIS del espectro electromagnético de estas bebidas y análisis estadísticos multivariables. Se trata de discriminar entre diferentes clases de tequilas (blancos y reposados) e incluso identificar y clasificar diferentes marcas de Tequila. A diferencia de los métodos tradicionales usados para la discriminación y clasificación de tequilas, el método propuesto no demanda instrumentación tan especializada e incluso no se requiere una preparación de la muestra previa a la medición, lo cual involucra un consumo de tiempo mucho menor comparado con el tiempo de técnicas empleadas hoy en día.

Introducción

En el mercado existe una gran variedad de marcas de tequilas, algunas de ellas de gran prestigio a nivel mundial y algunas otras nuevas e incluso de dudosa procedencia. Hoy en día la fama del tequila ha trascendido fronteras, posicionándose como una bebida de gran reconocimiento a nivel mundial, y su consumo aumenta año con año. Para el 2007 la producción de Tequila alcanzó los 284 millones de litros (23% más que un año previo, de acuerdo al Consejo Regulador del Tequila). Lamentablemente, el gran auge de esta debida en los últimos años provocó el surgimiento de pseudo tequilas y adulteraciones degradando la imagen y calidad de este producto genuino y provocando pérdidas millonarias a la industria tequilera.

Los métodos usados tradicionalmente para caracterizar bebidas alcohólicas consisten generalmente en la separación previa de los compuestos químicos puros antes de analizarlos. Estas técnicas garantizan la autentificación y clasificación de estas bebidas, sin embargo, estos métodos presentan costes elevados debido a la necesidad de disponer de instrumentación analítica especializada, personal de laboratorio además de ser técnicas que requieren una inversión de tiempo considerable para la separación de los componentes químicos y su análisis posterior.

La elaboración de esta tesis se enfoca principalmente en describir un método de clasificación para marcas de Tequila basado en técnicas no destructivas como lo es la espectroscopia por absorción en la región UV-Visible del espectro electromagnético y análisis estadísticos multivarientes. Previo a la descripción

del método de clasificación el capitulo uno introduce conceptos relacionados con el proceso de elaboración del Tequila así como algunas de sus características más distintivas. También se introducen conceptos relacionados con la espectroscopia de absorción y con espectroscopia en la región UV-Visible y se describe de manera breve las componentes de un espectrofotómetro utilizado para la obtención de los espectros de absorción. Para una mejor comprensión del problema de clasificación del que estamos interesados, en el capítulo dos se describen los métodos estadísticos de Análisis de Componentes Principales (PCA) y Maquinas de Soporte Vectorial (SVM) de manera general. PCA es un método estadístico que se utilizó para reducir la dimensionalidad del problema y proyectar los datos en un espacio de dos dimensiones. SVM es un método de clasificación binario que nos permitió discriminar entre Tequilas blancos y Reposados. Se utilizan otras técnicas estadísticas en la clasificación y validación del modelo generado pero no se describen en esta tesis ya que se usaron solo de manera complementaria en el método descrito. En el capítulo tres se muestran los resultados obtenidos de la espectroscopia de absorción de los Tequilas y se aplican las técnicas de análisis estadístico descritas en el capitulo dos. En el capitulo tres también se hace una discusión de los resultados y se describe el modelo de clasificación generado así como su validación. El modelo descrito en este trabajo discrimina de manera objetiva entre Tequilas blancos y reposados y agrupa a las marcas analizadas en regiones características para cada marca en un espacio generado por dicho modelo.

1. Principios básicos

En este capítulo se definen algunos conceptos relacionados con el Tequila y se describe de manera general el proceso de producción de esta bebida alcohólica; también se introducen conceptos generales de espectroscopia de absorción así como la descripción general de los componentes básicos del instrumento con que se realizan las mediciones. El capítulo uno tiene como objetivo principal brindar al lector una visión general del producto que se analiza y los principios ópticos usados para solucionar el problema de identificación y clasificación de los Tequilas.

1.1. El Tequila

El tequila es una bebida alcohólica de origen mexicano asociada históricamente al estado de Jalisco. Esta bebida alcohólica se deriva de la destilación de jugo fermentado proveniente exclusivamente de una planta conocida como Agave Tequilana Weber, de variedad azul. En México existen otras variedades de agave y también son utilizadas para la producción de bebidas alcohólicas regionales como el mezcal, sotol y bacanora.

En 1974, el gobierno mexicano emitió una declaración para la Protección de la denominación del origen del Tequila (DOT), afirmando que por su origen geográfico, reputación y cualidades específicas esenciales, el Tequila ha sido considerado como un distintivo geográfico de México. Esto significa que México reclamará el uso exclusivo, en el mundo entero, de la palabra "Tequila", y que sólo las bebidas alcohólicas hechas de la planta de Agave azul (variedad azul de Agave Tequilana Weber) que crece en el área oficial demarcada dentro de México, y bajo las reglas de la Norma Oficial del Tequila, pueden ser etiquetadas como "Tequila". El área oficialmente demarcada para la producción del Tequila incluye todo el Estado de Jalisco y otras áreas específicas, dentro de cuatro estados: Nayarit, Tamaulipas, Michoacán y Guanajuato. De acuerdo con la ley, sólo estas áreas poseen el clima adecuado y las características del suelo, para el desarrollo de la planta del Agave azul y solo aquí se puede producir Tequila.

De acuerdo al porcentaje de los azúcares provenientes del agave que se utiliza en la elaboración del Tequila, la Norma Oficial Mexicana NOM-006-SCFI define las categorías para esta bebida de la manera siguiente: El "Tequila 100% agave" es el tequila no enriquecido con otros azucares distintos a los obtenidos del Agave tequilana weber variedad azul y debe ser envasado por el productor autorizado dentro de la región comprendida por la misma Declaración. El "Tequila mixto" o simplemente "Tequila" es el producto en el que el jugo fermentado es susceptible a ser enriquecido y mezclado con azucares provenientes de otras especies de agave en una proporción no mayor al 49% y puede ser envasado en plantas ajenas a productores autorizados.

De acuerdo a las características adquiridas en procesos posteriores a la destilación, el tequila se clasifica de la siguiente manera: Tequila blanco. Producto embotellado posterior a la destilación. Tequila reposado. Producto susceptible a ser abocado (coloreado, suavizado y/o endulzado), sujeto a un proceso de maduración de por lo menos dos meses en contacto directo con la madera en

recipientes de roble o encino. Tequila añejo. Producto susceptible de ser abocado, sujeto a un proceso de maduración de por lo menos un año en contacto directo con la madera de recipientes de roble o encino. El contenido alcohólico comercial de todas las clases debe ajustarse, en su caso, con agua en dilución [1].

Debido a la importancia que ha tomado el Tequila en el mercado mundial, se creó el Consejo Regulador del Tequila, organismo acreditado por el gobierno mexicano encargado de inspeccionar y certificar que la producción, envasado y etiquetado del Tequila se lleve a cabo de acuerdo con la Norma Oficial Mexicana.

Producción de Tequila

El proceso de elaboración del Tequila tiene una gran influencia en la calidad final del producto. Algunas compañías elaboran el Tequila de manera artesanal y algunas otras emplean técnicas de producción avanzadas con mejor eficiencia en su producción y un mejor control de calidad en sus destilaciones. Sin embargo, la producción de tequila en todas las compañías consta básicamente de cuatro etapas: Cocimiento, molienda, fermentación y destilación.

Cocimiento

Cuando el cocimiento se realiza de manera artesanal, se utilizan hornos rústicos de paredes de ladrillo donde las piñas permanecen de 36 a 48 horas. El cocimiento del agave sirve principalmente para hidrolizar la inulina y los demás componentes del agave y proporcionarles una consistencia más suave que facilite la molienda. También durante este proceso, algunos azucares son caramelizados y contribuyen de manera significativa al aroma y sabor del tequila. Hoy en día, muchas destilerías han reemplazado estos hornos tradicionales por autoclaves de acero que son más eficientes y tienen un mejor control de presión y

temperatura permitiendo un cocimiento homogéneo previniendo de manera más eficaz un sobrecocimiento, lo que daría un sabor humeado al tequila e incrementaría la concentración de furfural en el producto final.

Molienda

La siguiente etapa es la extracción de jugo de agave conocida como molienda, en donde el jugo de la piña cocida se extrae mediante el desgarramiento de la pulpa y luego es prensada en molinos de rodillos añadiendo un poco de agua, lo que facilita la extracción de los azúcares. El jugo obtenido en la molienda llamado "mosto" es mezclado con jarabe obtenido del cocimiento y normalmente con azúcar de caña cuando el tequila no es 100% agave.

Fermentación

La fermentación es un proceso biológico anaeróbico donde los azucares simples como la glucosa y fructosa son transformados a etanol y dióxido de carbono por acción de microorganismos del medio como levaduras y bacterias. Algunas compañías permiten procesos de fermentación naturales y algunas otras inoculan con levadura del género Saccharomyces cerevisiae o algún otro tipo de levaduras secas originalmente preparadas para la producción de pan, cerveza o wisky; por lo general la calidad del tequila obtenida utilizando estas levaduras no es satisfactoria y sus variaciones de sabor y aroma son muy amplias [2]. Los inóculos utilizados son crecidos en laboratorios de manera controlada para evitar contaminaciones bacteriales no deseadas. Sin inoculación, el tiempo de fermentación puede durar hasta siete días; si se utiliza un inóculo la fermentación se alcanza entre 20 horas si el proceso es rápido y tres días si el proceso es lento. La producción de alcohol etílico por levaduras está asociada con la formación de

compuestos que contribuyen al sabor final del tequila. Al igual que en otras fermentaciones alcohólicas, los alcoholes superiores son los compuestos producidos en mayor cantidad además del etanol, pero también se producen otros compuestos organolépticos como el metanol, aldehídos, ácidos orgánicos pequeños y esteres, a los que se atribuye una contribución muy importante de aroma y sabor del producto final.

Destilación

En el proceso de producción de tequila se realiza una doble destilación. La primera destilación involucra la separación y concentración de alcohol del mosto fermentado mediante un sistema que se conoce como destrozamiento, donde se separan las vinazas (levaduras muertas, azucares no fermentables y minerales) y otros componentes como aldehídos y cetonas y donde la concentración de alcohol alcanza un 20 o 30% por volumen. Después se realiza una segunda destilación llamada "rectificación" donde se concentra el alcohol etílico y se purifica de otros alcoholes. En esta etapa se obtienen dos fracciones, la primera se le llama "cabeza" y la ultima se le llama "cola". En general las cabezas son ricas en componentes con puntos de ebullición bajos como el acetaldehído, metanol, 1-propanol, 2-propanol, etc. los cuales proporcionan un sabor placentero al tequila. Las colas, por el contrario, contienen compuestos con puntos de ebullición más elevados como el iso-amyl alcohol, ácido acético, 2-furaldehído, etc. Estas sustancias proporcionan al tequila un sabor fuerte y desagradable, por esta razón las colas no son empleadas para la elaboración de tequila.

La destilación es la etapa final de la producción si se quiere tequila blanco. Para Tequilas reposados y añejos se tiene que llevar a cabo una maduración en barriles de 200 litros o mayores de roble o encino. El tiempo de maduración legalmente requerido es de dos meses para tequilas reposados y 12 meses para

tequilas añejos aunque por lo general los periodos son mayores y dependen de las características que cada compañía desea para su marca en particular. Al madurar el tequila en barriles, esta bebida está sujeta a cambios que determinarán su calidad final y que dependen incluso de las condiciones ambientales en que se encuentren los barriles y de la cantidad de veces que se han usado afectando en gran medida el sabor y aroma finales del tequila. Algunos componentes de la madera del barril son extraídos por el tequila dotándolo de un color y sabor muy particular, además se llevan a cabo procesos de oxidación que cambian algunos componentes del tequila y algunos otros extraídos de la madera dando como resultado incrementos en la concentración de ácidos, esteres y aldehídos. Después de la maduración y dilución con agua mineralizada (en caso de ser necesario) el color del tequila se puede ajustar agregando caramelo. Algunas compañías mezclan diferentes lotes de producción para obtener un producto final mas estandarizado.

La producción de tequila en cada una de sus etapas varía para todas las compañías y no se tiene un control sobre todas las variables que pueden afectar directamente en la calidad del producto final. Son varios los factores que intervienen en la calidad del tequila, por ejemplo, para que un proceso fermentativo tenga éxito, es indispensable utilizar materias primas que proporcionen a las levaduras todos los nutrimentos necesarios para un crecimiento óptimo, así pues, la composición de la tierra, el riego, las plagas y las condiciones ambientales influyen directamente en la composición química del agave y posteriormente repercute en la calidad del tequila. Otro factor evidente está en los sistemas de destilación, los cuales son diferentes para cada compañía productora lo cual dará como resultado diferencias en el producto obtenido de cada destilería aunque se utilice la misma materia prima. Se ha notado también que las barricas empleadas (en caso de tequilas reposados y añejos) de roble proporcionan características diferentes a las barricas de encino e incluso se ha

notado que en barricas nuevas las características del tequila son diferentes a las que proporcionan barricas viejas (usadas anteriormente). Para obtener información más detallada respecto a la producción de Tequila puede consultarse las referencias [2,3,4].

1.2. Espectroscopia de absorción molecular

Las propiedades espectroscópicas de los átomos están determinadas por sus estructuras electrónicas, pero en el caso de las moléculas, las propiedades espectroscópicas dependen, además de su estructura electrónica, de los enlaces químicos de sus componentes individuales y del movimiento vibracional y rotacional de la molécula. En general, cuando la radiación interactúa con la materia, pueden ocurrir diferentes procesos como la reflexión, esparcimiento, absorción, fluorescencia, fosforescencia y reacciones fotoquímicas. Cuando medimos espectros electromagnéticos en este trabajo consideraremos que solo ocurre el fenómeno de la absorción.

Normalmente, las bandas y líneas producidas por una molécula se acumulan en tres regiones características del espectro electromagnético entre 100nm y 1mm. La región UV-Visible corresponde al rango espectral más energético donde se encuentran líneas y bandas producidas por transiciones simultaneas electrónicas, vibracionales y rotacionales. El otro rango espectral característico corresponde a las bandas rotacionales-vibracionales presentes en el infrarrojo cercano y una tercera región corresponde a transiciones puramente rotacionales y se encuentran en la región infrarroja lejana (Figura 1.1).

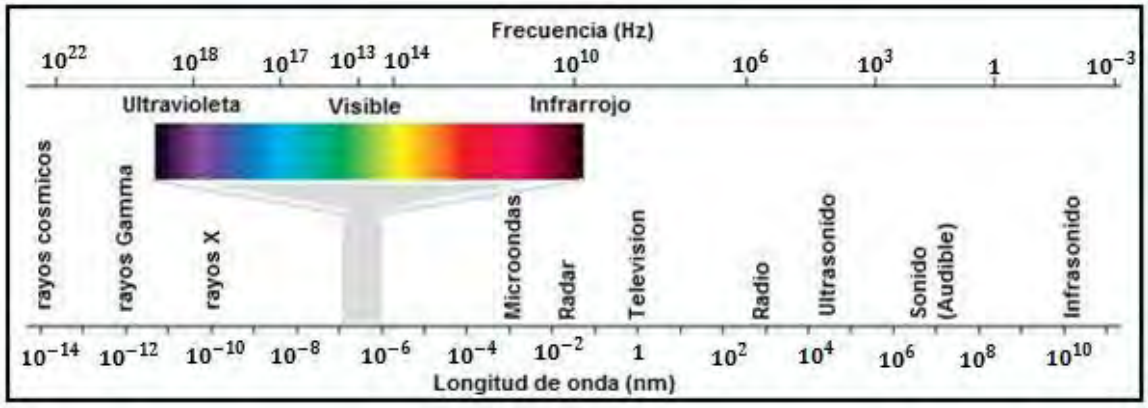


Figura 1.1. Espectro electromagnético

Así pues, la energía total en la región UV-Visible para una molécula está representada por la suma de su energía electrónica, vibracional y rotacional:

$$E_{total} = E_{electr\ ónica} + E_{vibracional} + E_{rotacional}$$

Para algunas moléculas y átomos, los fotones de luz UV y Visible son lo suficientemente energéticos que causan transiciones electrónicas entre diferentes niveles de energía en la materia. La luz con cierta longitud de onda es absorbida y su energía asociada promueve al electrón a un nivel más energético. Esta transición resulta en una banda de absorción estrecha asociada a la longitud de onda absorbida por los átomos. En el caso de las moléculas, los niveles de energía vibracionales y rotacionales se superponen en los niveles electrónicos y debido a que pueden ocurrir diversas transiciones las bandas se ensanchan (Figura 1.2).

