
**UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO**
**FACULTAD DE ESTUDIOS SUPERIORES CUAUTITLÁN**

# *Construcción y Aplicación de Modelos Quimiométricos Supervisados y No Supervisados para la Autentificación de Tequilas de acuerdo con su Clase a partir de Datos de Infrarrojo Medio*

**T E S I S**

QUE PARA OBTENER EL TÍTULO PROFESIONAL DE:
LICENCIADA EN QUÍMICA

PRESENTA:
ILEANA JIMÉNEZ RABADÁN

ASESORA:
DRA. MARÍA GUADALUPE PÉREZ CABALLERO

CUAUTITLÁN IZCALLI, ESTADO DE MÉXICO, 2019

VNIVERSIDAD NACIONAL AVPNOMA DE MEXICO

# FACULTAD DE ESTUDIOS SUPERIORES CUAUTITLÁN
## SECRETARÍA GENERAL
## DEPARTAMENTO DE EXÁMENES PROFESIONALES

FACULTAD DE ESTUDIOS SUPERIORES CUAUTITLÁ

**ASUNTO: VOTO APROBATORIO**

**M. en C. JORGE ALFREDO CUÉLLAR ORDAZ**
**DIRECTOR DE LA FES CUAUTITLAN**
**PRESENTE**

**ATN: I.A. LAURA MARGARITA CORTAZAR FIGUEROA**
**Jefa del Departamento de Exámenes Profesionales**
**de la FES Cuautitlán.**

Con base en el Reglamento General de Exámenes, y la Dirección de la Facultad, nos permitimos comunicar a usted que revisamos el: **Trabajo de Tesis**

**Construcción y Aplicación de Modelos Quimiométricos Supervisados y No Supervisados para la Autentificación de Tequilas de acuerdo con su clase a partir de Datos de Infrarrojo Medio.**

Que presenta la pasante: **Ileana Jiménez Rabadán**
Con número de cuenta: **309094506** para obtener el Título de la carrera: **Licenciatura en Química**

Considerando que dicho trabajo reúne los requisitos necesarios para ser discutido en el **EXAMEN PROFESIONAL** correspondiente, otorgamos nuestro **VOTO APROBATORIO**.

**ATENTAMENTE**
**"POR MI RAZA HABLARÁ EL ESPÍRITU"**
Cuautitlán Izcalli, Méx. a 10 de diciembre de 2018.

**PROFESORES QUE INTEGRAN EL JURADO**

| | NOMBRE | FIRMA |
|---|---|---|
| **PRESIDENTE** | Dr. René Miranda Ruvalcaba | |
| **VOCAL** | Dra. María Guadalupe Pérez Caballero | |
| **SECRETARIO** | Dra. María Inés Nicolás Vázquez | |
| **1er. SUPLENTE** | Dra. María Gabriela Vargas Martínez | |
| **2do. SUPLENTE** | Dra. Alma Luisa Revilla Vázquez | |

NOTA: los sinodales suplentes están obligados a presentarse el día y hora del Examen Profesional (art. 127).

LMCF/javg

## Agradecimientos

*A la vida, por ofrecerme la oportunidad de adquirir nuevos conocimientos, de conocer a cada persona con quien he cruzado mi camino y de disfrutar cada momento que ha dejado huella en mí.*

*A la Facultad de Estudios Superiores Cuautitlán, UNAM, por brindarme las instalaciones, los recursos y los profesores que contribuyeron a mi formación como Licenciada en Química.*

*A mi asesora, la Dra. Guadalupe Pérez Caballero y al Dr. José Manuel Andrade Garda, por guiarme en la escritura y revisión del presente trabajo.*

*A mis sinodales, por el tiempo que dedicaron a revisar mi escrito y hacerme notar las mejoras que éste podría tener.*

*A mis profesores de Química de la secundaria y preparatoria, quienes me mostraron lo maravillosa que es la Química.*

*A mis compañeros del laboratorio L10 de la UIM, que hicieron agradable el desarrollo de mi servicio social y del presente escrito.*

*¡¡¡ G R A C I A S !!!*

## Dedicatorias

*A mis padres, Martha y Héctor, que día con día me inspiran a ser una mejor persona, que son mis guías y mi ejemplo que seguir, quienes con todo su cariño y dedicación hicieron de mí la profesionista que comienzo a ser, por motivarme a soñar,*
*a caerme, a levantarme, a aprender y a nunca conformarme.*

*A mi hermana, Naomi, por ser la mejor compañera de este viaje llamado vida, porque, aunque seamos distintas en algunos aspectos, sabemos que estamos incondicionalmente la una para la otra.*

*A mi asesora, la Dra. Lupita, por el tiempo, el apoyo, la guía y la paciencia durante el desarrollo de esta investigación.*

*A cada persona con quien compartí momentos inolvidables durante la carrera, quienes dejaron algún aprendizaje en mí.*

*“Me parece haber sido sólo un niño jugando en la orilla del mar, divirtiéndose y buscando una piedra más lisa o una concha más bonita de lo normal, mientras el gran océano de la verdad yacía ante mis ojos con todo por descubrir”*

Isaac Newton

# Índice

## Abreviaturas y Acrónimos

| | |
|---|---|
| 5-HMF | 5-(Hidroximetil)furfural |
| A | Añejo (s) |
| Asig. | Asignación |
| B | Blanco (s) |
| CA | *Cluster Analysis*: Análisis Cluster |
| CATR | Corrección de ATR (Reflexión Total Atenuada) |
| CILB | Corrección Interactiva de la Línea Base |
| CRT | Consejo Regulador del Tequila |
| CV# | *Canonic Variable*: Variable Canónica 1, 2 o 3 |
| Cv | *Cross Validation*: Validación cruzada |
| Dist. | Distancia |
| EtOH | Etanol |
| FT–MIR | Infrarrojo Medio con Transformadas de Fourier |
| J | Joven (es) |
| k–NN | *k-Nearest Neighbor*: Vecino más Cercano |
| LV | *Latent Variable*: Variable latente |
| Met. | Método |
| NIPALS | *Nonlinear estimation by Iterative Partial Least Squares* |
| PA | Producto alcohólico (formulación química del tequila) |
| PC# o Comp | *Principal Component*: Componente Principal 1, 2 o 3 |
| PC | *Potential Curves*: Curvas de Potencia |
| PCA | *Principal Component Analysis*: Análisis de Componentes Principales |
| PCA–DA | *Principal Component–Discriminant Analysis*: Análisis Discriminante por Componentes Principales |
| PF | *Potential Functions*: Funciones de Potencia |
| PLS–DA | *Partial Least Squares–Discriminant Analysis*: Análisis Discriminante por el Método de Mínimos Cuadrados Parciales |
| R | Reposado (s) |
| RV | Rotación Varimax |
| SIMCA | *Soft Independent Modelling of Class Analogy*: Modelado Suave e Independiente por Analogía de Clase |

| | |
|---|---|
| SVD | *Single Value Decomposition*: Descomposición en Valores Singulares |
| SVM | *Support Vector Machines* |

Teq. o TTequila (s)

# Glosario

| | |
|---|---|
| ATR | Por sus siglas en inglés: *Attenuated Total Reflection* (Reflexión Total Atenuada), se refiere al accesorio utilizado para llevar a cabo la medición de un haz de luz infrarroja atenuado después de ser incidido sobre una muestra a través de un cristal ópticamente denso. Cuando se dice que se hace corrección de ATR, se refiere a la conversión de datos de reflexión a datos de transmitancia. |
| Autovalores (*eigenvalues)* | Se les denomina autovalores a las varianzas explicadas de un vector que modela una dispersión de puntos en una técnica quimiométrica de clasificación. |
| Autovectores (*eigenvectors)* | Se denominan autovectores a aquellos valores numéricos resultantes de un escalado (centrado en la media o autoescalado) que son proyectados en un modelo quimiométrico. También pueden llamarse factores que, en el caso de PCA son componentes principales y en PLS son variables latentes. |
| Categoría (Descrito con mayor profundidad en la página 10) | Término utilizado en la industria tequilera, referido al porcentaje de azúcares obtenidas del *Agave tequilana Weber* var. azul, empleado en la producción del tequila. Puede ser 100% agave o mixto (hasta 49% de azúcares de otra fuente). |
| Clase (Descrito con mayor profundidad en la página 9) | Término utilizado en la industria tequilera para describir el tiempo de añejamiento (o reposo en barrica) que tuvo un tequila. De acuerdo con la NOM-006- SCFI-2012, se definen cinco clases: Blanco o plata, Joven u oro, Reposado, Añejo y Extra Añejo. |
| Conjunto de calibración | En una técnica quimiométrica no supervisada, es el grupo de muestras empleadas para la construcción de un modelo de clasificación, mientras que, en una técnica supervisada, hace referencia a una parte del entrenamiento (*training*) del algoritmo en cuestión (la contraparte es la validación interna). |

| | |
|---|---|
| Conjunto de validación | En una técnica quimiométrica no supervisada, es el grupo de muestras que se proyectan en un modelo construido con el conjunto de calibración (y diferentes a este), para conocer el error que podría generarse durante su aplicación. Por otro lado, en una técnica supervisada, esta etapa es llamada de validación interna y se realiza con el mismo conjunto de calibración (*cross validation*). |
| Conjunto de predicción | En una técnica supervisada, es el grupo de muestras introducidas como prueba (*test*) en un modelo de clasificación, con el objetivo de aplicarlo para concretar si pertenecen o no a una clase dada. |
| Correlación | Término empleado para definir las relaciones interdependencia de un grupo de datos con respecto a otro en función de sus variables. Esta mide la fuerza y la dirección de estas relaciones. |
| Covarianza | Es aquella medición que revela la relación que existe entre dos variables, que puede adquirir valores numéricos de - ∞ hasta + ∞.<br>Se emplea para comprender la dirección de la relación entre las variables. |
| Distancia | Es la trayectoria que se sigue para llegar de un elemento a otro, por lo que señala cuán lejos se ubican estos. También se denomina disimilitud y se expresa mediante un número real. |
| Espectro bruto | Aquel espectro IR resultante de la corrección de línea base. |
| Hiperplano o hipersuperficie | Término usado en *Support Vector Machines* (SVM) referido a las líneas o fronteras que dividen las clases que conforman un modelo. |
| Margen | Término empleado en SVM que indica la distancia del hiperplano a una muestra determinada. |
| *Outlier* | Aquella muestra que no cumple las características usuales |

| | |
|---|---|
| | de un tequila de acuerdo con su clase. También llamada anómala. Este término puede ser aplicado en cualquier analito. |
| Penalización del error (C) | Concepto utilizado en SVM que define la tolerancia al error que tiene el modelo. Representado por una letra "C". |
| Repetibilidad | Precisión de un método analítico que evalúa la concordancia obtenida entre determinaciones independientes realizadas por un solo analista, usando los mismos instrumentos y métodos. |
| Reproducibilidad | Precisión de un método analítico que evalúa la concordancia entre determinaciones independientes realizadas por diferentes laboratorios, analistas o instrumentos. |
| Robustez | Capacidad del método analítico de mantener su desempeño al presentarse variaciones pequeñas pero deliberadas, en los parámetros normales de operación del método. |
| Similitud | Indica la semejanza existente entre una muestra y otra. Puede ser explicada con valores entre 0 y 1. |
| Validación | Conjunto de operaciones a partir de las cuales se indica que el método tiene la capacidad de satisfacer los requisitos necesarios para su aplicación analítica deseada. |
| Validación interna | También llamada "validación cruzada" (*cross validation*), se basa en la formación de subgrupos denominados grupos de cancelación que permiten calcular el modelo dejando fuera a uno de ellos; proceso que se repite n veces y suele llamarse *full-cross-validation* o *leave-one-out*. |

# Índice de figuras, tablas, gráficas y ecuaciones

# 1. Resumen

Para la elaboración de la presente tesis, se desarrollaron nueve modelos adecuados de clasificación de tequilas de acuerdo con su tiempo de añejamiento (clase), empleando técnicas quimiométricas supervisadas y no supervisadas. Las técnicas no supervisadas utilizadas fueron: Análisis de Componentes Principales (PCA) y Análisis de Conglomerados o Análisis Cluster (CA), mientras que, las técnicas supervisadas fueron: Curvas de Potencia (PC), *Support Vector Machines* (SVM), Funciones de Potencia (PF), Vecino más Cercano (k–NN), Análisis Discriminante por Componentes Principales (PCA–DA), Análisis Discriminante por el Método de Mínimos Cuadrados Parciales (PLS–DA) y Modelación Suave e Independiente por Analogías de Clase (SIMCA).

El fundamento de cada técnica se describe en el marco teórico, de tal manera que este trabajo puede constituir un material de estudio para quienes deseen familiarizarse con estas técnicas, todavía no consideradas en los planes de estudio de la carrera.

La parte experimental aborda la obtención de los espectros de infrarrojo medio (4000–450 $cm^{-1}$), su pretratamiento y el uso del software especializado para la construcción de los modelos de clasificación (*GenEx*© y *Matlab*©).

Cabe mencionar que, dado que en el presente estudio no se tuvo como objetivo diferenciar las muestras por categoría, se utilizaron muestras tanto 100 % agave como mixtos, así como contenidos alcohólicos entre 35 y 55 % v/v. Además, la representatividad de los modelos de clasificación aquí construidos se vio favorecida por tratarse de muestras con diferentes procesos de elaboración, distintos lotes e, incluso, diferentes casas tequileras. Tales muestras fueron proporcionadas por el Consejo Regulador del Tequila (CRT).

En la sección de análisis de resultados, se discuten y comparan los aciertos y limitaciones obtenidos en cada modelo mediante el análisis gráfico y sus correspondientes parámetros de desempeño, a partir de lo cual fue posible concluir que los objetivos se cumplieron satisfactoriamente.

Posterior a la construcción de los modelos, se introdujeron dos muestras apócrifas para tener una idea de sus características, aunque no con resultados concluyentes.

La utilización de los modelos requerirá de su actualización mediante un conjunto de muestras de tequilas de clase desconocida, puesto que, en la presente etapa se aborda su desarrollo y validación (procedimientos independientes).

# 2. Introducción

Actualmente, el Consejo Regulador del Tequila (CRT), cuenta con dos metodologías de verificación en la clase de un tequila:

1) De forma física, en la que el verificador de la empresa coloca sellos (fracción de hoja) a las barricas para garantizar que los litros de tequila estén el tiempo de reposo estipulado en la misma hoja. Él lo coloca y lo reporta al CRT, posteriormente vuelve en la fecha estipulada y rompe el sello, por consiguiente, la empresa embotella. En caso de estar violado el sello, el verificador sabrá que la barrica podría no cumplir las características de la clase a la que iba destinada y lo hace saber al CRT.
2) El verificador toma una muestra de la barrica y la lleva al laboratorio para examinarla por relación isotópica de C y O, de tal forma que sea posible establecer la autenticidad de la muestra según los compuestos que generan el aroma y sabor al producto. No obstante, este análisis es útil para determinar únicamente la categoría del tequila y no su clase (finalidad de esta tesis).

Cabe mencionar que la primera técnica no es completamente confiable, puesto que no contempla el factor importante que representa si la barrica es nueva o vieja (otorga mayor/menor cantidad de compuestos aromáticos volátiles característicos del reposo) y, aunque la segunda técnica puede proporcionar esta información, es muy costosa.

La pregunta principal planteada durante el desarrollo de esta investigación fue: ¿Existen métodos eficientes analítica y económicamente para la autentificación de un tequila por su clase?
Además, se respondieron los cuestionamientos: ¿Por qué es importante autentificar y determinar la calidad de un tequila?, ¿Cuáles son los beneficios de utilizar técnicas quimiométricas para su autentificación?, ¿Qué matriz de datos permite obtener los mejores modelos quimiométricos?, ¿Qué es la varianza explicada? ¿Cómo se relaciona con los modelos? ¿Es importante? ¿Y la varianza residual?, sin olvidar: ¿Cuál es la influencia de los grados de alcohol de cada espectro de tequila?

El presente estudio busca establecer nuevas técnicas de análisis químico de tequilas del CRT que permitan determinar su autenticidad de clase mediante la elaboración de modelos matemáticos que no requieran una inversión económica (ni temporal) fuerte pero que, sin embargo, cumplan eficientemente su objetivo.

# 3. Objetivos

## 3.1 Objetivo general

Desarrollar métodos alternativos de análisis de muestras del Consejo Regulador del Tequila, de acuerdo con su clase, que sean confiables, eficaces, rápidos y de bajo costo mediante la aplicación de técnicas quimiométricas supervisadas y no supervisadas, a partir de datos de espectrofotometría de Infrarrojo Medio con transformadas de Fourier con el propósito de ser aplicados por las diferentes instancias involucradas en la comercialización del tequila.

## 3.2 Objetivos particulares

1. Establecer modelos adecuados de clasificación de tequilas del CRT de acuerdo con su tiempo de añejamiento mediante la aplicación del Análisis de Componentes Principales y el Análisis Cluster como técnicas quimiométricas no supervisadas en *GenEx©*.
2. Establecer modelos adecuados de clasificación de tequilas del CRT de acuerdo con su tiempo de añejamiento mediante las técnicas quimiométricas supervisadas: Curvas de Potencia y *Support Vector Machines* en *GenEx©*.
3. Establecer modelos adecuados de clasificación de tequilas del CRT de acuerdo con su tiempo de añejamiento mediante las técnicas quimiométricas supervisadas: Análisis Discriminante por el Método de Mínimos Cuadrados Parciales, Análisis Discriminante por Componentes Principales, Vecino más Cercano, Funciones de Potencia y Modelación Suave e Independiente por Analogía de Clase en *Matlab©*.
4. Proyectar en los modelos de técnicas no supervisadas dos muestras externas que se conocen como no auténticas.
5. Predecir en los modelos de técnicas supervisadas dos muestras externas que se conocen como no auténticas.
6. Ofrecer alternativas eficientes y económicas al Consejo Regulador del Tequila para la clasificación y autentificación de muestras.
7. Coadyuvar a una estrategia de mejora en el control de calidad del tequila.

# 4. Marco teórico

## 4.1 Generalidades del tequila

El tequila es una bebida de origen mexicano producida en las regiones especificadas por su Denominación de Origen, conformado por algunos municipios de los estados Guanajuato, Jalisco, Michoacán, Nayarit y Tamaulipas (*figura 1*), producido a partir de la extracción de azúcares del *Agave Tequilana weber* variedad azul.


***Figura 1.*** **Mapa de la República Mexicana con la región de Denominación de Origen del Tequila (DOT).** Fuente: https://www.crt.org.mx/ recuperado el 28/Agosto/2018

El proceso de elaboración del tequila varía dependiendo de la casa tequilera, incluso un solo producto dentro de una misma casa. Sin embargo, se trata en general de la siembra del *Agave Tequilana weber* var. azul, 7 años mínimos de crecimiento, cosecha, jima (proceso de eliminación de las hojas o pencas del agave), cocción, posteriormente se trituran las piñas y con ayuda de un poco de agua caliente se extraen los carbohidratos o, comúnmente llamados, azúcares (*figura 2*).


***Figura 2*. Diagrama de flujo para la elaboración del Tequila.** (Imagen propia)

En este punto, hay que mencionar que para obtener un *tequila 100 % agave*, en su formulación se incorpora únicamente jugo cocido de agave y agua, mientras que, si se desea obtener tequila mixto, al mosto se le adicionan otros azúcares como la extraída de la caña, jarabe de maíz, piloncillo, glucosa, melaza, entre otros, en una proporción no mayor al 49 %.

Posteriormente, se adicionan levaduras para llevar a cabo la fermentación del mosto, que una vez fermentado, el producto sufre una doble destilación para dar lugar al tequila [22].

En caso de querer un tequila blanco, simplemente se agrega agua de dilución y se embotella; si en cambio se desea un tequila joven, reposado, añejo o extra añejo, es en este punto donde el producto de la doble destilación puede ser susceptible a la adición de edulcorantes, colorantes, aromatizantes y

saborizantes [22]; proceso llamado “abocamiento”, cuyo fin es suavizar o intensificar su color, aroma y sabor. Estos ingredientes pueden ser color caramelo, extracto de roble o encino, glicerina y jarabe de azúcar. Cabe mencionar que abocar un tequila es permitido en una proporción no mayor al 1 % con respecto al volumen final embotellado [22].

#### 4.1.1 Clases de tequila

Con base en la NOM-006-SCFI-2005 [22] se distinguen cinco clases de tequila según las características adquiridas a posteriori de la doble destilación:

1. Tequila Blanco o Plata: resultante de la doble destilación del producto de fermentación, sin ningún tipo de maduración, ajustado solamente con agua para dilución del alcohol.
2. Tequila Joven u Oro: producto de la doble destilación con menos de dos meses de reposo en barricas de encino o roble, o bien, aquel resultante de la combinación de un tequila blanco, con tequilas reposados, añejos o extra añejos.
3. Tequila Reposado: sujeto a maduración mínima de dos meses, y menor a un año en barricas, aunque también puede ser resultado de la combinación de un tequila reposado con tequilas añejos o extra añejos.
4. Tequila Añejo: éste sufre un proceso de maduración mínimo de un año, y menor a tres años, en contacto directo con barricas cuya capacidad máxima sea de 600 L; o puede también resultar de la combinación de un tequila añejo con un extra añejo.
5. Tequila Extra Añejo**:** producto sujeto a maduración mínima de tres años en barricas de roble o encino con capacidad máxima de 600 L.

Todos los tequilas son susceptibles a un proceso de abocamiento y deben ser ajustados en su contenido alcohólico con agua de dilución [22].

#### 4.1.2 Categorías

Por otro lado, esta misma norma (NOM-006-SCFI-2005, [22]) define dos categorías de tequila:

1) *100 % Agave***:** cuyo mosto es producido con azúcares obtenidas exclusivamente del *Agave Tequilana weber* variedad azul.

2) *Tequila mixto*, o simplemente, *Tequila*: cuyas azúcares pueden provenir de fuentes distintas al *Agave Tequilana weber* var. azul hasta en un 49 %.

### 4.1.3 Características químicas del tequila

En la composición química del tequila podemos encontrar alcoholes, ésteres, aldehídos, ácidos, cetonas, furanos, lactonas y carbohidratos que contribuyen a su perfil en aroma y sabor de esta bebida [40]; sin embargo, los furfurales (aldehídos aromáticos que se forman a partir de reacciones secundarias durante el cocimiento de las piñas de agave), son compuestos no deseables en el perfil del tequila [13], puesto que son irritantes para el cuerpo humano, provocando dolor de cabeza, cansancio, picor de garganta, entre otros efectos. Además, se sabe que estos pueden afectar negativamente la eficiencia del proceso de fermentación [13]. Durante el proceso de reposo del tequila, los compuestos volátiles presentes en el tequila se ven modificados; donde, entre muchos otros, el furfural (2-furaldehído) y el 5-(hidroximetil)furfural aumentan su concentración [40].

Según las especificaciones fisicoquímicas del tequila, contenidas en la NOM-006-SCFI-2005, los tequilas pueden contener entre 33–55% de etanol, 30–300 mg de metanol por cada 100 mL de etanol anhidro y máximo 4 mg de furfural en 100 mL de etanol anhidro [22].

## 4.2 Consejo Regulador del Tequila

El Consejo Regulador del Tequila A. C. es la institución interprofesional encargada de verificar y certificar el cumplimiento de la NOM-006-SCFI-2005, así como de promover la calidad, cultura y el prestigio del tequila como bebida nacional por excelencia [23].

Éste, se dedica a salvaguardar la Denominación de Origen del Tequila (DOT), tanto en México como en el extranjero, garantizando al consumidor la autenticidad de la bebida [24].

De acuerdo con su reglamento interno, para el desempeño de sus funciones, el CRT cuenta con la estructura organizacional, comités técnicos, empleados,

capacitaciones, lineamientos y documentaciones necesarios. Entre sus funciones destacan la inscripción de las tequileras (firma de contrato de prestación de servicios), así como la certificación, verificación y exportación del tequila [25].

### 4.2.1 Producción y exportación de tequila

El tequila es una bebida nacional producida a gran escala, lo cual implica un ingreso importante a la economía mexicana. A continuación, se presentan las estadísticas de producción de esta bebida.

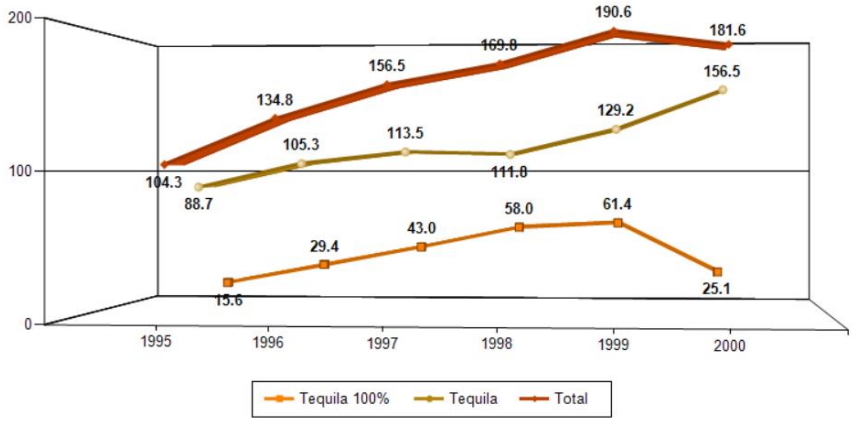



***Figura 3*. Millones de litros de tequila producidos en el periodo 1995-2000 [33].**

En la *figura 3* se puede observar que en el periodo 1995-2000, la producción de tequila oscilaba entre 100 y 200 millones de litros producidos anualmente.

A partir del año 2005 se comenzaron a producir más de 200 millones de litros, siendo en el 2008 la mayor producción de tequila (312.1 millones de litros, *figura 4*), para los cuales se ocuparon 1,125.1 toneladas de agave [33].

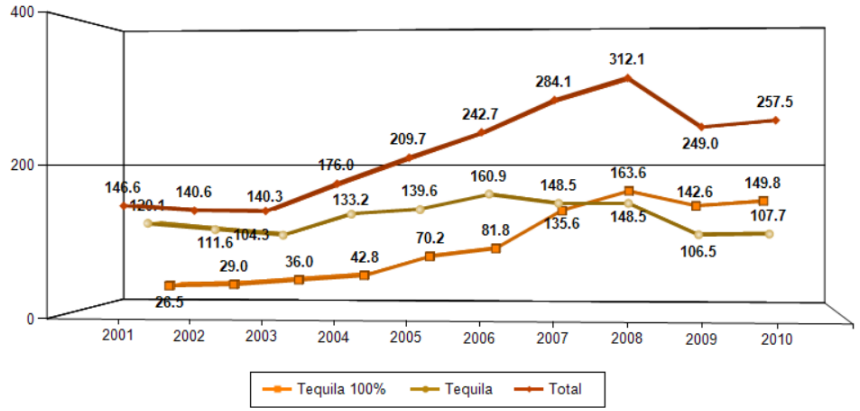



***Figura 4*. Millones de litros de tequila producidos en el periodo 2001-2010 [33].**

De estos millones de litros producidos, entre el 40-80% es exportado anualmente [33]. En la *figura 5*, se aprecia un gráfico de millones de litros de tequila exportados en el periodo 2008-2017.

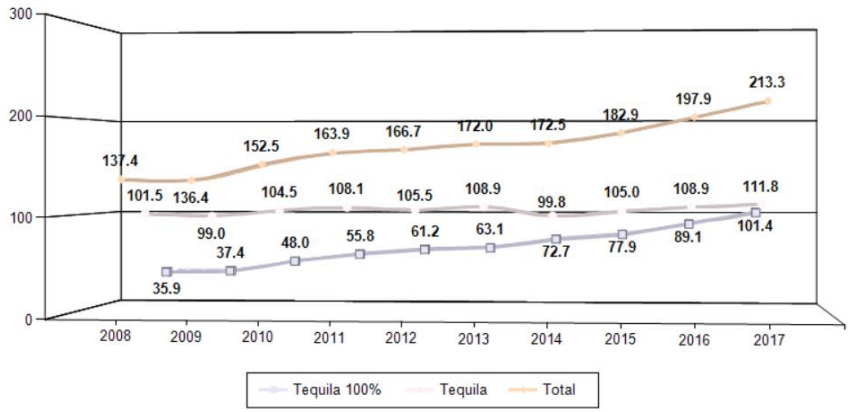



***Figura 5*. Millones de litros de tequila exportados en el periodo 2008-2017 [33].**

En la *tabla 1* se muestran los primeros 10 países con mayor importación de tequila en el periodo 2015-2017.

***Tabla 1*. Principales países a los que se exporta el tequila.**

| | País | Millones de Litros Exportados |
|---|---|---|
| 1 | Estados Unidos de América | 472,912,096.59 |
| 2 | España | 12,319,989.17 |
| 3 | Alemania | 11,743,812.71 |
| 4 | Francia | 8,721,225.47 |
| 5 | Japón | 5,246,958.11 |
| 6 | Sudáfrica | 4,357,027.45 |
| 7 | Reino Unido | 4,283,750.12 |
| 8 | Letonia | 3,741,328.91 |
| 9 | Canadá | 3,583,191.69 |
| 10 | Brasil | 3,259,906.22 |

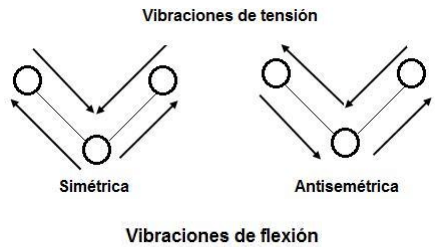
### 4.3 Espectrofotometría de infrarrojo

Esta técnica de análisis tiene alrededor de 140 años de existencia y es ampliamente usada en la identificación de estructuras moleculares de acuerdo con los grupos funcionales reconocidos en un espectro [26].