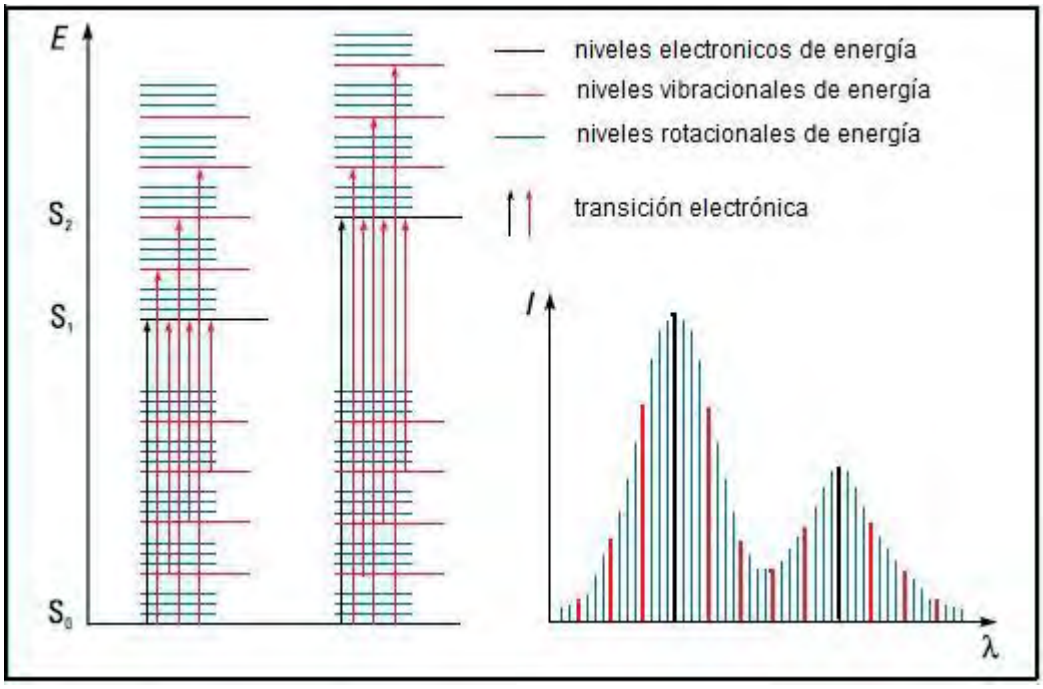


Figura 1.2. Transiciones electrónicas, vibracionales y rotacionales producidas en moléculas.

Los análisis espectroscópicos son usados en muchas áreas del ámbito científico, y aunque no realizan un análisis molecular preciso ni dan información detallada sobre estructuras de macromoléculas, pueden usarse para la identificación de compuestos y determinar en algunas ocasiones las concentraciones de los mismos.

1.3. Ley de Beer-Lambert-Bouguer

El objetivo principal de la espectroscopia de absorción es determinar qué tanta luz se transmite en una muestra para una longitud de onda específica. En 1729 y 1970 Bouguer y Lambert respectivamente descubrieron que la atenuación de luz que pasa por un medio claro es proporcional a la intensidad de la luz I y a la longitud de la muestra que es atravesada dx.

$$dI \propto I \cdot dx \tag{1.3.1}$$

Introduciendo un coeficiente de absorción o extinción podemos escribir la ecuación anterior como:

$$dI = \alpha(\lambda) \cdot I \cdot dx \tag{1.3.2}$$

La Ley de Bouguer-Lambert solo se aplica bajo las siguientes condiciones.

- La luz incidente sobre la muestra a analizar debe ser monocromática.
- Las moléculas que absorben deben estar distribuidas en una base de manera homogénea y no deben exhibir esparcimiento ni interacciones con otras moléculas.
- El esparcimiento y la reflexión de las superficies de la muestra causan atenuación en la luz lo cual no está considerado en dicha ley.

En 1852 Beer mostró que en la mayoría de las soluciones el factor de atenuación α es también proporcional a la concentración c de la(s) molécula(s) absorbente(s). Así, podemos escribir la ecuación 1.3.2 como:

$$dI = -\alpha(\lambda) \cdot c \cdot I \cdot dx \tag{1.3.3}$$

Integrando la ecuación 1.3.3 sobre el camino total que recorre la luz sobre la muestra obtenemos la ley de Beer-Lambert-Bouguer:

$$I = I_0 e^{-\alpha(\lambda)cx} \tag{1.3.4}$$

Donde la constante de integración I_0 describe la intensidad de luz incidente e I la intensidad de la luz en cualquier posición x de la muestra. La ecuación 1.3.4 también se puede escribir de la siguiente manera:

$$-log\frac{I}{I_0} = A(\lambda) = \varepsilon(\lambda)cx \tag{1.3.5}$$

Donde $\varepsilon(\lambda)$ representa el coeficiente de extinción molar $\varepsilon(\lambda) = \alpha(\lambda) \cdot 0.4343[M^{-1}cm^{-1}]$ y donde $A(\lambda)$ está definida como absorción o absorbancia para cada longitud de onda. Como el logaritmo es una función adimiensional, es evidente que $A(\lambda)$ no es una unidad física sino un número, y si se cumplen las condiciones de la ley Bouguer-Lambert, la densidad óptica $OD(\lambda)$ es idéntica a la absorbancia.

Las desviaciones de esta ley son provocadas principalmente cuando las concentraciones del material absorbente son muy altas o por cambios en las constantes de disociación, entre otros efectos como distribuciones no homogéneas de la sustancia absorbente.

1.4. Espectrofotómetro

Los métodos analíticos basados en medidas especroscópicas dependen de la interacción de la luz con la materia y podemos distinguir entre espectroscopias de absorción, reflexión, esparcimiento y emisión. En todos los casos, un monocromador selecciona luz de una longitud específica proveniente de una fuente de iluminación adecuada dirigida a la muestra que va a analizarse. Con este mecanismo, es posible determinar experimentalmente la cantidad de luz que es reflejada (espectroscopia por reflexión), transmitida (espectroscopia de transmisión o absorción) o esparcida (espectroscopia por esparcimiento).

Un espectrofotómetro es un instrumento que mide la transmitancia o absorción de una muestra como función de la longitud de onda de la radiación electromagnética. Los componentes básicos de un espectrofotómetro son una fuente de luz, un monocromador y un fotodetector que transforma la señal de luz recibida en una señal eléctrica, la cual es procesada por computadora (Figura 1.3). La selección particular de cada elemento y la configuración del arreglo en general dependerán del tipo de espectroscopia que se lleve a cabo, sin embargo, nos enfocamos solo a espectroscopia por absorción.

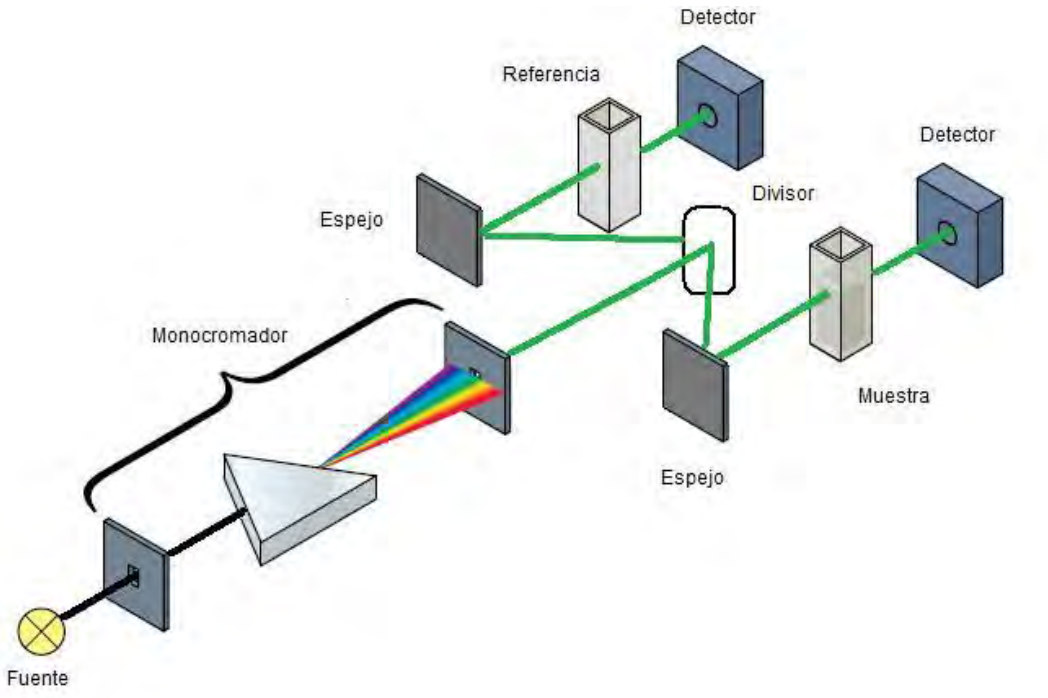


Figura 1.3. Esquema de un espectrofotómetro de doble haz.

Fuentes de iluminación

Una fuente de luz ideal sería aquella que mantuviese constante su intensidad sobre todas las longitudes de onda deseadas con bajo ruido y una gran estabilidad térmica. Desafortunadamente, una fuente ideal no existe. Las fuentes de iluminación usadas en espectroscopia pueden ser lámparas incandescentes, lámparas de descarga, diodos emisores de luz (LEDs por su acrónimo en inglés) e incluso láseres. En el caso específico de espectroscopia en el rango UV-Visible comúnmente se utilizan dos tipos de lámparas (se describen a continuación) aunque algunos espectrofotómetros recientes han incorporado fuentes de iluminación basados en LEDs.

La lámpara de deuterio permite una intensidad continua en la región UV y parte del Visible del espectro electromagnético (Figura 1.4). Aunque las lámparas

de deuterio modernas producen bajo ruido, el ruido sigue siendo un factor que limita el desempeño del espectrofotómetro. Otro inconveniente es que la intensidad de estas lámparas disminuye con el tiempo. El tiempo medio de vida útil de estas lámparas es de 1000 horas aproximadamente.

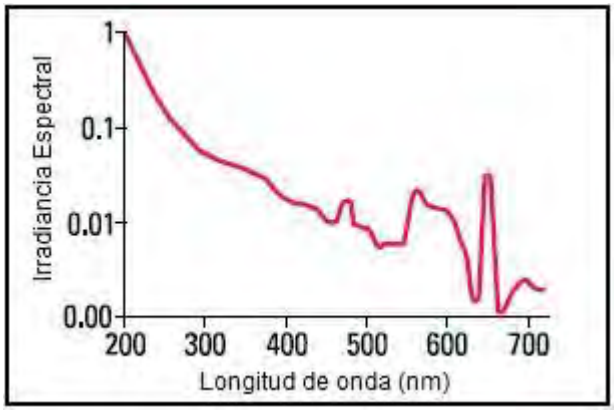


Figura 1.4. Emisión espectral de una lámpara de deuterio.

La segunda fuente de iluminación es la lámpara de tungsteno-halógeno y su intensidad se extiende desde la región UV y abarca toda la región Visible. Estas lámparas tienen niveles de ruido muy bajas y su tiempo medio de vida es de 10,000 horas (Figura 1.5).

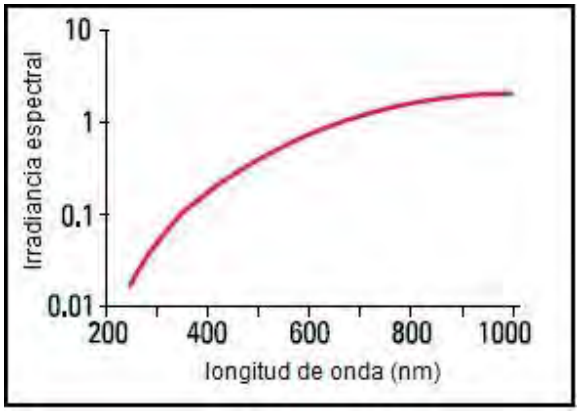


Figura 1.5. Emisión espectral de una lámpara de tungsteno-halógeno.

La mayoría de los espectrofotómetros diseñados para medir en la región UV-Visible usan los dos tipos de lámparas mencionados y utilizan un interruptor que selecciona entre ambas lámparas, otros espectrofotómetros utilizan la luz proveniente de ambas lámparas para producir una banda ancha del espectro electromagnético.

Monocromadores

Por lo general, se utilizan elementos dispersivos para monocromatizar la luz en los espectrofotómetros. Estos elementos dispersan la luz a diferentes ángulos dependiendo de la longitud de onda. Combinados con rendijas de entrada y salida (figura 1.3), estos elementos pueden usarse para seleccionar longitudes de onda muy definidas, o en otras palabras, una banda muy estrecha del espectro electromagnético. Por lo general se utilizan dos tipos de dispositivos para la dispersión de la luz, prismas y rejillas holográficas para espectrofotómetros usados en la región UV-Visible.

Detectores

Los fotodetectores se clasifican generalmente en base al efecto producido que emite la señal, esto es, efecto fotoeléctrico, reacción fotoquímica o en su caso detección de calor. La sensibilidad de los detectores es muy variado y depende del rango específico del espectro electromagnético en que se genera la señal. Los fotodetectores que describimos a continuación son aquellos basados en el efecto fotoeléctrico externo. Básicamente, los fotones son capaces de “desprender” electrones de un fotocátodo que consiste en aleaciones de metales alcalinos con lo que el fotocátodo queda cargado positivamente. Si el cátodo es alimentado

por electrones que a su vez son desprendidos y alcanzan un ánodo con la ayuda de un voltaje adecuado, entonces observamos una corriente eléctrica que llamaremos fotocorriente. Con el objetivo de evitar interacciones con moléculas en el aire este proceso se lleva a cabo a un alto vacío. Generalmente esta fotocorriente se amplifica con ayuda de un fotomultiplicador. Esto es una serie de dinodos ordenados en serie que aumentan el potencial eléctrico. Una cascada de electrones es arrancada de los dinodos y la señal puede amplificarse hasta un orden de 10^7. Los fotomultiplicadores son los fotodetectores más sensibles por lo que son muy utilizados en el rango UV-Visible.

Celdas

Las muestras líquidas se miden en contenedores especiales llamados celdas. Dependiendo de la aplicación, existen diferentes tipos y calidades de celdas. Para mediciones en el rango UV visible por lo general se utilizan celdas de cuarzo que son transparentes en esta región del espectro. Las celdas estándar utilizadas en la mayoría de los espectrofotómetros se colocan de manera que la luz pase horizontalmente sobre la celda. La desventaja esencial es que solo una fracción de la celda es iluminada por el haz de medición (menos del 10% en celdas estándar) lo que impide la medición cuando el volumen de la muestra es muy pequeño. Para una información más detallada respecto a las secciones 1.1, 1.2 y 1.3 véase referencias [5,6].

2. Métodos Multivariantes

El campo multidisciplinario donde se desenvuelven los métodos multivariantes es muy extenso debido a su gran aplicabilidad en muchas situaciones que presenten un análisis estadístico ya que pueden simplificar la estructura o representación de los datos de estudio. También pueden usarse para clasificar, es decir, ubicar observaciones dentro de grupos o concluir que las muestras o individuos están dispersos en un espacio dado. Además pueden emplearse en el análisis de dependencias de variables con regresiones o análisis de correlaciones e incluso, a partir de un conjunto de datos pueden predecir o encontrar modelos que permitan formular hipótesis en función de los resultados de estas técnicas estadísticas.