#### 4.3.1 Fundamento

La espectrofotometría infrarroja se fundamenta en la absorción de radiación electromagnética de longitud entre 800 nm y 1000 µm, por los enlaces en vibración que conforman una molécula igual a la necesaria para llevar a cabo su transición vibracional [27].

Se definen dos tipos de vibraciones: tensión y flexión (*figura 6*). Las primeras representan cambios en la distancia interatómica a lo largo del eje del enlace entre dos átomos, mientras que las segundas se originan por el cambio en el ángulo de dos enlaces [27].



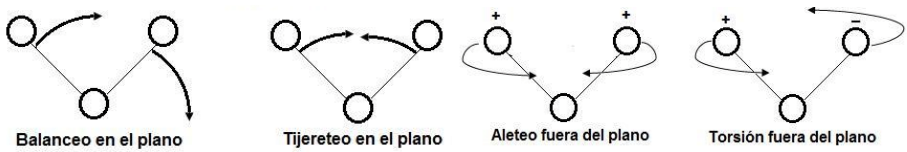



***Figura 6*. Tipos de vibraciones de enlace en Infrarrojo (IR).**
Fuente: imágenes Google recuperado el 28/Agosto/2018

Existe la certeza de que cada molécula tiene enlaces di- o tri- atómicos que generan una vibración y esta a su vez, una absorción a determinado número de onda. Así pues, se generan bandas típicas (*tabla 2*), de diferentes grupos funcionales y con localización e intensidad específica, que permiten identificar un compuesto químico [26].

Dichas intensidades en las bandas son proporcionales a las concentraciones de los componentes, por lo tanto, es posible cuantificar una sustancia y realizar un análisis de varias muestras [26].

***Tabla 2*. Tabla general de identificación de vibración de enlaces en Infrarrojo Medio (MIR).**

| Intervalo de frecuencia ($cm^{-1}$) | Enlace | Tipo de vibración |
|---|---|---|
| 3600-3200 | O-H | Tensión |
| 3500-3200 | N-H | Tensión |
| 3000-2800 | C-H | Tensión |
| 1600-1700 | O-H | Flexión |
| 1640-1550 | N-H | Flexión |
| 1400-1200 | C-H | Flexión |
| 1350-1000 | C-N | Flexión |
| 900-800 | As-O | Tensión (simétrica) |
| 700-750 | As-O | Tensión (antisimétrica) |
| 500-400 | As-O | Flexión |

### 4.3.2 Instrumentación

Los instrumentos empleados para la medición de absorción en el infrarrojo requieren una fuente de radiación continua y un detector. Entre las fuentes, se pueden encontrar el emisor de Nernst, el filamento nicromo, el arco de mercurio, la lámpara de filamento de Wolframio, láser de dióxido de carbono, etc. Además, existen los detectores térmicos, piroeléctricos y fotoconductores [28].

En este sentido, se encuentran tres tipos de instrumentos para la medición de IR [28]:

- Espectrofotómetros dispersivos de red, que nos ayudan al análisis cualitativo.
- Multiplex, que emplean la transformada de Fourier para medición cuantitativa y cualitativa.
- Fotómetros no dispersivos, para la determinación cuantitativa. Pueden ser de absorción, emisión o reflectancia.

El uso de espectrofotómetros dispersivos presenta algunas desventajas importantes, entre ellas que la respuesta suele retrasarse en regiones donde la señal cambia rápidamente, provocando una lectura inadecuada. Además, en regiones donde la respuesta se aproxima a cero, la radiación casi no llega al detector y disminuye la exactitud de la medición. Por otra parte, los instrumentos no dispersivos son menos complicados, más resistentes, de fácil mantenimiento y más baratos; sin embargo, suelen tener efectos de atenuación por el uso de rendijas [28].

Por el contrario, los espectrofotómetros con aplicación de la transformada de Fourier tienen alto rendimiento, ya que poseen pocos elementos ópticos y carecen de rendijas, lo cual provoca la observación de una relación señal/ruido muy grande, puesto que se disminuye la atenuación de señales. Éstos tienen, además, un elevado poder de resolución y reproducibilidad, facilitando la interpretación de los espectros de muestras complejas [28].

### 4.3.3 Zonas espectrales

Existen tres intervalos de interpretación para los espectros de infrarrojo [29]:

- Infrarrojo Cercano (NIR): 780–2500 nm, que es equivalente a 12800–4000 $cm^{-1}$.
- Infrarrojo Medio (MIR): 2500 – 25000 nm, equivalente a 4000–400 $cm^{-1}$.

- Infrarrojo Lejano (FIR): 25000 – 40000 μm, equivalente a 400 – 25 $cm^{-1}$.

De estas regiones, la más comúnmente empleada para la identificación molecular, es Infrarrojo Medio o MIR (por sus siglas en inglés), ya que el espectro resultante es único para cada compuesto [26].

La región de MIR permite la identificación de grupos funcionales, mediante la vibración del esqueleto molecular: -C=O, -C=N, -N-H, -C-H, etcétera. La región de FIR provoca la rotación de las moléculas en estado gaseoso. Finalmente, la región del NIR hace vibrar principalmente los átomos del enlace -O-H [29].

### 4.3.4 Obtención de espectros

El resultado de medir una muestra en un espectrofotómetro de IR es llamado espectro. Su representación gráfica está dada por las ecuaciones:

$$Absorbancia = f\ (número\ de\ onda) \quad \text{.....Ec. (1)}$$

o

$$\%Transmitancia = f(número\ de\ onda) \quad \text{.....Ec. (2)}$$

es decir, en el eje x (variable independiente) están los valores de número de onda en $cm^{-1}$, mientras que en el eje y (variable dependiente) están los valores de absorbancia o porcentaje de transmitancia.

La absorbancia o transmitancia se debe a movimientos intramoleculares, es decir, se genera por la formación de espectros vibracionales denominados *bandas* que están en función de la estructura interna de la molécula [26].

## 4.4 Análisis multivariante

El análisis multivariante es aquella rama de la estadística encargada del estudio de dos o más variables obtenidas de un conjunto de objetos. Éste resulta de la combinación de dos aspectos importantes: la estadística y el análisis de datos (cuantitativos o cualitativos) [2].

Comúnmente, se realiza mediante el uso de una matriz de datos y puede ser de tipo descriptivo/inductivo con un gran número de variables que dan lugar a una representación multidimensional [2].

El análisis multivariante o multivariado genera nuevas técnicas y modelos cuyo objetivo es proporcionar información vasta y fructífera de una serie de datos que puede dar lugar a propiedades aplicables a una población [6].

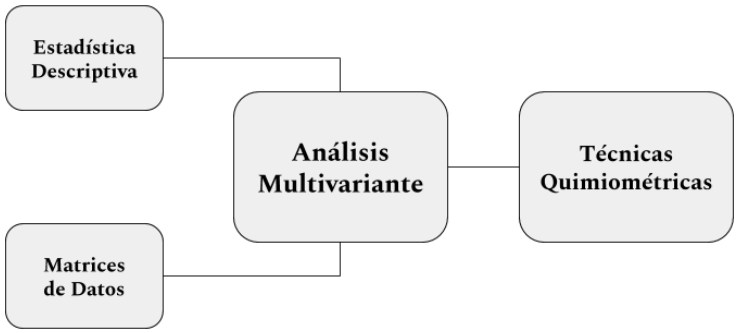



***Figura 7.*** **Origen de las técnicas quimiométricas.** (imagen propia)

Partiendo de dicho análisis, surgen las técnicas quimiométricas (*figura 7*), las cuales permiten obtener mucha información de los objetos con que se busca trabajar, siendo entonces, muy útiles aun cuando estas son relativamente complejas debido al uso de matrices de datos [1].

Una de sus principales aplicaciones nace de su capacidad de determinar si los objetos de estudio son similares o distintos. Esto, de manera general, es representado por diagramas bi- o tri- dimensionales que, dependiendo de qué tan cercanos o alejados estén una muestra de otra, indican el grado de relación que poseen entre sí [1].

En este tipo de análisis, se pueden encontrar diferentes tipos de **escalado** (centrado en la media y autoescalado), **métodos de acoplamiento** (completo, promedio, individual y algoritmo de Ward), **distancias** (euclidiana, euclidiana cuadrada, Manhattan -también llamada *city block*-, *mahalanobis* y *Minkowski*), variedad de **criterios de asignación** (bayes, máx, *class modelling* y dist), diversos **métodos de validación** (*cross validation* -por *venetian blinds* o por

*contiguous blocks*-, *Montecarlo 20% out* y *bootstrap*) que nos permiten, de acuerdo con la técnica en estudio, construir modelos de clasificación para la obtención de información de un grupo de muestras.

4.4.1 Tipos de escalado

4.4.1.1 Centrado en la media (*mean center*/*meancentering*)

Este escalado genera una nueva matriz de datos de acuerdo con la ecuación siguiente:

$$x' = x_n - \bar{x} \quad \text{.....Ec. (3)}$$

donde x' es el dato centrado, $x_n$ el valor numérico correspondiente a la variable y $\bar{x}$ la media estadística de cada variable [11].

4.4.1.2 Autoescalado (*autoscale*/*autoscalling*)

El autoescalado se basa en el escalado centrado en la media seguido de una estandarización, es decir, los valores obtenidos de x' se dividen entre la desviación estándar de cada variable [11]. Esto se expresa con la siguiente ecuación:

$$x'' = \frac{x_n - \bar{x}}{s} \quad \text{.....Ec. (4)}$$

donde x'' es el dato autoescalado, $x_n$ el valor numérico correspondiente a la variable, $\bar{x}$ la media estadística de cada variable y $\boldsymbol{s}$ la desviación estándar de la variable.

4.4.2 Métodos de acoplamiento

4.4.2.1 Acoplamiento simple o individual (*single linkage*)

Se genera un acoplamiento con base en la distancia mínima (vecino más próximo), que existe entre dos elementos, un elemento y un grupo o dos grupos, donde, se forman los conglomerados considerando la menor distancia existente [35].

4.4.2.2 Acoplamiento promedio (*average linkage*)

Este tipo de acoplamiento calcula las distancias entre todos los pares de muestras posibles, además de la distancia promedio para cada grupo, y junta aquellas que tienen una distancia promedio menor [45].

4.4.2.3 Acoplamiento completo (*complete linkage*)

Contraria al acoplamiento simple, este considera la máxima distancia existente entre dos elementos, por lo que se aplica la regla del vecino más lejano. Sin embargo, también contempla la menor distancia a la hora de formar los conglomerados [35].

#### 4.4.2.4 Algoritmo de Ward (*Ward's algorithm*)

El algoritmo de Ward toma en cuenta la varianza intra-grupos, es decir, calcula las medias de todas las variables y la distancia euclidiana cuadrada de cada muestra con respecto al centroide del conglomerado [45]. Dicho de otra forma, el método de Ward consiste en la unión de muestras (o grupos) que posea un incremento menor en la suma de cuadrados (disimilitud), con lo que se logra obtener grupos más homogéneos [39].

### 4.4.3 Distancias

#### 4.4.3.1 Euclidiana o euclídea (*euclidean*)

La distancia euclídea o euclidiana (*figura 8*) representa la trayectoria (en línea recta) de menor tamaño que separa a dos elementos; si se relaciona esto con el teorema de Pitágoras, se dice que la distancia euclídea es representada por la hipotenusa y viene dada por la ecuación [7, 45]:

$$D_{Euclídea} = \sqrt{\sum (x_i - x_j)^2} \qquad \text{.....Ec. (5)}$$

#### 4.4.3.2 Euclidiana cuadrada (*euclidean squared*)

La distancia euclidiana cuadrada es básicamente igual a la euclídea, sólo de cálculo más rápido, por lo que suele utilizarse más. Suponiendo que el término "cuadrada" se refiera a elevar al cuadrado la distancia euclídea, tendríamos la siguiente ecuación:

$$D_{euclídea\ cuadrada} = \sum (x_i - x_j)^2 \qquad \text{.....Ec. (5.1)}$$

#### 4.4.3.3 Manhattan o *city block*

Este tipo de distancia traza una trayectoria entre dos elementos más larga con respecto a la euclidiana y puede explicarse como la suma de los catetos (*figura 8*), según el teorema de Pitágoras, e implica la siguiente ecuación [7, 45]:

$$D_{Manhattan} = \sum | x_i - y_i | \qquad \text{.....Ec. (6)}$$

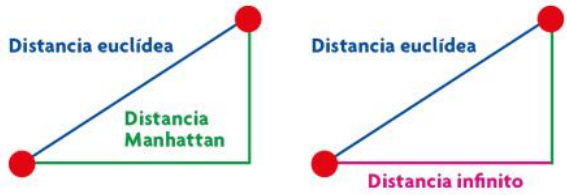



***Figura 8.*** **Distancias Euclidiana y Manhattan.**
Fuente: Imágenes de Google recuperado el 28/Agosto/2018

4.4.3.4 Mahalanobis

Esta distancia, introducida en 1936, toma en cuenta la correlación entre variables independientemente de su escala, por lo tanto, permite diferenciar grupos de elementos mediante el cálculo de variables aleatorias [34]. Esta se recomienda cuando la magnitud de las variables es muy distinta o cuando existen más de dos grupos [45]. Se expresa bajo la ecuación:

$$D_{Mahalanobis}{}^{2} = (x_i - x_j)' \emptyset^{-1} x_i - x_j \text{.....Ec. (7)}$$

donde Ø es la matriz de varianza entre las variables. Cuando la varianza de las variables es independiente, la distancia resultante sería similar a la euclidiana cuadrada [34].

4.4.3.5 Minkowski

Esta distancia está conformada a su vez, por diferentes distancias, como Manhattan o Euclídea, incluso Mahalanobis [45], puesto que es capaz de resumir sus magnitudes en un espacio con p dimensiones. Esta puede ser calculada de acuerdo con la ecuación:

$$D_{Minkowski} = \sqrt[p]{\sum_{i=1}^{n} [(\,|\, x_{ai} - x_{bi} \,|\,)^p + (\,|\, y_{ai} - y_{bi} \,|\,)^p \,]} \text{.....Ec. (8)}$$

donde q debe ser un valor numérico real entre 1 e ∞ [34].
Describiendo la *figura 9*:
p=1 implica la distancia Manhattan
y p=2 la distancia euclidiana

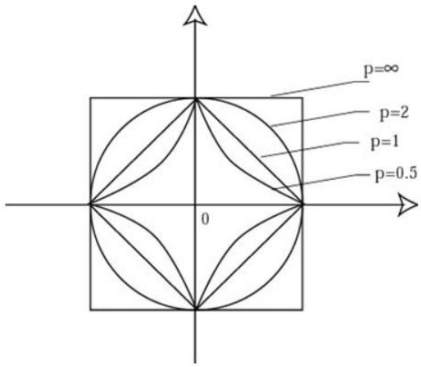



***Figura 9*. Distancias Minkowski.** Fuente: https://bit.ly/2T2JoNU recuperado el 28/08/18

4.4.4 Criterios de asignación (*assignation criterion*)

Los criterios de asignación son reglas que permiten determinar la clase a la que pertenece una muestra. Estos pueden ser:

- Discriminantes

Este criterio, o *regla discriminante*, ubica la muestra en cuestión a la n-ésima clase de acuerdo con su vector p-dimensional, para lo cual, este último debe formar parte de la región construida a partir de otras muestras o elementos [6]. Una de las desventajas de los métodos discriminantes es que definen un límite matemático que podría asignar muestras a una clase a la que no corresponden, o bien, no asignarla, aunque esta pertenezca a ella [45].

Este criterio es utilizado por la técnica quimiométrica supervisada PLS-DA, que permite utilizar los criterios Bayes y Máx.

- De Modelación

Un *criterio de modelación*, o modelado, define una región de dominio experimental a partir de la cual se puede caracterizar una clase de forma individual, es decir, sin tomar en cuenta ninguna información de otra clase [45]. La ventaja que ofrece este tipo de criterio es que sus límites de clase no son afectados por la distribución que estas tengan, a diferencia de las reglas discriminantes [45].

#### 4.4.5 Métodos de validación

Después de construir modelos de clasificación, es importante conocer el error que este puede generar en la predicción de muestras, este proceso se denomina validación [5].

Para esto, existen diferentes técnicas, como son la validación simple (*holdout*), validación cruzada, Montecarlo 20% out, bootstrap, etc. [5].

##### 4.4.5.1 Validación cruzada (*cross validation*)

Este tipo de validación emplea las muestras del grupo de calibración y se basa en la formación de subgrupos denominados grupos de cancelación que permiten calcular el modelo dejando fuera a uno de ellos. Este proceso se repite n veces y suele llamarse *full-cross-validation* o *leave-one-out* [11].

Empleando la validación cruzada, es posible determinar el error del modelo en la predicción de muestras desconocidas a partir de la construcción y evaluación de clasificadores *k* x *t*, siendo la media de los errores obtenidos en cada entrenamiento [5].

Esta puede llevarse a cabo mediante persianas venecianas (*venetian blinds*) o bloques contiguos (*contiguous blocks*).

##### 4.4.5.2 *Montecarlo 20% out*

Tiene como objetivo calcular la incertidumbre de la contribución de dos variables aleatorias independientes con los mismos valores de media y varianza [34]. Emplea un estimador matemático de una función. Esta considera un intervalo de confianza del 95%, la varianza de los elementos, su media, así como las iteraciones que siga la técnica [36].

##### 4.4.5.3 *Bootstrap*

Esta validación estima los errores de los modelos predictivo, para lo cual emplea n cantidad de elementos y genera el conjunto de entrenamiento de forma aleatoria, siendo el grupo de prueba las muestras no seleccionadas [5].

#### 4.4.6 Quimiometría

La quimiometría es un área de la química analítica que hace uso tanto de la química como de las matemáticas para extraer el máximo conocimiento de un grupo de datos. Su principal objetivo es optimizar cada fase del análisis químico basándose en modelos matemáticos fundados en el análisis multivariante [1].

De acuerdo con la Sociedad Internacional de Quimiometría (1975):

*"La quimiometría es la disciplina química que utiliza métodos matemáticos y estadísticos para diseñar y seleccionar procedimientos de medida y experimentos óptimos, para extraer la máxima información química mediante el análisis de datos químicos."*

El análisis de datos por métodos quimiométricos, puede aprovecharse en dos direcciones principales:

- Cuantificación: Cuyo objetivo es determinar la cantidad, sea peso o volumen, o bien, la concentración de uno o más analitos existentes en una muestra.
- Clasificación: Que busca lograr la diferenciación entre los analitos de una muestra, atendiendo a diversas variables [8].

Las técnicas de clasificación son metodologías multivariantes que tienen la finalidad de encontrar modelos matemáticos capaces de reconocer la pertenencia de una muestra a su clase. Dentro de ellas, encontramos dos tipos: las técnicas supervisadas y las no supervisadas, descritas más adelante.

La construcción de modelos que reducen la dimensionalidad de los datos implica la descomposición de la matriz original [45], para lo cual, existen diferentes metodologías:

- Cálculo de eigenvalores y eigenvectores

Se lleva a cabo mediante una matriz X de orden m x n (donde m corresponde al número de filas o muestras y n al número de columnas o variables), con un vector $\mathbf{v}$ y un escalar $\lambda$ que satisface la ecuación [42, 45]:

$$X \times \mathbf{v} = \lambda x \mathbf{v} \qquad \text{.....Ec. (9)}$$

Cuya resolución se lleva a cabo mediante la determinante:

.....**Ec. (10)**

$$\det (X - \lambda \, x \, \mathbf{I}) = 0$$

donde ***I*** es la matriz identidad y $\lambda$ es el valor propio asociado al vector [42]. Esto, da lugar a una ecuación polinómica que se resuelve mediante factorización, donde, además, es posible despejar a los vectores [45].

Ejemplo [43]:

Sea la matriz cuadrada: $\begin{pmatrix} 1 & 2 & 1 \\ 6 & -1 & 0 \\ -1 & -2 & -1 \end{pmatrix}$

con determinante: $\det\left[\begin{pmatrix} 1 & 2 & 1 \\ 6 & -1 & 0 \\ -1 & -2 & -1 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right]$

$-\lambda^3 - \lambda^2 + 12\lambda = 0$

cuya factorización es: $-\lambda\,(\lambda - 3)(\lambda + 4) = 0$

y valores propios son: $\lambda = 3$, $\lambda = 3$ y $\lambda = -4$

- <u>Descomposición en Valores Singulares (SVD)</u> [44, 45]

Llevado a cabo a partir de la fragmentación de la matriz original X en tres nuevas matrices expresadas por las letras U, S y V. La última (V) corresponde a los eigenvectores en forma traspuesta que, al estar relacionada con las variables, se denota *loadings*. Las matrices U y S, al ser multiplicadas, dan lugar a los *scores*, que se relacionan con las muestras, donde S es la matriz diagonal de las raíces cuadradas de los eigenvalores (valor singular, $\sigma = \sqrt[2]{\lambda}$ ).

Esta relación, viene dada por la ecuación:

$$X = U\,S\,V^T \quad \text{.....Ec. (11)}$$

Este tipo de construcción se aplica en la obtención de los factores denominados componentes principales (PC's) de un PCA.

- <u>Algoritmo NIPALS</u>

*Nonlinear estimation by Iterative Partial Least Squares*, NIPALS por sus siglas en inglés, es una aproximación al método anterior, desarrollada por Wold en 1966, que tiene como objetivo extraer dos nuevas matrices a partir de la original (X) dadas por la ecuación [45]:

$$X = TP^T \quad \text{.....Ec. (12)}$$

donde **T** es la matriz de *scores* (relacionada con las muestras) y $\mathbf{P}^{\mathbf{T}}$ es la matriz traspuesta de los *weights* (relacionada con las variables).

Conceptualmente, los *loadings* y *weights* son similares, salvo la forma en que se calcula uno y otro. Los segundos, y en general el algoritmo NIPALS, se asocia con la obtención de variables latentes (LV) en un análisis PLS discriminante.

### 4.5 Técnicas quimiométricas no supervisadas

Una técnica quimiométrica no supervisada es aquella en la que no se dispone de un grupo de muestras previamente clasificadas, sino que su clasificación se lleva a cabo a partir de sus propiedades medidas y son agrupadas por su similitud [30].

Los métodos no supervisados descubren la tendencia de las variables y las muestras, por lo cual suelen llamarse de reconocimiento de pautas. Éstos tienen la capacidad de representar de manera gráfica y resumida una gran cantidad de datos que actualmente generan diversas técnicas analíticas [20]. Un estudio de tipo no supervisado es importante realizarlo como punto de partida (método exploratorio) para llevar a cabo técnicas supervisadas, especialmente cuando la naturaleza de las muestras es desconocida [20].

#### 4.5.1 Análisis de Componentes Principales (PCA)

Esta técnica tiene como objetivo obtener los llamados Componentes Principales (PC's) que resultan de combinaciones lineales de acuerdo con las variables originales, aportando el primero de ellos (PC1) la mayor cantidad de información [6]. Estas son funciones ortogonales, lo que implica que sus productos escalares son nulos; es decir, que no están correlacionados entre sí.
De forma más detallada, se dice que los coeficientes que conforman a PC1 definen una dirección en la cual las proyecciones de los datos tienen la mayor dispersión. Posteriormente el PC2 recoge la mayor información que no ha rescatado el primer componente, y así sucesivamente hasta generar tantos componentes principales como variables o muestras se tengan (siempre el de menor numeración) [6, 45].

Comúnmente, los primeros componentes principales reúnen la mayor cantidad de información de la relación muestra-variable, lo cual permite desarrollar un

espacio de menor dimensión para el análisis de datos y de esta forma, se facilita la interpretación [6].
Uno de los criterios para elegir el número de componentes a considerar en un modelo es que estos expresen entre el 70-90% de *varianza total* y posean una representación gráfica de clasificación adecuada [6].

Cuando se construye un vector que modela una dispersión de puntos, la varianza total se divide en dos: *varianza explicada* por el vector y *varianza residual* [4].

$$s^2{}_T = s^2{}_{explicada} + s^2{}_{residual} \quad \text{.....Ec. (13)}$$

Si el vector es un componente principal, su varianza explicada se denomina "autovalor" (*eigenvalue*) y es la varianza del vector o de las proyecciones de puntos sobre el vector [4].

Se dice que un vector describe adecuadamente una tendencia cuando las muestras están distribuidas en la dirección de este. En ese caso, se dice que se obtiene un buen modelo multidimensional, que da como resultado una varianza explicada alta y, en consecuencia, una varianza residual pequeña [4].

A grosso modo, PCA es una técnica de disminución dimensional, mediante la reducción de variables, que pretende visualizar las principales tendencias entre ellas (a través de un estudio detallado de los *loadings*) y en las muestras (visualización de los *scores*).

### 4.5.2 Varimax

Varimax es una técnica que pertenece al análisis factorial, el cual estudia las relaciones de interdependencia (también llamada correlación) existente entre diferentes grupos de elementos, donde las variables permiten generar conceptos subyacentes denominados *factores* [6]. La extracción de factores se realiza en función de la técnica que se desee aplicar, siendo la más conocida el Análisis de Componentes Principales (PCA), para el cual estos factores se denominan "componentes principales" [6]. La siguiente técnica más utilizada para dicho fin es el método de mínimos cuadrados, donde este factor se denomina "variable latente" [6]. En ambas técnicas, se obtienen *scores* y *loadings*; sin embargo, hay ocasiones en que los *loadings* no son muy claros en

cuanto a su representación gráfica. Es entonces donde es posible aplicar el análisis Varimax para facilitar su interpretación.

El nombre completo de esta técnica es Rotación Varimax, propuesta por Kaiser (1985), que busca maximizar los *loadings* de mayor importancia (más elevados) y minimizar (reducir a cero) los de menor importancia, lo cual permite obtener factores con correlaciones altas entre las primeras variables y correlaciones nulas en el resto. Esto se aprecia mejor en los valores de varianza de los factores [6].

La ventaja de esta técnica es que los factores siguen siendo ortogonales y que minimiza el número de variables relevantes para cada factor. Además, facilita la interpretación de estos sin alteración de los datos [6].

#### 4.5.3 Análisis Cluster (CA)

El análisis de conglomerados es una técnica de análisis exploratoria (comprende el estudio de muestras anómalas, transformaciones, etc.), que precisa el objetivo de clasificar un grupo de individuos en función de sus características que no son definidas previamente sino por el resultado del mismo análisis [6].

Algunas veces, esta técnica requiere una reducción en la dimensionalidad de los datos, puesto que sugieren la distribución de los conglomerados finales [6].

El algoritmo general de la técnica es [6]:

- Cálculo de distancias entre n elementos o muestras.
- Formación de grupos entre aquellas más cercanas.
- Generación de conglomerados a partir de los grupos más cercanos.
- A partir de estos conglomerados, se crean tantos más sean necesarios hasta obtener un macro grupo conformado por todas las muestras.

El análisis cluster, o de conglomerados, da como resultado una representación gráfica denominada dendrograma.

### 4.6 Técnicas quimiométricas supervisadas

Las técnicas no supervisadas PCA y CA suelen ser eficientes en el reconocimiento de pautas en un conjunto de muestras cuando no se dispone de ninguna información previa; sin embargo, no son las más adecuadas cuando nos enfrentamos a problemas más avanzados [45].
Cuando se cuenta con varios grupos de elementos para los cuales se conocen sus variables numéricas y se tiene la tarea de identificar a aquellas que diferencian mejor a los elementos en clases, se encuentran las técnicas de clasificación supervisadas, tales como k-NN, Análisis Discriminante (DA), *Artificial Neural Networks*, etc. [6, 45].

Una técnica quimiométrica supervisada es aquella en la que se involucra un conjunto de muestras, llamado de entrenamiento (calibración-validación interna) o *training*, que son útiles en la generación de grupos iniciales bien clasificados para posteriormente introducir un segundo conjunto, llamado de predicción o prueba (*test*) que puedan ser asignadas a estos [30].

Una vez que se tiene el conjunto de calibración, se debe medir la eficiencia del modelo ya que esta se comporta de formas distintas de un algoritmo a otro. Si el conjunto de calibración es capaz de generalizar el comportamiento aprendido para clasificar el conjunto de predicción correctamente, decimos que se trata de un modelo eficiente [30]. Esta eficiencia es determinada por el conjunto de validación interna (*cross validation*).