Cada situación requiere una evaluación particular para elegir el método multivariante más adecuado, que permita extraer la máxima información posible del conjunto de datos, pero que a su vez garantice la validez de su aplicabilidad. En este capítulo se describen brevemente dos métodos estadísticos multivariantes básicos empleados para la elaboración de este trabajo. Análisis de Componentes Principales (PCA) utilizado para reducir la dimensionalidad del problema y Maquinas de Soporte Vectorial (SVM) para la clasificación de los tipos de los tequilas.

2.1. Análisis de Componentes Principales (PCA)

En 1901 Karl Pearson publicó un trabajo sobre ajustes ortogonales por mínimos cuadrados de un multiespacio a una línea o a un plano. Este enfoque fue retomado por Hotelling en 1933 quien fue el primero en formular el análisis por componentes principales tal y como se ha difundido hasta nuestros días. El trabajo original de Pearson se centraba en aquellos componentes, o combinaciones lineales que generan un plano, función de las variables originales, en el cual el ajuste del sistema es "el mejor" por ser mínima la suma de las distancias de cada punto al plano de ajuste. El enfoque de Hotelling se centraba en el análisis de las componentes que sintetizan la mayor variabilidad del sistema de puntos, ello explica quizás el calificativo de "principal".

Un problema central en el análisis de datos multivariantes es la reducción de la dimensionalidad, es decir, determinar si es posible describir con precisión los valores de p variables por un subconjunto $r < p$ de ellas, con ello se habrá reducido la dimensión del problema a costa de una pérdida de información que, en principio, debe ser pequeña. Los objetivos principales de ésta técnica, además de la reducción de dimensionalidad del problema, son encontrar nuevas variables incorreladas que expresen de la mejor manera el sistema original y además, eliminar, cuando sea posible, las variables originales que aporten muy poca o nula información del sistema en cuestión.

Las nuevas variables generadas se denominan componentes principales, y poseen algunas características estadísticas deseables, tales como ortogonalidad, por lo tanto no correlación entre ellas. Esto significa que si las variables originales están incorreladas, el análisis por componentes principales no ofrece ventaja alguna. La literatura acerca de la construcción de los componentes principales, de su uso y de sus propiedades es muy amplia, sin embargo, en este documento solo

se hará una descripción breve que ayude a entender el problema que nos concierne. En esta sección planteamos el problema y describimos el cálculo de las primeras dos componentes principales y la generalización de este cálculo.

Planteamiento de problema

Supongamos que se dispone de los valores de p-variables en n elementos de una población dispuestos en una matriz $\mathbf{X}$ de dimensiones $n \times p$, donde las columnas contienen las variables y las filas los elementos. Supondremos en este capítulo que previamente se ha restado a cada variable su media, de manera que las variables de la matriz $\mathbf{X}$ tienen media cero y su matriz de covarianzas vendrá dada por $\frac{1}{n}\mathbf{X}'\mathbf{X}$.

El problema que se desea resolver es encontrar un subespacio de dimensión más reducida que represente los datos. Puede abordarse desde tres perspectivas diferentes.

a) Enfoque desccriptivo

Se desea encontrar un subespacio de dimensión menor que p tal que al proyectar sobre él los puntos conserven su estructura con la menor distorsión posible. Veamos cómo convertir esta noción intuitiva en un criterio matemático operativo. Consideremos primero un subespacio de dimensión uno, una recta. Se desea que las proyecciones de los puntos sobre esta recta mantengan, lo más posible, sus posiciones relativas. Para concretar, consideremos el caso de dos dimensiones $(p = 2)$.

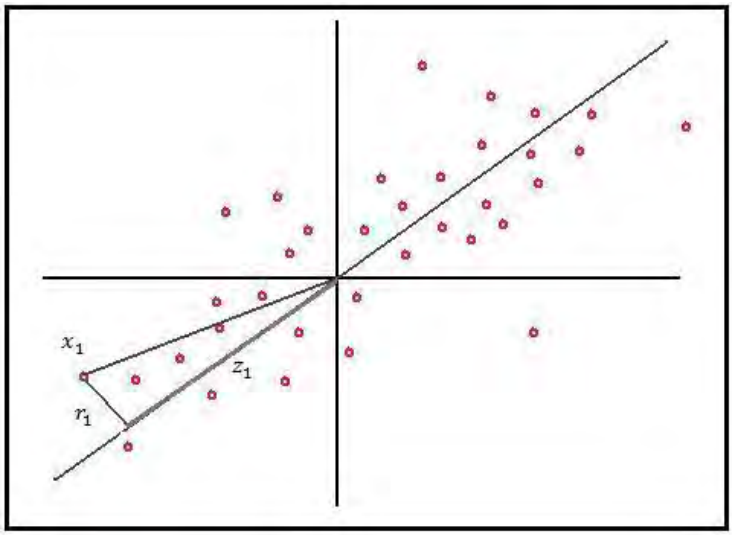


Figura 2.1. Ejemplo de una recta donde la proyección de los datos conserva la mayor información.

La figura 2.1 indica el diagrama de dispersión y una recta que, intuitivamente, proporciona un buen resumen de los datos, ya que la recta pasa cerca de todos los puntos y las distancias entre ellos se mantienen aproximadamente en su proyección sobre la recta. La condición de que la recta pase cerca de la mayoría de los puntos puede concretarse exigiendo que las distancias entre los puntos originales y sus proyecciones sobre la recta sean lo más pequeñas posibles. En consecuencia, si consideramos un punto $\boldsymbol{x_i}$ y una dirección $\boldsymbol{a_1} = (a_{11}, \dots a_{1p})'$, definida por un vector $\boldsymbol{a}_1$ de norma unidad, la proyección del punto $\boldsymbol{x}_i$ sobre esta dirección es el escalar:

$$z_i = a_{11}, x_{i1} + \cdots + a_{1p} x_{ip} = \boldsymbol{a}_1 \boldsymbol{x}_i \tag{2.1.1}$$

Y el vector que representa ésta proyección será $z_i \boldsymbol{a}_1$. Llamando r_i a la distancia entre el punto $\boldsymbol{x}_i$, y su proyección sobre la dirección $\boldsymbol{a}_1$, este criterio aplica:

$$minimizar \ \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} |\mathbf{x}_i - z_i \boldsymbol{a}_1|^2, \tag{2.1.2}$$

donde $|\boldsymbol{u}|$ es la norma euclídea o modulo del vector **u**.

La figura 2.1 muestra que al proyectar cada punto sobre la recta se forma un triángulo rectángulo donde la hipotenusa es la distancia del punto al origen $(\mathbf{x}_i{}'\mathbf{x}_i)^{1/2}$, y los catetos la proyección del punto sobre la recta (z_i) y la distancia entre el punto y su proyección (r_i). Por el teorema de Pitágoras, podemos escribir:

$$\mathbf{x}_i^{'}\mathbf{x}_i = z_i^2 + r_i^2 \tag{2.1.3}$$

Y sumando esta expresión para todos los puntos se obtiene:

$$\sum_{i=1}^{n} \mathbf{x}_i^{'}\mathbf{x}_i = \sum_{i=1}^{n} z_i^2 + \sum_{i=1}^{n} r_i^2. \tag{2.1.4}$$

Como el primer miembro es constante, minimizar $\sum_{i=1}^{n} r_i^2$, la suma de las distancias a la recta de todos los puntos, es equivalente a maximizar $\sum_{i=1}^{n} z_i^2$, la suma al cuadrado de los valores de las proyecciones. Como las proyecciones z_i son, por la ecuación (2.1.1) variables de media cero, maximizar la suma de sus cuadrados equivale a maximizar su varianza, y obtenemos el criterio de encontrar la dirección de proyección que maximice la varianza de los datos proyectados. Este resultado es intuitivo: la recta de la Figura 2.1 parece adecuada porque conserva lo más posible la variabilidad original de los puntos. El lector puede convencerse considerando una dirección de proyección perpendicular a la de la recta en esta figura: los puntos tendrían muy poca variabilidad y perderíamos la información sobre sus distancias en el espacio. Si en lugar de buscar la dirección que pasa cerca de los puntos buscamos la dirección tal que los puntos proyectados sobre ella conserven lo más posible sus distancias relativas llegamos al mismo criterio. En efecto, si llamamos $d_{ij}^2 = \mathbf{x}_i{}'\mathbf{x}_j$ a los cuadrados de las distancias originales entre los puntos y $\hat{d}_{ij}^2 = (z_i - z_j)^2$ a las distancias entre los puntos proyectados sobre una recta, deseamos que

$$D = \sum_i \sum_j (d_{ij}^2 - \hat{d}_{ij}^2)$$

Sea mínima. Como la suma de las distancias originales es fija, minimizar D requiere minimizar $\sum_i \sum_j \hat{d}_{ij}^2$, las distancias entre los puntos proyectados.

b) Enfoque estadístico

Representar puntos p dimensionales con la mínima pérdida de información en un espacio de dimensión uno es equivalente a sustituir las p variables originales por una nueva variable, z_1, que resuma óptimamente la información. Esto supone que la nueva variable debe tener globalmente máxima correlación con las originales o, en otros términos, debe permitir prever las variables originales con la máxima precisión. Esto no será posible si la nueva variable toma un valor semejante en todos los elementos, y puede demostrarse que la condición para que podamos prever con la mínima pérdida de información los datos observados, es utilizar la variable de máxima variabilidad.

Volviendo a la Figura 2.1 se observa que la variable escalar obtenida al proyectar los puntos sobre la recta sirve para prever bien el conjunto de los datos. La recta indicada en la figura no es la línea de regresión de ninguna de las variables con respecto a la otra, que se obtienen minimizando las distancias verticales u horizontales, sino la que minimiza las distancias ortogonales entre los puntos y la recta. Este enfoque puede extenderse para obtener el mejor subespacio resumen de los datos de dimensión 2. Para ello, calcularemos el plano que mejor aproxima a los puntos. El problema se reduce a encontrar una nueva dirección definida por un vector unitario, a_2, que, sin pérdida de generalidad, puede tomarse ortogonal a a_1, y que verifique la condición de que la proyección de un punto sobre este eje maximice las distancias entre los puntos proyectados. Estadísticamente esto equivale a encontrar una segunda variable z_2, incorrelada con la anterior, y que tenga varianza máxima. En general, la componente $\mathrm{z}_\mathrm{r}(\mathrm{r} < p)$ tendrá varianza máxima entre todas las combinaciones lineales de las p

variables originales, con la condición de estar incorrelada con las $z_1, \dots z_{r-1}$ previamente obtenidas.

c) Enfoque geométrico

El problema puede abordarse desde un punto de vista geométrico con el mismo resultado final. Si consideramos la nube de puntos de la Figura 2.1 vemos que los puntos se sitúan siguiendo una elipse y podemos describirlos por su proyección en la dirección del eje mayor de la elipse. Puede demostrarse que este eje es la recta que minimiza las distancias ortogonales, con lo que volvemos al problema que ya hemos resuelto. En tres dimensiones tendremos elipsoides, y la mejor aproximación a los datos es la proporcionada por su proyección sobre el eje mayor del elipsoide. Intuitivamente la mejor aproximación en dos dimensiones es la proyección sobre el plano de los dos ejes mayores del elipsoide esto se generaliza a más dimensiones. Considerar los ejes del elipsoide como nuevas variables originales supone pasar de variables correladas a variables ortogonales o incorreladas como veremos a continuación.

Calculo de los componentes

El primer componente principal se define como la combinación lineal de las variables originales que tiene varianza máxima. Este primer componente está representado por un vector z_1, dado por:

$$z_1 = Xa_1 \tag{2.1.5}$$

Como las variables originales tienen media cero también z_1 tendrá media nula. Su varianza será:

$$\frac{1}{n}z_1'z_1 = \frac{1}{n}a_1'X'Xa_1 = a_1'Sa_1 \qquad (2.1.6)$$

Donde S es la matriz de varianzas y covarianzas de las observaciones. Es obvio que podemos maximizar la varianza sin límite aumentando el módulo del vector a_1. Para que la maximización de (2.1.6) tenga solución debemos imponer una restricción al módulo del vector a_1, y, sin pérdida de generalidad, impondremos que $a_1'a_1 = 1$. Introduciremos esta restricción mediante los multiplicadores de Lagrange:

$$M = a_1'Sa_1 - \lambda(a_1'a_1 - 1) \qquad (2.1.7)$$

Y maximizaremos esta expresión derivando respecto a los componentes de $\mathbf{a}_1$ e igualando a cero. Esto es:

$$\frac{\partial M}{\partial a_1} = 2Sa_1 - 2\lambda a_1 = 0, \qquad (2.1.8)$$

Cuya solución es:

$$Sa_1 = \lambda a_1 \qquad (2.1.9)$$

Que implica que $a_1$es un vector propio de la matriz S, y λ su correspondiente valor propio. Para determinar qué valor propio de S es la solución de (2.1.9), multiplicando por la izquierda por a_1' esta ecuación tenemos:

$$a_1'Sa_1 = \lambda a_1'a_1 = \lambda \qquad (2.1.10)$$

Y concluimos por (2.1.7) que λ es la varianza de z_1. Como esta es la cantidad que deseamos maximizar, λ será el mayor valor propio de la matriz S. Su vector asociado, a_1, define los coeficientes de cada variable en el primer componente principal.

Vamos a obtener el mejor plano de proyección de las variables $\mathbf{X}$. Lo calcularemos estableciendo como función objetivo que la suma de las varianzas de $z_1 = Xa_1$ y $z_2 = Xa_2$ sea máxima, donde a_1 y a_2 son los valores que definen el plano. La función objetivo será:

$$\phi = \mathbf{a}_1^{'}\mathbf{S}\mathbf{a}_1 + \mathbf{a}_2^{'}\mathbf{S}\mathbf{a}_2 - \lambda(\mathbf{a}_1^{'}\mathbf{a}_1 - 1) - \lambda(\mathbf{a}_2^{'}\mathbf{a}_2 - 1) \quad (2.1.11)$$

Que incorpora las restricciones de que las direcciones deben tener módulo unitario $(\mathbf{a}_i{}'\mathbf{a}_i) = 1$, $i = 1{,}2$. Derivando e igualando a cero:

$$\frac{\partial\phi}{\partial\mathbf{a}_i} = 2\mathbf{S}\mathbf{a}_i - 2\lambda_i\mathbf{a}_i = 0 \quad (2.1.12)$$

La solución de este sistema es:

$$\mathbf{S}\mathbf{a}_1 = \lambda\mathbf{a_1} \quad (2.1.13)$$

$$\mathbf{S}\mathbf{a}_2 = \lambda\mathbf{a_2} \quad (2.1.14)$$

Lo cual indica que $\mathbf{a_1}$ y $\mathbf{a}_2$ deben ser vectores propios de $\mathbf{S}$. Tomando los vectores propios de norma uno y sustituyendo en (2.11), se obtiene que, en el máximo, la función objetivo es:

$$\phi = \lambda_1 + \lambda_2 \quad (2.1.15)$$

Es claro que λ_1 y λ_2 deben ser los valores propios mayores de la matriz $\mathbf{S}$ y $\mathbf{a}_1$ y $\mathbf{a}_2$ sus correspondientes valores propios. Observamos que la covarianza entre $\mathbf{z}_1$ y $\mathbf{z}_2$, dada por $\mathbf{a}_1{}'S\mathbf{a}_2$ es cero ya que $\mathbf{a}_1^{'}\mathbf{a}_2 = 0$, y las variables $\mathbf{z}_1$ y $\mathbf{z}_2$ estarán incorreladas. Si en lugar de maximizar la suma de las varianzas, que es la traza de la matriz de covarianzas de la proyección, se maximiza la varianza generalizada (el determinante de la matriz de covarianzas) se obtiene el mismo resultado.

Generalización

Puede demostrarse análogamente que el espacio de dimensión r que mejor representa a los puntos viene definido por los vectores propios asociados a los r mayores valores propios de $\mathbf{S}$. Estas direcciones se denominan direcciones principales de los datos y a las nuevas variables por ellas definidas *componentes principales*. En general, la matriz $\mathbf{X}$ (y por tanto la $\mathbf{S}$) tiene rango p, existiendo entonces tantas componentes principales como variables que se obtendrán calculando los valores propios o raíces características $\lambda_1, \dots, \lambda_p$ de la matriz de varianzas y covarianzas de las variables $\mathbf{S}$ mediante:

$$|\mathbf{S} - \lambda\mathbf{I}| = 0 \tag{2.1.16}$$

Y sus vectores asociados son:

$$(\mathbf{S} - \lambda_i\mathbf{I})\mathbf{a_i} = 0 \tag{2.1.17}$$

Los términos λ_i son reales, al ser la matriz $\mathbf{S}$ simétrica, y positivos, ya que $\mathbf{S}$ es definida positiva. Por ser $\mathbf{S}$ simétrica si λ_j y λ_h son dos raíces distintas sus vectores asociados son ortogonales. Si $\mathbf{S}$ fuese semidefinida positiva de rango $r < p$, lo que ocurriría si $p - r$ fuesen combinación lineal de las demás, habría solamente r raíces características positivas y el resto serían ceros.

Llamando $\mathbf{Z}$ a la matriz cuyas columnas son los valores de los p componentes en los n individuos, estas nuevas variables están relacionadas con las originales mediante:

$$\mathbf{Z} = \mathbf{XA} \tag{2.1.18}$$

Donde $\mathbf{A}'\mathbf{A} = \mathbf{I}$. Calcular los componentes principales equivale a aplicar una transformación ortogonal $\mathbf{A}$ a las variables $\mathbf{X}$ (ejes originales) para obtener unas nuevas variables $\mathbf{Z}$ incorreladas entre sí. Esta operación puede interpretarse como elegir unos nuevos ejes coordenados, que coincidan con los "ejes naturales" de los datos. Para una descripción más amplia del Análisis de Componentes Principales se recomiendan las lecturas [7,8,9,10 y 11].

2.2. Support Vector Machines (SVM)

El objetivo de esta sección no es proporcionar una descripción explicita y detallada del método, sino una descripción básica que permita entender de manera sencilla la clasificación binaria de una base de datos a través de planos de separación. Máquinas de Soporte Vectorial (SVM) es un sistema de aprendizaje que usa como hipótesis un espacio de funciones lineales en un espacio de alta dimensionalidad, entrenado con un algoritmo basado en teorías de optimización que permita separar óptimamente dos clases. Este método de aprendizaje estadístico fue introducido por Vapnik en 1995. El clasificador toma una parte de los datos a analizar para entrenar el modelo y otra parte de los datos son empleados para verificar el modelo o clasificar.

La clasificación de un objeto se puede describir asignándolo a una clase u otra por medio de un plano se separación (hiperplano en general). Cuando se habla de clases que se sobreponen, una buena aproximación es permitir a algunos objetos estar en el lado equivocado del margen. A continuación describimos un clasificador binario. Considérese un hiperplano como el siguiente:

$$f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{w} + b \tag{2.2.1}$$

Donde $\boldsymbol{w}$ denota un vector de peso y b representa el corrimiento en el eje de las abscisas. Una interpretación geométrica de esta hipótesis es que si se considera al espacio de entrada $\boldsymbol{x}$ ser dividido por un hiperplano definido por la ecuación $\boldsymbol{x}^T \boldsymbol{w} + b = 0$, donde este hiperplano es de dimensiones $n - 1$ el cual divide al espacio en dos partes que corresponde a las dos distintas clases. El hiperplano está representado por la figura 2.2 como la línea que divide al plano en dos regiones, una región positiva que está por encima del hiperplano y otra negativa por debajo del mismo. El vector $\boldsymbol{w}$ define la dirección normal al hiperplano, mientras que una variación del parámetro b mueve al hiperplano de forma paralela a él.

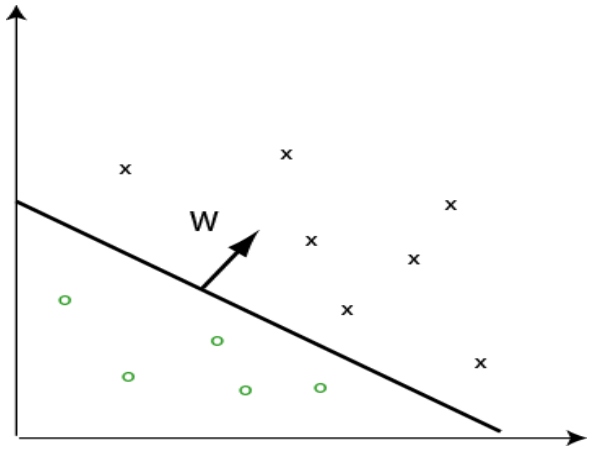


Figura 2.2. Interpretación geométrica del hiperplano de separación en dos dimensiones.

Las reglas de decisión se pueden definir por:

$$G(\boldsymbol{x}) = sign(\boldsymbol{x}^T \boldsymbol{w} + b) \tag{2.2.2}$$

Dando un vector de etiquetas o de clasificación, $\boldsymbol{y}$, en el intervalo $[-1,1]$. Para un problema de dos clases, una máquina de soporte vectorial es entrenada de manera que la función de decisión y los datos de entrada $\boldsymbol{x}$ en un espacio m sean

mapeados a otro espacio l donde $l \geq m$ llamado $\boldsymbol{z}$. Así, en $\boldsymbol{z}$, se resuelve el problema de programación cuadrática y se separan las dos clases con el hiperplano óptimo. Se debe encontrar una función como (2.2.1) con $y_i f(x_i) > 0$ para toda i y se debe encontrar el hiperplano que genere el margen más grande entre los puntos de entrenamiento para las clases definidas como 1 y -1. El problema de optimización lo podemos escribir como:

$$minimizar\left(\frac{1}{2}\|w\|^2\right) \tag{2.2.3}$$

sujeto a la condición $y_i(\boldsymbol{x}_i^T \boldsymbol{w} + b) \geq 0$ para $i = 1, \ldots, n$. El problema se puede visualizar en la figura 2.3, donde la frontera de decisión está representada por una línea sólida. El margen máximo es el doble de la distancia C de la línea de decisión. C es justo el valor recíproco de la norma de los pesos, esto es $\frac{1}{\|w\|}$.

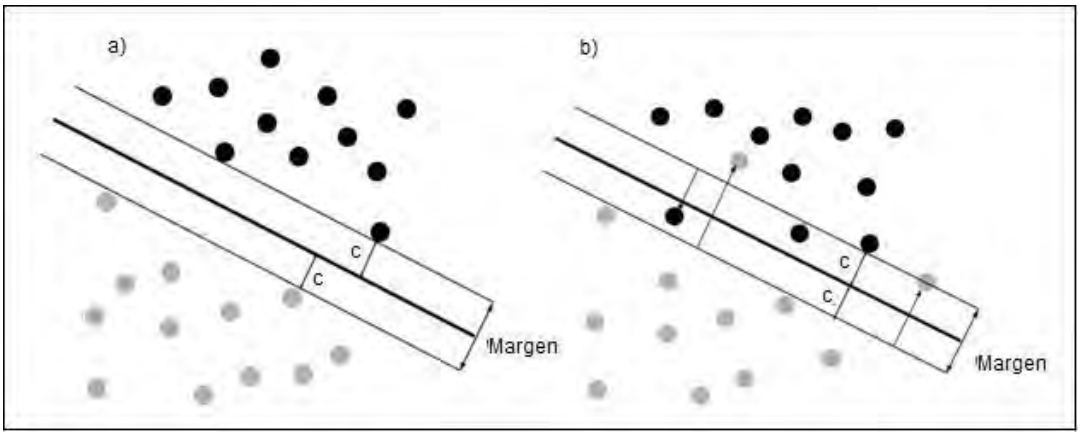


Figura 2.3. SVM. a) Caso separable y b) Caso no separable del problema de máximo margen.

Ahora consideremos el caso más general donde las clases se traslapan. Aun se puede tratar de encontrar el hiperplano si permitimos que algunos puntos estén

en la clase que no les corresponde, es decir, del otro lado del margen. Podemos definir una nueva variable ξ y modificar la ecuación 2.2.3 de la siguiente manera:

$$minimizar\left\{\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i\right\} \quad (2.2.4)$$

Sujeta a la condición:

$$y_i(\boldsymbol{x}_i^T\boldsymbol{w} + b) \geq 1 - \xi_i \quad (2.2.5)$$

para todo $i, \xi_i \geq 0$. El primer término de la ecuación (2.2.4) maximiza el margen de separación mientras que el segundo término penaliza los objetos que se encuentran del lado equivocado de la separación en casos linealmente no separables.

El problema en la ecuación (2.2.4) es una situación estándar de optimización, la minimización de una función cuadrática con restricciones lineales. Este problema se puede resolver aplicando la teoría de multiplicadores de Lagrange. Resolviendo el problema sigue que el vector de peso de la función de decisión está dado por una combinación lineal de los datos empleados en el entrenamiento y por el multiplicador de Lagrange α de la siguiente forma:

$$\boldsymbol{w} = \sum_{i=1}^{n}\alpha_i y_i \boldsymbol{x}_i \quad (2.2.6)$$

Los vectores de entrenamiento $\boldsymbol{x}_i$ con eigenvalores diferentes de cero que corresponden a las fronteras de cada clase determinarán la función de decisión y son los llamados *vectores de soporte*. En contraste con los métodos de clasificación tradicionales como *discriminant analysis* no se asume nada acerca de las distribuciones de las clases con este método de *SVM* [12,13].

3. Resultados y Discusión

El proceso de elaboración de Tequila termina con el envasado del producto, el cual está limitado por cierto volumen conocido como lote. Cada lote representa un ciclo de producción de Tequila y es envasado en un mismo lapso de tiempo por lo que garantiza la identificación de la producción, es decir, una botella de Tequila proveniente de un lote debe contener características idénticas a otra botella procedente de un mismo lote. Las botellas de Tequila llevan grabadas o marcadas la identificación del lote al que pertenecen. Todas las marcas de Tequila poseen un control de calidad similar, sin embargo algunas de ellas, ya consagradas en el mercado, han logrado una producción donde se controla rigurosamente cada fase de la elaboración del producto. El análisis de Tequilas realizado en este trabajo requiere de muestras de Tequila provenientes todas ellas de distintos lotes, para garantizar un análisis más general para cada marca y evitar comparar muestras que, de antemano, sabemos deben ser idénticas. Las muestras de Tequila se adquirieron en establecimientos comerciales dedicados a la venta de bebidas alcohólicas y algunos bares en las ciudades de León Guanajuato y Zacatecas, Zac., teniendo especial cuidado en no repetir el número de lote de ninguna muestra.

El análisis de este trabajo se basa en dos clases de Tequila: blanco y reposado. Para evitar confusiones con la terminología empleada en el texto se utilizará indistintamente ***clase*** ó ***tipo***, haciendo referencia a las clases de Tequilas que define la Norma Oficial Mexicana. El criterio de selección de las marcas de Tequila que se analizan se basa en el prestigio y popularidad que gozan en la

actualidad. Se analizan un total de 80 muestras, de las cuales 39 son muestras de Tequila blanco de 4 marcas distintas y 41 más para el tipo reposado procedentes de 4 marcas distintas. La tabla 3.1 indica el número de muestras analizadas y se asigna un nombre o símbolo a cada marca. MB1 y MR1 son los símbolos para la misma marca analizada a la que llamaremos marca 1 para los tipos blanco y reposado respectivamente. Las marcas 1 y 3 se analizan en las dos clases de Tequila. Todas las marcas, a excepción de MR3 corresponden a Tequilas 100% agave.

Tabla 3.1

Adquisición de muestras			
Tequila Blanco		Tequila Reposado	
Marca	Número de muestras	**Marca**	Número de muestras
MB1	9	**MR1**	11
MB3	13	**MR2**	10
MB5	10	**MR3**	10
MB6	7	**MR4**	10

3.1. Espectroscopia UV-Visible

Las muestras de Tequila mencionadas en la sección anterior se analizaron a través de espectroscopia de absorción en la región Ultravioleta-visible del espectro electromagnético. Se empleó un espectrofotómetro Perkin Elmer

modelo Lambda 900 para medir la absorción de las muestras en el rango de 250nm a 500nm para ambos tipos de Tequilas. Se utilizaron celdas de cuarzo con una longitud de camino óptico de 1cm para los Tequilas blancos. Los Tequilas de tipo reposado presentan una mayor absorción en el rango de análisis lo que provocó una saturación en la detección de la señal. Para contrarrestar la absorción y evitar esta saturación se redujo el camino óptico de 1cm a 2mm para estos Tequilas. La diferencia de caminos ópticos entre cada tipo de Tequila se iguala con la ayuda de la ley de Beer-Lambert-Bouguer (sección 1.3) con el fin de analizar los espectros de ambos tipos bajo las condiciones lo más idénticas posibles. Para obtener el espectro de absorción es suficiente una cantidad aproximada de 3.5ml en el caso de los Tequilas blancos y 0.7ml para los reposados. En el canal de referencia del espectro se colocaron celdas de cuarzo vacías de 10mm y 2mm respectivamente.

La resolución mínima del espectrofotómetro es de 0.5nm, esto es, podemos discernir entre picos de absorción separados por media unidad de nanómetro. Para fines prácticos, los espectros de absorción de los Tequilas medidos con resoluciones de un nanómetro arrojan la misma información que los medidos a 0.5 nanómetro. Con el objetivo de optimizar el tiempo en las mediciones se optó por medir los espectros con pasos de un nanómetro. Asociamos una variable a cada longitud de onda en los análisis posteriores de este capítulo, es decir, a cada espectro de Tequila se le asocian 251 variables, una para cada nanómetro en el intervalo 250-500nm del espectro electromagnético (figura 3.1).

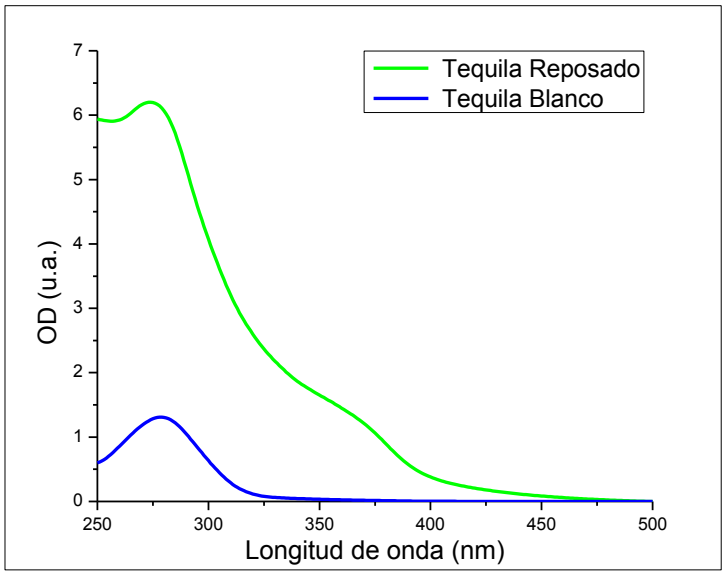


Figura 3.1. Espectros de absorción característicos de Tequilas tipo blanco y reposado. El rango de absorción comprende de 250nm a 500 nm en la región UV-Visible del espectro electromagnético. Para nuestro estudio se asocia una variable a cada nanómetro del intervalo comprendido.

Se seleccionó una absorción de valor cero en 500nm con el fin de asegurar la misma línea base antes de realizar cualquier análisis sobre los datos. Se realizó sin dificultad alguna la reproducibilidad de las mediciones en ambos tipos de Tequilas. De acuerdo al proceso de elaboración, los Tequilas reposados se someten a un proceso de maduración en barricas de madera, lo que podría añadir partículas de orden microscópico que favorecen el esparcimiento de la luz afectando la medición del espectro de manera significativa. Con el objetivo de verificar esta hipótesis se midió la absorbancia para una muestra de Tequila reposado previa y posteriormente a un filtrado de partículas de tamaño mayor a los 200nm; los espectros no mostraron variaciones notorias. Este resultado se corroboró aplicando la primera derivada mostrando que, para fines prácticos, es

idéntica en ambos espectros en todo el intervalo a excepción de una región del espectro que se encuentra entre los 320-380nm aproximadamente (figura 3.2).