En este punto, es importante definir tres etapas que se presentan durante el desarrollo de un modelo. La primera, es la **calibración** (también llamada de ajuste o *fitting*), donde se cuenta con un conjunto de muestras cuya clase fue previamente conocida (mediante un modelo exploratorio: PCA o CA). La segunda etapa, es la **validación interna**, que tiene el objetivo de optimizar el modelo construido en la calibración. Comúnmente, es realizada mediante la validación cruzada (*cross validation*), que utiliza las mismas muestras de calibración dejando siempre una cantidad de ellas fuera para aplicarlas en el modelo como grupos de cancelación [41]. Un grupo de cancelación está formado por cierta cantidad de muestras (establecido en el algoritmo durante la etapa de entrenamiento) y es el que queda fuera del modelo en construcción. La tercera y última, es la **predicción**, donde se emplea un conjunto de muestras (no consideradas en las dos etapas anteriores) que son proyectadas en

el modelo validado [45]. Cada modelo, genera una serie de parámetros de desempeño que no son necesariamente idénticos para estas etapas, tales como: precisión (*precision*), sensibilidad (*sensitivity*), especificidad (*specificity),* exactitud (*accuracy),* además reporta la proporción de muestras no asignadas (*ratio of not assigned samples*), la tasa de no error (*non error rate, ner*) y la tasa de error (*error rate, er*) [31]. A continuación, se definen cada uno de ellos [31]:

- **Precisión (*precision, pr*):** Es la capacidad de un modelo para excluir muestras de otras clases en la clase en estudio. Dada por la relación de las muestras que pertenecen a una clase correctamente asignadas, y el número total de muestras asignadas a ella. Expresado con la siguiente ecuación:

$$Pr_g = \frac{n_{gg}}{n'_g} \quad \text{.....Ec. (14)}$$

  donde $n'_g$ es el número de muestras asignadas a una clase y $n_{gg}$ es el número de muestras pertenecientes a una clase y correctamente asignadas.
- **Sensibilidad (*sensitivity, sn*):** Representa la capacidad del modelo para reconocer correctamente la pertenencia de las muestras a su clase. Se define como:

$$Sn_g = \frac{n_{gg}}{n_g} \quad \text{.....Ec. (15)}$$

  donde $n_g$ es el total de muestras pertenecientes a una clase, es decir, las muestras que no fueron asignadas, no se consideran en el cálculo.
- **Especificidad (*specificity, sp*):** Capacidad de la clase para rechazar las muestras de las otras clases. Viene dada por la ecuación:

$$\text{Ec. (16)}.....Sp_g = \frac{\sum_{k=1}^{G}(n'_k - n_{gk})}{n - n_g} \quad for \quad k \neq g \qquad n'_k = \sum_{g=1}^{G} n_{gk} \text{.....Ec. (17)}$$

  donde $n'_k$ es el total de muestras asignadas a la clase y $n_{gk}$ son las muestras pertenecientes a la clase y asignadas a otra. Para calcular este parámetro, las muestras no asignadas, no se consideran.

- **Exactitud (*accuracy, ac*):** Es la proporción de muestras asignadas correctamente. Dada por la ecuación:

$$AC = \frac{\sum_{g=1}^{G} n_{gg}}{n} \quad \text{.....Ec. (18)}$$

donde n es el total de muestras. Tampoco considera muestras no asignadas.

- **Proporción de muestras no asignadas (*ratio of not assigned samples*)**: son aquellas muestras que no son reconocidas y clasificadas en ninguna clase. Se calcula mediante la suma de la última columna de la matriz de confusión y se divide por el total de muestras.
- **Tasa de no error (*non error rate, ner*)** es el promedio de la sensibilidad de las clases. Definida por la ecuación:

$$NER = \frac{\sum_{g=1}^{G} Sn_g}{G} \quad \text{.....Ec. (19)}$$

- **Tasa de error (*error rate, er*)** se define por:

$$ER = 1 - NER \quad \text{.....Ec. (20)}$$

Por último, en estas técnicas, se puede hablar de una matriz de confusión que tiene la función de indicar para cada una de las clases, cuántos miembros se clasifican en ellas. Cada columna representa la clase asignada y cada fila la clase real (*tabla 3*), además en la última columna se enlista el número de muestras no asignadas a su clase [30, 31].

***Tabla 3*. Composición de una matriz de confusión.**

| | | assigned class | | | | | not assigned |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | ... | $G$ | $G+1$ |
| true class | 1 | $n_{11}$ | $n_{12}$ | $n_{13}$ | ... | $n_{1G}$ | $n_{1G+1}$ |
| | 2 | $n_{21}$ | $n_{22}$ | $n_{23}$ | ... | $n_{2G}$ | $n_{2G+1}$ |
| | 3 | $n_{31}$ | $n_{32}$ | $n_{33}$ | ... | $n_{3G}$ | $n_{3G+1}$ |
| | ... | ... | ... | ... | ... | ... | ... |
| | $G$ | $n_{G1}$ | $n_{G2}$ | $n_{G3}$ | ... | $n_{GG}$ | $n_{GG+1}$ |

4.6.1 Curvas de Potencia (PC)

Curvas de Potencia pertenece a un método conocido como *funciones de densidad* (o de probabilidad), similar a la técnica Funciones de Potencia (PF), salvo que PC se aplica de forma más sencilla [32]. Esta técnica, propuesta por el profesor Xavier Tomás, tiene aplicación predictiva del Análisis de Componentes Principales (PCA) en el conjunto de calibración para posteriormente utilizar sus *scores* en los elementos de prueba (*test*) y proyectarlos en el modelo, lo cual, hace a PC gráficamente similar a PCA, dado que arroja un diagrama PC1 vs. PC2 conformado por elipses de iso-probabilidad que ayudan a la clasificación de los elementos [32].

Al emplear funciones gaussianas en tres dimensiones (3D), se genera un gráfico en 2D con tantas elipses de iso-probabilidad como clases haya, siendo la zona de mayor probabilidad la que está al centro, en el caso de las elipses, o bien, en la parte más alta de la función de densidad (*figura 10*).

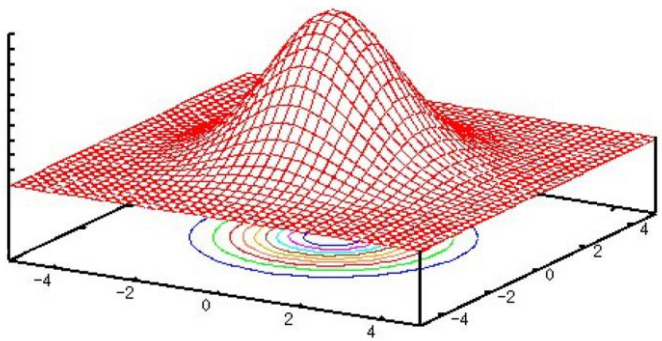


***Figura 10*. Función gaussiana (arriba, rojo), Elipses de iso-probabilidad (abajo, colores).**
Imagen del Dr. José Manuel Andrade Garda [45].

Cabe mencionar que, con este análisis, se elimina la necesidad de evaluar las varianzas de cada componente y se presentan los resultados en probabilidad, ya que, aunado al gráfico, PC ofrece una tabla donde se expresan en porcentaje la probabilidad de pertenencia de cada elemento a cierta clase [32].

### 4.6.2 Support Vector Machines (SVM)

Técnica desarrollada a finales de la década de 1970 por Corinna Cortes y Vladimir Vapnik. Ésta tiene su fundamento en las redes neuronales, salvo que no requiere de un entrenamiento previo, como es el caso de ANN (*Artificial Neural Networks*) [11]. SVM busca aumentar la dimensionalidad de los datos (variables que describen a cada elemento) y así mejorar la separación de las clases, lo cual, resulta de gran utilidad cuando existe comportamiento no lineal o ruidoso en los grupos de muestras.
Los datos que se podrían utilizar para llevar a cabo este análisis son: mediciones en técnicas analíticas o una perspectiva simplificada de estas, por ejemplo, los *scores* obtenidos en un Análisis de Componentes Principales previo [20].

Aquí se deben distinguir dos objetivos fundamentales de la técnica [18]:

1. Proyección de los datos en un espacio de mayor dimensión (*espacio de caracterización*) mediante el cálculo de las dimensiones adicionales (proceso denominado *mapeo no lineal* o *mapeo de caracterización*).
2. Construcción de una superficie óptima de separación a partir de la generación de un *hiperplano* que maximice el margen entre clases, comúnmente, representado por un color distinto al resto de clases.

Las dimensiones originales (*espacio de entrada*) son proyectadas por diferentes funciones matemáticas en un espacio que se denomina *función de núcleo*; entre las más utilizadas se encuentran las funciones lineales, de base radial y polinomios [14]. De forma general, los núcleos o *kernel*, pueden ser interpretados como combinaciones de las variables originales que abarcan la dimensionalidad donde las muestras son caracterizadas [20].

Al modificar el valor de C (penalización del error), se modifica el margen de acuerdo con las *figuras 11 y 12*, donde se observa que a mayor valor de C más pequeño es el margen entre las clases.

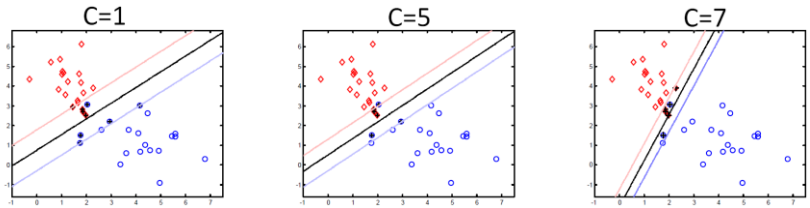



***Figura 11*. Penalización en un análisis SVM con kernel lineal.**
Imagen del Dr. José Manuel Andrade Garda [45].

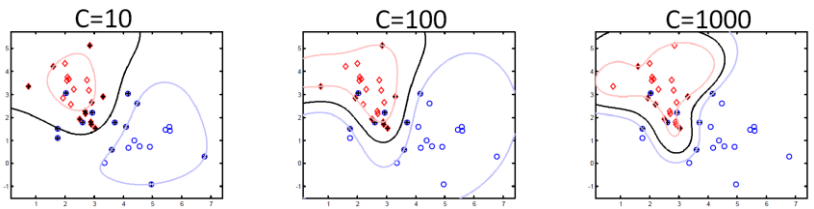



***Figura 12*. Penalización en un análisis SVM con kernel gaussiano (radial).**
Imagen del Dr. José Manuel Andrade Garda [45].

### 4.6.3 Vecino más cercano (k-NN)

k-NN es una técnica de clasificación supervisada simple y no paramétrica que permite estimar una función de densidad o de probabilidad de que un elemento pertenezca a una clase de acuerdo con el entrenamiento del modelo [7]. La distancia en este algoritmo se define como el mínimo de las distancias entre los elementos A y B [6], asignando los miembros del conjunto de predicción al conjunto de muestras (calibradas y validadas) más cercanas de acuerdo con sus características y cumpliendo la regla de que este contenga la mayor cantidad de miembros [30].

La determinación del valor óptimo para k (vecino más cercano) es la parte más importante de este proceso. Para ello, se lleva a cabo una etapa de optimización donde se hace uso de la validación cruzada para buscar aquel valor que minimice el error de clasificación [9].

Cabe mencionar que esta técnica supone que los vecinos más cercanos dan la mejor clasificación para las muestras de prueba, lo cual genera la probabilidad de que exista un error significativo de clasificación [7].
k-NN, a diferencia de las demás técnicas supervisadas desarrolladas en *Matlab*©, no genera ningún resultado gráfico, sino simplemente los parámetros de desempeño de la modelación.

#### 4.6.4 Análisis Discriminante (DA)

Desarrollado por Mahalanobis y Fisher en 1936, maximiza la relación de varianzas entre clases y minimiza la proporción de varianzas dentro de la clase [20]. Este resulta útil cuando ya tenemos perfectamente identificados diversos grupos de muestras, puesto que su objetivo es encontrar una dirección que alcance la máxima separación entre las clases en una dimensionalidad menor [20]. El análisis discriminante puede ser lineal (LDA) o cuadrático (QDA) y el uso de uno u otro depende de si la separación de clases es lineal o no, además de la confiabilidad de las matrices de covarianza de las clases [20]. Haciendo una comparación con PCA, donde los factores maximizan la varianza total explicada, LDA maximiza la relación entre las varianzas dentro y fuera de las clases (*ratio between-class* y *within-class variances*) [20].

Uno de los requisitos para que DA funcione es que, si hay n clases, deben existir n-1 funciones discriminantes. Y uno de los limitantes es que asigna todas las muestras desconocidas, o de predicción, a uno de los grupos formados durante el entrenamiento del algoritmo en cuestión [20]. El objetivo principal de DA es generar reglas de clasificación, de acuerdo con las características de cada grupo, que permitan predecir el grupo al que pertenecen nuevas muestras [6].

##### 4.6.4.1 Por Componentes Principales (PCA-DA)

Esta técnica, como su nombre lo indica, es una combinación entre el análisis discriminante y análisis de componentes principales. El primero de ellos, como ya se dijo, maximiza la varianza entre clases y minimiza la existente en cada una [20], mientras que el segundo reparte la varianza total en los componentes principales obtenidos (factores).

Uno de los requisitos del análisis discriminante es que el número de variables no exceda el número de muestras en cada clase, aunque, comúnmente, esto no

ocurre. Es entonces cuando se procede a aplicar DA en los *scores* generados a partir de un PCA [20].

Los factores obtenidos en un análisis discriminante se denominan variables canónicas, cuyo objetivo es definir las propiedades o características comunes entre las clases y suelen aplicarse a un modelo de PCA (puesto que en él la varianza explicada de las variables suele presentarse mayoritariamente en las primeras más que en el resto), para eliminar los problemas que puedan resultar de la multicolinealidad de las variables [19].

Como resultado de este modelo, se calculan los *Q residuals*, que implican un cálculo estadístico del tipo *lack-of-fit* (fuera de ajuste). Los residuos Q nos indican qué tan bien se ajusta una muestra al modelo construido mediante el desarrollo de una medida de diferencia, o residual, entre ésta y su proyección [38].

#### 4.6.4.2 Por el Método de Mínimos Cuadrados Parciales (PLS-DA)

El método de mínimos cuadrados parciales (PLS) es uno de los algoritmos de regresión mayormente utilizados en el desarrollo de modelos predictivos multivariantes; sin embargo, éste fue modificado para ser aplicado en clasificación. Esta modificación es aplicar el Análisis Discriminante (DA) y recibe el nombre de PLS–DA [20].

El análisis discriminante por el método de mínimos cuadrados parciales genera las llamadas *variables latentes* (LV), que resultan de la combinación lineal de las discriminantes lineales (variables originales) [20].

De acuerdo con R.A. Fisher, a partir de una población cuyos elementos se dividen en dos grupos, donde para cada muestra se tienen n variables y se conoce la clase a la que pertenece, se busca una combinación lineal de las variables originales que mejor expliquen la separación de dichos elementos de un grupo con respecto del otro. Dicha combinación se denomina *discriminante lineal de Fisher* [6]. Cuando existen solamente dos clases a distinguir, el gráfico de respuesta codifica en +1 a una clase y en -1 a la otra [21].

PLS-DA es capaz de predecir diferentes variables simultáneamente, por lo cual es ampliamente utilizada en clasificación supervisada. No obstante, en ocasiones resulta importante definir el mismo número de muestras de calibración en cada clase con el fin de evitar que el modelo presente inclinación hacia una en particular [20].

Para identificar *outliers* (muestras anómalas), PLS-DA ofrece un gráfico de residuales Q (que incluso puede trazarse por clase), un gráfico de "*leverages*", o bien, de "*Hotelling* $T^2$".
"*Leverage*" es una medida que expresa la influencia de cierta muestra sobre un modelo de regresión (como lo es originalmente PLS), que puede ser interpretada como aquella distancia entre esta y su centroide vectorial [16].

### 4.6.5 Funciones de Potencia (PF)

Funciones de Potencia, desarrollado por Forina et.al. en 1991, se basa en la densidad potencial de los elementos que pertenecen al grupo de calibración [17]. A partir de ellos se calcula una función de densidad (o de probabilidad) que considera la suma de las contribuciones individuales de dichos elementos, las cuales son de tipo gaussiano y contemplan un factor de suavizado (*smoothing*) que determina su grosor [11].

De acuerdo con Coomans, D. et.al., este algoritmo considera cada muestra del conjunto de entrenamiento como un punto en el espacio alrededor del cual, hay un campo potencial que disminuye con la distancia de la muestra [31].

La clasificación de una nueva muestra está dada por el potencial acumulativo de la clase obtenido mediante la suma de los potenciales individuales de las muestras de calibración-validación interna (etapa de entrenamiento), posteriormente, la muestra de predicción se asigna a la clase que da lugar al mayor potencial acumulativo [31]. Además, la forma del campo potencial depende de la *función kernel* (gaussiano/triangular) que se emplee y del parámetro de suavizado (0.1-1.2) seleccionado para cada clase. Este último, se define mediante la validación cruzada siendo el óptimo, aquel que dé el

mínimo error de clasificación [31]. Cabe mencionar que la función *kernel* (k), es una densidad de probabilidad suave y simétrica [10].

En esta técnica, es importante definir *percentil* ya que, con base en él, se asignan las muestras de predicción a la clase que pertenecen. Éste es una medida de posición no central (tal como los cuartiles, deciles, etc.) útil para la comparación de resultados e implica un nivel de confianza estadístico [3]. Un percentil se conforma de 99 valores que dividen una serie de datos en 100 partes iguales, teniendo entonces, que un percentil equivale al 1% del total de observaciones y que, el percentil 50 ($P_{50}$) es igual a la mitad de las observaciones [3].

4.6.6 Modelación Suave e Independiente por Analogías de Clase (SIMCA)

SIMCA, introducida por Svante Wold en 1976, es una técnica que pertenece a los métodos de modelado de clase independiente, los cuales, se distinguen de los anteriores puesto que sólo consideran una clase por modelo [11].

El área de control de calidad de un producto es el principal campo de aplicación de esta técnica puesto que es capaz de discernir si las muestras pertenecen a una clase, a varias o a ninguna [15].

La técnica SIMCA, desarrolla un modelo PCA para cada clase, obteniendo los *eigenvectores* con centrado en la media o autoescalado, y posteriormente la integra calculando sus fronteras de acuerdo con una probabilidad, comúnmente del 95% [15].

Como su nombre lo indica, el estudio gráfico de cada clase se realiza de forma independiente, es decir, sin información adicional del resto de clases [11].

El algoritmo de SIMCA emplea el criterio de asignación de modelado, el cual, define una región de dominio experimental a partir del cual se puede caracterizar una clase de forma individual [45].

# 5. Experimentación

## 5.1 Diseño

Una vez establecidos el proceso de producción, el contexto socioeconómico del tequila y sentadas las bases fundamentales de esta tesis, se presentan las actividades realizadas para su elaboración.

### 5.1.1 Material

1 micropipeta Finnpipette de 50 µL
Puntas para micropipeta
1 vaso de precipitados de 50 mL
1 vaso de precipitados de 100 mL

### 5.1.2 Instrumento

Espectrofotómetro Perkin Elmer Frontier FT-IR con UATR de diamante/KRs5 (TlBr-TlI) de una reflexión, también de la marca Perkin Elmer.

### 5.1.3 Reactivos

En la tabla 4 se presentan las propiedades químicas y características físicas de los reactivos empleados.

***Tabla 4*. Propiedades de los reactivos empleados.**

| Nombre | Fórmula | Propiedades físicas y químicas. |
|---|---|---|
| Etanol absoluto |  $C_2H_6O$ | Líquido incoloro<br>Peso molecular: 46.07 g/mol<br>Pureza: 99.91 % |

### 5.1.4 Tequilas

Muestras de Tequila: 236 muestras proporcionadas por el CRT: 71 Blancos, 29 Jóvenes, 74 Reposados, 42 Añejos y 20 Extra añejos.

Con el fin de facilitar la interpretación de los modelos, se asignó una nomenclatura compuesta por dos letras mayúsculas y 4 dígitos. La primera letra fue siempre "T" correspondiente a Tequila; la segunda letra indica la clase del tequila y fue representada por las letras "B" (Blanco), "J" (Joven), "R" (Reposado), "A" (Añejo), "EA" (Extra añejo); finalmente, la clave de cuatro dígitos que fue asignada de acuerdo con la etiqueta de la muestra,

seleccionando los últimos cuatro números de esta. P. ej., si un tequila estaba etiquetado como "Tequila Joven 164723", su clave de identificación fue "TJ 4723".

**Tabla 5. Información conocida para las muestras.**

| Cantidad | Clase | Categoría | % v/v Etanol |
|---|---|---|---|
| 71 | Blancos (TB) | 18 mixtos<br>48 100 % agave<br>5 sin dato (S/D) | 35.12–54.94<br>2 S/D |
| 29 | Jóvenes (TJ) | 18 mixtos<br>1 100 % agave<br>10 S/D | 35.206–54.97 |
| 74 | Reposados (TR) | 5 mixtos<br>47 100 % agave<br>22 S/D | 35.158–54.74<br>2 S/D |
| 42 | Añejos (TA) | 1 mixto<br>21 100 % agave<br>20 S/D | 39.92–54.014 |
| 20 | Extra Añejos (TEA) | 11 100 % agave<br>9 S/D | 38–54.768 |

5.1.5 Softwares

*Spectrum*© (Perkin Elmer, Waltham, MA, U.S.A.)

*GenEx*© version 6 (MultiD Analysis AB, Goteborg, Sweden).

*Matlab*© (The MathWorks, Massachusetts, U.S.A.) con la herramienta *Classification Toolbox* (v.3.1) [12].

5.1.6 Medición de tequilas en FT-IR

1) Prender la computadora y abrir el programa *Spectrum*©.
2) Limpiar el ATR con etanol absoluto y dejar volatilizar.
3) Realizar el espectro "blanco" (aire). Al mismo tiempo, enjuagar tres veces la punta a utilizar con el tequila que se pretende medir.
4) Colocar en el diamante del ATR 50 ?L de la muestra, limpiar con etanol absoluto, colocar nuevamente y leer de 4000-400 $cm^{-1}$, 16 barridos, resolución de 4 $cm^{-1}$, y utilizando la nomenclatura descrita en el apartado 5.1.4 (Tequilas).
5) Repetir los pasos 2-4 hasta leer la última muestra.
6) Guardar los espectros en formato **.sp* y **.ascii*.

#### 5.1.7 Pretratamiento de datos (para todos los modelos construidos)

Después de obtener los espectros de infrarrojo en unidades de transmitancia para las 236 muestras de tequila, estos se transformaron a valores de absorbancia y, como acto seguido, se llevó a cabo su pretratamiento (en *Spectrum*©), descrito a continuación.

##### 5.1.7.1 Corrección de ATR (CATR)

La corrección de ATR resulta necesaria, dado que al realizar la medición de las muestras se obtienen sus espectros en valores de reflexión. Para llevarla a cabo se usó el software *Spectrum*© de la siguiente forma: Menú > Proceso > Corrección de ATR > Aceptar. Consecutivamente, los espectros se guardaron en los formatos **.sp* y **.ascii* en una nueva carpeta, para continuar con la corrección de la línea base.

##### 5.1.7.2 Corrección Interactiva de Línea Base (CILB)

El objetivo de esta corrección es evitar valores de absorbancia negativos en los espectros, además de establecer valores de 0 al inicio y al final de éstos (4000 y 450 $cm^{-1}$), para que estos tengan la misma magnitud y puedan ser comparados en condiciones de inicio.

Existen dos formas de hacer corrección de línea base, la primera la hace el programa automáticamente (Menú > Proceso > Corrección de la Línea Base); sin embargo, de esta forma suelen quedar muchos valores de absorbancia negativos, lo cual, no es muy adecuado. Por ello, se realizó de la segunda forma: Menú > Proceso > Corrección Interactiva de la Línea Base (CILB), eligiendo los siguientes puntos ($cm^{-1}$): Blancos: 4000, 3895.82, 2590.67, 1878.38, 1173.54, 958.91 y 450. Jóvenes: 4000, 3885.62, 1872.47, 1172.24, 952.07 y 450. Reposados: 4000, 3908.81, 2587.63, 1950.75, 1868.51, 957.38 y 450. Añejos: 4000, 3913.12, 1851.72, 950.65 y 450. Extra añejos: 4000, 3903.57, 1861.39, 957.93 y 450. Una vez tratados, los espectros se guardaron en los formatos **.sp* y **.ascii*.

Aquí, vale la pena mencionar, que el conjunto de puntos depende de cada clase, puesto que la complejidad química de cada una aumenta en función del tiempo de reposo, donde, surgen nuevos compuestos (al estar en contacto con las barricas) y algunos otros tienden a desaparecer. Sin mencionar que las muestras

provienen de diferentes casas tequileras, lotes y diferentes procesos de elaboración.

#### 5.1.7.3 Ajuste a cero

Este tratamiento implicó ajustar los valores de absorbancia de modo tal que en 4000 y 450 $cm^{-1}$ fuera igual a cero, con la finalidad de que los espectros puedan ser comparados.

Se hace en el archivo **ascii* (compatible con hojas de cálculo) correspondiente a cada tequila antes de construir las matrices de datos.

### 5.1.8 Elaboración de Matrices

Después de hacer el pretratamiento de los espectros, se extrajeron en un sólo documento los datos de absorbancia de todas las muestras como aparece en la *tabla 6*. En la primera columna de la hoja de cálculo se colocó la clave del tequila (TX ####, que son las muestras), en la primera fila los valores de número de onda de los espectros (4000-450 $cm^{-1}$, que son las variables), y en las celdas restantes los valores de absorbancia correspondientes a cada muestra y cada variable.

***Tabla 6*. Hoja de cálculo para la construcción de una matriz típica de datos.**

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 10 | | Tequila | 4000 | 3999 | 3998 | 3997 | 3996 |
| 11 | | TB 9 sin | 0 | -0.000011 | -0.000004 | 0.000017 | 0.000042 |
| 12 | | TB 1397 | 0 | 0.000003 | 0 | -0.000007 | -0.000013 |
| 13 | | TB 2767 | 0 | 0.000026 | 0.000069 | 0.0001 | 0.000119 |
| 14 | | TB 2858 | 0 | 0.000016 | 0.000022 | 0.000019 | 0.000012 |
| 15 | | TB 2908 | 0 | -0.000004 | 0.000016 | 0.000048 | 0.000073 |
| 16 | | TB 2916 | 0 | 0.000013 | 0.000029 | 0.000045 | 0.000059 |
| 17 | | TB 2968 | 0 | 0.000021 | 0.000034 | 0.000036 | 0.000031 |
| 18 | | TB 3052 | 0 | -0.000002 | -0.000002 | -0.000003 | -0.000001 |
| 19 | | TB 3138 | 0 | -0.000061 | -0.000109 | -0.000109 | -0.000051 |
| 20 | | TB 3382 | 0 | 0.000038 | 0.000066 | 0.000073 | 0.000066 |

La matriz para SVM en *GenEx*© fue un poco distinta a la expresada en la *tabla 6* dado que es una técnica de clasificación supervisada. En ella, se agregaron una serie de columnas en las que se define el número de clase y unos códigos de

identificación (#Texto) en las muestras de entrenamiento para indicarle al programa qué muestra corresponde a qué grupo (*tabla 7*).

***Tabla 7.*** **Columnas que deben agregarse a la matriz típica de SVM para *GenEx*©.**

| 451 | 450 | #Clase | #TB_resto | #TJ_resto | #TR_resto | #TA_resto |
|---|---|---|---|---|---|---|
| 0.000342 | 0 | 1 | 1 | 2 | 2 | 2 |
| 0.00035 | 0 | 1 | 1 | 2 | 2 | 2 |
| 0.00045 | 0 | 1 | 1 | 2 | 2 | 2 |
| 0.000411 | 0 | 2 | 2 | 1 | 2 | 2 |
| 0.000277 | 0 | 2 | 2 | 1 | 2 | 2 |
| 0.000315 | 0 | 3 | 2 | 2 | 1 | 2 |
| 0.000211 | 0 | 3 | 2 | 2 | 1 | 2 |
| 0.000274 | 0 | 4 | 2 | 2 | 2 | 1 |
| 0.000331 | 0 | 4 | 2 | 2 | 2 | 1 |

En la columna de #Clase, se colocaron los números del 1 al 4, según la siguiente dicotomía: 1-TB, 2-TJ, 3-TR y 4-TA/TEA). En las siguientes columnas (#TB_resto, #TJ_resto, etc.), se asignó el número 1 a las muestras que correspondían a la clase con que comenzaba el código de la columna y el número 2 al resto.

Además, se separaron las muestras de calibración-validación interna (entrenamiento: *training*) de las de predicción (*test*), para ello, se colocaron las de prueba hasta abajo de la matriz puesto que son minoría. Como adelanto de lo descrito en el apartado 6.1.3, las muestras tipo *test* fueron seleccionadas al azar (a partir del modelo más adecuado de PCA), de forma que abarcaran todos los vectores de la clase y representaran aproximadamente el 10% de las muestras empleadas en el modelo. En el caso de las técnicas supervisadas en *Matlab*©, no se pueden colocar las etiquetas tanto para muestras como para variables sino solamente los datos de absorbancia. Por lo cual, es necesario definir antes los números de las muestras en los modelos (primera columna de la *tabla 8*).