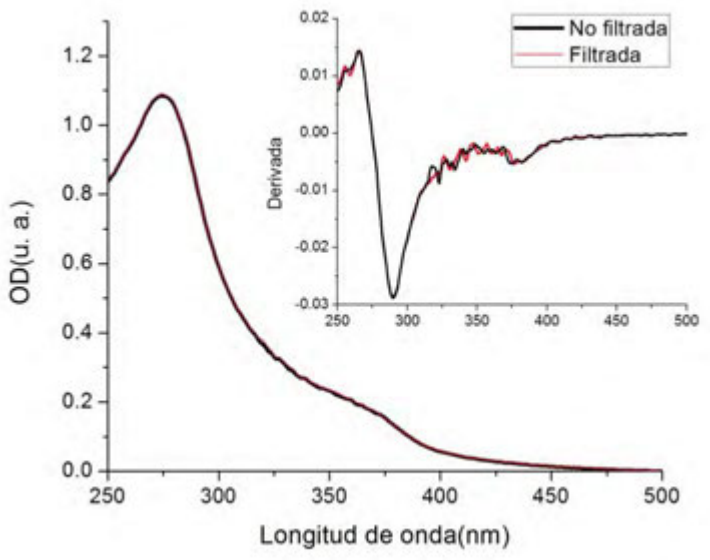


Figura 3.2. Espectro de absorción y su primera derivada para una muestra de Tequila antes y después de un filtrado de partículas mayores a los 200nm.

Con base a este resultado despreciamos el esparcimiento que pudiese existir en el experimento debido a micro partículas presentes en los Tequilas reposados y consideramos que no es un factor determinante en el análisis posterior de este trabajo.

Tequilas Blancos

La absorción de los Tequilas blancos en la región UV-Visible del espectro electromagnético está caracterizada por una banda que se encuentra aproximadamente entre los 250nm y 350nm y absorción nula de los 375 a los

500nm Para fines prácticos se considera la absorción nula a partir de los 350nm. Las bandas de absorción de los Tequilas revelan la presencia de compuestos orgánicos originados en la cocción y fermentación del agave [14,15]; y se han reportado espectros muy similares en otras bebidas alcohólicas como brandis, coñacs y whiskies [16]. La figura 3.3 muestra la densidad óptica para las 39 muestras de Tequila blanco en cuestión.

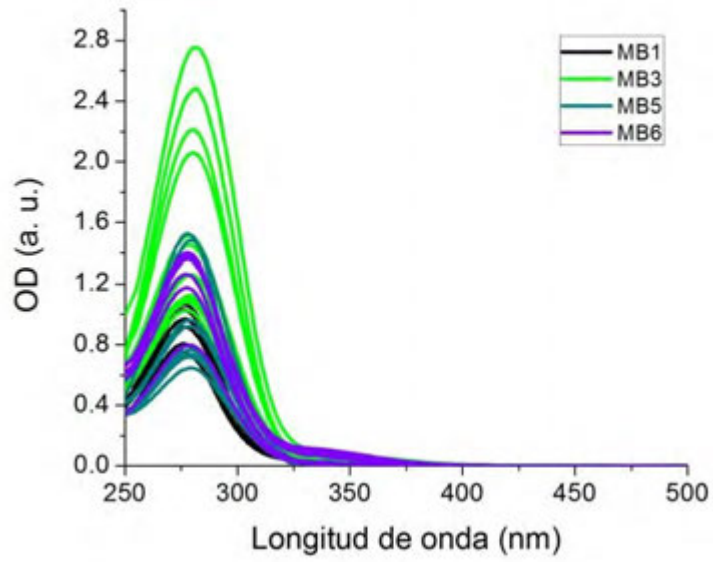


Figura 3.3. Espectros de absorción en la región UV-Visible de 39 muestras de Tequila blanco correspondientes a las 4 marcas analizadas.

La densidad óptica de los Tequilas blancos analizados se encuentra entre 0.5 y 2.8, sin embargo la mayoría de las muestras no sobrepasan el valor de 1.5 con excepción de MB3; esta marca presenta una absorción mayor que cualquiera de las marcas estudiadas y sus muestras presentan mayor variación en su máximo de absorbancia. En contraste a MB3, MB1 está representada por muestras muy similares con máximos de absorción muy cercanos a un valor de 0.9. Estas variaciones se pueden asociar de manera intuitiva al control de calidad que posee

cada compañía productora. Se han reportado estudios de cromatografía, RP-HPLC (Reverse Phase High Performance Liquid Chromatography), para determinar la naturaleza de la banda de absorción en la región de 250-400nm de los tequilas blancos. Son tres componentes orgánicos los que dan origen a esa banda ancha de absorción y corresponden a Furfural, 5-Methylfurfural y 2-Acetylfuran. Es la suma de los espectros de absorción de estos componentes, ver Fig. 3.4, quienes generan no sólo el espectro de absorción de tequilas sino también de mezcales [17].

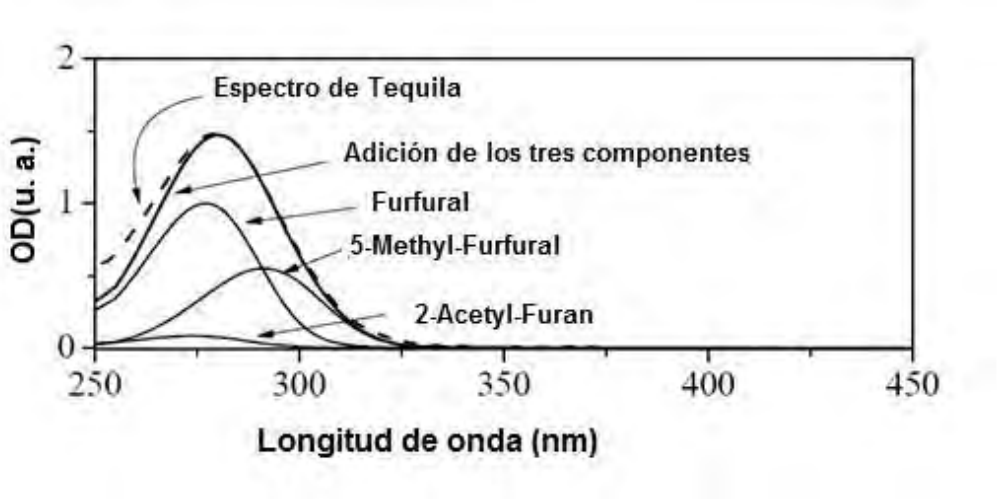


Figura 3.4. Espectros de absorción de soluciones sintéticas de componentes individuales disueltos a ciertas concentraciones para una muestra de Tequila blanco. La línea punteada representa el espectro de absorción de un Tequila blanco. Imagen obtenida de la referencia [17].

De acuerdo a la concentración de esas componentes en cada tequila se genera el espectro característico correspondiente; esto da explicación a las pequeñas diferencias del pico de absorción máximo observado en nuestros espectros reportados en la figura 3.5; esto es, la posición del máximo no es igual para todos los tequilas debido a que contienen distintas concentraciones de sus componentes. Sin embargo, la curva generada por la suma de esos espectros no genera completamente el espectro medido. Esta diferencia, reportan los autores

de esa referencia, corresponde a la presencia de otros componentes orgánicos no determinados en el estudio realizado. También es de importancia hacer notar que de acuerdo a la norma mexicana y para la certificación de cualquier tequila el contenido de Furfural tiene una cota máxima de 4 ppm (partes por millón); en el caso de los tequilas estudiados ninguno sobrepasa este valor y en la misma referencia se demuestra que los tequilas blancos 100% agave pueden distinguirse de los mixtos con base a la concentración de Furfural; en el mismo artículo se discrimina entre tequilas y mezcales.

Con el fin de mostrar una manera más clara el comportamiento de las cuatro marcas analizadas, se promedian los espectros de absorción para cada marca y se grafica cada promedio con barras de error estándar de cada marca (Figura 3.5). Para poder distinguir las barras de error entre cada nanómetro se muestra un subconjunto menor de longitudes de onda comprendido entre los 250nm y los 350nm. En el caso de los Tequilas blancos, este intervalo de 100nm contiene aproximadamente el 100% de la información obtenida por absorción en el rango de análisis.

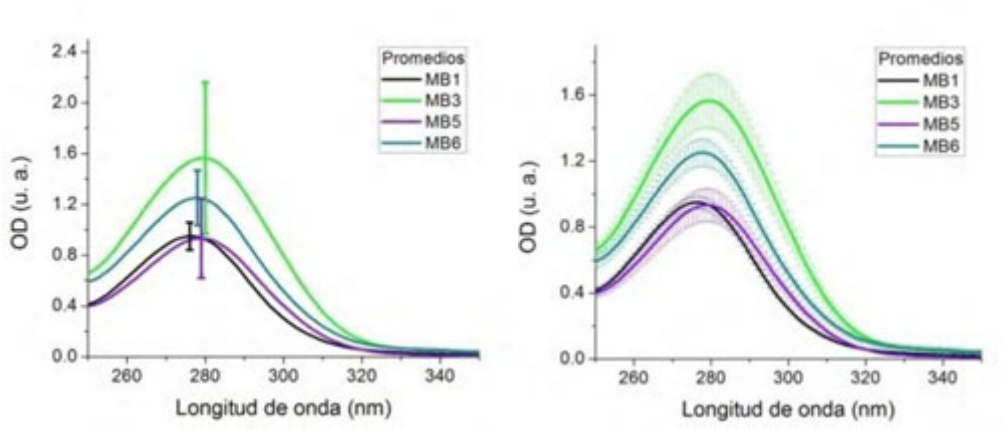


Figura 3.5. a) Espectros de absorción promedio y su desviación estándar para el máximo de absorción para las cuatro marcas de Tequila blanco analizadas. b) Promedio de los espectros de absorción para las cuatro marcas de Tequila blanco analizadas y su error estándar a cada longitud de onda.

La Figura 3.5 muestra de manera más evidente que cada marca está representada por una banda característica en el rango de análisis de la región UV-Visible de espectro electromagnético. Los máximos de absorción de las bandas promedio de cada marca se encuentran en 276nm, 280nm, 279nm y 278nm para las marcas MB1, MB2, MB5 y MB6 respectivamente. También se muestra la desviación estándar correspondiente a los máximos del promedio de cada marca. Los máximos de absorción y su respectiva desviación estándar también son diferentes para cada marca y tienen valores de 0.95±0.10817, 1.55±0.59686, 0.93±0.31309, y 1.25±0.2169 para MB1, MB3, MB5 y MB6 con el mismo orden. La desviación estándar indica la dispersión de muestras provenientes de distintos lotes para cada marca en particular y calculando el error estándar para cada longitud de onda podemos definir una cota de error en la absorción para cada marca. Sin embargo este análisis no es suficiente para discriminar entre marcas a pesar de que se observan las diferencias ya mencionadas. Por ejemplo, se puede distinguir claramente entre espectros de absorción que pertenecen a las marcas MB3 y MB6 pero no podemos decir lo mismo entre espectros de las marcas MB1 y MB5.

Tequilas Reposados

En el caso de los Tequilas reposados, la absorción en la región UV-Visible está caracterizada por una banda ancha con un máximo cercano a los 278nm y absorción prácticamente nula después de los 450nm. La figura 3.6 muestra los espectros de absorción de las 41 muestras de Tequila analizadas correspondientes a las marcas MR1, MR2, MR3 y MR4.

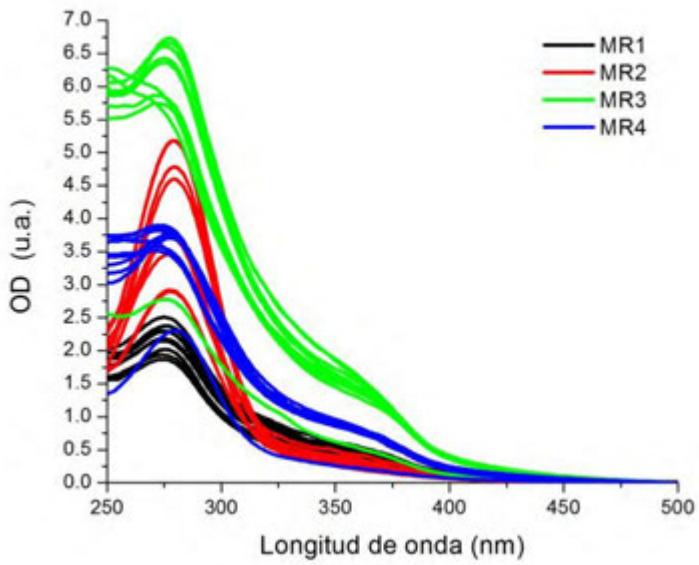


Figura 3.6. Espectros de absorción en la región UV-Visible del espectro electromagnético de las 41 muestras de Tequilas reposados.

Los espectros de absorción de la figura 3.6 presentan densidades ópticas mayores a los Tequilas blancos y oscilan entre los 6.75 y 1.75 unidades para sus respectivos máximos de absorción. Para cada marca, existen diferencias de amplitud en las bandas correspondientes a las muestras analizadas como ocurre en los Tequilas blancos. Incluso cuando el control de calidad sea adecuado, no es extraño encontrar diferencias en el espectro de absorción para muestras de una marca ya que cada muestra proviene de destilaciones distintas. Así mismo, se observa un rango definido en OD para cada marca, como podemos ver, por ejemplo, para las muestras MR3 oscilan entre los 5.5 y los 6.75 aproximadamente, mientras que en el caso de las muestras MR1 oscilan entre 1.75 y 2.5 unidades arbitrarias. Con el fin de mostrar las diferencias entre cada marca de manera más clara se grafican los espectros de absorción promedio de

las 4 marcas de Tequilas reposados con su respectivo error estándar y su desviación estándar en el máximo de absorción de cada marca (figura 3.7).

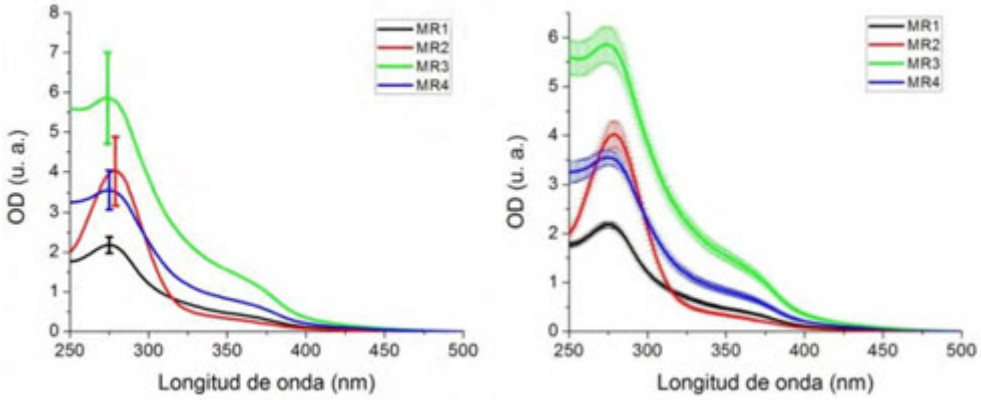


Figura 3.7. a) Espectros de absorción promedio y su desviación estándar para el máximo de absorción para las cuatro marcas de Tequila reposado analizadas. b) Promedio de los espectros de absorción para las cuatro marcas de Tequila reposado analizadas y su error estándar a cada longitud de onda.