***Tabla 8.*** **Matriz de datos típica para *Matlab*©.**

| No. de Muestra | Tequila | 4000 | 3999 | 3998 | 3997 | 3996 | 3995 |
|---|---|---|---|---|---|---|---|
| 1 | TB 9 sin | 0 | -0.000011 | -0.000004 | 0.000017 | 0.000042 | 0.000064 |
| 2 | TB 1397 | 0 | 0.000003 | 0 | -0.000007 | -0.000013 | -0.000009 |
| 3 | TB 2767 | 0 | 0.000026 | 0.000069 | 0.0001 | 0.000119 | 0.000127 |
| 4 | TB 2858 | 0 | 0.000016 | 0.000022 | 0.000019 | 0.000012 | 0.00001 |
| 5 | TB 2908 | 0 | -0.000004 | 0.000016 | 0.000048 | 0.000073 | 0.000084 |
| 6 | TB 2916 | 0 | 0.000013 | 0.000029 | 0.000045 | 0.000059 | 0.000068 |
| 7 | TB 2968 | 0 | 0.000021 | 0.000034 | 0.000036 | 0.000031 | 0.000018 |
| 8 | TB 3052 | 0 | -0.000002 | -0.000002 | -0.000003 | -0.000001 | 0.000005 |
| 9 | TB 3138 | 0 | -0.000061 | -0.000109 | -0.000109 | -0.000051 | 0.00004 |
| 10 | TB 3382 | 0 | 0.000038 | 0.000066 | 0.000073 | 0.000066 | 0.000023 |
| 216 | TB 4743 | 0 | 0.000006 | -0.000004 | -0.000024 | -0.000048 | -0.000059 |
| 217 | TB 5429 | 0 | 0.000005 | 0.000024 | 0.000055 | 0.000089 | 0.000127 |
| 218 | TB 8650 | 0 | 0.00003 | 0.000057 | 0.000078 | 0.000079 | 0.000079 |
| 219 | TB 8680 | 0 | 0.000021 | 0.000043 | 0.00006 | 0.000072 | 0.000082 |
| 220 | TB 8692 | 0 | 0.000001 | -0.000004 | -0.000001 | -0.000009 | -0.000017 |

**Muestras de entrenamiento (*fitting* y *cross validation*)**

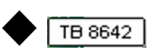
**Muestras de predicción (*test*)**

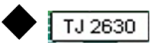
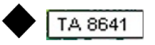
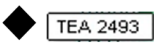
Por último, es necesario agregar también la columna #Clase en una matriz de *Matlab*© y, por única ocasión, se colocan los números de clase en todas las muestras (entrenamiento y predicción).

5.2 Construcción de Modelos

La construcción de modelos varía de acuerdo con la técnica que se desee usar y depende de si se trata de una técnica supervisada o de una no supervisada. Las técnicas no supervisadas, como PCA y CA, se realizaron en el software *GenEx*© versión 6, mientras que las técnicas supervisadas se llevaron a cabo en *GenEx*© y *Matlab*© (con la herramienta Classification Toolbox v.3.1 [12]). Las técnicas supervisadas en *GenEx*© fueron SVM y PC, y en *Matlab*©, PF, PLS-DA, PCA-DA, k-NN y SIMCA. La *tabla 9* muestra la simbología empleada para desarrollar los modelos. Cabe mencionar, que los tequilas Añejos (TA) y Extra añejos (TEA), se consideraron parte de un sólo grupo denominado "Anejos" (TA).

***Tabla 9*. Simbología de los tequilas *training*/*test* en los modelos.**

| | ***Training* *GenEx*© y *Matlab*©** | ***Test* *GenEx*©** | | ***Test* *Matlab*©** |
|---|---|---|---|---|
| | | **SVM** | **PC** | |
| **TB** |  |  |   |  |

| | | | | |
|---|---|---|---|---|
| **TJ** |  |  |   |  |
| **TR** |  |  |   |  |
| **TA** |  |  |   |  |
| **TEA** | | |   | |

Partiendo de los 236 tequilas, se formaron dos grupos que se emplearon en las técnicas supervisadas. Estos son:

- **Calibración y validación interna o *training*:** 65 TB, 26 TJ, 68 TR, 38 TA y 18 TEA.
- **Predicción o *test*:** 6 TB, 3 TJ, 6 TR, 6 TA (4 TA y 2 TEA).

Por comportamiento anómalo (*outlier*), el TJ2593 se eliminó de todos los modelos, construyéndose entonces con 235 muestras.

### 5.2.1 Introducción de datos en los softwares

#### 5.2.1.1 GenEx ©

Una vez elaboradas las matrices, se creó en GenEx© el *file* cuya extensión es **.mdf*, para lo cual se copió la matriz construida en la hoja de cálculo y se pegó en la ventana "Data editor", se guardó y se cerró dicha ventana (*figuras 13* y *14*).

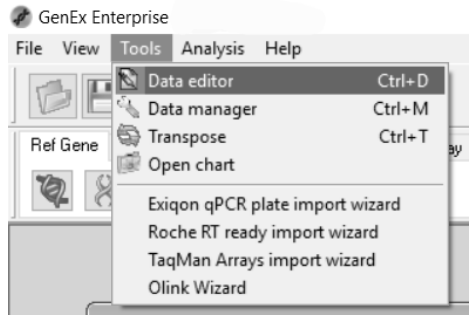



***Figura 13*. Procedimiento para abrir "Data editor".**

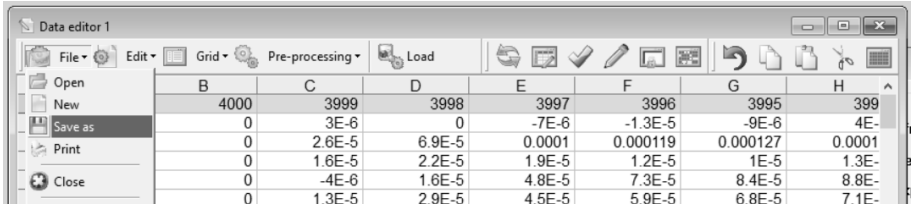



***Figura 14*. Datos en la ventana "Data editor", procedimiento para guardar el *file *.mdf*.**

Posteriormente, se abrió el *file *.mdf* como se observa en la *figura 15* que, una vez cargado, aparece como en la *figura 16*.

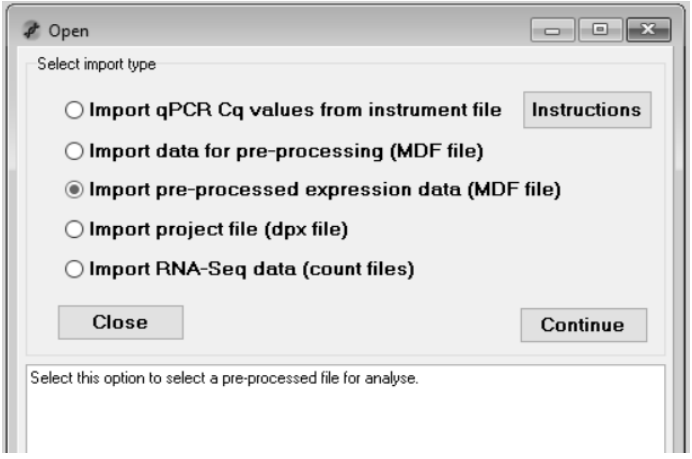



***Figura 15*. Procedimiento para abrir el *file* con extensión **.mdf*.**

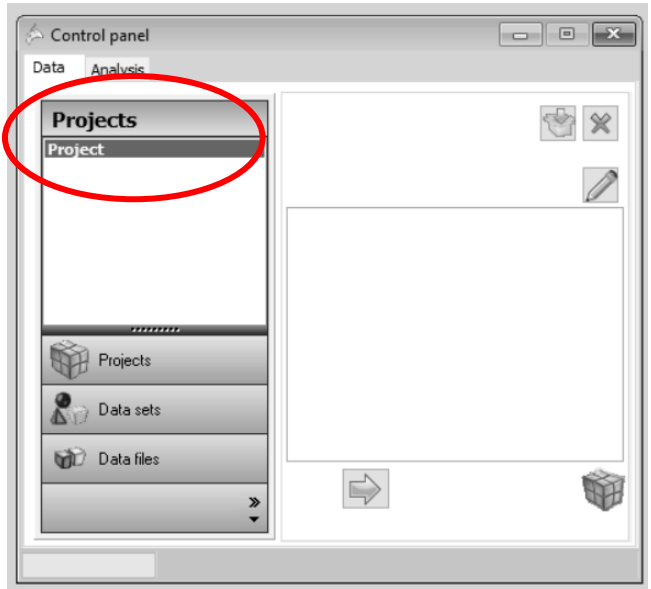



***Figura 16*. Archivo *.*mdf* cargado en *GenEx*©.**

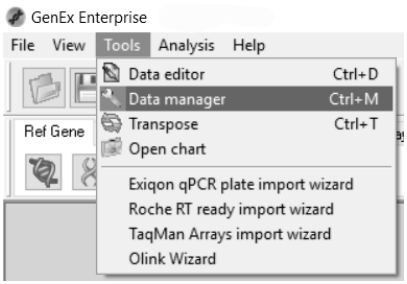



***Figura 17*. Procedimiento para abrir la ventana "*Data Manager*".**

Una vez cargado el archivo **.mdf*, se procedió a abrir "*Data manager*" (*figura 17*), para realizar los cambios pertinentes. Estos son:

- Creación de grupos para facilitar la asignación de simbología a cada clase (*figura 18*). Para lo cual se debe ir a la pestaña "*Groups*" y agregar los grupos TB, TJ, TR y TA dando clic en el signo de "+" cada que se escribe un grupo nuevo.

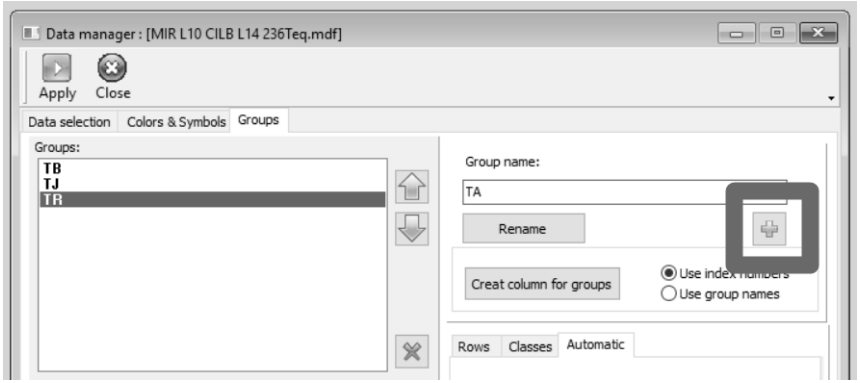



***Figura 18.*** **Fisionomía de la pestaña "*Groups*".**

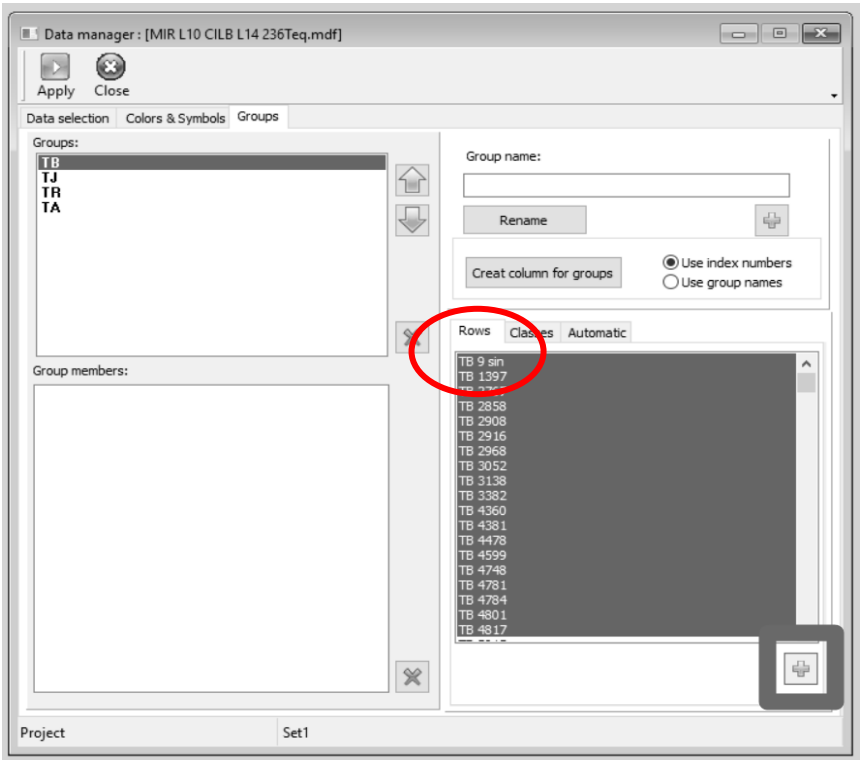



***Figura 19.*** **Procedimiento para agregar muestras a un determinado grupo.**

- Posterior a esto, en la pestaña "*Rows*" se seleccionan las muestras que pertenecen a un grupo particular (teniendo este seleccionado en color azul) y se da clic en el signo "+", como se indica en la *figura 19*.

Cabe mencionar que cuando se trata de una técnica supervisada (PC y SVM), las muestras seleccionadas para validación, no se deben incluir en los grupos. Posteriormente, se configuró la simbología asignada a cada grupo (*figura 20*):


***Figura 20*. Procedimiento de configuración de la simbología de cada tequila.**

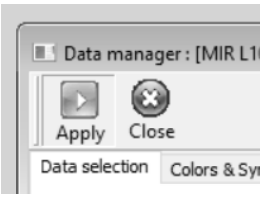



***Figura 21*. Cómo aplicar los cambios realizados al *file*.**

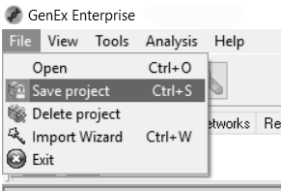



***Figura 22*. Cómo guardar un proyecto en *GenEx©*.**

Una vez asignada la simbología poniendo especial atención en los recuadros de la *figura 20*, se aplicaron los cambios dando clic en el botón con la leyenda "*Apply*" (*figura 21*), y se guardó el proyecto (*figura 22*) cuya extensión es **.dpx*.

5.2.1.2 Matlab©

Para introducir las matrices de datos en Matlab, primero debemos conocer las ventanas de la interfaz de Matlab (*figura 23*):

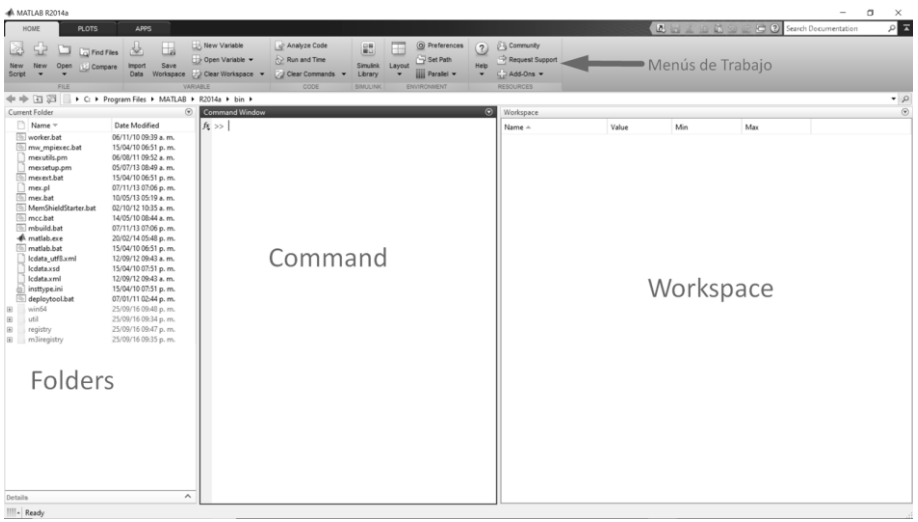



***Figura 23*. Secciones que conforman el software *Matlab*©.**

El primer paso fue dar clic en la ventana "*Command*" y escribir los comandos para crear los archivos con extensión **.mat* que se utilizaron como matrices. Hay que destacar, que se construyeron dos matrices por grupo (dos para calibración-validación interna y dos para predicción). Con más detalle, fue necesario construir una matriz exclusivamente con los datos de absorbancia medidos y otra con las clases, ambas para el conjunto *training*, pero esto mismo para el conjunto *test*.

Al dar clic en la sección "*Command Window*" se procedió a escribir *training_data=[ ];*, seguido de un *enter*, lo cual, creó automáticamente el archivo que aparece en la sección "*Workspace*" tal como se observa en la *figura 24*. Posteriormente, al dar doble clic sobre este nuevo archivo, se abrió otra sección con aspecto de una hoja de cálculo ordinaria (*figura 25*). En esta nueva ventana se pegaron las lecturas de absorbancia (sin etiquetas) de las muestras de entrenamiento.

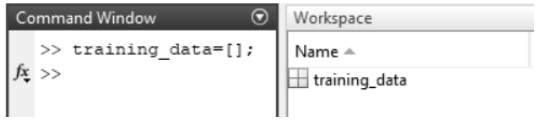



***Figura 24*. Procedimiento de creación del archivo "training_data" en *Matlab*©.**

21x3551 double

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 6.0000e-06 | -4.0000e-06 | -2.4000e-05 | -4.8000e-05 | -5.9000e-05 | -7.7000e-05 | -9.0000e-05 | -9.3000e-05 | -8.0000e-05 |
| 2 | 0 | 5.0000e-06 | 2.4000e-05 | 5.5000e-05 | 8.9000e-05 | 1.2700e-04 | 1.4000e-04 | 1.3200e-04 | 1.1000e-04 | 8.8000e-05 |
| 3 | 0 | 3.0000e-05 | 5.7000e-05 | 7.8000e-05 | 7.9000e-05 | 7.9000e-05 | 7.4000e-05 | 6.6000e-05 | 6.4000e-05 | 7.0000e-05 |
| 4 | 0 | 2.1000e-05 | 4.3000e-05 | 6.0000e-05 | 7.2000e-05 | 8.2000e-05 | 8.6000e-05 | 8.1000e-05 | 6.7000e-05 | 5.3000e-05 |
| 5 | 0 | 1.0000e-06 | -4.0000e-06 | -1.0000e-06 | -9.0000e-06 | -1.7000e-05 | -2.4000e-05 | -2.2000e-05 | -6.0000e-06 | 2.2000e-05 |
| 6 | 0 | 1.2000e-05 | 2.6000e-05 | 4.1000e-05 | 4.9000e-05 | 4.7000e-05 | 3.4000e-05 | 1.7000e-05 | 8.0000e-06 | 1.2000e-05 |
| 7 | 0 | 2.3000e-05 | 4.7000e-05 | 8.0000e-05 | 1.1400e-04 | 1.2600e-04 | 9.6000e-05 | 3.8000e-05 | -4.0000e-06 | 3.0000e-06 |
| 8 | 0 | -1.3000e-05 | -2.3000e-05 | -1.1000e-05 | 2.4000e-05 | 5.2000e-05 | 9.7000e-05 | 1.2300e-04 | 1.2300e-04 | 1.0100e-04 |
| 9 | 0 | 1.3000e-05 | 2.3000e-05 | 2.4000e-05 | 1.6000e-05 | 6.0000e-06 | 2.0000e-06 | 8.0000e-06 | 2.7000e-05 | 4.8000e-05 |
| 10 | 0 | -1.0000e-05 | -1.6000e-05 | -1.8000e-05 | -1.4000e-05 | -5.0000e-06 | 5.0000e-06 | 1.0000e-05 | 1.0000e-05 | 1.0000e-05 |
| 11 | 0 | 8.0000e-06 | 1.8000e-05 | 3.2000e-05 | 5.0000e-05 | 6.6000e-05 | 7.3000e-05 | 6.4000e-05 | 4.4000e-05 | 2.5000e-05 |
| 12 | 0 | -1.0000e-05 | -9.0000e-06 | -4.0000e-06 | 4.0000e-06 | 1.1000e-05 | 1.2000e-05 | 7.0000e-06 | 0 | -6.0000e-06 |
| 13 | 0 | 7.0000e-06 | -1.0000e-06 | -1.5000e-05 | -2.3000e-05 | -2.5000e-05 | -2.7000e-05 | -1.6000e-05 | -2.7000e-05 | -3.3000e-05 |
| 14 | 0 | 1.4000e-05 | 2.5000e-05 | 3.2000e-05 | 6.5000e-05 | 7.3000e-05 | 7.6000e-05 | 7.6000e-05 | 7.2000e-05 | 6.3000e-05 |
| 15 | 0 | -7.0000e-06 | -2.3000e-05 | -4.0000e-05 | -4.9000e-05 | -4.2000e-05 | -2.2000e-05 | 0 | 1.3000e-05 | 1.3000e-05 |
| 16 | 0 | 0 | 5.0000e-06 | 1.0000e-05 | 1.0000e-05 | 7.0000e-06 | 1.1000e-05 | 2.8000e-05 | 6.0000e-05 | 9.7000e-05 |
| 17 | 0 | -3.5000e-05 | -7.0000e-05 | -9.3000e-05 | -9.5000e-05 | -7.0000e-05 | -1.6000e-05 | 5.5000e-05 | 1.2500e-04 | 1.6800e-04 |
| 18 | 0 | 3.0000e-06 | 7.0000e-06 | 8.0000e-06 | 4.0000e-06 | 0 | 1.0000e-06 | 9.0000e-06 | 2.6000e-05 | 4.9000e-05 |
| 19 | 0 | 1.4000e-05 | 3.2000e-05 | 5.5000e-05 | 7.9000e-05 | 9.3000e-05 | 9.0000e-05 | 7.3000e-05 | 5.4000e-05 | 4.2000e-05 |

***Figura 25*. Matriz de datos típica en la interfaz de *Matlab*©.**

Se repitió el proceso de escritura de los comandos para crear una nueva matriz, pero esta vez con el texto "training_class" (que contuvo en una sola columna los valores numéricos de #Clase), otra con "test_data" (igual a training_data, salvo que aquí estuvieron los valores de absorbancia del conjunto de prueba) y otra con "test_class".

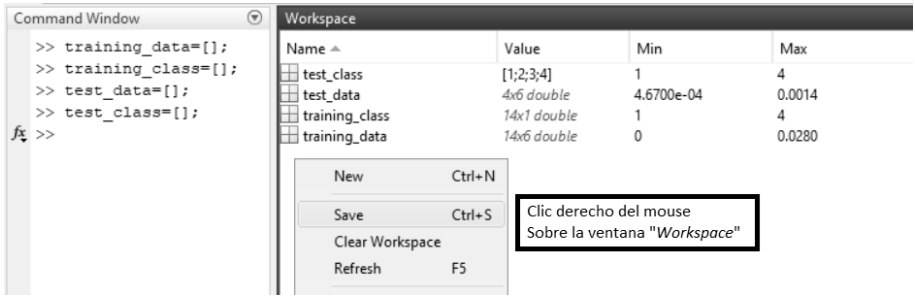



***Figura 26*. Comandos, matrices y cómo guardar los archivos en la interfaz de *Matlab*©.**

En la *figura 26* se observan los comandos escritos, las matrices creadas y el procedimiento para guardar los archivos en extensión **.mat* (conjunto de datos compatible con *Matlab*©).

Si se tiene un archivo **.mat* previamente construido y se desea trabajar con él, lo único que hay que hacer es ubicar dicho archivo en la ventana "*Folders*" y dar doble clic sobre él (*figura 27*).

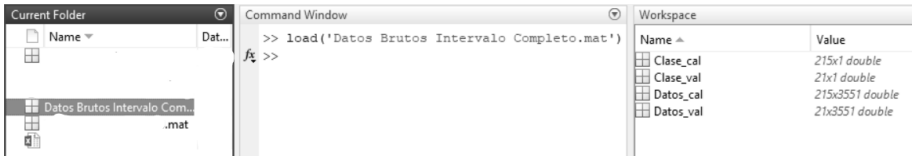



***Figura 27*. Cómo abrir un archivo **.mat* ya establecido.**

# 6. Resultados y discusión

Esta sección se dividió en cuatro capítulos con la intención de facilitar su lectura, por lo que va aumentando gradualmente su complejidad.

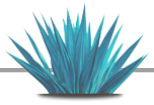
Comienza con la exploración de técnicas no supervisadas (Capítulo I), continua con la construcción de modelos de clasificación mediante técnicas supervisadas de *GenEx*© (Capítulo II), seguidos de la obtención de los desarrollados en *Matlab*© (Capítulo III).

Finalmente, se muestran los resultados de la proyección y predicción de dos muestras identificadas como formulaciones fisicoquímicas de tequila en los modelos seleccionados como satisfactorios (Capítulo IV).

# C A P Í T U L O  I
# Exploración de técnicas no supervisadas en *GenEx*©

En este capítulo se plantea el procedimiento de obtención de modelos adecuados de clasificación de las técnicas no supervisadas: Análisis de Componentes Principales (**PCA**), Análisis Cluster (**CA**) y Rotación Varimax (**RV**) construidos con el programa *GenEx*©.

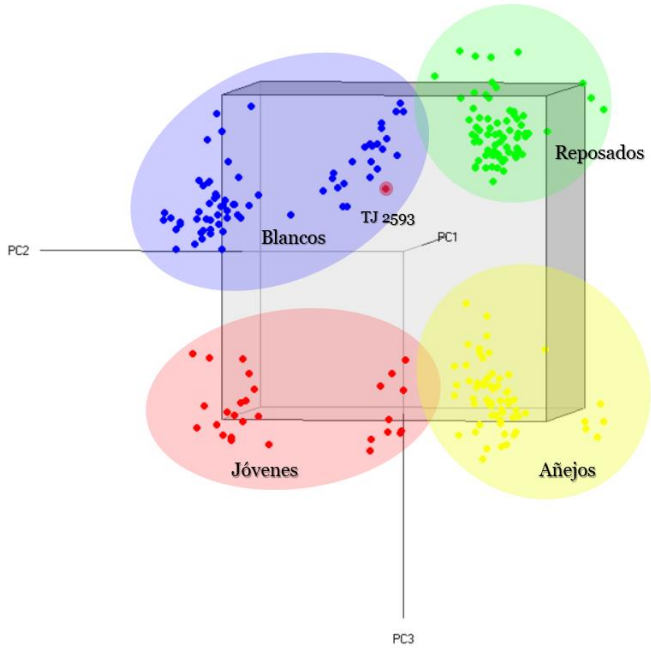
Se decidió iniciar con estas técnicas puesto que, como se dijo en el marco teórico, son técnicas de interpretación sencilla, bastante relacionadas entre sí que, además, arrojan información de gran utilidad para la construcción de modelos quimiométricos supervisados.



6.1 Análisis de Componentes Principales (PCA)

El análisis de componentes principales permitió distinguir pautas entre las clases de tequila empleadas, estas son: Blancos, Jóvenes, Reposados y Añejos.

Es importante mencionar que no es el primer estudio realizado con este propósito [8]; sin embargo, el actual considera una cantidad alta de muestras auténticas de diferentes casas tequileras, lotes y procesos de producción. Además, considera los tequilas jóvenes, utiliza el autoescalado y define un intervalo óptimo de estudio.



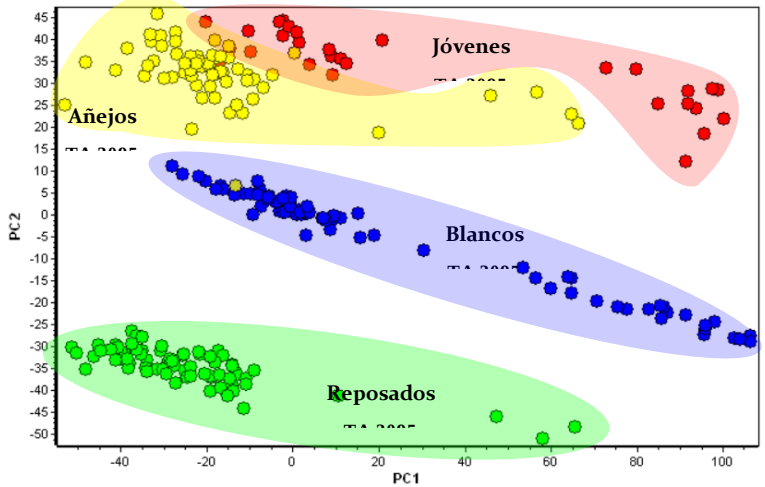

***Gráfica 1.*** **PCA: espacio 3D (PC1-PC2-PC3), centrado en la media, 236 tequilas, intervalo completo (4000–450 $cm^{-1}$).**

Al utilizar centrado en la media se observó que el modelo era útil sólo en el espacio de tres dimensiones PC1-PC2-PC3 (*gráfica 1*), donde, la varianza explicada era del 94.55 % para PC1, para PC2 del 3.37 % y del 1.45 % para PC3.

Con este mismo escalado, el modelo en 2D (PC2-PC3) permitió la divergencia de las clases con una varianza explicada acumulada del 4.82 % y, por tanto, una varianza residual del 95.18 %. Esto es importante puesto que únicamente el 4.82

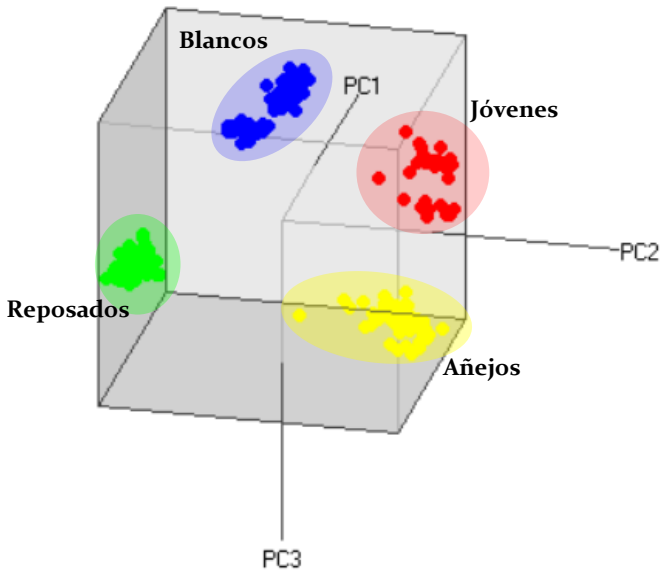
% de la varianza del modelo justifica la formación de los vectores que diferencian las clases, y el 95.18 % restante corresponde a información "basura", lo cual resulta poco conveniente.

En cambio, al utilizar autoescalado, se observó una buena tendencia de clasificación en el subespacio PC1-PC2 (cuyas varianzas son las más elevadas); sin embargo, no se trataba de un modelo satisfactorio puesto que hay traslape entre tequilas añejos, blancos y jóvenes (*gráfica 2*).




***Gráfica 2*. PCA: subespacio 2D (PC1 vs. PC2), autoescalado, 236 tequilas, intervalo completo.**

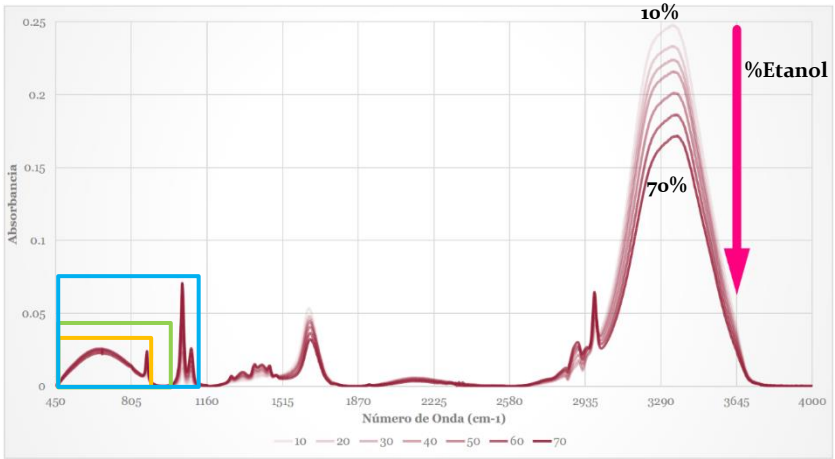
Al comparar los resultados obtenidos con ambos escalados, se encontró mejor el construido con datos autoescalados, donde, la varianza explicada acumulada con dos componentes (modelo de la *gráfica 2*) fue del 70.92 %, y del 87.95 % con tres componentes (*gráfica 3*).




***Gráfica 3*. PCA: Subespacio 3D (PC1 vs. PC2 vs. PC3), autoescalado, 236 tequilas, intervalo completo (4000–450 $cm^{-1}$).**

Tras obtener una varianza residual del 12.05 %, se procedió a optimizar el modelo de forma que la varianza explicada con tres componentes aumentara y que el componente involucrado con la divergencia de las clases tuviera el mayor valor de varianza. Lo cual, implicó explorar diferentes intervalos de número de onda.

Una primera propuesta, fue eliminar el conjunto de variables de 4000–3000 $cm^{-1}$, dado que en esta zona se encontró la contribución más importante de los enlaces –OH de los alcoholes y del agua. Información que fue corroborada mediante la medición del espectro IR de diferentes proporciones EtOH-$H_2O$ (*gráfica 4*).




***Gráfica 4*. Espectros de absorbancia de diferentes disoluciones EtOH-$H_2O$.**

La tendencia general observada fue que cuanto más alcohol tenía una mezcla, la banda entre 3000 y 3700 tenía menor absorbancia.

Esto se justificó por la influencia de los enlaces –OH de las moléculas de agua que podrían presentar fuerzas intermoleculares fuertes (puentes de H), provocando una vibración de mayor intensidad y, por lo tanto, mayor absorbancia en aquellas disoluciones con menor cantidad de etanol (y mayor de agua). En el presente estudio, al tratarse de muestras de tequila con diferentes grados de alcohol, esta banda no nos arrojó más información que lo descrito anteriormente.

En primera instancia, se propuso eliminar la región de números de onda entre 3000–4000 $cm^{-1}$, a partir de lo cual, el modelo construido con tres PC's tuvo una varianza explicada del 93.51% (y, por tanto, 6.49% de varianza residual).

Al no tener aún un modelo por clases de tequila satisfactorio, se buscó eliminar una segunda región del espectro que pudiera significar información no relevante para el objetivo del modelo. Estas regiones fueron de 850-450, 950-450 y 1100-450 (*gráfica 4*).

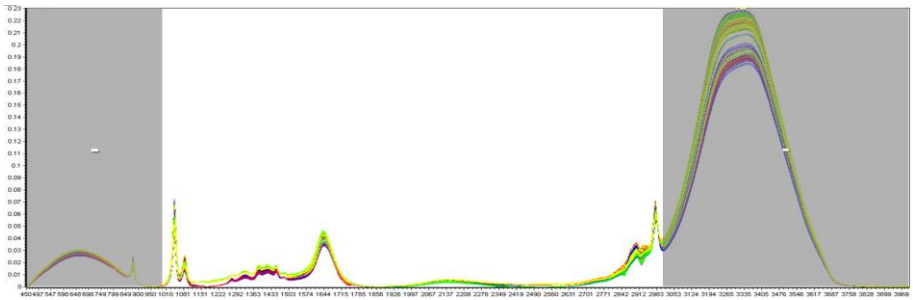
En la *tabla 10*, se enlistan las varianzas por intervalo analizado y los componentes involucrados con la separación por clases.

***Tabla 10*. Estudio de los modelos con diferentes intervalos de número de onda.**

| Intervalo de trabajo ($cm^{-1}$) | Varianza explicada acumulada (3 PC's) | Varianza explicada por componente | Componentes involucrados en la separación por clase | Desventajas o inconvenientes del modelo |
|---|---|---|---|---|
| **4000-450 (completo)** | 87.95% | PC1: 46.53%<br>PC2: 24.39%<br>PC3: 17.03% | PC2, pero no separa TA's y TJ's. | En todos los subespacios hay pautas internas y traslapes. |
| **3000-450** | 93.51% | PC1: 42.83%<br>PC2: 31.40%<br>PC3: 19.28% | PC1, PC2 y PC3 | Hay traslapes entre las clases. |
| **3000-850** | 93.18% | PC1: 40.33%<br>PC2: 30.33%<br>PC3: 22.52% | PC1 y PC2 | Los TA's con mayor grado de alcohol se traslapan con TJ's. |
| **3000-950** | 93.59% | PC1: 41.85%<br>PC2: 29.99%<br>PC3: 21.75% | PC1 y PC2 | Los TA's con mayor grado de alcohol se traslapan con TJ's. |
| **3000-1100** | 94.36% | PC1: 45.07%<br>PC2: 31.95%<br>PC3: 17.34% | PC1 y PC2 | Ninguna. Ventaja: identifica a TJ2593 como outlier. |

Con base en la tabla anterior, se decidió que el intervalo más adecuado de trabajo era de 3000–1100 $cm^{-1}$ (área en blanco de la *gráfica 5*), ya que tiene la mayor varianza explicada con tres componentes (94.36%), además, PC1 y PC2 también tienen la mayor varianza y PC3 la menor. Esto es sumamente importante puesto que PC1 y PC2 son los que permiten la divergencia de las clases.

En la *gráfica 5*, se indican en color gris traslúcido las partes del espectro IR que se eliminaron en el modelo final (4000–3001 $cm^{-1}$ y 1099–450 $cm^{-1}$).

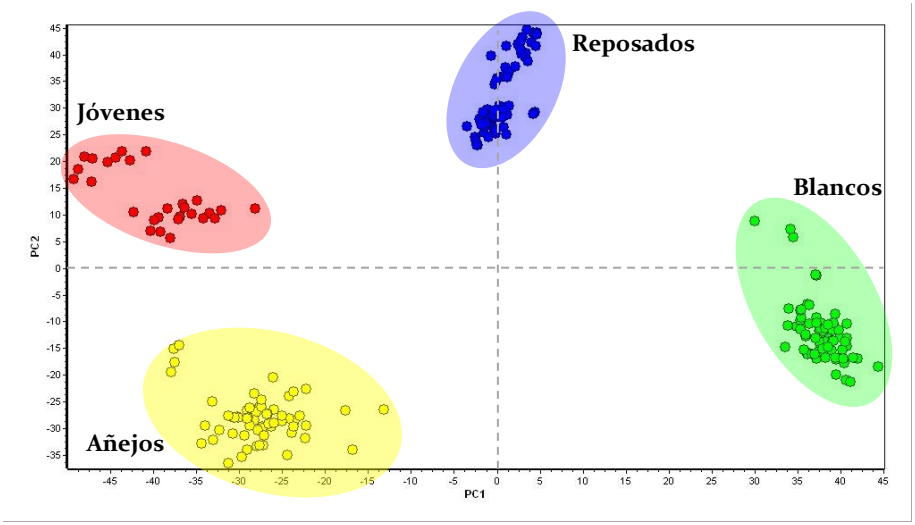


***Gráfica 5*. Espectros de 236 tequilas con datos brutos y sin escalado.**

### 6.1.1 Modelo final

El modelo más adecuado de PCA fue construido con las variables de los números de onda 3000-1100 $cm^{-1}$, datos autoescalados y 235 tequilas. Éste, descartó el TJ2593 como miembro de los tequilas jóvenes puesto que presentó comportamiento de *outlier*.

En la *gráfica 6*, se observó la formación de cuatro grupos muy bien separados que correspondieron a las clases estipuladas al principio (tequilas blancos, jóvenes, reposados y añejos). PC1 y PC2 fueron los componentes responsables de la divergencia de las clases.

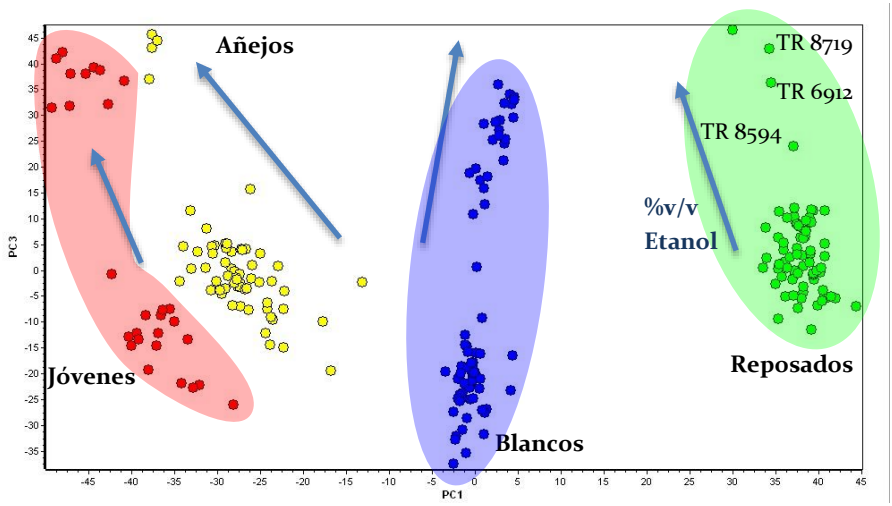



***Gráfica 6*. PCA: Subespacio PC1-PC2, intervalo 3000-1100 $cm^{-1}$, autoescalado, 235 tequilas (varianza explicada del 94.61 %).**

Si se observa esta misma gráfica y se trazan líneas imaginarias en los valores de PC1= 0 y PC2= 0, se dice que los TJ's se encuentran en valores negativos de PC1 y positivos de PC2, los TA's en valores de PC1 y PC2 negativos, TR's en PC1 positivos pero PC2 negativos (salvo los tequilas TR5590, TR8719 y TR6912, que entran al cuadrante de los TB's) y, finalmente, los TB's que están en valores de PC1 y PC2 positivos aunque, por su pauta interna, invaden el cuadrante de los TJ's.

El modelo tuvo una varianza explicada para el PC1 igual al 45.12 %, del 32.12 % para PC2 y de 17.37 % para PC3. Lo cual, da lugar a una varianza acumulada del 94.61 % y una varianza residual del 5.39 %. Esto quiere decir que el modelo que considera los tres primeros PC's (*gráfica 9*) explica el 94.61 % de la información de las muestras e ignora u omite el resto.

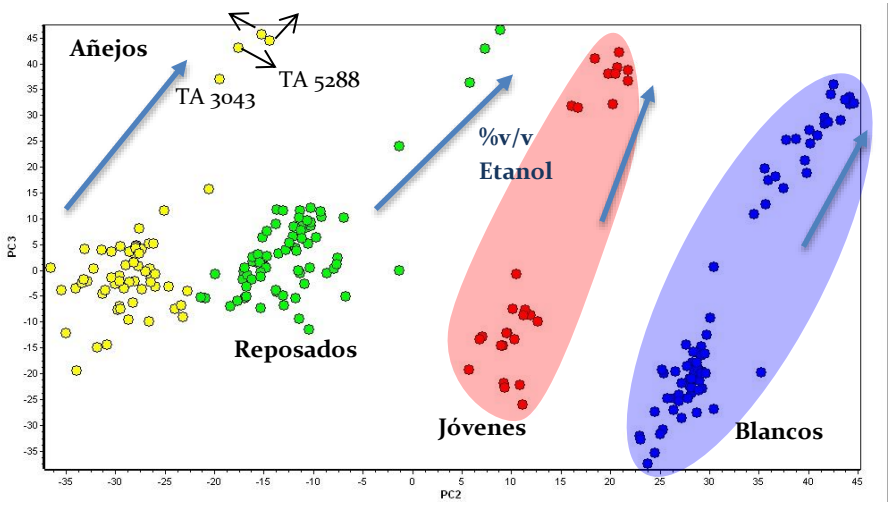
Tras graficar PC1 vs. PC3 (*gráfica 7*) y PC2 vs. PC3 (*gráfica 8*), fue posible aseverar que el componente 3 fue el encargado de generar pautas internas más marcadas en cada clase. Las cuales, se produjeron por el porcentaje de alcohol contenido en cada muestra. Es decir, aquellas que se ubicaron en valores de *scores* de PC3 mayores a 15 (en el caso de los TJ's y TA's), eran las que tenían por lo menos cincuenta grados de alcohol.



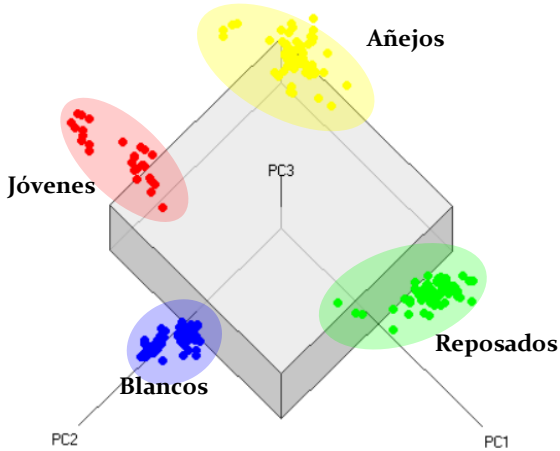

***Gráfica 7.*** **PCA: Subespacio PC1-PC3, intervalo 3000-1100 cm$^{-1}$, autoescalado, 235 tequilas.**

Las muestras TR5590, TR6912, TR8594 y TR8719, que son las de la coletilla de las *gráficas 7* y *8*, oscilaban entre 49.59 y 54.74% v/v de etanol. Los TA3043, TA5288, TA5427 y TEA2591, que son los ubicados en *scores* de PC3> 30, tenían más de 50 %v/v de alcohol.

En el caso de los tequilas blancos y jóvenes, el grupo de tequilas, con mayor volumen de alcohol (49.39-54.93% y 54.88-54.95%, respectivamente), fue un poco más grande: TB2968, TB3052, TB4360, TB4748, TB4817, TB5164, TB5412, TB5428, TB5429, TB5451, TB5546, TB6848, TB6872, TB6908, TB6910, TB8606, TB8624, TB8627, TB8638, TB8642, TB8650, TB8692, TB8718, TJ2551, TJ2650, TJ4778, TJ4945, TJ6001, TJ6870, TJ8244, TJ8605, TJ8733 y TJ8765.




***Gráfica 8.*** **PCA: Subespacio PC2-PC3, intervalo 3000-1100 cm$^{-1}$, autoescalado, 235 tequilas.**

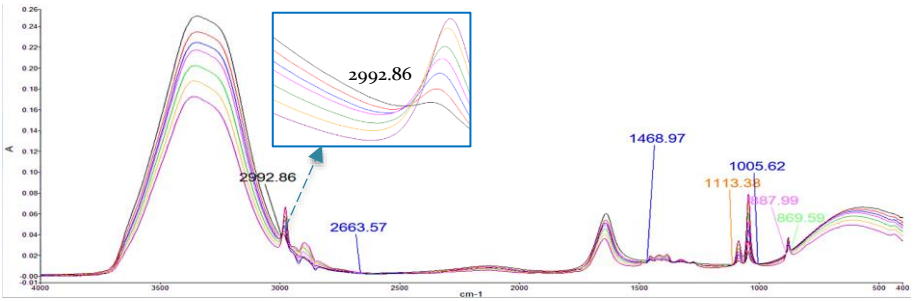



***Gráfica 9.*** **PCA: Subespacio PC1-PC2-PC3, intervalo 3000-1100 cm$^{-1}$, autoescalado, 235 tequilas.**

Con la finalidad de no restringir el alcance del modelo, o bien, para no omitir información de las clases, ninguna de estas muestras se eliminó del modelo.

En la *gráfica 9* se presenta un esquema de PCA en tres dimensiones (PC1-PC2-PC3) donde se puede ver satisfactoriamente la divergencia de todas las clases, aun con la pauta interna que presenta cada una.

Con base en la *gráfica 10*, fue posible identificar siete puntos isosbésticos: 2992.86, 2663.57, 1468.97, 1113.38, 1005.62, 887.99 y 869.59 $cm^{-1}$.



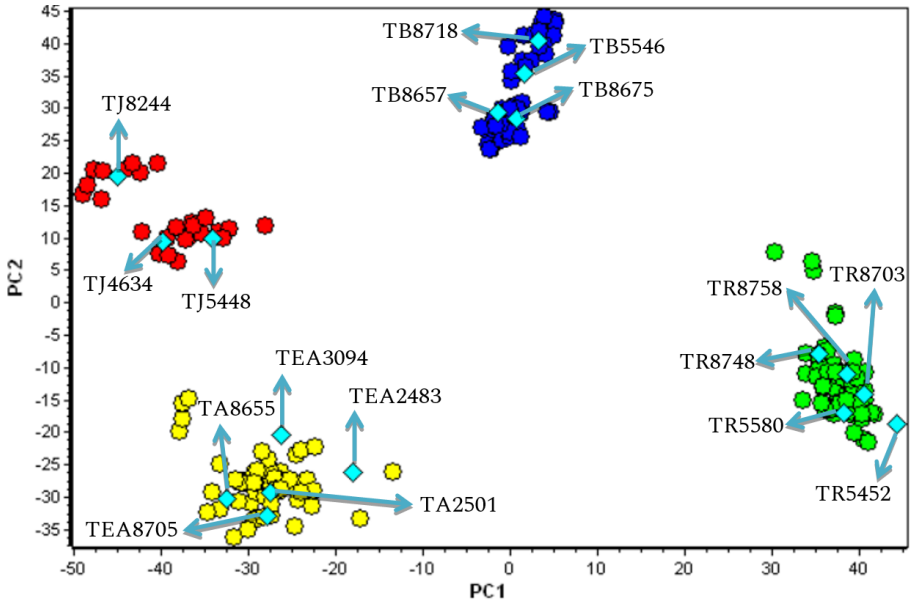

***Gráfica 10*. Espectros IR de disoluciones EtOH-H2O con rótulos de los puntos isosbésticos.**

En este punto, cabe mencionar que, aunque la banda más relacionada con las vibraciones del enlace –OH (etanol y agua) está en la región de 3700-3000 $cm^{-1}$, hay influencia de él en todo el espectro.

### 6.1.2 Selección de muestras de prueba o predicción (*test*)

Una vez obtenido el modelo final de PCA (235 tequilas, 1901 variables por muestra y datos autoescalados), se seleccionaron las muestras a emplear como *test* en modelos supervisados, las cuales deben ubicarse lo más dispersas posible y abarcar todas las direcciones vectoriales de la clase, además, deben representar aproximadamente el 10% del total.

Subsecuentemente, se hizo una matriz "nueva" en la que las muestras seleccionadas se ubicaron hasta abajo en la hoja de cálculo (proceso descrito en el apartado 5.1.8). Las muestras seleccionadas fueron: cuatro tequilas blancos (TB5546, TB8657, TB8675 y TB8718), tres tequilas jóvenes (TJ4634, TJ5448, TJ8244), cinco tequilas reposados (TR5452, TR5580, TR8703, TR8748 y TR8758), dos tequilas añejos (TA2501 y TA8655) y tres tequilas extra añejos (TEA2483, TEA3094 y TEA8705). En la *gráfica 11* se señalan las muestras de predicción de acuerdo con la simbología asignada en la *tabla 6* (rombo azul aqua).




***Gráfica 11.*** **PCA: subespacio PC1-PC2 con 235 tequilas y rótulos en tequilas para *test*.**

### 6.2 Rotación Varimax (RV)

Hasta ahora, se ha planteado el proceso de obtención del modelo más adecuado de PCA observando las muestras (*scores*); sin embargo, otra parte importante del Análisis de Componentes Principales es la interpretación de los *loadings* para identificar las variables (números de onda) con mayor peso. La importancia de estos números de onda es que pueden ser vinculados con la estructura química de los compuestos que determinan la diferencia entre las clases de tequila.

Para empezar, se graficaron los obtenidos por PCA, contra un espectro de cada clase de tequilas bien comportados (*gráfica 12*), donde se encontró que el *loading 1* se asocia principalmente al intervalo 2950–1880 $cm^{-1}$, el *loading 2* se centra en los intervalos 1850-1110 y 2800–2200 $cm^{-1}$; el *loading 3* de 2990–2800, 2300-1950 y las bandas en 1640 y 1390 $cm^{-1}$; posteriormente, el *loading 4* está definido por la banda de 1880 $cm^{-1}$ y, finalmente, el *loading 5* por las bandas en 2930, 2850 y 1750 $cm^{-1}$.

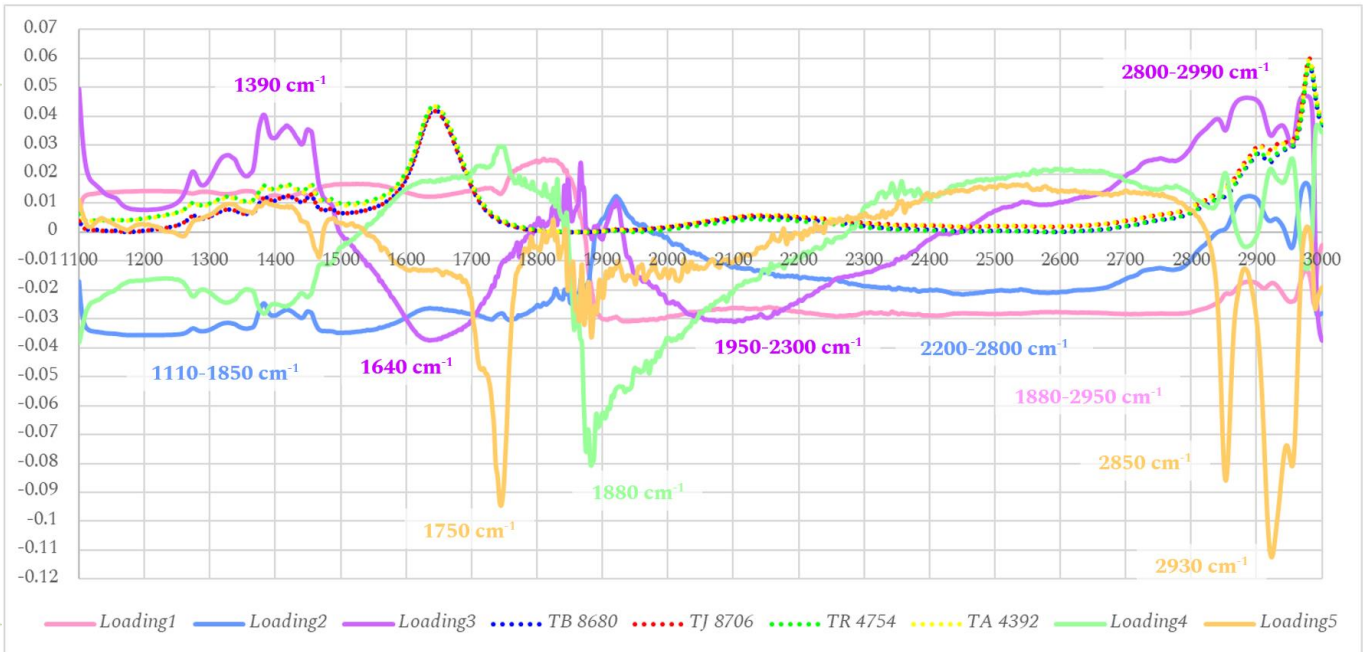
Hay que enfatizar que el análisis de *loadings* por PCA es complicado puesto que la escala que estos tienen es muy cercana a cero (de -0.115 a 0.05), lo cual, no permite apreciar completamente los intervalos o las bandas que estudia cada componente principal. Esto se debe a la normalización interna del algoritmo y al número tan elevado de variables espectrales, no obstante, lo importante es el perfil observado de cada *loading*.

Se conoce, de acuerdo con el marco teórico, que Rotación Varimax es útil en casos como este donde la interpretación de los *loadings* resulta un poco confusa. Por ello, se decidió aplicarlo al modelo de la gráfica 6 (definido como óptimo) y comparar los primeros cinco factores no rotados con los obtenidos por PCA.

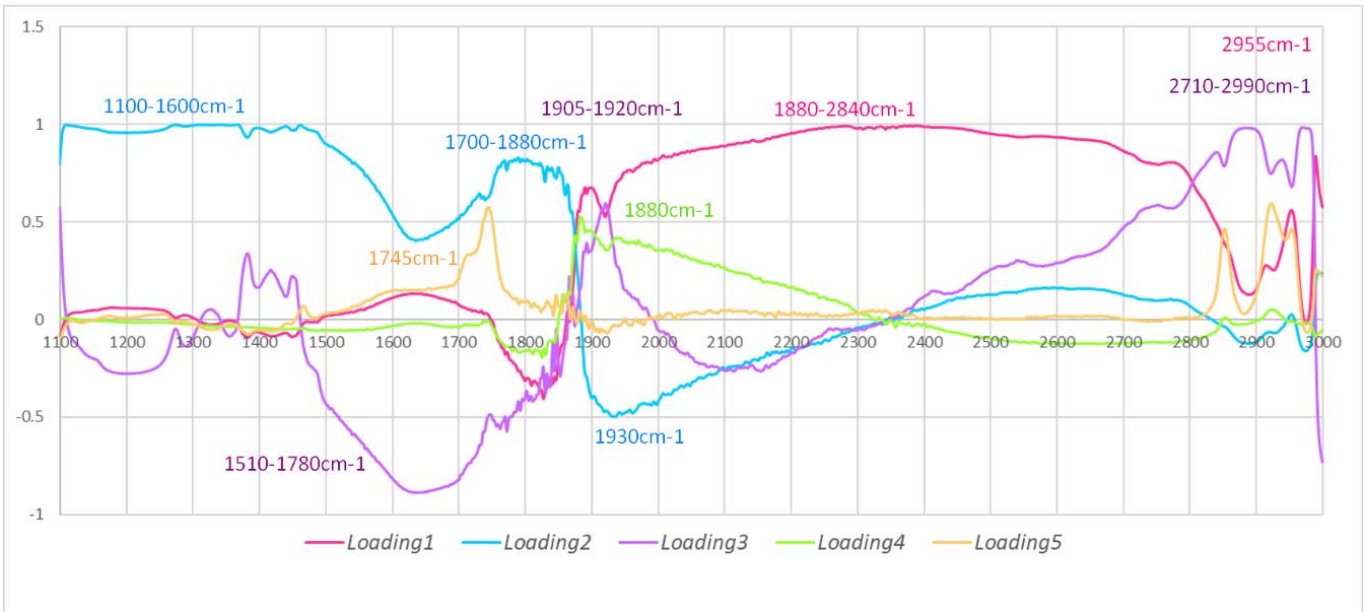
En la *gráfica 13*, son más evidentes las bandas o intervalos de número de onda a las que estos dan importancia, ya que los *loadings* relevantes se han maximizado en valores de -1 a 1 por normalización interna del algoritmo. Se observó que el *loading 1* está asociado al intervalo 2840–1880 $cm^{-1}$ y, en menor proporción, a la banda en 2955 $cm^{-1}$; el *loading 2*, en orden decreciente, de 1600–1100, 1880–1700 $cm^{-1}$ y la banda en 1930 $cm^{-1}$ y el *loading 3*, también de mayor a menor relevancia, de 2990–2710, 1780–1510 y 1920–1905 $cm^{-1}$. Por otro lado, el

*loading 4* destaca la señal de 1880 $cm^{-1}$ y en el *loading* 5 las de 1745, 2850 y 2930 $cm^{-1}$.