En la figura 3.7 se puede observar que cada marca en particular tiene una banda de absorción promedio definida con máximos de absorción en diferentes longitudes de onda al igual que los Tequilas blancos. Las desviaciones estándar y los máximos de absorción de las bandas mostradas en la figura 3.7 son 2.16±0.20487, 3.98±0.85838, 5.87±1.1524 y 3.53±0.48231 para las marcas MR1, MR2, MR3 y MR4 respectivamente. El error estándar de cada marca define una cota de absorción y se puede distinguir a las cuatro marcas claramente. Estas diferencias entre marcas son de esperarse ya que las compañías productoras elaboran sus Tequilas bajo condiciones diferentes y los procesos de producción, el equipo de destilación usado y los tiempos de reposo y maduración de sus productos dependen de cada compañía [2]. Además de los compuestos orgánicos

originados en la cocción y en la fermentación de un Tequila, los Tequilas reposados son sometidos a un proceso de maduración en contacto directo con barricas de encino o roble por determinado tiempo añadiéndole nuevos compuestos orgánicos. Una diferencia notoria en la absorbancia de los Tequilas blancos y reposados es que estos últimos presentan un hombro de absorción entre 350 y 450nm del espectro electromagnético prolongando la banda de los 250nm a los 450nm. Para visualizar de una manera más clara las diferencias ya mencionadas, se muestra la figura 3.8 con espectros de Tequilas blancos y reposados para dos marcas analizadas en este trabajo.

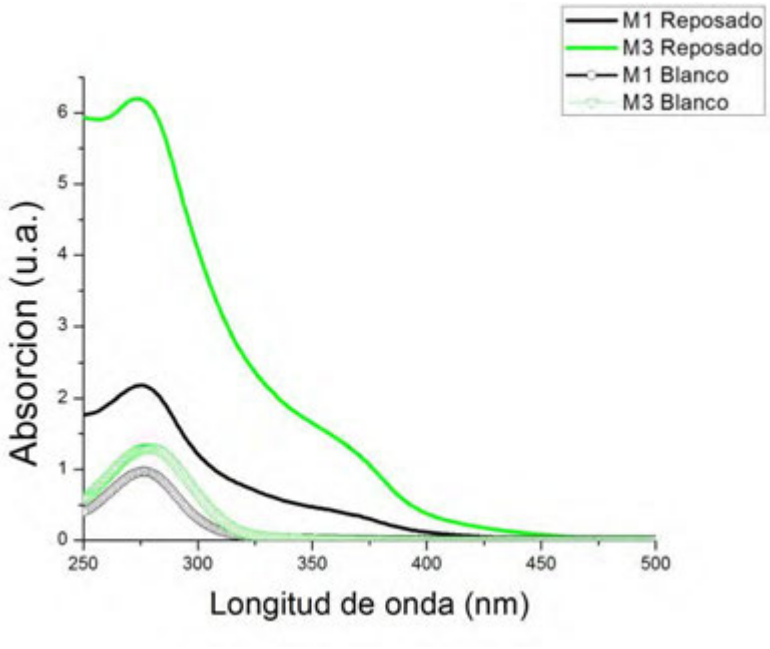


Figura 3.8. Espectros de absorción promedio de dos marcas de Tequila en los tipos blanco y reposado.

Los máximos de absorción para Tequilas blancos y reposados están ubicados sobre un mismo rango de longitudes de onda. Por otro lado, la absorción en los Tequilas blancos prácticamente es nula para longitudes de onda mayores a los 350nm mientras que en los Tequilas reposados se extiende una banda de

absorción desde los 250nm hasta la región visible cercana a los 450nm. La banda de absorción de los Tequilas reposados es más compleja que la banda de los Tequilas blancos en el rango analizado, lo que corrobora la presencia de otros componentes orgánicos adicionales originados probablemente en el proceso de maduración. Algunos autores reportan la presencia de más de 200 componentes orgánicos en la composición de un Tequila; algunos de estos compuestos como los aldehídos se detectaron espectroscópicamente en la región UV entre 365nm y 390nm [18]. El análisis realizado en este trabajo no nos permite determinar ni cuantificar los componentes de los Tequilas blancos y reposados, sin embargo usamos las diferencias y similitudes presentes en sus bandas de absorción para una posible clasificación e identificación de marcas a través de técnicas quimiométricas, así como la discriminación entre estos dos tipos de Tequilas.

3.2. Análisis multivariante

Hasta el momento, hemos representando la absorción en función de la longitud de onda y observamos ciertas diferencias y similitudes entre las marcas analizadas pero para generar un modelo práctico y confiable y para predecir si una nueva muestra de Tequila pertenece a una de las cuatro marcas recurrimos al análisis multivariante.

Observamos que debido al gran contenido de información en los espectros de absorción y al número de muestras empleadas, el problema se torna estadístico. Cada muestra de Tequila está representada en 251 nanómetros y la información presente en cada nanómetro se encuentra correlacionada con información en

otras longitudes de onda. Si empleamos una notación matricial y asociamos a cada fila una muestra de Tequila y a cada columna una variable representada por un nanómetro del rango 250-500nm, obtenemos una matriz mulivariada. Análisis de Componentes Principales (PCA) es uno de los tantos métodos que existen en la actualidad creados con la intención de identificar patrones presentes en matrices multivariadas. Otra tarea fundamental de PCA es reducir la dimensionalidad del problema, es decir, representar las 251 variables que se obtienen de los espectros a un número menor de variables incorreladas sin perder información de manera significativa. Esta idea se representa de manera esquemática en la figura 3.9.

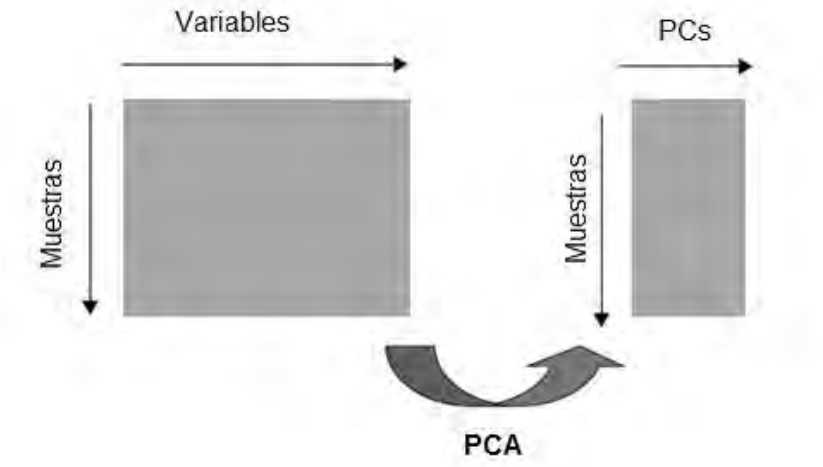


Figura 3.9. Diagrama ilustrativo de la reducción de dimensiones por PCA.

Todos los espectros de absorción se suavizaron de la misma manera previamente al análisis estadístico con el fin de mejorar la razón señal a ruido. Debido al bajo número de muestras en este análisis estadístico, empleamos la técnica de validación cruzada completa, donde todas las muestras son usadas

para calibrar y validar el modelo. La validación cruzada completa se utilizó en todos los análisis de PCA presentados en este trabajo.

El cálculo realizado por PCA depende directamente de la matriz de varianza o covarianza (como es el caso) que contiene la información de muestras de Tequilas, y por lo tanto, una matriz con información diferente producirá un modelo diferente. También es importante hacer énfasis en que el método empleado es no supervisado, es decir, las muestras de Tequila empleadas en cualquier análisis de componentes principales no están etiquetadas y el mapeo de las variables originales a las nuevas variables incorreladas conocidas como componentes principales está basado solo en la información contenida en la matriz de covarianza de la base de datos. El análisis de componentes principales nos permitió reducir la dimensionalidad del problema sin una pérdida de información significativa ya que las primeras dos componentes representan el 100% de la varianza de acuerdo al modelo predicho. Los resultados obtenidos de PCA para la base de datos de 80 muestras de Tequila predicen un modelo de 2 componentes principales con una varianza de 97% para la componente principal y 3% para la segunda componente. En el análisis se centraron los datos, es decir, la media para todas las variables originales usadas es cero. La figura 3.10 muestra una gráfica de PC1 contra PC2 de dicho modelo.

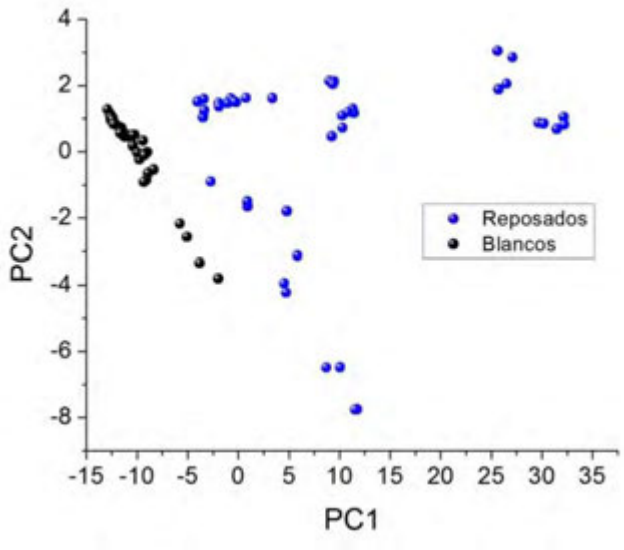


Figura 3.10. Gráfica de PC1 vs PC2 generada por el modelo de PCA para 80 muestras de Tequila, las esferas negras representan 39 muestras de Tequila blanco y las esferas azules representan 41 muestras de Tequila reposado, todas ellas proyectadas en el espacio PC1-PC2.

La idea de este cálculo es encontrar diferencias entre Tequilas blancos y reposados y se utilizó aproximadamente el mismo número de muestras para cada tipo (39 para Tequilas blancos y 41 para los reposados). La figura 3.10 revela una buena agrupación de Tequilas blancos y otra agrupación más extensa pero bien definida de Tequilas reposados. Incluso, para el caso de los Reposados, se observan cuatro pequeños subconjuntos que corresponden a las 4 marcas analizadas. Este Resultado se discutirá más adelante. Una vez reducida la dimensionalidad del problema, utilizamos una técnica que nos permita delimitar el nuevo espacio con la intención de clasificar a los Tequilas en sus tipos blanco y reposado. Utilizamos un método supervisado conocido como Linear Discriminant analysis (LDA), el cual ha sido empleado en trabajos anteriores relacionados con clasificación de Tequilas [17]. Básicamente, LDA encuentra una frontera de

decisión que separa dos grupos. Si se desea separar a más de dos grupos, se deben calcular más funciones discriminantes.

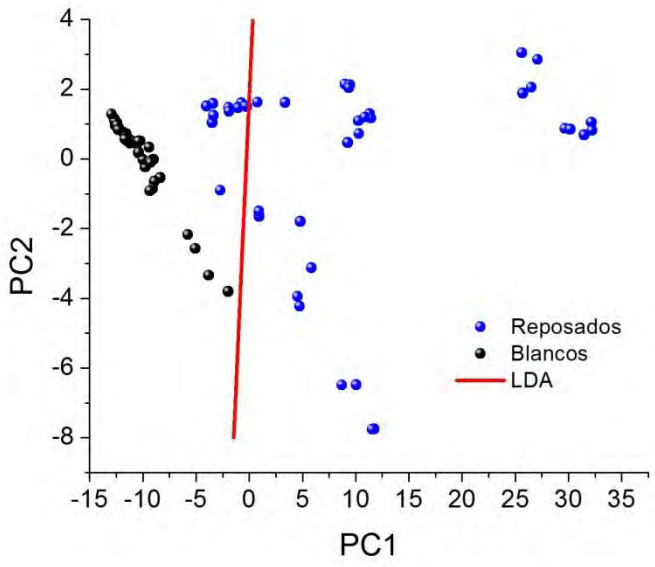


Figura 3.11. Gráfica de PC1 vs PC2 generada por el modelo de PCA para 80 muestras de Tequila, las esferas negras representan 39 muestras de Tequila blanco y las esferas azules representan 41 muestras de Tequila reposado, todas ellas proyectadas en el espacio PC1-PC2. La frontera de separación representada como una línea roja se calcula con el método Linear Discriminant analysis (LDA).

La frontera de decisión obtenida con LDA separa los dos tipos de Tequilas (Figura 3.11), aunque a simple vista se esperaría una mejor separación. Algunas muestras de Tequila reposado se encuentran clasificadas como Tequilas blancos. LDA es un método supervisado que depende directamente de la distribución normal de la información, y en nuestro caso, es evidente que la distribución en cada tipo es diferente, así pues, para este caso en particular, LDA no ofrece la frontera óptima de separación entre los dos tipos de Tequila analizados.

Otro método supervisado de clasificación binaria es Support Vector Machines (SVM), descrito en la sección 2.2. SVM, a diferencia de LDA, no basa el análisis de decisión de fronteras en ninguna distribución de las muestras a clasificar. Básicamente busca un hiperplano de separación óptimo basado en la maximización de los márgenes que delimitan las dos clases en cuestión, tomando como soporte, o base de estos márgenes, a las muestras más cercanas a dicho hiperplano. El análisis es estadístico, y requiere una fase de entrenamiento para generar el hiperplano y seleccionar muestras llamadas vectores de soporte y una fase de clasificación. Es evidente que entre más extenso sea el conjunto de datos disponibles para el análisis los resultados serán mucho mejores. Utilizamos aleatoriamente el 50% de la información para entrenar el modelo y el otro 50% de la información para clasificar.

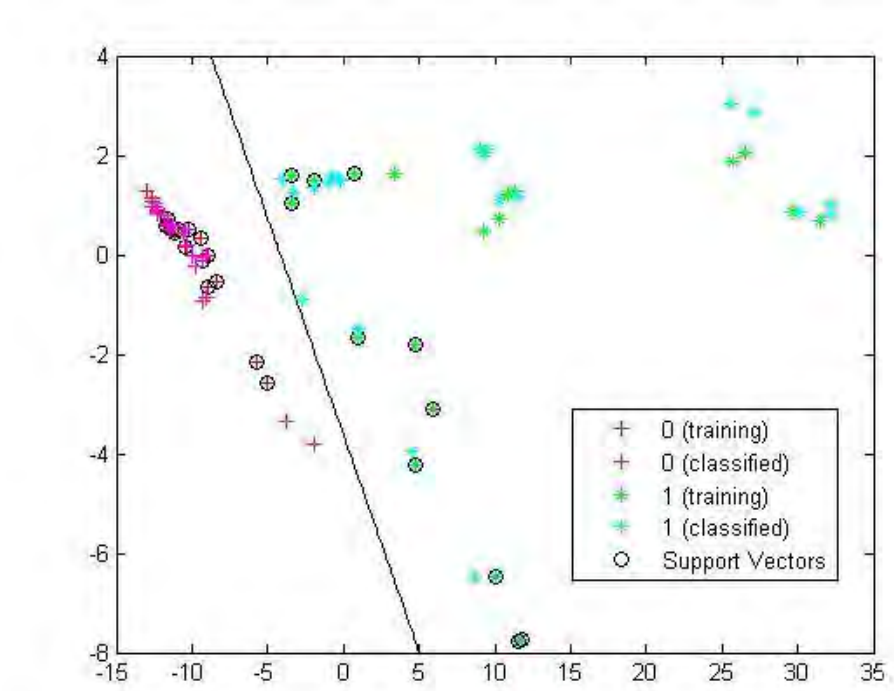


Figura 3.12. Grafica de PC1 vs PC2 generada a partir de un análisis de componentes principales para 80 muestras de Tequila, 39 blancos y 41 reposados. Se etiquetan las muestras de acuerdo a su tipo y clasifican con SVM. Se selecciona aleatoriamente el 50% de la información empleada

para el entrenamiento y selección de vectores de soporte; el 50% restante valida dicha clasificación. Las muestras representadas por círculos corresponden a los vectores de soporte.

Para el caso específico que nos concierne, el análisis de SVM construye una frontera de decisión óptima ya que clasifica correctamente el total la información entre los dos tipos de Tequila en cuestión. Se obtiene un modelo de 24 vectores de soporte.