Al realizar una comparación de los valores con tablas de MIR para la identificación de grupos funcionales orgánicos, se encontró que el PC1 rotado, en el intervalo entre 1880-2840 $cm^{-1}$ puede asociarse a compuestos aromáticos con heteroátomos (derivados furfurales) mientras que la banda en torno a 2950-2990 $cm^{-1}$ corresponde a la tensión simétrica del enlace C-H en grupos $CH_3$ y/o $CH_2$; en relación al PC2 rotado, la región entre 1100 y 1600 $cm^{-1}$, incluye varios grupos funcionales tales como de alcoholes (primarios y secundarios) que aparecen entre 1000-1100 $cm^{-1}$. Alrededor de 1000 $cm^{-1}$ se ubica una banda correspondiente a la vibración C-H en compuestos con heteroátomos típica que podría relacionarse con los furfurales presentes en los tequilas. La banda de 1930 $cm^{-1}$ podría deberse también a la aromaticidad de compuestos con heteroátomos. El intervalo de 1700-1880 $cm^{-1}$ corresponde esencialmente a compuestos con grupo carbonilo de todo tipo; p.ej. ácidos alifáticos: 1700-1725, éteres y cetonas: 1600-1700 $cm^{-1}$, aldehídos alifáticos: 1720-1740 $cm^{-1}$, ésteres: 1650-1790 $cm^{-1}$. Los grupos funcionales mencionados pertenecen a los diversos componentes de un tequila, por lo que podemos constatar que la diferencia entre las clases de tequila está relacionada con los componentes químicos presentes en las bebidas.

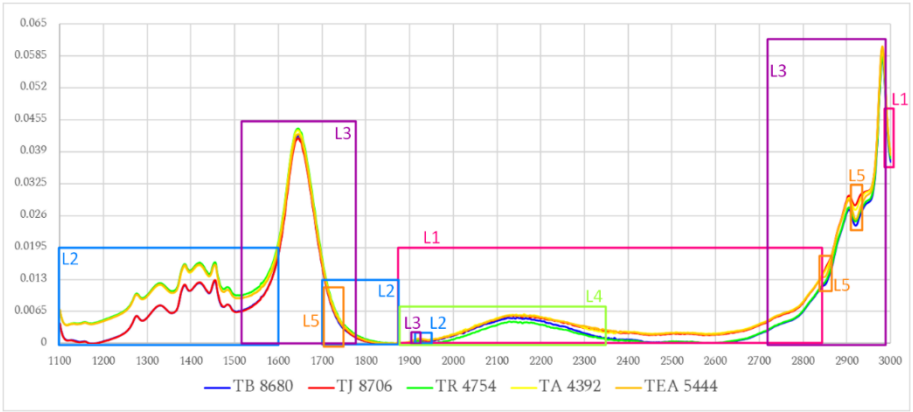



***Gráfica 12*. Análisis de *loadings* a partir del PCA en condiciones óptimas.**




***Gráfica 13. Loadings* obtenidos mediante la aplicación de Rotación Varimax al modelo óptimo.**

En la *gráfica 14*, se señalaron las zonas que estudia cada *loading*. Vale la pena recordar que los factores (o *loadings* del PC# rotado) 1 y 2 son los responsables de la separación por clase en los modelos de PCA.




***Gráfica 14*. Perfiles de un tequila bien comportado de cada clase en el intervalo 1100-3000 $cm^{-1}$ y relación con los *loadings* más relevantes de cada PC rotado (factor).**

En forma de resumen, con rotación Varimax, fue posible determinar que el PC1 rotado involucra las variables entre 2840-1880 y 2955 $cm^{-1}$ con una varianza explicada del 46.26 %; PC2 rotado aquellas entre 1600-1100, 1880-1700 y 1930 $cm^{-1}$ con una varianza del 32.85 % y que PC3 rotado, con una varianza del 15.54 %, las de 2710-2990, 780-1510 y 1905-1920 $cm^{-1}$. Aunque con varianzas muy bajas (1.68 % y 1.49 %, respectivamente) y pese a que no fueron usados en ningún modelo, PC4 rotado se debe a la banda de 1880 y PC5 rotado a las variables en 1745, 2850 y 2920 $cm^{-1}$.

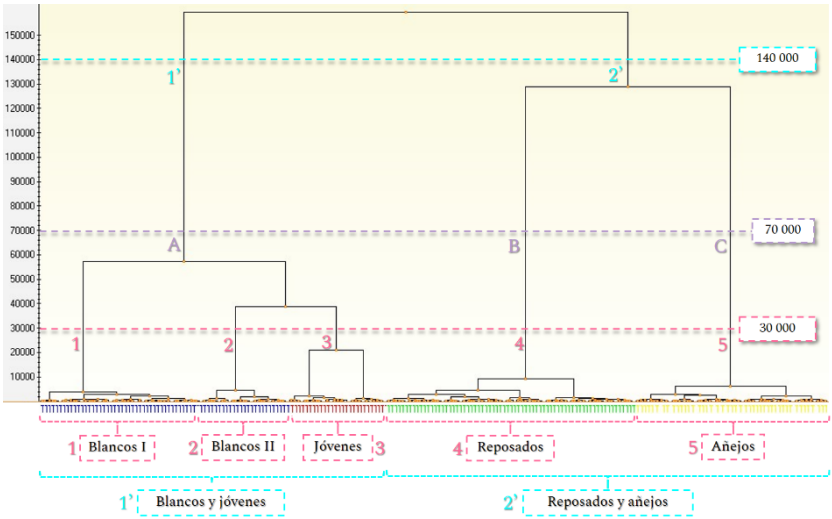
### 6.3 Análisis Cluster (CA)

Para empezar, hay que decir que el análisis cluster es una técnica no supervisada de tipo exploratoria, o sea, capaz de señalar pautas en un grupo de datos.

Este análisis puede realizarse de dos formas: con la matriz de datos original en las condiciones pertinentes (autoescalado, 235 tequilas e intervalo 3000-1100 $cm^{-1}$), o con la matriz de *scores* que resulta del modelo óptimo de PCA. En este

caso, se ocupó la matriz original, donde el método de agrupamiento (o de acoplamiento) que arrojó mejores resultados fue el Algoritmo de Ward.

A continuación, se estudiaron los dendrogramas con algoritmo de Ward y distancias Manhattan, euclídea y euclidiana cuadrada (*gráficas 15-17*).




***Gráfica 15*. Dendrograma de 235 tequilas con autoescalado, mét. de Ward y dist. Manhattan.**

En el caso del dendrograma obtenido con método de Ward y distancia Manhattan (*gráfica 15*), se encontraron tres distancias importantes (en las unidades propias de las métricas utilizadas): la primera de ellas en 140 000, que distinguió dos grupos (1': blancos/jóvenes y 2': reposados/añejos); la segunda en 70 000, que permitió identificar tres grupos (A: blancos y jóvenes, B: reposados y C: añejos) y la tercera en 30 000, donde se identificaron cinco grupos (1: blancos I, 2: blancos II, 3: jóvenes, 4: reposados y 5: añejos).

Posteriormente, se identificaron y estudiaron los miembros de los grupos "blancos I" y "blancos II" para tratar de entender su separación (*tabla 11*).

***Tabla 11*. Tequilas blancos conglomerados en las distancias Manhattan y euclídea.**

| Tequila | Manhattan | | Euclídea | |
|---|---|---|---|---|
| | Gpo. I | Gpo. II | Gpo. I | Gpo. II |
| TB9 | ✓ | | ✓ | |
| TB1397 | ✓ | | ✓ | |
| TB2767 | ✓ | | ✓ | |
| TB2858 | ✓ | | ✓ | |

| Tequila | Manhattan | | Euclídea | |
|---|---|---|---|---|
| | Gpo. I | Gpo. II | Gpo. I | Gpo. II |
| TB5738 | ✓ | | ✓ | |
| TB5972 | ✓ | | ✓ | |
| TB6838 | ✓ | | ✓ | |
| TB6848 | | ✓ | | ✓ |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| TB2908 | | ✓ | ✓ | | TB6872 | | ✓ | | ✓ |
| TB2916 | ✓ | | ✓ | | TB6873 | ✓ | | ✓ | |
| TB2968 | | ✓ | | ✓ | TB6900 | ✓ | | ✓ | |
| TB3052 | | ✓ | | ✓ | TB6907 | ✓ | | ✓ | |
| TB3138 | ✓ | | ✓ | | TB6908 | | ✓ | | ✓ |
| TB3382 | ✓ | | ✓ | | TB6910 | | ✓ | | ✓ |
| TB4360 | | ✓ | | ✓ | TB6913 | ✓ | | ✓ | |
| TB 4381 | ✓ | | ✓ | | TB6917 | ✓ | | ✓ | |
| TB 4478 | ✓ | | ✓ | | TB6930 | ✓ | | ✓ | |
| TB4743 | ✓ | | ✓ | | TB6933 | ✓ | | ✓ | |
| TB4748 | | ✓ | | ✓ | TB6940 | ✓ | | ✓ | |
| TB4781 | ✓ | | ✓ | | TB6941 | ✓ | | ✓ | |
| TB4784 | ✓ | | ✓ | | TB8598 | ✓ | | ✓ | |
| TB4801 | ✓ | | ✓ | | TB8606 | | ✓ | | ✓ |
| TB4817 | | ✓ | | ✓ | TB8609 | ✓ | | ✓ | |
| TB5015 | ✓ | | ✓ | | TB8624 | | ✓ | | ✓ |
| TB5164 | | ✓ | | ✓ | TB8627 | | ✓ | | ✓ |
| TB5318 | ✓ | | ✓ | | TB8632 | ✓ | | ✓ | |
| TB5401 | ✓ | | ✓ | | TB8635 | ✓ | | ✓ | |
| TB5403 | ✓ | | ✓ | | TB8638 | | ✓ | | ✓ |
| TB5405 | ✓ | | ✓ | | TB8642 | | ✓ | | ✓ |
| TB5406 | ✓ | | ✓ | | TB8650 | | ✓ | | ✓ |
| TB5412 | | ✓ | | ✓ | TB8657 | ✓ | | ✓ | |
| TB5428 | | ✓ | | ✓ | TB8675 | ✓ | | ✓ | |
| TB5429 | | ✓ | | ✓ | TB8680 | ✓ | | ✓ | |
| TB5451 | | ✓ | | ✓ | TB8687 | ✓ | | ✓ | |
| TB5501 | ✓ | | ✓ | | TB8692 | | ✓ | | ✓ |
| TB5546 | | ✓ | | ✓ | TB8718 | | ✓ | | ✓ |
| TB5558 | | | ✓ | | TB8739 | ✓ | | ✓ | |
| TB5588 | ✓ | | ✓ | | TB 8762 | ✓ | | ✓ | |

Tras analizar una lista de tequilas con sus grados de alcohol, así como su categoría (proporcionada por el CRT) y realizar una comparación con las muestras en cada grupo, se encontró que las pertenecientes al grupo II tenían entre 49.39 y 54.93 % de alcohol, mientras que las ubicadas en el grupo I, de 35.12 a 42.71 % volúmenes de alcohol.

Al igual que con la distancia Manhattan, se examinó el grupo II de tequilas blancos para la distancia euclídea (o euclidiana), con lo cual se encontró que la única diferencia entre ambas distancias es que una coloca la muestra TB2908 en el grupo II y la otra no (*tabla 11*).

En el caso del dendrograma construido por el método de Ward y distancia euclidiana, se establecieron tres distancias representativas (*gráfica 16*):

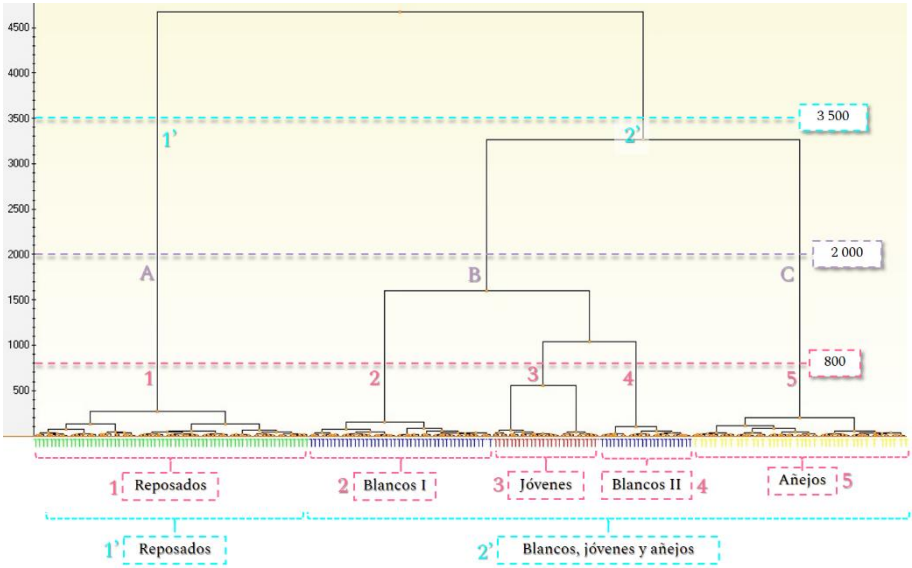
1. 3 500, donde se observaron dos grupos: 1'-reposados y 2'-blancos/jóvenes/añejos.

2. 2 000 con formación de tres grupos: A-reposados, B-blancos/jóvenes y C-añejos.
3. 800 con cinco grupos: 1-reposados, 2-blancos I, 3-jóvenes, 4-blancos II y 5-añejos.

En este punto, es importante recalcar que las escalas de distancia en cada dendrograma construido por una distancia diferente no son comparables entre sí (en cuanto a su valor grande o pequeño), puesto que cada una tiene lugar por su métrica. Es decir, su análisis debe ser independiente.

La mejora que ofrece este dendrograma con respecto al de la distancia Manhattan es que fue posible separar inmediatamente reposados y añejos del resto (en una distancia de aproximadamente 3500) lo cual podría tener importancia comercial ya que son los de mayor costo.

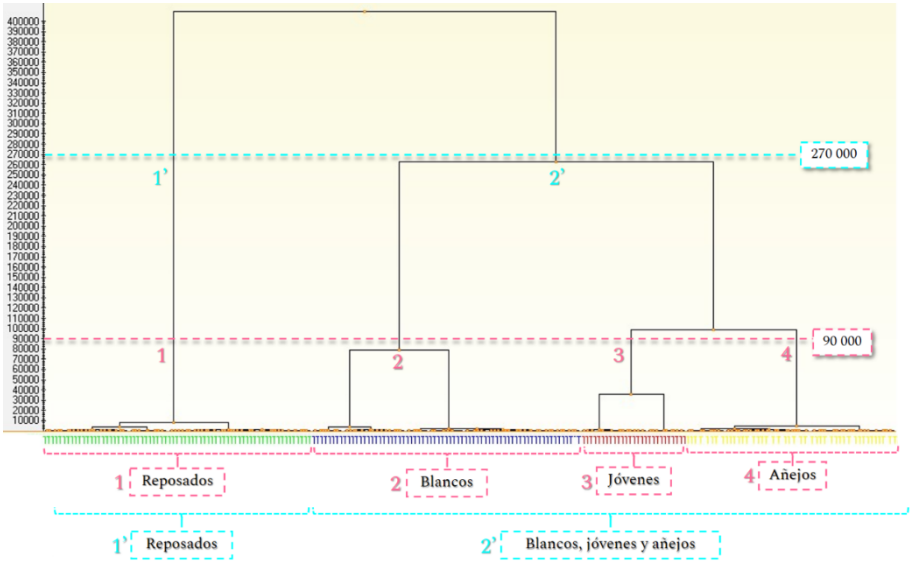
Más detalladamente, los tequilas reposados son completamente agrupados en una distancia euclídea de 273.69 y los añejos en 206.13. Mientras que, con Manhattan, los reposados se conglomeran en la distancia 9 170.42 y los añejos en 6 173.89. Sin embargo, lo realmente importante fue que aún no era posible separar todas las clases adecuadamente.




***Gráfica 16.*** **Dendrograma de 235 tequilas con autoescalado, mét. de Ward y dist. euclidiana.**

Por último, se propuso realizar el análisis cluster con el método de Ward y la distancia euclidiana cuadrada. A continuación, se exhiben los resultados.

Al emplear distancia euclidiana cuadrada (con algoritmo de Ward) fue posible separar satisfactoriamente todas las clases de tequila involucradas en el modelo en una distancia de 90 000 a partir de una matriz de 235 tequilas y datos autoescalados (*gráfica 17*).




***Gráfica 17.*** **Dendrograma con 235 teq., autoescalado, met. de Ward y dist. euclidiana cuadrada.**

De este gráfico, hay que destacar las siguientes distancias euclidianas cuadradas:

1. Aquella en 270 000 que engloba los grupos 1' (correspondiente a tequilas reposados) y 2', que considera a tequilas blancos, jóvenes y añejos.
2. La de magnitud igual a 90 000 (que expresa la diferencia entre clases), donde se encontraron cuatro conglomerados: 1-reposados, 2-blancos, 3-jóvenes y 4-añejos.

Cabe mencionar que, aunque se utilice la distancia de 90 000 para describir el modelo óptimo, en realidad, la distancia en la que se unen todos los tequilas blancos es equivalente a 79 206.55, para los jóvenes de 36 140.74, igual a 7 772.54 para los reposados y de 4 632.84 para los añejos.

Finalmente, se propone como modelo a utilizar, mediante la técnica de análisis cluster, el obtenido con método de Ward y distancia euclidiana cuadrada, dado que fueron las condiciones que lograron la separación entre todas las clases y

un solo conjunto de tequilas blancos. No obstante, se presentan los tres resultados puesto que también son excelentes modelos.

**Resumen Capítulo I** – PCA, RV y CA

Para el Análisis de Componentes Principales (PCA), las condiciones óptimas de trabajo fueron: 235 tequilas (sin TJ2593), intervalo de variables (número de onda) 3000 – 1100 $cm^{-1}$ y datos autoescalados; lo cual, se definió a partir de la revisión del escalado centrado en la media y de la exploración de diferentes intervalos con autoescalado. Este modelo arrojó resultados bastante buenos para generar una idea del comportamiento de las muestras (ya que es una técnica exploratoria no supervisada), tales como la identificación de pautas internas debidas al contenido alcohólico de las muestras, la observación de puntos isosbésticos y la selección de muestras de predicción. Cabe mencionar que, aunque la banda más relacionada con las vibraciones del enlace –OH (etanol y agua) está en la región de 3700-3000 $cm^{-1}$, hay influencia de él en todo el espectro. Por ello, la observación de puntos isosbésticos.

Rotación Varimax (RV) se utilizó en las mismas condiciones de PCA con el objetivo de interpretar más adecuadamente los *loadings*, lo cual, resultó de gran utilidad puesto que permitió identificar las zonas del espectro que estudia cada componente principal de forma más satisfactoria (en comparación con PCA). Se determinó que el PC1 rotado involucra las variables entre 2840-1880 y 2955 $cm^{-1}$ con una varianza explicada del 46.26 %; PC2 rotado aquellas entre 1600-1100, 1880-1700 y 1930 $cm^{-1}$ con una varianza del 32.85 % y que PC3 rotado, con una varianza del 15.54 %, las de 2990-2710, 780-1510 y 1920-1905 $cm^{-1}$. Aunque con varianzas muy bajas (1.68 % y 1.49 %, respectivamente) y pese a que no fueron usados en ningún modelo, PC4 rotado se debe a la banda de 1880 y PC5 rotado a las variables en 2920, 1745 y 2850 $cm^{-1}$.
Por Análisis de Conglomerados o Análisis Cluster (CA), se encontró que el método de Ward, en combinación con la distancia euclidiana cuadrada, permitía una excelente diferenciación de las cuatro clases (TB, TJ, TR y TA). Sin embargo, también con las distancias Manhattan y euclidiana simple se obtenían buenos resultados, siendo su único inconveniente, la separación de los tequilas blancos en dos subgrupos (debidos a los grados de alcohol contenidos en cada muestra). Para desarrollar este análisis, se ocupó la misma matriz que PCA, en las mismas condiciones.

# CAPÍTULO II
# Modelos supervisados con *GenEx©*

En este capítulo se exhiben los modelos de clasificación construidos en el programa *GenEx©* mediante las técnicas supervisadas: Curvas de Potencia (**PC**) y *Support Vector Machines* (**SVM**).

Estos fueron construidos a partir de las condiciones establecidas en el capítulo anterior; estas son: 218 tequilas en *training*, 17 tequilas *test*, intervalo de números de onda (variables) 3000–1100 $cm^{-1}$, autoescalado previo. Las muestras de predicción fueron: TB5546, TB8650, TB8675, TB8718, TJ4634, TJ5448, TJ8244, TR5452, TR5580, TR8703, TR8748, TR8758, TA2848, TA4769, TA8640, TA8655 y TEA8705.

### 6.4 *Support Vector Machines* (SVM)

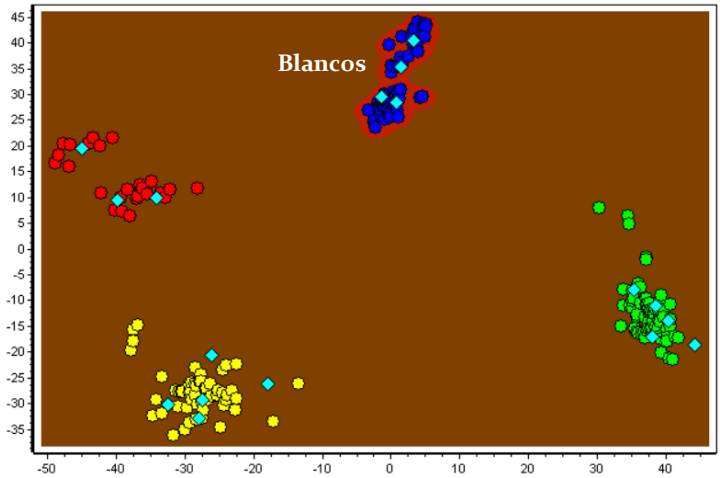
Esta técnica, no tiene una traducción formal todavía, por ello, se presenta su nombre en inglés. Vale la pena recordar, que SVM admite utilizar un *kernel* lineal, gaussiano o polinomial, dependiendo del tipo de datos con que se trabaje.

Dado que los grupos tienen un comportamiento gaussiano, el presente estudio se llevó a cabo con una *función kernel* RBF (Radial Basis Function) tipo gaussiana y en la modalidad *one-vs-all*, o sea, una clase contra las demás.

Para generar los modelos de clasificación por tiempo de añejamiento, se utilizó un valor de $\sigma = 2$ para los tequilas blancos y jóvenes, un valor de $\sigma = 3$ para los tequilas reposados y añejos, una penalización del error de clasificación (C) de 100 y 1000 iteraciones; partiendo de la matriz con 235 tequilas, en el intervalo 3000–1100 $cm^{-1}$, con datos previamente autoescalados. Cabe mencionar, que los valores empleados en los parámetros de construcción de los modelos fueron elegidos después de realizar diversas pruebas.

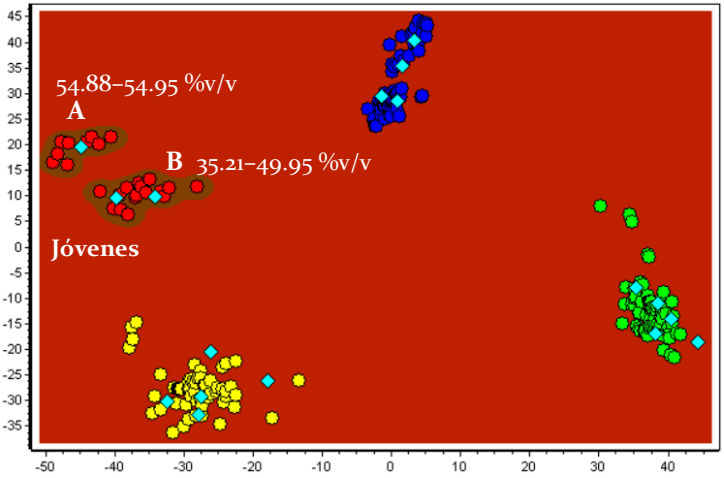
Como se observa en la *gráfica 18*, todas las muestras de tequila blanco del grupo de validación fueron correctamente asignadas a su clase. Estas fueron representadas con un rombo azul aqua.

Cabe mencionar que el algoritmo de *Support Vector Machines* genera un hiperplano (o hipersuperficie, representada por el área de color rojo en la *gráfica 18*), en el que se ubican las muestras que corresponden a la clase, además, la zona de color café representa el resto de las clases, es decir, todo lo que no corresponde al grupo en estudio.




***Gráfica 18.*** **Esquema *one-vs-all* de los TB's contra el resto, σ= 2, C= 100 y 1000 iteraciones.**

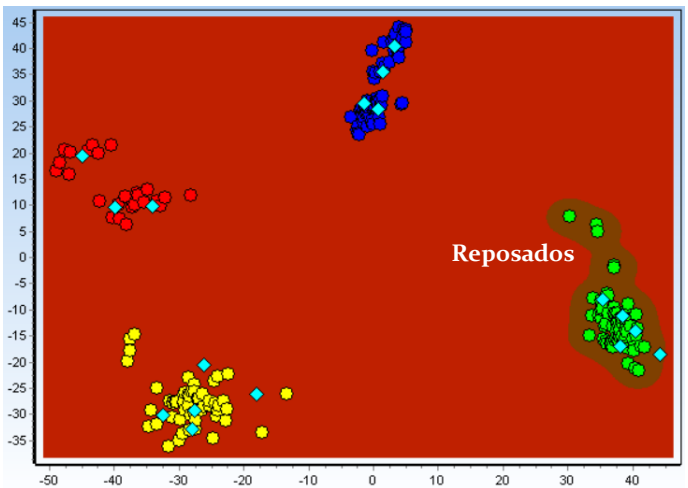
En el caso de los tequilas jóvenes (*gráfica 19*), las muestras de validación también fueron determinadas satisfactoriamente; sin embargo, el hiperplano (región de color café) se dividió en dos subgrupos. Por lo cual, se procedió a identificar los tequilas (*tabla 12*).




***Gráfica 19.*** **Esquema *one-vs-all* de los TJ's contra el resto, σ= 2, C= 100 y 1000 iteraciones.**

Cabe destacar que, en el gráfico del análisis de tequilas jóvenes contra el resto mediante SVM, el hiperplano fue de color café y la zona de las demás clases de color rojo (al inverso de los tequilas blancos). Esto no implica nada, se interpretan de la misma forma: diferencia entre regiones por su color.

Después de identificar las muestras de cada subgrupo y hacer un cotejo de las ubicadas en el subgrupo A con una lista de grados de alcohol, se encontró que los miembros del subgrupo A corresponden a porcentajes en volumen de alcohol mayores del 54.7 %v/v, mientras que, los del subgrupo B tenían entre 35.2 y 49.96 %v/v (*tabla 12*).




***Gráfica 20*. Esquema *one-vs-all* de los TR's vs. el resto, σ= 3, C= 100 y 1000 iteraciones.**

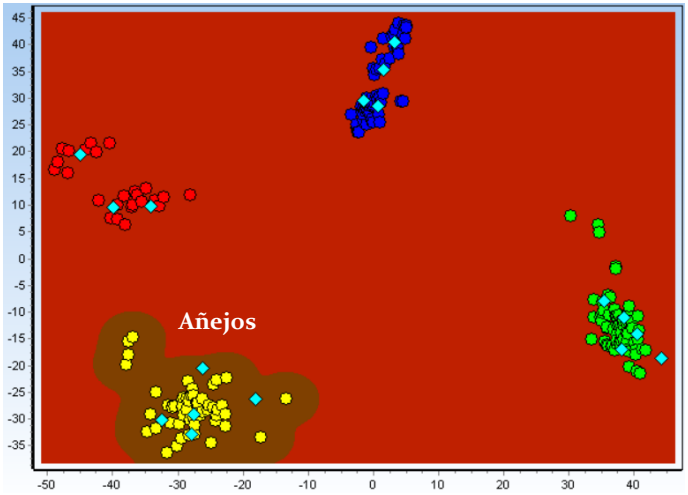
El análisis mediante *Support Vector Machines* (SVM) para los tequilas reposados se llevó a cabo con un σ= 3, C= 100 y 1000 iteraciones, dando excelentes resultados, puesto que las cinco muestras de validación quedaron dentro del hiperplano construido (*gráfica 20*).

***Tabla 12*. Tequilas jóvenes del subgrupo A con sus %v/v de alcohol.**

| Subgrupo A | | Subgrupo B | |
|---|---|---|---|
| **Tequilas** | **% v/v EtOH** | **Tequilas** | **% v/v EtOH** |

| | | | |
|---|---|---|---|
| **TJ4778, TJ4945, TJ6001, TJ6870, TJ8244, TJ8605, TJ8733, TJ 8765.** | 54.73–54.97 | **TJ4634, TJ4757, TJ4764, TJ4872, TJ5400, TJ5448, TJ6862, TJ6871, TJ6914.** | 35.2–49.96 |
| **TJ2521, TJ2650** | S/D | **TJ2371, TJ2550, TJ2576, TJ2636, TJ2656, TJ2829, TJ2839, TJ5466, TJ8706.** | S/D |

De igual forma, el modelo para los tequilas añejos (con □= 3, C= 100 y 1000 iteraciones) fue satisfactorio puesto que todas las muestras de validación fueron correctamente asignadas (*gráfica 21*).

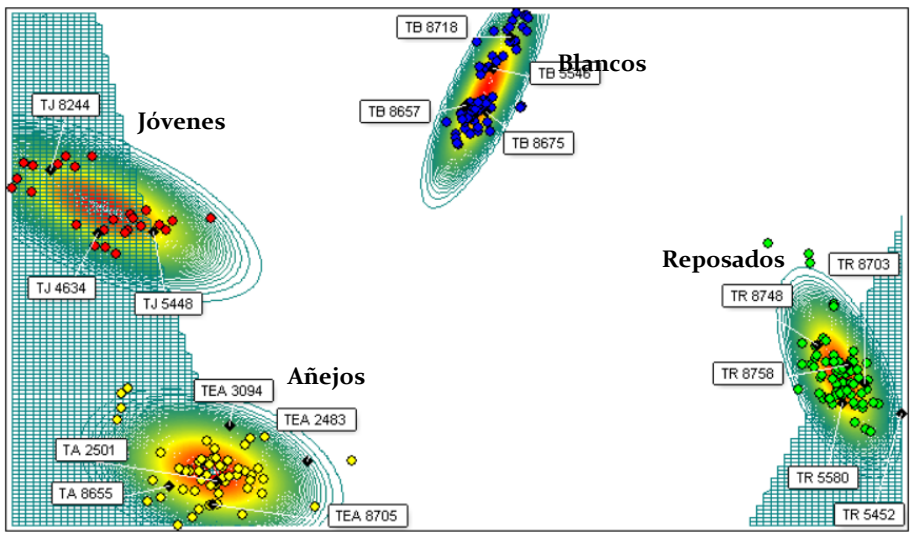



***Gráfica 21*. Esquema *one-vs-all* de los TA's contra el resto, σ= 3, C= 100 y 1000 iteraciones.**

6.5 Curvas de Potencia (PC)

Curvas de potencia se llevó a cabo con la misma matriz que SVM (235 tequilas, intervalo 3000–1100 $cm^{-1}$ y datos previamente autoescalados), en el subespacio PC1 vs. PC2 que por *default* genera *GenEx*©, aunque también puede hacerse con PC1-PC3, PC2-PC3, etcétera. Éste se construyó con 100 niveles y con un valor de 130 en grosor (*fitness*), donde, la zona de menor probabilidad se representó con el color ■, la probabilidad media con color ■ y finalmente, la de mayor probabilidad con color ■.

En la *gráfica 22*, se sitúan cuatro elipses de iso-probabilidad (una por clase) establecidas con el conjunto de calibración, a partir de las cuales, es posible confirmar o descartar la pertenencia de cierta (s) muestra (s) a una clase.




***Gráfica 22*. Elipses de iso-probabilidad en el subespacio PC1-PC2 mediante Curvas de Potencia (PC).**

Como puede apreciarse en la *gráfica 22*, todas las muestras de validación fueron correctamente asignadas, puesto que se encuentran dentro de las elipses de iso-probabilidad, a excepción del tequila TR5452 que se ubicó fuera.

Algunas muestras de entrenamiento cayeron fuera de las elipses de iso-probabilidad (TB4599, TB8710, TR5590, TR6912, TR8719, TA5427, TEA2591 y TEA2847), que, como se vio en el análisis por PCA, son muestras con contenido alcohólico mayor a 50 %.

Tras haber realizado el análisis PCA, ya sabíamos que el TR5452 podría ser anómalo, lo cual, a través de la *tabla 13* se confirma que se trata de un *outlier* puesto que tuvo 0 % de probabilidad de pertenencia a todas las clases, incluso la que se supone ser suya.

***Tabla 13*. Tabla de los tequilas de validación con porcentajes de iso-probabilidad.**

| | Blancos | Jóvenes | Reposados | Añejos |
|---|---|---|---|---|
| **TB5546** | 81.65% | 0% | 0% | 0% |
| **TB8657** | 49.44% | 0% | 0% | 0% |
| **TB8675** | 60.93% | 0% | 0% | 0% |
| **TB8718** | 35.57% | 0% | 0% | 0% |
| **TJ4634** | 0% | 52.95% | 0% | 0% |
| **TJ5448** | 0% | 60.57% | 0% | 0% |
| **TJ8244** | 0% | 45.75% | 0% | 0% |
| **TR5452** | 0% | 0% | 0.0% | 0% |
| **TR5580** | 0% | 0% | 63.97% | 0% |
| **TR8703** | 0% | 0% | 38.96% | 0% |
| **TR8748** | 0% | 0% | 53.66% | 0% |
| **TR8758** | 0% | 0% | 79.60% | 0% |
| **TA2501** | 0% | 0% | 0% | 94.52% |
| **TA8655** | 0% | 0% | 0% | 41.23% |
| **TEA2483** | 0% | 0% | 0% | 4.41% |
| **TEA3094** | 0% | 0% | 0% | 14.12% |
| **TEA8705** | 0% | 0% | 0% | 46.47% |

Por otro lado, las probabilidades más bajas de pertenencia a su clase fueron para los tequilas TA2483 (4.41%) y TEA3094 (14.12%). A pesar de ello, estas se consideraron correctamente asignadas puesto que tienen 0% de probabilidad de pertenencia a otras clases.

En este punto, se consideró que el alejamiento de cada tequila al núcleo de las elipses podría deberse a las características naturales que estos adquieren durante todo el proceso de producción. Por ejemplo, puede afectar que la barrica sea nueva o vieja, que haya sido quemada o no previo al proceso de reposo, el contenido alcohólico de la muestra, que haya sufrido abocamiento o no, incluso el tiempo que haya permanecido en barrica y del tamaño de esta.

Finalmente, se afirma que Curvas de Potencia fue una buena técnica de clasificación supervisada y que el modelo construido (*gráfica 22*) fue satisfactorio, aunque comparado con algunas técnicas del capítulo III, no es el mejor.

**Resumen Capítulo II** – SVM y PC

Como primera técnica supervisada, se construyó *Support Vector Machines* (SVM), cuyas condiciones óptimas de trabajo fueron: 235 tequilas (sin TJ2593), intervalo 3000–1100 $cm^{-1}$, con autoescalado previo, es decir, la matriz introducida a *GenEx*©, ya tenía autoescalado y en el programa no se colocó ningún escalado. Además, todas las clases requirieron un *kernel* gaussiano, con valores de □= 2 para TB y TJ y □= 3 para TR y TA, una penalización del error de clasificación (C) de 100 y 1000 iteraciones. Todos los modelos, en modalidad *one-vs-all* fueron satisfactorios, salvo que en el de TJ's se observó la formación de dos subgrupos debidos al contenido alcohólico de las muestras que no se pudo evitar.

Posteriormente, para Curvas de Potencia (PC), se empleó la misma matriz que SVM con 100 niveles y grosor (*fitness*) de 130. Lo cual, generó excelentes resultados puesto que todas las muestras de validación fueron asignadas correctamente (a excepción del TR5452), aunque existieron muestras del conjunto de calibración que no fueron asignadas a la clase conocida previamente (debido a limitaciones que tiene el algoritmo de esta técnica). Estas muestras son las que se encuentran fuera de su elipse de iso-probabilidad, sus etiquetas son: TB4599, TB8710, TR5590, TR6912, TR8719, TA5427, TEA2591, TEA2847; y tienen en común que su contenido alcohólico es muy cercano a 55 %v/v.

# C A P Í T U L O  III
# Modelos supervisados con *Matlab©*

En este capítulo se describen los modelos elaborados en *Matlab*© con la herramienta *Classification Toolbox* (v.3.1) [14], mediante el desarrollo de las técnicas supervisadas: Vecino más cercano (**k-NN**), Funciones de Potencia (**PF**, similar a PC del capítulo II), Análisis Discriminante por el Método de Mínimos Cuadrados (**PLS-DA**), Análisis Discriminante por Componentes Principales (**PCA-DA**) y Modelación Suave e Independiente por Analogía de Clases (**SIMCA**).

La mejora que ofrecen estos modelos con respecto a los construidos en los dos capítulos anteriores es que permiten la identificación de muestras no asignadas a ninguna clase (limitaciones del algoritmo) y la caracterización más detallada de cada una. Además, cada modelo reporta parámetros de desempeño, tales como: especificidad, sensibilidad, precisión, exactitud y tasa de error; para realizar una comparación de condiciones de modelado bajo una misma técnica o entre técnicas y decidir cuáles son los mejores.

## 6.6 Parámetros de desempeño

Los parámetros de desempeño de un modelo permiten evaluar de forma rápida y sencilla (pero no gráfica), el comportamiento de una técnica supervisada dada. Éstos a su vez, ayudan a decidir cuál de ellas y bajo qué condiciones cumplen el objetivo de la presente tesis, mediante una comparación.

En la *tabla 14*, se exhiben los parámetros de desempeño de todas las técnicas supervisadas construidas en *Matlab*©. Cabe mencionar que estos se colocaron al principio del capítulo para anticipar los resultados obtenidos a partir de cada técnica; no obstante, los juicios de utilidad de los modelos de clasificación construidos dependen también de las representaciones gráficas obtenidas en las secciones 6.7-6.10.

La primera técnica abordada en la sección 6.7 (pág. 98) fue Funciones de Potencia, donde se encontró que existían 13 muestras no asignadas durante la calibración del modelo, veinte durante la validación cruzada y dos durante la predicción de muestras. Lo cual, representa una mejora con respecto a Curvas de Potencia (sección 6.5, pág. 90), puesto que esta última supone que las muestras fuera de las elipses de iso-probabilidad son las únicas no asignadas en el modelo, pero podría no ser así. Sin embargo, a pesar de las muestras no asignadas a ninguna clase, de acuerdo con los parámetros de desempeño se dice que Funciones de Potencia es un buen modelo de clasificación, aunque no el mejor con respecto a otras técnicas.

***Tabla 14*. Parámetros de desempeño por los modelos supervisados construidos en *Matlab*©.**

| Técnica | Condiciones de Modelado | Calibración (*fitting*) – Validación interna (*cross validation*) | | | | | | | Predicción (*test*) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Clase | Parámetros | | | AC | ER | NER | Clase | Parámetros | | | AC | ER | NER |
| | | | SP | SN | PR | | | | | SP | SN | PR | | | |
| **PCA-DA**<br>Datos autoescalados<br>Discriminante lineal | 235 tequilas<br>Intervalo completo | **B** | 1.00 | 1.00 | 1.00 | 1.00 | 0.00 | 1.00 | **B** | 1.00 | 1.00 | 1.00 | 1.00 | 0.00 | 1.00 |
| | | **J** | 1.00 | 1.00 | 1.00 | | | | **J** | 1.00 | 1.00 | 1.00 | | | |
| | | **R** | 1.00 | 1.00 | 1.00 | | | | **R** | 1.00 | 1.00 | 1.00 | | | |
| | | **A** | 1.00 | 1.00 | 1.00 | | | | **A** | 1.00 | 1.00 | 1.00 | | | |
| | 235 tequilas<br>Intervalo<br>3000–1100 $cm^{-1}$ | **B** | 1.00 | 1.00 | 1.00 | 1.00 | 0.00 | 1.00 | **B** | 1.00 | 1.00 | 1.00 | 1.00 | 0.00 | 1.00 |
| | | **J** | 1.00 | 1.00 | 1.00 | | | | **J** | 1.00 | 1.00 | 1.00 | | | |
| | | **R** | 1.00 | 1.00 | 1.00 | | | | **R** | 1.00 | 1.00 | 1.00 | | | |
| | | **A** | 1.00 | 1.00 | 1.00 | | | | **A** | 1.00 | 1.00 | 1.00 | | | |
| **k-NN**<br>Distancia euclídea<br>k= 1 | 235 tequilas<br>Intervalo completo<br>Sin escalado | **B** | 0.99 | 1.00 | 0.97 | 0.99 | 0.01<br>0.02 | 0.99<br>0.98 | **B** | 1.00 | 1.00 | 1.00 | 0.94 | 0.05 | 0.95 |
| | | **J** | 1.00 | 0.96 | 1.00 | | | | **J** | 1.00 | 1.00 | 1.00 | | | |
| | | **R** | 1.00 | 0.97 | 1.00 | | | | **R** | 0.92 | 1.00 | 0.83 | | | |
| | | **A** | 0.99 | 1.00 | 0.98 | | | | **A** | 1.00 | 0.80 | 1.00 | | | |
| | 235 tequilas<br>Intervalo completo<br>Autoescalado | **B** | 1.00 | 1.00 | 1.00 | 1.00 | 0.00 | 1.00 | **B** | 1.00 | 1.00 | 1.00 | 1.00 | 0.00 | 1.00 |
| | | **J** | 1.00 | 1.00 | 1.00 | | | | **J** | 1.00 | 1.00 | 1.00 | | | |
| | | **R** | 1.00 | 1.00 | 1.00 | | | | **R** | 1.00 | 1.00 | 1.00 | | | |
| | | **A** | 1.00 | 1.00 | 1.00 | | | | **A** | 1.00 | 1.00 | 1.00 | | | |
| **PLS-DA**<br>Datos autoescalados<br>Criterio de asig. Bayes | 219 tequilas<br>3000–1100 $cm^{-1}$ | **B** | 1.00 | 1.00 | 1.00 | 1.00 | 0.00 | 1.00 | **B** | 1.00 | 1.00 | 1.00 | 1.00 | 0.00 | 1.00 |
| | | **J** | 1.00 | 1.00 | 1.00 | | | | **J** | 1.00 | 1.00 | 1.00 | | | |
| | | **R** | 1.00 | 1.00 | 1.00 | | | | **R** | 1.00 | 1.00 | 1.00 | | | |
| | | **A** | 1.00 | 1.00 | 1.00 | | | | **A** | 1.00 | 1.00 | 1.00 | | | |

Abreviaturas: **PR** (Precisión), **SN** (Sensibilidad), **SP** (Especificidad), **AC** (Exactitud), **NER** (Tasa de no error), **ER** (Tasa de error).

***Tabla 14*. (continuación)**

| Técnica | Condiciones de Modelado | Calibración (*fitting*) – Validación interna (*cross validation*) | | | | | | | Predicción (*test*) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Clase | Parámetros | | | AC | ER | NER | Clase | Parámetros | | | AC | ER | NER |
| | | | SP | SN | PR | | | | | SP | SN | PR | | | |
| **PF**<br>Datos autoescalados<br>*Kernel* gaussiano<br>235 tequilas<br>Int. 3000–1100 $cm^{-1}$ | Percentil 95<br>2 PC's<br>*Smoothing*:<br>TB: 0.5, TJ: 0.7,<br>TR: 0.6, TA:0.6 | **B** | 1.00 | 1.00 | 1.00 | 1.00 | 0.00 | 1.00 | **B** | 1.00 | 1.00 | 1.00 | 1.00 | 0.00 | 1.00 |
| | | **J** | 1.00 | 1.00 | 1.00 | | | | **J** | 1.00 | 1.00 | 1.00 | | | |
| | | **R** | 1.00 | 1.00 | 1.00 | | | | **R** | 1.00 | 1.00 | 1.00 | | | |
| | | **A** | 1.00 | 1.00 | 1.00 | | | | **A** | 1.00 | 1.00 | 1.00 | | | |
| | | 13 muestras no asignadas en calibración (0.06)<br>20 muestras no asignadas en validación (0.10) | | | | | | | 2 muestras no asignadas (0.12) | | | | | | |
| **SIMCA**<br>Matriz de 235 *scores*<br>Int. 3000–1100 $cm^{-1}$ | Matriz de scores mediante 235 tequilas<br>Sin escalado adicional<br>Criterio *class modelling*<br>TB: 2PC's, TJ: 3PC's,<br>TR: 4PC's, TA: 3PC's | **B** | 1.00 | 1.00 | 1.00 | 1.00 | 0.00 | 1.00 | **B** | 1.00 | 1.00 | 1.00 | 1.00 | 0.00 | 1.00 |
| | | **J** | 1.00 | 1.00 | 1.00 | | | | **J** | 1.00 | 1.00 | 1.00 | | | |
| | | **R** | 1.00 | 1.00 | 1.00 | | | | **R** | 1.00 | 1.00 | 1.00 | | | |
| | | **A** | 1.00 | 1.00 | 1.00 | | | | **A** | 1.00 | 1.00 | 1.00 | | | |
| | | 2 muestras no asignadas en calibración (0.01)<br>6 muestras no asignadas en validación (0.03) | | | | | | | 1 muestra no asignada (0.06) | | | | | | |

Abreviaturas: **PR** (Precisión), **SN** (Sensibilidad), **SP** (Especificidad), **AC** (Exactitud), **NER** (Tasa de no error), **ER** (Tasa de error).

Las siguientes técnicas son k-NN, PCA-DA y PLS-DA que, a grandes rasgos, fueron modelos muy satisfactorios de clasificación de tequilas de acuerdo con su clase, dado que se lograron parámetros de desempeño iguales a 1 (especificidad, sensibilidad y precisión) y tasa de error de 0.

Por otro lado, los modelos obtenidos por SIMCA tuvieron parámetros de desempeño satisfactorios, aunque también existieron muestras no asignadas a ninguna clase (dos durante la calibración, seis en la validación cruzada y uno en la predicción).

Finalmente, de acuerdo con los parámetros de desempeño, los modelos más adecuados de clasificación según su tiempo de añejamiento fueron los obtenidos mediante las técnicas PCA-DA y PLS-DA.

### 6.7 Funciones de Potencia (PF)

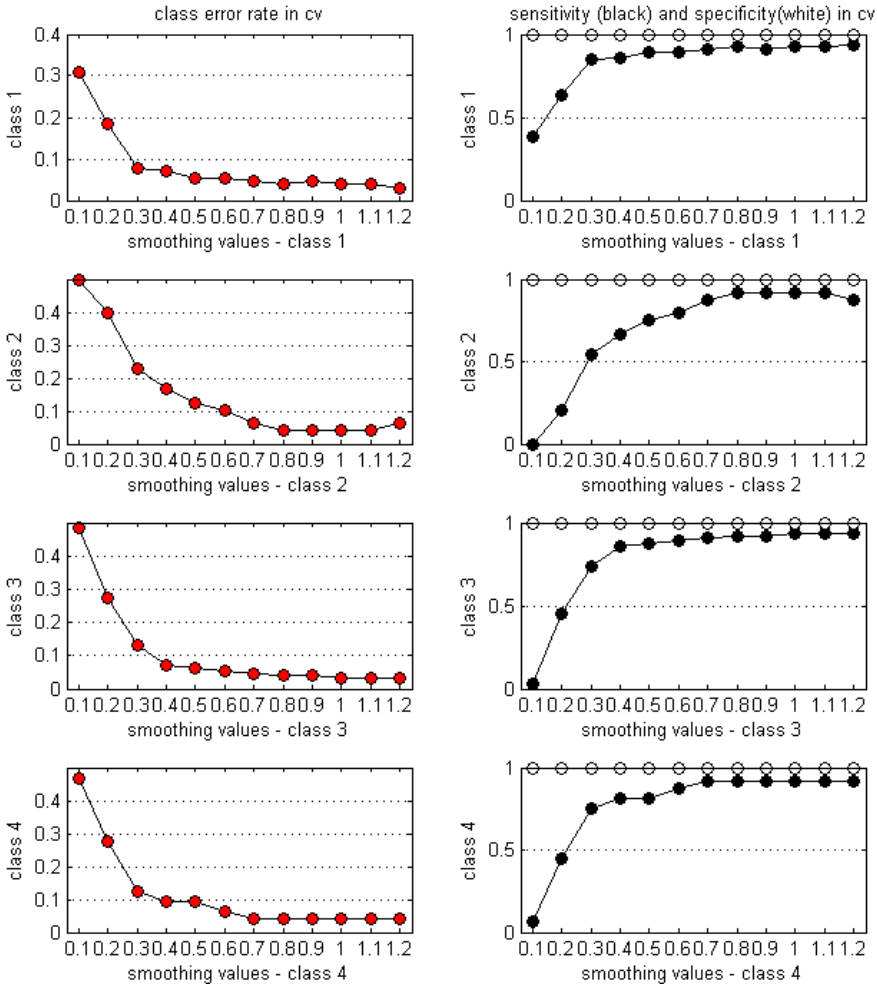
Las condiciones de trabajo para esta técnica fueron: matriz de 235 tequilas con datos previamente autoescalados, intervalo 3000–1100 $cm^{-1}$, percentil (o nivel de confianza) del 95%, 2 componentes principales y validación interna cruzada mediante persianas venecianas con diez grupos de cancelación. Aunado a esto, se pueden modificar tres parámetros: el tipo de *kernel* (o distribución asumida: gaussiano o triangular) y el factor de suavizado del *kernel* (entre 0.1 y 1.2).

Para empezar, dado que PF tiene un algoritmo similar a Curvas de Potencia (PC, capítulo I), ambos basados en el Análisis de Componentes Principales (PCA) sin rotar, se solicita el número de componentes a aplicar en el modelo. Para ello, es necesario recordar que los PC's 1 y 2 eran los responsables de la separación por clases de los tequilas. Por lo cual, es sugerible utilizar esa cantidad, recordando que esto se relaciona con la varianza explicada.
Es decir, podemos utilizar sólo un componente, pero si a este le corresponde apróx. el 46% de la varianza, dejaríamos en residuales el 54% de ella y no es lo más adecuado. Por ello, se decidió utilizar dos componentes. En seguida, hay que definir la función *kernel* (o de probabilidad) con la cual trabajar, que puede ser gaussiana o triangular. La primera resultó ser la más adecuada, pues con *kernel* triangular la tasa de error en la validación por clase era de: 0.2-0.5 en tequilas blancos y reposados, de 0.5 en tequilas jóvenes y de 0.38 a 0.5 para tequilas añejos; dependiendo del factor de suavizado que se utilizara.

Utilizando *kernel* tipo gaussiano, se encontró que para los tequilas blancos era razonable utilizar un factor de suavizado (*smoothing*) entre 0.3 y 0.5, para los tequilas jóvenes entre 0.6 y 0.8, para reposados entre 0.4 y 0.6, y para añejos entre 0.5 y 0.7 para obtener un error menor a 0.1, lo cual puede apreciarse en la *gráfica 23*.

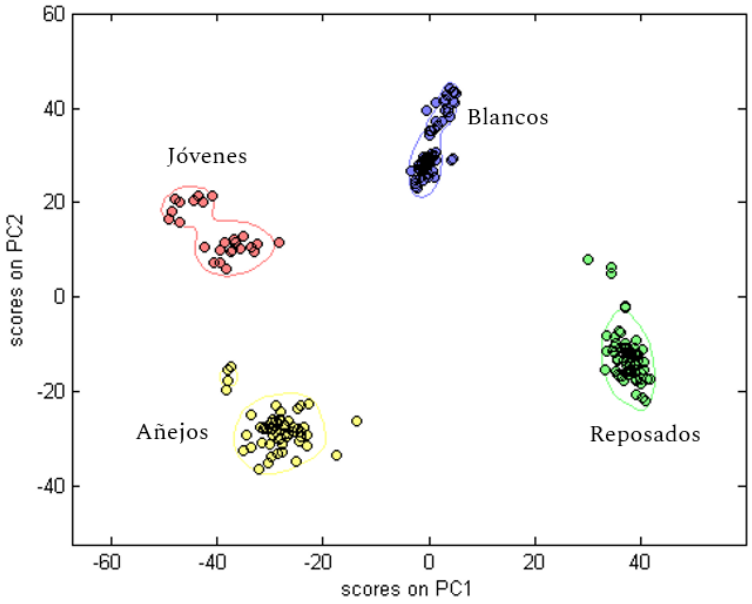
Cabe mencionar, que una de las ventajas que ofrece PF sobre Curvas de Potencia (que, de hecho, es una simplificación de las PF) es precisamente el uso del factor de suavizado en la función de clasificación, ya que este puede ser ajustado por clase. Además, el uso del percentil (o nivel de confianza) que funciona como indicador de *outliers*, permite a PF tener más capacidades y ni hablar de la retroalimentación que permite la validación cruzad. Finalmente, el análisis de los modelos gráficos puede ser en dos modalidades: *one-vs-all* o por límites de clase.

Tras explorar diferentes condiciones, se encontró que la mejor se presentaba con dos componentes principales y con factores de suavizado de 0.5 para tequilas blancos, 0.7 para jóvenes y de 0.6 para reposados y añejos (*gráfica 23*).




***Gráfica 23*. Tasa de error y parámetros de desempeño mediante PF con *kernel* gaussiano y dos componentes principales.**

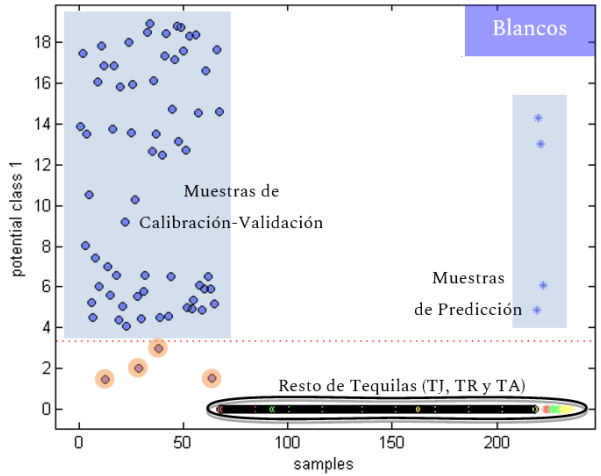
En la *gráfica 24*, se presenta el subespacio PC1 vs. PC2 construido a partir de las muestras de calibración (y validación interna), con límites de clase (líneas sólidas en color azul, rojo, verde y amarillo), donde se pudieron distinguir trece muestras no asignadas (4 TB's, 2 TJ, 4 TR's y 3 TA's).




***Gráfica 24*. Diagrama de *scores* en el subespacio PC1-PC2 con límites de clase empleando PF.**

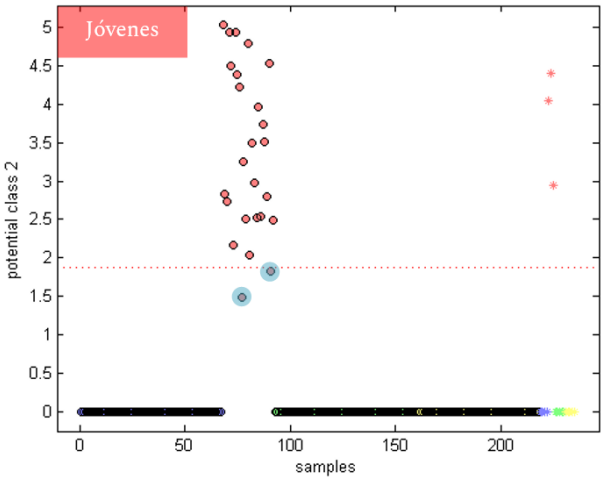
Posteriormente, se estudiaron los esquemas de potencial vs. número de muestra obtenidos para cada clase (modalidad *one-vs-all*, *gráficas 25-28*), donde el potencial debe entenderse como el resultado de aplicar la función de cálculo (en este caso, gaussiana bidimensional -al haber elegido 2 PC's-, con los parámetros arriba indicados) a los valores asociados a cada muestra.

Las líneas punteadas en las *gráficas 25-28* representan el percentil de 95 % (el cual indica que las muestras debajo no fueron asignadas a ninguna clase), los asteriscos representan las muestras de predicción y la "línea gruesa" (debajo de la línea punteada) está conformada por los tequilas de clase diferente a aquella en cuestión.



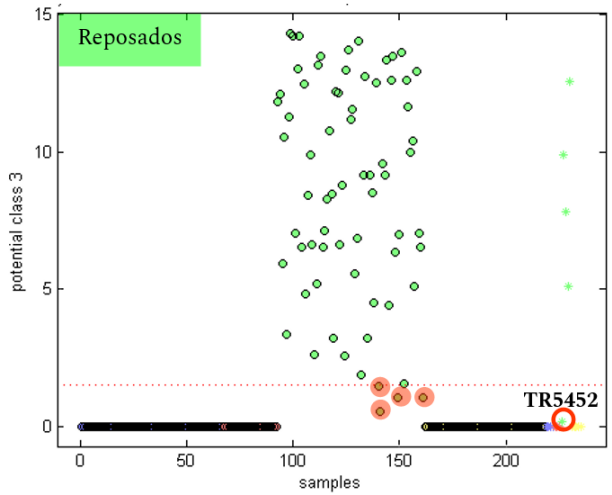

***Gráfica 25*. Potencial (x10$^{-3}$) de los TB's vs. el resto con percentil de 95 % y *smoothing* de 0.5.**

Las muestras no asignadas, en el caso de los TB's, fueron los tequilas TB4599, TB5428, TB6848 y TB8710, mientras que, los TJ's fueron el TJ4757 y TJ8733; todos del conjunto de calibración. Cabe mencionar que esto ocurre por limitaciones del propio algoritmo al realizar el modelado.




***Gráfica 26*. Potencial (x10$^{-3}$) de los TJ's contra el resto con percentil de 95 % y *smoothing* de 0.7.**

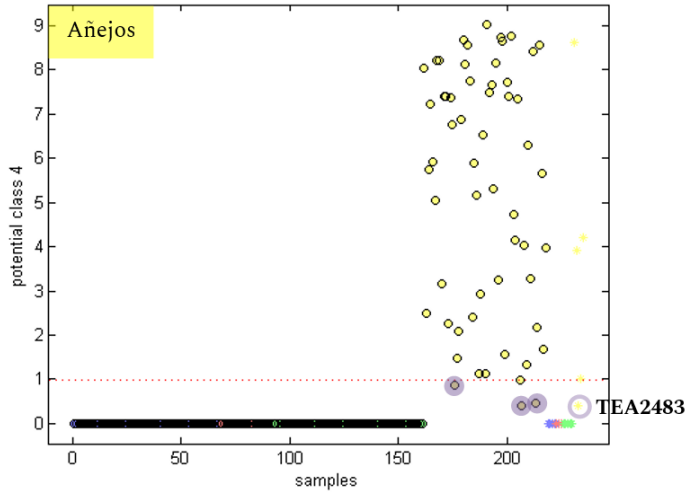
De igual forma, en las *gráficas 27* y *28* se exhiben los modelos construidos para las clases TR y TA, donde se identificaron, para el conjunto de calibración, ocho muestras debajo del percentil. Estas fueron, cuatro TR's: 5577, 5590, 6912, 8719 y tres TA's: TA3043, TEA2847 y TEA5426.