El objetivo principal de este trabajo es identificar las marcas que se analizan y obtener una buena agrupación y clasificación de las mismas. Sin embargo, se realizó un análisis previo para discriminar entre los dos tipos de Tequila que se analizaron. Se ha reducido la dimensionalidad del problema con ayuda de PCA y se ha clasificado el 100% de la información en su respectiva clase utilizando SVM. De esta manera, basando el análisis en pruebas estadísticas, podemos conocer el tipo (blanco o reposado) de Tequila que se analiza y podemos predecir nuevas muestras correspondientes a Tequilas blancos o reposados de manera objetiva utilizando el método presentado en este trabajo. Los análisis posteriores se llevaran a cabo sobre muestras de un solo tipo de Tequila con el fin de identificar las marcas que se analizan. Los cálculos de PCA y SVM se realizaron en software comerciales The Unscrambler® y Mathlab ® respectivamente.

Tequilas Blancos

Al igual que en la sección anterior, se utiliza PCA para reducir la dimensionalidad del problema y se intentará clasificar por marca a las muestras de Tequilas blancos. El problema se torna un poco diferente. En la sección anterior se deseaba discriminar entre las clases o tipos de Tequila blancos y reposados y se utilizó una base de datos con aproximadamente un 50% de

información correspondiente a cada clase. El objetivo en esta sección es encontrar las diferencias y similitudes que se perciben en las muestras de Tequila blanco por lo que solo se utilizarán Tequilas de este tipo con la finalidad de realizar una posible clasificación.

El modelo predicho por PCA en el caso de los Tequilas blancos sugiere dos componentes principales, el primero contiene el 98% de la información y el segundo componente solo el 1%. Así, con estos dos componentes principales conservamos el 99% de la varianza total. La figura 3.13 muestra el plano PC1-PC2 del modelo generado por las 39 muestras de Tequilas blancos.

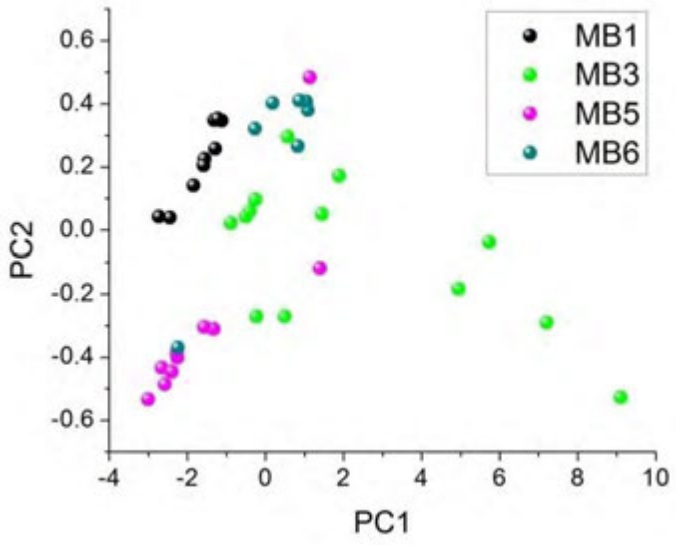


Figura 3.13. Gráfica de PC1 vs PC2 generada por el modelo de PCA para 39 muestras de Tequila blanco, los colores de las esferas indican las marcas de acuerdo a la etiqueta de la figura, todas las muestras están proyectadas en el espacio PC1-PC2.

De acuerdo al análisis de PCA, la figura 3.13 muestra una agrupación de marcas. En el caso de la marca MB1 se observa una baja dispersión de sus muestras comparada con las otras marcas y todas ellas se proyectan en una región compacta en el plano PC1-PC2. Para el caso de otras marcas, existen muestras que no se comportan como el resto de sus correspondientes como es el

caso de MB5, donde la mayoría de las muestras se agrupan en la parte inferior izquierda de la grafica y dos de ellas se alejan hacia la parte central de dicha grafica. A estas muestras que presentan un comportamiento inconsistente a las demás muestras se les conoce en estadística como outliers y su presencia se puede asociar a varios factores, que van desde la recolección de la muestra, errores en la medición o incluso una muestra alterada En el caso de Tequilas blancos, podemos considerar un total de 3 outilers. Dos para la marca MB5 y uno para la marca MB6. Sin embargo, aunque la marca MB3 contiene la dispersión más grande de las cuatro muestras analizadas, evitamos considerar outliers en MB3 debido a que la distribución en general para esta marca presenta dispersión muy alta para todas sus muestras.

El paso natural a seguir, una vez reducida la dimensionalidad del problema, es clasificar las marcas de Tequila blanco. Cuando se clasificaron los tipos de Tequila en la sección anterior, solo se tenían dos diferentes clases y se utilizaron métodos binarios (LDA y SVM) para tal tarea. En el caso de n-clases se emplea una extrapolación del mismo método y se deben encontrar n – 1 fronteras. Para las cuatro marcas de Tequila blanco, en teoría, debemos encontrar tres fronteras para delimitar cada marca de las demás. Sin embargo, el problema no solo difiere en el número de clases, sino también en el conjunto de muestras para cada marca. Esto se convierte en un inconveniente ya que las técnicas estadísticas requieren de un conjunto amplio de datos para generar un modelo confiable y en el caso de clasificar Tequilas blancos por marca se cuenta con un número realmente bajo de muestras para cada marca (en el caso de MB6 solo se tienen 7 ejemplares). Sin embargo, a pesar de que el número de muestras es pequeño podemos usar aproximaciones que nos permitan agrupar o delimitar regiones para agrupar las marcas de Tequila. Aunque los datos reales nunca presentan una distribución normal multivariable exacta, podemos aproximarlos a una densidad

normal. En el caso p-dimensional la densidad normal de distribución para un vector $\boldsymbol{X} = \left[x_1, x_2 x_3 \ldots x_p\right]$ tiene la forma:

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-(\boldsymbol{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})/2}$$

La expresión anterior para una densidad normal variable en p-dimensiones representa elipsoides de densidad que encierran pesos constantes de información para ciertos valores de $\boldsymbol{x}$. Los ejes de las elipsoides están en dirección de los eigenvectores de la matriz de covarianzas $\boldsymbol{\Sigma}^{-1}$ y sus longitudes son proporcionales a los valores cuadrados recíprocos de los eigenvalores de la misma matriz $\boldsymbol{\Sigma}^{-1}$. Así pues, podemos describir contornos de forma elipsoidal para distribuciones normales p-dimensionales de acuerdo a la siguiente expresión:

$$(\boldsymbol{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) = c^2$$

El valor de la constante determinará el porcentaje de la información que se desea delimitar [19]. En nuestro caso específico, generamos elipses de confiabilidad que encierran el 95% de la información de cada marca excluyendo además los tres outiers mencionados. La figura 3.14 muestra las elipses de confiabilidad para MB1, MB3, MB5 y MB6.

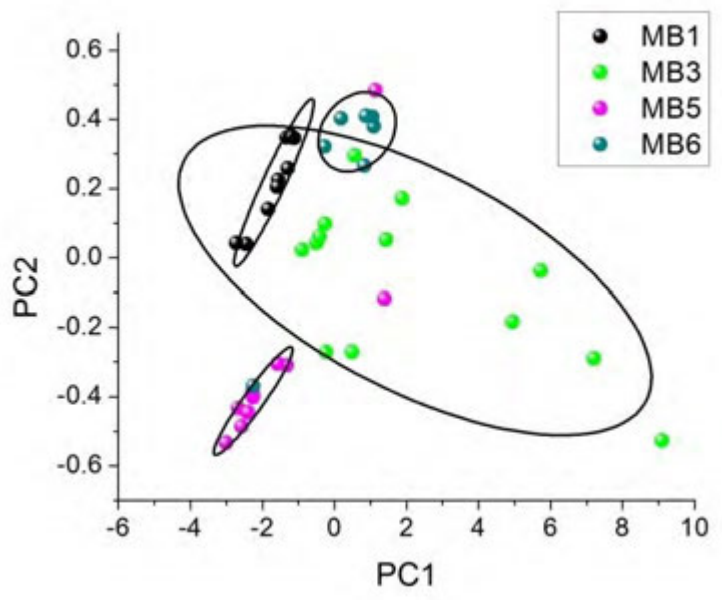


Figura 3.14. Grafica de PC1 vs PC2 para 39 muestras de Tequila blanco de 4 marcas diferentes. Cada marca está representada por un color diferente como lo indica la etiqueta. Las elipses de confiabilidad se generaron con un 95% de probabilidad de encontrar una muestra dentro de la elipse correspondiente de acuerdo a su marca.

A pesar de que las elipses de confiabilidad no representan técnicas de clasificación, son de gran ayuda para delimitar fronteras y agrupar las muestras que presentan alguna correlación. La figura 3.14 representa un mapeo al plano PC1-PC2 de los espectros de absorción de 39 muestras de Tequila blanco correspondientes a cuatro marcas distintas, dichas marcas se agrupan en regiones definidas y se utilizan las elipses para delimitar estas zonas. Se observa en la gráfica que las marcas (excluyendo los outliers mencionados anteriormente) MB1, MB5 y MB6 se agrupan en regiones bien definidas, mientras que la marca MB3 está proyectada en un área más extensa en comparación con las otras marcas. Relacionando los espectros de absorción en función de la longitud de onda (Figuras 3.3 y 3.5) con las componentes principales se observa que la marca MB3 es la que presenta una propagación más extensa de sus muestras en el

espacio PC1-PC2, en contraste con MB1, que contiene una distribución más compacta y ningún outlier. Comparando éstas distribuciones mencionadas para las marcas MB1 y MB3 con su comportamiento respectivo en el espacio OD-longitud de onda observamos que existe una correlación y ambos espacios conservan esta variación entre muestras de una misma marca. Esto es, si una marca presenta muestras muy distintas en su amplitud de absorbancia, las muestras se mapean al plano PC1-PC2 de manera muy dispersa. En cambio, si la absorción para muestras de una marca es similar, estas muestras se proyectan en el plano PC1-PC2 en una región muy definida.

Tequilas Reposados

El análisis de Tequilas reposados se llevó a cabo de una manera similar al de los Tequilas blancos. Se analizan un total de 41 muestras de Tequila provenientes de cuatro marcas diferentes como lo indica la tabla 3.1. Los resultados obtenidos del análisis de PCA para Tequilas reposados predicen un modelo de 2 componentes principales PC1 y PC2 con una varianza de 93.2% y 6.2% respectivamente para la base de datos de los Tequilas reposados. PC1 y PC2 constituyen el 99.4% de la varianza total. La figura 3.15 muestra la grafica de los dos componentes principales PC1-PC2 y la proyección de las 41 muestras en dicho plano.

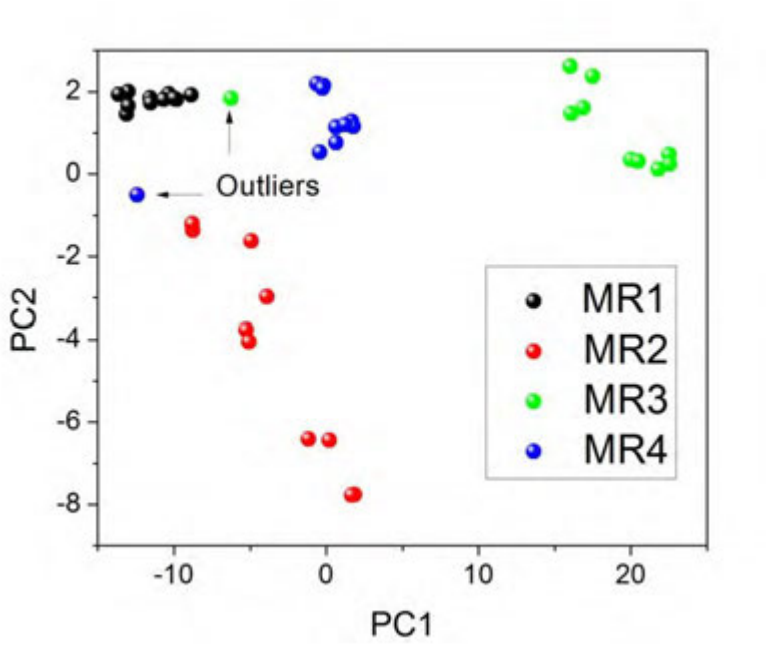


Figura 3.15. Gráfica de PC1 vs PC2 generada por el modelo de PCA para 41 muestras de Tequila reposado, los colores de las esferas indican las marcas de acuerdo a la etiqueta de la figura, todas las muestras están proyectadas en el espacio PC1-PC2.

De acuerdo a la proyección de las muestras de Tequila (figura 3.15) en el plano PC1-PC2, las 41 muestras de Tequila se agrupan en regiones definidas para cada marca. Excluyendo los outliers señalados de la misma gráfica podemos considerar que cada marca está definida por un área específica en este plano. Se analizan tres marcas de Tequila reposado 100% agave y una marca de Tequila mixto, sin embargo, el modelo parece no reconocer entre estas categorías de Tequila y solo agrupa por marca a las muestras analizadas. MR4 representa una marca de Tequila mixto. A partir de la proyección mostrada en la figura 3.15 generamos las elipses de confiabilidad para cada marca sin considerar dentro del análisis a los outliers mencionados (figura 3.16).

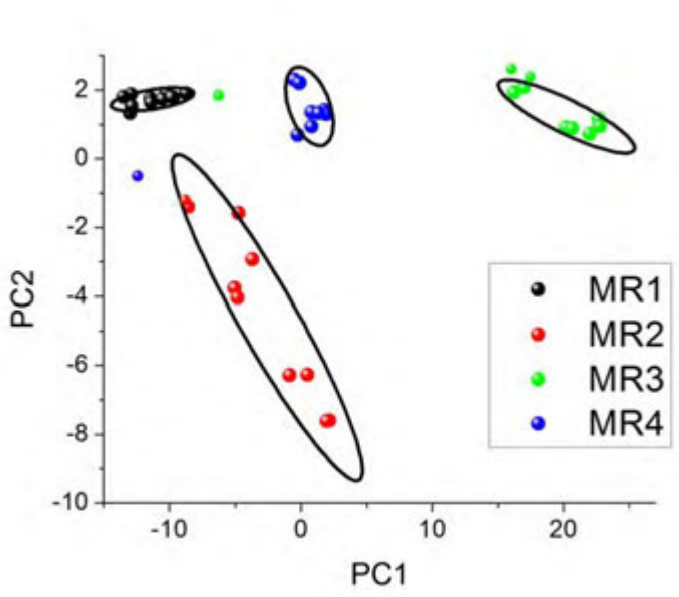


Figura 3.16. Grafica de PC1 vs PC2 para 41 muestras de Tequila reposado de 4 marcas diferentes. Cada marca está representada por un color diferente como lo indica la etiqueta. Las elipses de confiabilidad se generaron con un 95% de probabilidad de encontrar una muestra dentro de la elipse correspondiente de acuerdo a su marca.

Las muestras se distribuyen en el plano PC1-PC2 de manera distinta para cada marca, y como sucede en los Tequilas blancos, MR1 representa la marca con una agrupación más definida en un área pequeña de dicho plano. En contraste a MR1, MR2 genera una elipse de confiabilidad de mayor área que podría asociarse con la variación de la amplitud de OD de sus muestras correspondientes. La variación en amplitud de OD de las muestras provenientes de distintos lotes para cada marca (figuras 3.6 y 3.7) parece conservarse en el plano de componentes principales para ambos tipos de Tequilas. El análisis presentado arroja resultados satisfactorios y agrupa de manera adecuada las marcas que se estudiaron a través de técnicas espectroscópicas y modelos estadísticos no supervisados.

3.3. Validación

Una manera de validar los modelos generados para agrupar las marcas en ambos tipos de Tequilas es analizar nuevas muestras de Tequila de las marcas en cuestión y esperar que se agrupen o queden delimitadas dentro de las elipses de confiabilidad. Lamentablemente para este trabajo no se cuenta con un número mayor de muestras para MB1, MB3, MB5, MB6 en el caso de Tequilas blancos ni muestras de las marcas MR1, MR2, MR3, MR4 en el caso de Tequilas reposados. Una manera alternativa de validar el modelo es utilizar muestras de Tequila que no correspondan a dichas marcas esperando que se proyecten fuera de las elipses de confiabilidad en el plano PC1-PC2. Afortunadamente se cuenta con una base de datos de 178 bebidas alcohólicas que incluyen 33 mezcales, 82 Tequilas blancos, 54 reposados y 9 añejos. La base de datos de las bebidas alcohólicas contiene muestras de Tequila de diversas marcas, 33 de ellas corresponde a muestras usadas para generar los modelos de Tequilas blancos y no se utilizan para validar, así, se cuenta con un total de 145 muestras de validación.

Analizamos la validación del modelo PCA-4marcas de Tequilas blancos. La figura 3.17 muestra el modelo generado que corresponde a la figura 3.16 y la base de datos de 145 bebidas mapeadas al plano PC1-PC2 del mismo modelo.

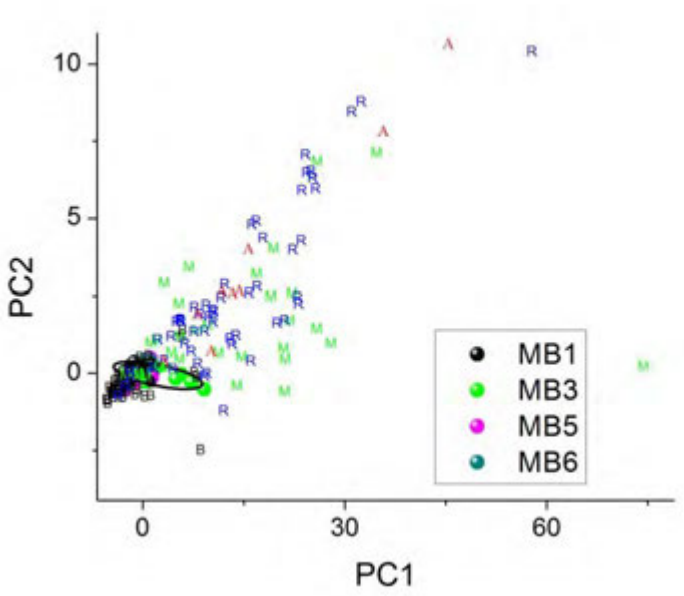


Figura 3.17. Validación del modelo PCA-4 marcas de Tequilas Blancos. Se utiliza un total de 145 bebidas alcohólicas, entre mezcales, Los Tequilas blancos, reposados y añejos se representan por M, B, R y A respectivamente.

La figura 3.17 muestra la distribución general de las muestras correspondientes a 145 bebidas alcohólicas. Los Tequilas blancos empleados para validar este modelo se encuentran distribuidos en una región muy cercana a las elipses y presentan poca dispersión en comparación con las otras bebidas utilizadas en la misma validación. Se simbolizan como M, B, R y A a los mezcales, Tequilas blancos, reposados y añejos respectivamente en todas las graficas de validación. Para hacer un análisis más concreto sobre las marcas de interés se muestra una amplificación sobre la zona que contiene las cuatro elipses de confiabilidad.

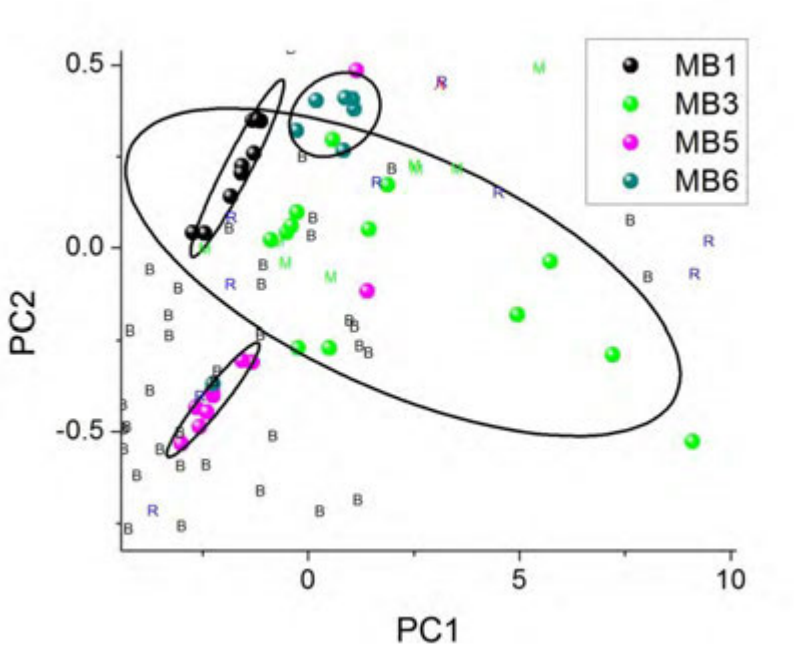


Figura 3.18. Validación del modelo PCA-Tequilas blancos por marca. La gráfica muestra el intervalo correspondiente a las elipses de confiabilidad de las cuatro marcas.

La figura 3.18 muestra una distribución cercana a las cuatro elipses de confiabilidad. Para el caso de las marcas MB1 y MB6, no se observan muestras de validación dentro de sus elipses correspondientes. Para MB5 aparecen tres muestras de validación muy cercanas a la frontera, mientras que para la elipse correspondiente a MB3 delimita una región ocupada por 19 muestras de validación.

Existen medidas basadas en el análisis estadístico que determinan el desempeño o rendimiento de los modelos de predicción y clasificación binarios. La sensibilidad mide la proporción de muestras positivas que son identificadas como tales.

$$Sensitivity = \frac{number\ of\ true\ positives}{number\ of\ true\ positives + number\ of\ false\ negatives}$$

La especificidad mide la proporción de negativos que son identificados como tales.

$$Specificity = \frac{number\ of\ true\ positives}{number\ of\ true\ positives + number\ of\ false\ negatives}$$

Debido al déficit de muestras positivas (muestras de validación correspondientes a las marcas usadas para generar el modelo) solo se calcula la especificidad para ambos modelos. Para el modelo de Tequilas blancos le corresponde una especificidad de 0.8482. Esto significa que reconoce aproximadamente el 85% de bebidas alcohólicas como no pertenecientes a las cuatro marcas analizadas.

En la validación del modelo PCA-4marcas de Tequilas reposados utilizamos un total de 164 muestras de bebidas alcohólicas correspondientes a 33 mezcales, 82 Tequilas blancos, 40 Tequilas reposados y 9 Tequilas añejos. La figura 3.19 muestra la distribución general de las muestras de validación proyectada sobre el modelo de PCA para Tequilas reposados.

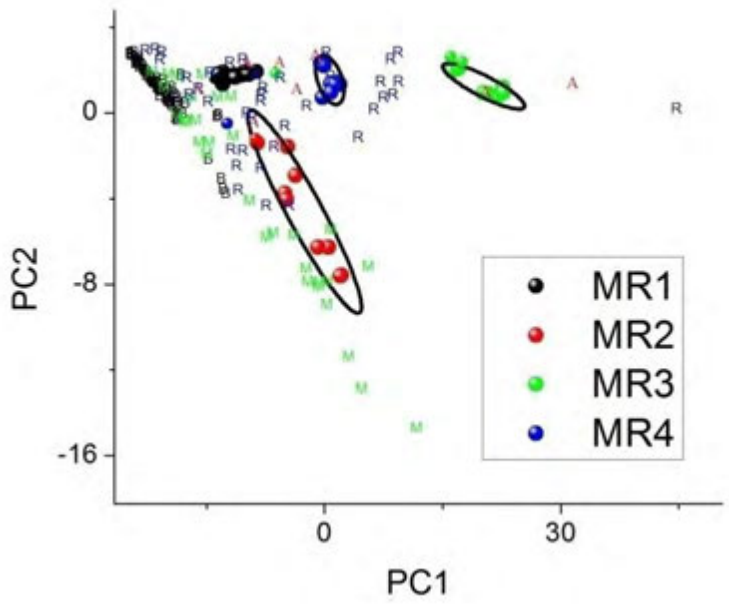


Figura 3.19. Validación del modelo PCA-4 marcas de Tequilas reposados. Se utiliza un total de 163 bebidas alcohólicas, entre mezcales y Tequilas blancos, reposados y añejos.

Las muestras correspondientes a cada bebida alcohólica preservan un comportamiento bien definido y se agrupan en regiones específicas del plano PC1-PC2. La figura 3.20 muestra una amplificación del Plano PC1-PC2 que contiene a las elipses de confiabilidad para las marcas de Tequila reposado MR1-4.

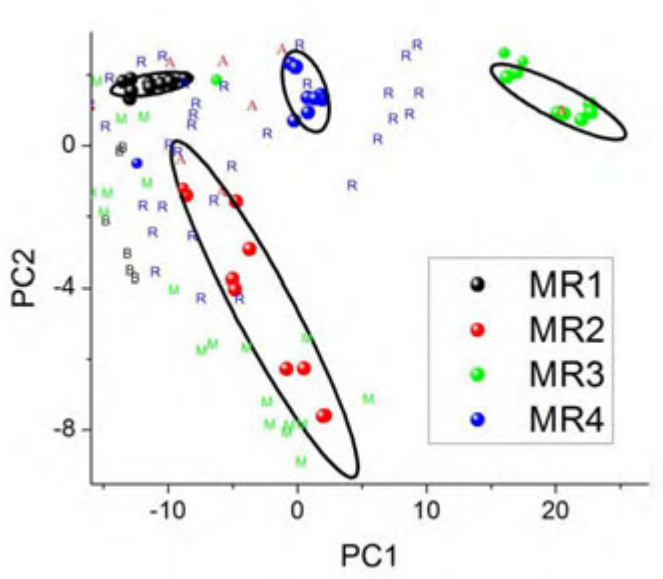


Figura 3.20. Validación del modelo PCA-4 marcas de Tequilas reposados. Se utiliza un total de 164 bebidas alcohólicas, entre mezcales y Tequilas blancos, reposados y añejos. Se grafica la sección correspondiente a las elipses de confiabilidad de las cuatro marcas.

Para el modelo de Tequilas reposados, solo nueve muestras de validación caen dentro de las elipses de confiabilidad, lo que produce un factor de especificidad de 0.9447. En otras palabras, del 100% de muestras usadas para validar el modelo, aproximadamente el 95% de estas muestras se clasifican correctamente como muestras no pertenecientes a las marcas analizadas.

En base a los resultados presentados en este trabajo, se ha generado un modelo basado en espectroscopia uv-visible y análisis multivariante capaz de predecir si una muestra de Tequila es de tipo blanco o reposado y se puede predecir si corresponde o no a una de las marcas utilizadas para generar el modelo.

4. Conclusiones

Se analizaron un total de 80 muestras de Tequila de los tipos blanco y reposado pertenecientes a ciertas marcas de prestigio y tradición a nivel nacional. Se utilizaron técnicas espectroscópicas en la región UV-Visible del espectro electromagnético para obtener información de dichas muestras. Los espectros de absorción presentan características únicas que dependen de las marcas de Tequila en cuestión y por medio del análisis multivariado se agruparon las muestras en sus marcas correspondientes. Se clasificó un total de 80 muestras en su respectivo tipo de Tequila 100% agave y mixto. Se ha generado un modelo capaz de predecir si un Tequila pertenece a los de tipo blanco o reposado y además si pertenece o no a una de las marcas que se analizan con un alto grado de confiabilidad. Consideramos un total de 5 de 80 muestras como outliers y las 75 restantes presentan un agrupamiento bien definido de acuerdo a la marca a la que pertenecen. El modelo obtenido en este trabajo representa una técnica potencial alternativa para la identificación y clasificación de tequilas y puede extrapolarse a una cantidad mayor de marcas e incluso a otras bebidas alcohólicas. Para mejores resultados en la identificación y agrupación de marcas es necesaria una base de datos más amplia debido a que el análisis se basa en técnicas estadísticas. La metodología empleada no requiere de alguna preparación de las muestras ni exige costes elevados ni gran demanda de tiempo para su análisis comparada con técnicas actualmente usadas para la identificación y caracterización de estas bebidas. Además de las ventajas ya señaladas, no existe la necesidad de realizar las mediciones en laboratorios especializados y una vez generado un modelo confiable de predicción la técnica puede utilizarse en

cualquier lugar y obtener resultados en cuestión de segundos si se cuenta con el equipo adecuado.

5. Trabajo a futuro

Los modelos generados para la predicción y clasificación de marcas de Tequila están basados sobre resultados estadísticos de muestras analizadas. Una mejor predicción y clasificación se puede realizar si la base de datos es más amplia. Se propone como trabajo a futuro ampliar la base de datos que corresponde a las muestras analizadas para generar un modelo de mayor confianza y extenderlo a un número mayor de marcas de Tequila. También se propone estudiar con estas técnicas el comportamiento de los Tequilas añejos para su posible clasificación y agrupación. Además de generalizar este método a los tres tipos de Tequila (blanco, reposado y añejo) y generar un modelo de mayor confianza se propone explorar el análisis espectroscópico en otras regiones como el infrarrojo e incluso indagar sobre otras características peculiares de estas bebidas con técnicas diferentes como espectroscopia Raman, todo esto con el objetivo final ofrecer una técnica alternativa capaz de autentificar estas bebidas alcohólicas.

6. Bibliografía

[1] Norma Oficial Mexicana NOM-006-SCFI-2005, Bebidas-Alcohólicas-Tequila-Especificaciones.

[2] Cedeño Cruz M. Tequila production. Crit. Rev. Biotechnol., núm. 15, pp. 1 -11.

[3]Aguilar-Cisneros B. O., M. G. Lopez, W. Freank, P. Shreier, J. Agric. Fod Chem. 50(2002) 7520.

[4] Bautista-Justo M., L. Garcia-Oropeza, J. E. Barbosa-Corona, L. A. Parra-Negrete. "El agave Tequilana Weber y la producción de Tequila". Acta Universitaria, agosto, año/vol. 11, numero 002. Universidad de Guanajuato. Pp 26-34. [14].

[5] Owen, T. "Fundamentals of modern UV-Visible spectroscopy". Copyright Agilent Technologies 2000.

[6] Schmidt Werner. "Optical Spectroscopy in Chemestry and Life Sciences" WILLEY-VCH Verlag GmbH & Co. KGaA 2005.

[7] Brereton G. Richard. "Chemometrics. Data Analysis for the Laboratory and Chemical Plant". John Wiley & Sons LTD England, 2003.

[8] Otto, Mathias. "Chemometrics. Statistics and Computer Application in Analytical Chemestry". WILLEY-VCH, 2007.

[9] Jolliffe I. T., "Principal Compoent Analysis". Second Edition. Springer Series in Statistics.

[10]Johnson Richard A., Applied Multivariate Statitistical Analysis. Pretice Hall, Third Edition, 1992.

[11]Smith Lindsay, “A tutorial on Principal Components”, 2002 (no es publicación ni libro).

[12] Abe, Shigeo. “Suport Vector Machines for Pattern Classification”. Springer-Verlag London Limited 2005.

[13]Cristiannini Nello, Shawe-Taylor John. “Support Vector Machines and another kernel-based learning methods” Cambridge University Press 2000.

[14]Benn S. M., T.L. Peppard, J. Agric. Food Chem. 44 (1996) 557.

[15] Vallejo-Cordoba B., A.F. González-Córdova, M.C. Estrada-Montoya, J. Agric. Food Chem. 52 (2004) 5567.

[16] Savchuk S.A., V.N. Vlasov, S. A. Appolonova, V.N. Arbuzov, A.N. Venedin, A. B. Mesinov, B. R. Grigor’yan, J. Anal. Chem. 56 (2001) 214.

[17] Ana Celia Muñoz-Muñoz, O. Barbosa-García, G. Ramos-Ortiz, J.L. Maldonado, J.L. Pichardo-Molina, M.A. Meneses-Nava, Pedro Luis López de Alba. “UV-vis spectroscopy and multivariate calibration (PLS) as a tool for identification and clasification of tequilas”. *Enviado a pubilcación.*

[18] Muñoz R. David, Wrobel K. Wrobel K., Determination of aldehydes in tequila by high-performance liquid chromatography with 2,4-dinitrophenylhydrazine derivatization. European Food Research and Technology, Vol. 221, No. 6, Springer – Verlag, Noviembre 2005.

[19] Jobson J. D., Applied Multivariate Data Analysis, Regresion and Experimental Design, vol 1, Springer-Verlag, New York, 1991.