***Gráfica 27*. Potencial (x10$^{-3}$) de los TR's vs. el resto con percentil de 95 % y *smoothing* de 0.6.**

En el modelo de clasificación de los tequilas reposados (*gráfica 27*), además de las muestras de calibración, un tequila del grupo de predicción no fue asignado: TR5452. Esto es análogo con lo obtenido en PCA y, como ahí se expresó, este tequila tenía características distintas a los demás, con lo que es posible definirlo como *outlier*.

En el caso de los tequilas añejos, también se encontró una muestra no asignada para el conjunto de predicción, ésta fue el TEA2483 (*gráfica 28*).




***Gráfica 28.*** **Potencial (x10$^{-3}$) de los TA's vs. el resto con percentil de 95 % y *smoothing* de 0.6.**

En este punto, es necesario recordar que se llevan a cabo tres etapas por cada modelo supervisado desarrollado en *Matlab*©: La primera es la calibración (o ajuste, *fitting*), que corresponde a la etapa de entrenamiento del algoritmo, en donde hubo una tasa de error del 0%, parámetros de desempeño del 100% (1.0), pero también, el 6% (0.06) de muestras no asignadas. Durante la validación cruzada, se tuvo un error del 0%, parámetros del 100% pero 10% de muestras no asignadas. Y durante la predicción, hubo el 0% de error, 100% en los parámetros de desempeño y 12% de muestras no asignadas (véase *tabla 14,* págs. 71-72).

Con base en lo expresado en el párrafo anterior y en la *tabla 15*, se expone que fueron veinte muestras no asignadas en el entrenamiento del modelo (calibración y validación interna) y dos durante la predicción. No obstante, fueron quince las que presentaron mayor diferencia en cuanto a sus características (ubicadas debajo del percentil en las *gráficas 25-28*).

***Tabla 15.*** **Tequilas no asignados en los modelos de clasificación realizados mediante Funciones de Potencia con 219 teq.**

| No. | Tequila | Etapa del modelado | | |
|---|---|---|---|---|
| | | Calibración | Validación cruzada | Predicción |
| 13 | TB4599 | ✓ | ✓ | |

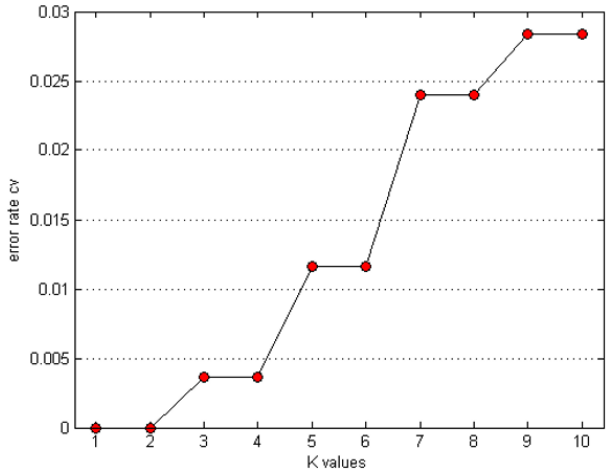
| | | | | |
|---|---|---|---|---|
| 23 | TB5401 | | ✓ | |
| 29 | TB5428 | ✓ | ✓ | |
| 38 | TB6848 | ✓ | ✓ | |
| 64 | TB8710 | ✓ | ✓ | |
| 77 | TJ4757 | ✓ | ✓ | |
| 81 | TJ4945 | | ✓ | |
| 91 | TJ8733 | ✓ | ✓ | |
| 132 | TR5397 | | ✓ | |
| 140 | TR5577 | ✓ | ✓ | |
| 141 | TR5590 | ✓ | ✓ | |
| 149 | TR6912 | ✓ | ✓ | |
| 152 | TR8594 | | ✓ | |
| 161 | TR8719 | ✓ | ✓ | |
| 176 | TA3043 | ✓ | ✓ | |
| 187 | TA5288 | | ✓ | |
| 206 | TEA2591 | | ✓ | |
| 207 | TEA2847 | ✓ | ✓ | |
| 209 | TEA2941 | | ✓ | |
| 213 | TEA5426 | ✓ | ✓ | |
| T8 (*test* 8) | TR5452 | | | ✓ |
| T15 (*test* 15) | TEA2483 | | | ✓ |

### 6.8 Vecino más cercano (k-NN)

Bajo la técnica k-NN se construyeron dos modelos, descritos a continuación.

El primer análisis se erigió con los espectros de 235 tequilas con todas las variables (4000-450 $cm^{-1}$), autoescalado, distancia euclídea y se llevó a cabo también una validación cruzada por persianas venecianas con diez grupos de cancelación.

A continuación, se muestra el gráfico de la tasa de error en función del valor de k (número de vecinos) bajo estas condiciones.

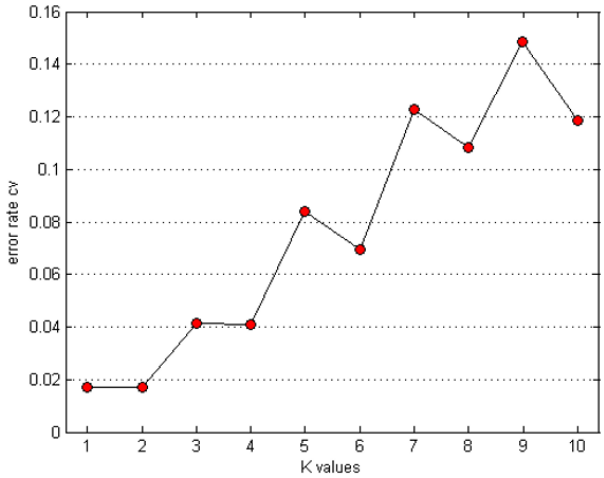



***Gráfica 29.*** **Tasa de error para 235 teq., datos autoescalados, intervalo completo y distancia euclídea.**

De acuerdo con la *gráfica 29*, el error mínimo se ubicó en un número mínimo de k= 1 o 2 vecinos, por lo que se decidió utilizar k igual a 1 obteniéndose excelentes parámetros de desempeño (véase *tabla 14*), donde todas las muestras de predicción fueron correctamente asignadas y ningún tequila del conjunto de calibración-validación interna fue ubicado fuera de su clase, por lo cual, sus parámetros de desempeño tuvieron un valor de 1.0 y sus tasas de error de 0.0.

Pensando en la posibilidad de un sobreajuste en el modelo, se decidió explorar la tasa de error con los 235 tequilas, distancia euclídea, validación cruzada con diez grupos de cancelación y todas las variables, pero esta vez, sin escalar los datos.

Al igual que con datos autoescalados, el menor error se encontró en k= 1 o 2 vecinos (*gráfica 30*), los cuales, daban resultados muy similares, por lo que se decidió emplear k de uno.




***Gráfica 30.*** **Tasa de error con 235 teq., intervalo completo, sin escalado y distancia euclídea.**

Este modelo, tuvo una tasa de error del 1% (0.01) y una exactitud del 99% (0.99) para el conjunto de calibración, donde un TJ y un TR fueron asignados como TB (TJ4757 y TR8594). Mientras que, el conjunto de validación interna tuvo una tasa de error del 2% (0.02) y una exactitud del 99% (0.99), donde, un TJ y un TR se identificaron con los TB's y otro TR como TA (TJ4757, TR8594 y TR5590, respectivamente).

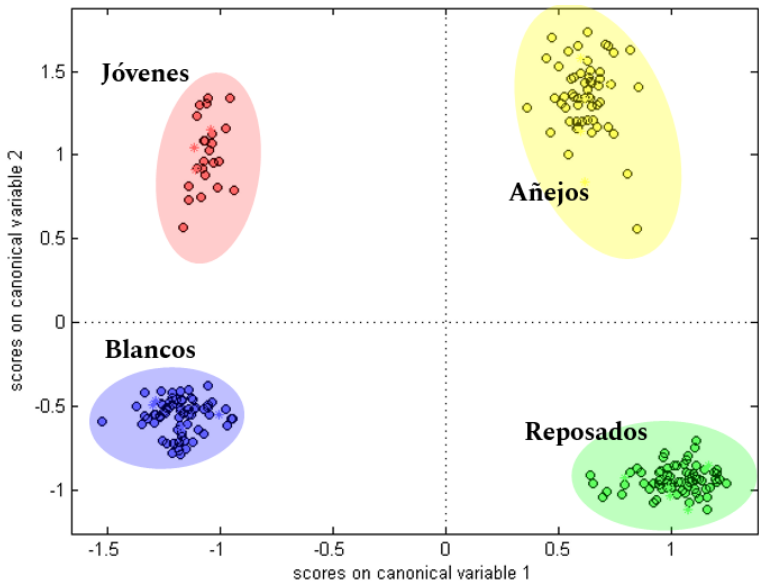
Finalmente, la tasa de error en el conjunto de predicción fue del 5% (0.05), con una exactitud del 94% (0.94). En este caso, fue un TA el que se ubicó en el grupo de TR's: TEA3094. Para mayor información sobre los parámetros de desempeño, véase la *tabla 14*.

## 6.9 Análisis Discriminante por Componentes Principales (PCA-DA)

PCA-DA, permite estudiar diagramas de *scores* de componentes principales (PC), o bien, de variables canónicas (CV), donde lo más adecuado, por tratarse de análisis discriminante, es realizar el estudio de las CV's, dado que, al aplicar análisis discriminante al PCA, en lugar de trabajar con componentes principales (factores que maximizan la varianza total acumulada), se trabaja con variables canónicas o factores discriminantes que maximizan la relación entre varianzas que permiten alcanzar la máxima separación de las clases.

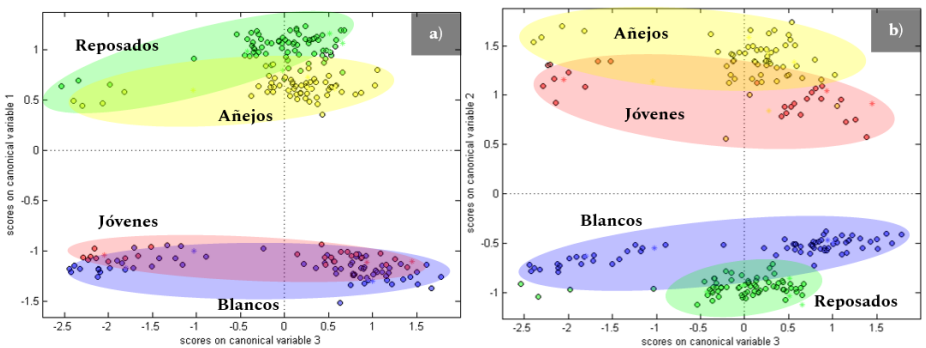
Para lo cual, se llevó a cabo un análisis con 235 muestras (sin TJ2593), de las cuales 218 fueron del conjunto de calibración-validación interna y 17 del conjunto de predicción. El primer modelo se desarrolló con autoescalado, intervalo completo (4000–450 $cm^{-1}$, lo cual implica 3551 variables), tres variables canónicas, discriminación lineal y validación cruzada mediante persianas venecianas con diez grupos de cancelación.

En la *gráfica 31*, se muestra el modelo seleccionado mediante este análisis, en el subespacio CV1–CV2. Donde se observa que cada clase se ubica en uno de los planos trazados por los ejes punteados en CV= 0. Si bien, esto no es algo que generalmente deba suceder, este modelo tuvo una tasa de error en la validación cruzada y en la predicción igual a cero usando tres CV's. Al ser, esta técnica, una combinación entre Análisis de Componentes Principales y Análisis Discriminante, permite explorar la varianza explicada de cada componente, donde, PC1 tuvo una varianza explicada del 46.39%, PC2 del 24.41% y PC3 del 17.34% dando lugar a una varianza acumulada por el modelo del 88.14%, lo cual da una varianza residual del 11.86%.




***Gráfica 31.*** **Diagrama de *scores* de CV1-CV2 con 3 componentes para 235 teq. usando PCA-DA.**

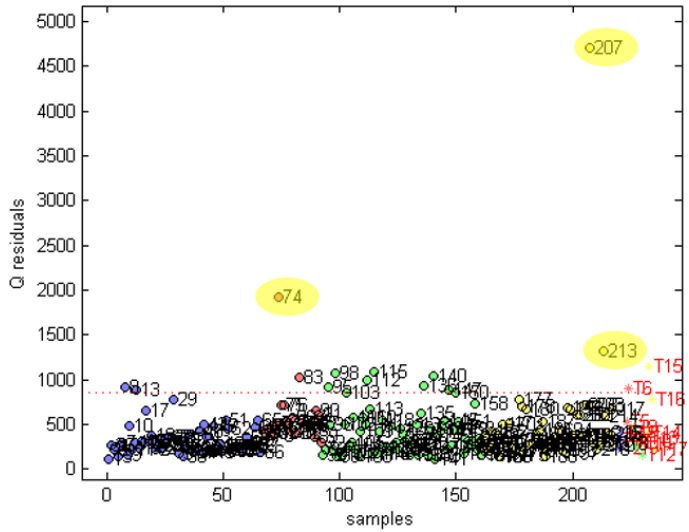
En la *gráfica 32 a*) y *b*), se exhiben los subespacios CV1-CV3 y CV2-CV3, respectivamente, donde se observan traslapes entre las clases, por lo que se descartaron como modelos adecuados de clasificación.




***Gráfica 32***. **Diagramas de *scores* de *a*) CV1-CV3 y *b*) CV2-CV3 para 235 teq., int. completo.**

Además de las gráficas de variables canónicas, y de componentes principales, el análisis PCA-DA permite estudiar el gráfico de residuales Q cuya función es destacar las muestras con comportamiento distinto a la población, es decir, a

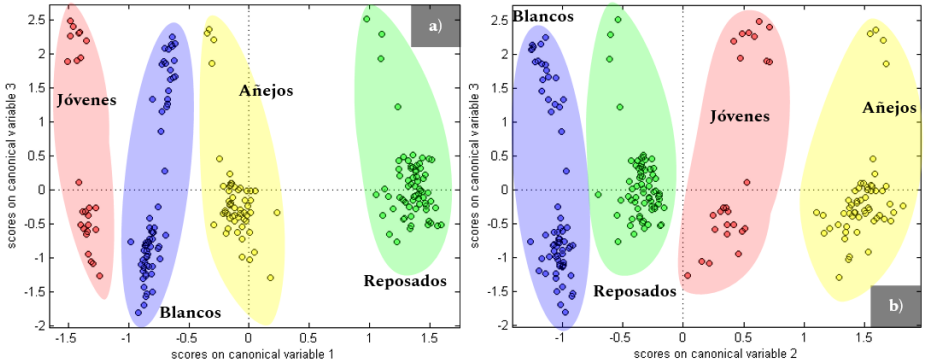
las más diferentes que, en cierto punto, podrían ser *outliers.* En la *gráfica 33*, se señalan tres muestras que, en orden decreciente, podrían ser *outliers*: 207, 74 y 213 (TEA2847, TJ2656 y TEA5426, respectivamente).




***Gráfica 33.*** **Residuales Q vs. Muestras por PCA-DA con 235 tequilas.**

Al existir traslape en CV1-CV3 y CV2-CV3 con el intervalo completo de variables (4000-450 $cm^{-1}$, *gráfica 32*), se trazaron los modelos de PCA-DA con el intervalo señalado como óptimo en PCA (3000-1100 $cm^{-1}$, *gráfica 34*).

Cabe mencionar, que estos modelos se trazaron con tres variables canónicas, autoescalado, discriminación lineal, validación interna cruzada por persianas venecianas y 10 grupos de cancelación. Por otro lado, la varianza explicada por el modelo aumentó a 94.67% (PC1= 44.81, PC2= 32.23 y PC3= 17.63).




***Gráfica 34.*** **Diagramas de *scores* de a) CV1-CV3 y b) CV2-CV3 con 235 teq., int. 3000-1100cm$^{-1}$.**

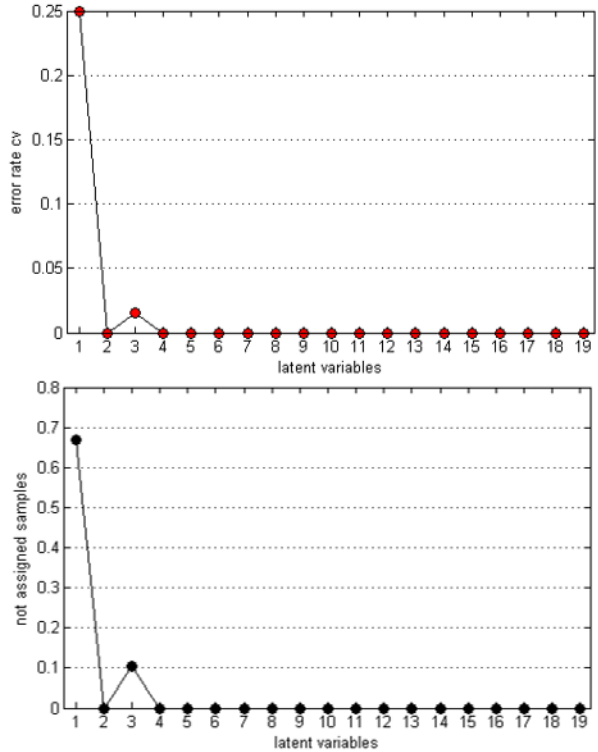
En resumen, el mejor modelo por la técnica PCA-DA fue el construido con 235 muestras, intervalo 3000-1100 cm$^{-1}$, tres factores, discriminación lineal y validación cruzada mediante persianas venecianas con diez grupos de cancelación (*gráficas 32* y *34*).

Finalmente, mediante la técnica PCA-DA también se determinaron los parámetros de desempeño y tasas de error (*tabla 14*, páginas 71-72), los cuales, fueron del 100% y su tasa de error fue el 0%.

### 6.10 Análisis Discriminante por el Método de Mínimos Cuadrados Parciales (PLS-DA)

Esta técnica, permite calcular los factores discriminantes teniendo en cuenta la correlación que existe entre los espectros y el tipo de muestra (usando como variable de predicción de clase la aproximación NIPALS). Además, es posible estudiar los *scores* de las muestras, los *loadings* y *weights* de las variables latentes, así como los valores estimados "y *calc*". Para construir el modelo de PLS-DA se introdujeron los datos (intervalo 3000–1100 cm$^{-1}$) de las 218 muestras de calibración-validación interna y 17 de predicción.

Éstos se autoescalaron y se llevó a cabo el análisis con tres factores, criterio de asignación bayes, validación interna cruzada mediante persianas venecianas y diez grupos de cancelación. A continuación, se muestra la tasa de error y las muestras no asignadas en función del número de variables latentes utilizadas (*gráfica 35*).

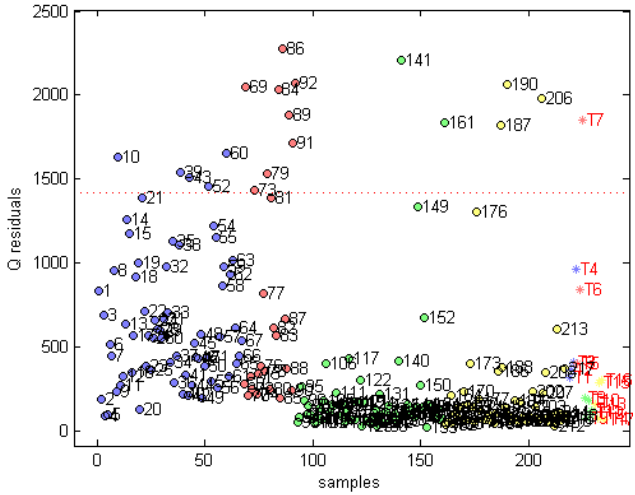



***Gráfica 35*. Tasa de error y muestras no asignadas en función de las variables latentes (LV).**

Primero, se decidió calcular el modelo con cuatro variables latentes (LV), ya que el mínimo error se presentó con dos o con cuatro variables. No obstante, la varianza explicada por el modelo con dos LV's fue de ≈76.81%, mientras que, con cuatro LV's fue de 96.27%.

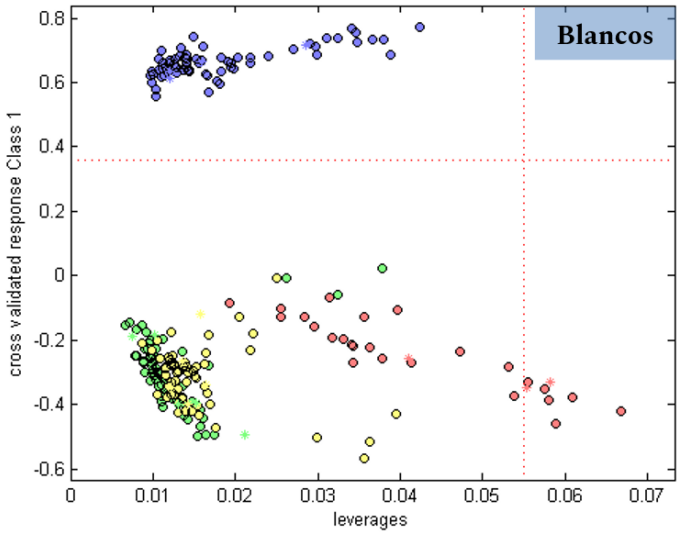
Descomponiendo la varianza acumulada: 44.78% para LV1, 32.03% para LV2, 15.61% para LV3 y 3.85% para LV4. Lo que da lugar a una varianza residual del 3.73%.

En la *gráfica 36*, se muestran los residuales Q, donde existen varias muestras que podrían considerarse *outliers* dependiendo de su clase; sin embargo, se observó que no eran realmente muestras anómalas sino muestras con más de 50 grados de alcohol.




***Gráfica 36*. Residuales Q vs. muestras para PLS-DA con 219 teq. en el int. 3000–1100 $cm^{-1}$.**

PLS-DA, además del gráfico de residuales, ofrece un gráfico de "*leverages*", que tiene también la función de encontrar muestras anómalas. Estos, pueden graficarse contra los valores estimados "y calc cv" (*gráfica 37*) para conocer específicamente las muestras *outliers* por clase. Cabe mencionar que, "y calc cv" se refiere al valor estimado (o predicho) para las muestras a partir de la validación interna cruzada de las muestras. En las *gráficas 37-40*, se exhiben los modelos obtenidos por PLS-DA para cada clase.

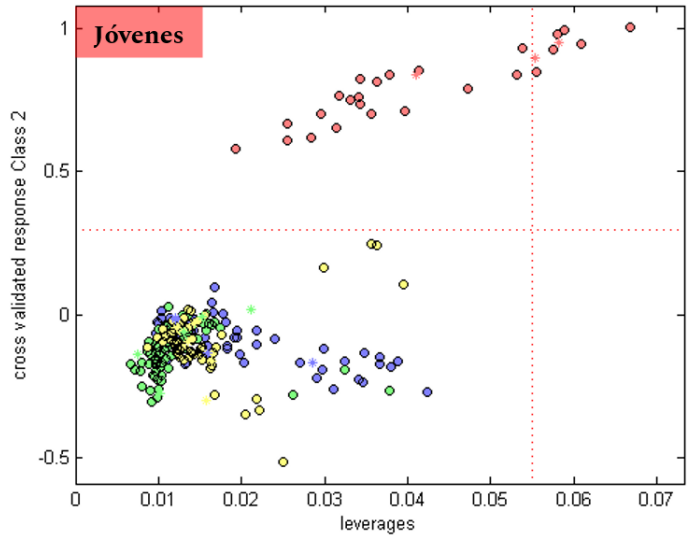



***Gráfica 37*. Estimación de grupo "y calc cv" vs. *leverages* para los TB's mediante PLS-DA.**

En las *gráficas 37-40*, la línea punteada verticalmente sobre el eje "leverages" implica el límite de pertenencia a la clase en cuestión (es decir, las muestras después de esta línea podrían ser *outliers)*; la línea punteada horizontalmente sobre el eje "y calc cv" representa el límite de clasificación del modelo (es decir, las muestras por debajo de esta línea no pertenecen a la clase en cuestión). Finalmente, los asteriscos revelan las muestras del conjunto de predicción.

Para el conjunto de tequilas blancos, todas las muestras de predicción fueron correctamente predichas, puesto que, todas se encuentran por arriba del límite de la respuesta "y calc cv". Además, las muestras de predicción son similares a las de calibración ya que se sitúan antes del límite calculado para los "*leverages*" (*gráfica 37*).

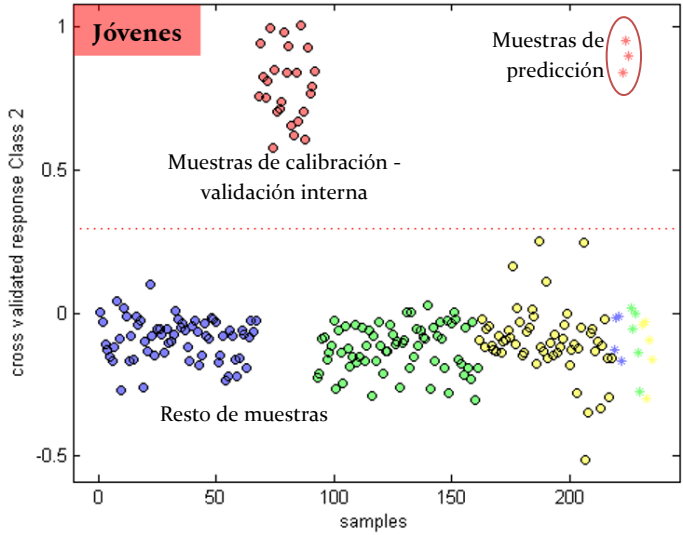
Profundizando en la estimación generada mediante la respuesta y calculada durante la validación cruzada (y calc cv) para el conjunto de calibración, se observó que los TB's se ubican en valores positivos, mientras las muestras restantes están en valores negativos, lo cual, es análogo al caso que se mencionó en el marco teórico, donde para dos clases, PLS codifica en +1 a una clase y en -1 a las demás.




***Gráfica 38*. Estimación de grupo "y calc cv" vs. *leverages* para los TJ's mediante PLS-DA.**

En la *gráfica 38*, si tomamos en cuenta ambos ejes, los tequilas de predicción de la clase "jóvenes", no fueron asignados correctamente sino sólo uno de ellos (TJ4634). El motivo de esta afirmación es que se observó que existían ocho muestras de calibración ubicadas fuera del límite de los *leverages*. Estas fueron (de derecha a izquierda): 86, 69, 73, 79, T6, 89, T7 y 92 (TJ6870, TJ2521, TJ2650, TJ4778, TJ5448, TJ8605, TJ8244 y TJ8765, respectivamente).

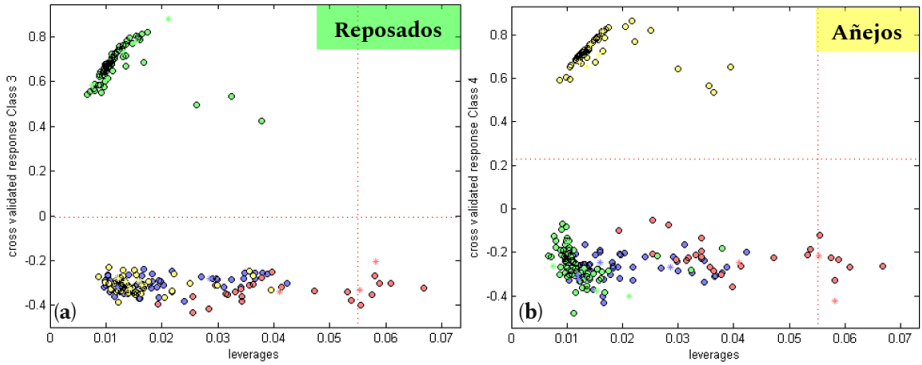
Si recordamos, estas muestras, también se distinguieron en los modelos desarrollados bajo la técnica SVM, donde se reportó que se debía a los grados de alcohol que cada una tenía: mientras el grupo que estaba en valores 0.055 ≤ *leverages* tiene entre 35.21 y 49.96 %v/v de etanol, aquellas en *leverages*≥ 0.055 tienen ≈55%. Puesto que se consideraron pocas muestras de tequila joven (26) y que el porcentaje de alcohol entre ellas era muy distinto (min. 35 vs. máx. 55), no se eliminaron ninguna de ellas, ya que se correría el riesgo de descartar información sobre la naturaleza de la clase. Otra forma de justificar el no eliminar estas muestras es con los *leverages* (otra medida de distancia entre los datos) donde estos tequilas, al tener ciertos grados de alcohol más, se distanciaron del resto.




***Gráfica 39*. Estimación de grupo "y calc cv" vs. muestras para los TJ's mediante PLS-DA.**

Por otro lado, si tomamos en cuenta sólo el eje de la respuesta y estimada mediante PLS (*gráfica 39*), se puede afirmar que los tres tequilas de predicción fueron correctamente asignados y que los *outliers* descritos anteriormente podrían no ser eliminados puesto que están por arriba del límite inferiores de la clase (línea punteada de color rojo en ≈0.3).

En la *gráfica 40*, se presentan los valores estimados para los tequilas reposados (a) y añejos (b) contra *leverages*, a partir de los cuales fue posible decir que no existieron muestras anómalas y que todos los tequilas fueron correctamente predichos. Finalmente, los parámetros de desempeño para esta técnica fueron del 100% y su tasa de error fue el 0% (*tabla 14*).

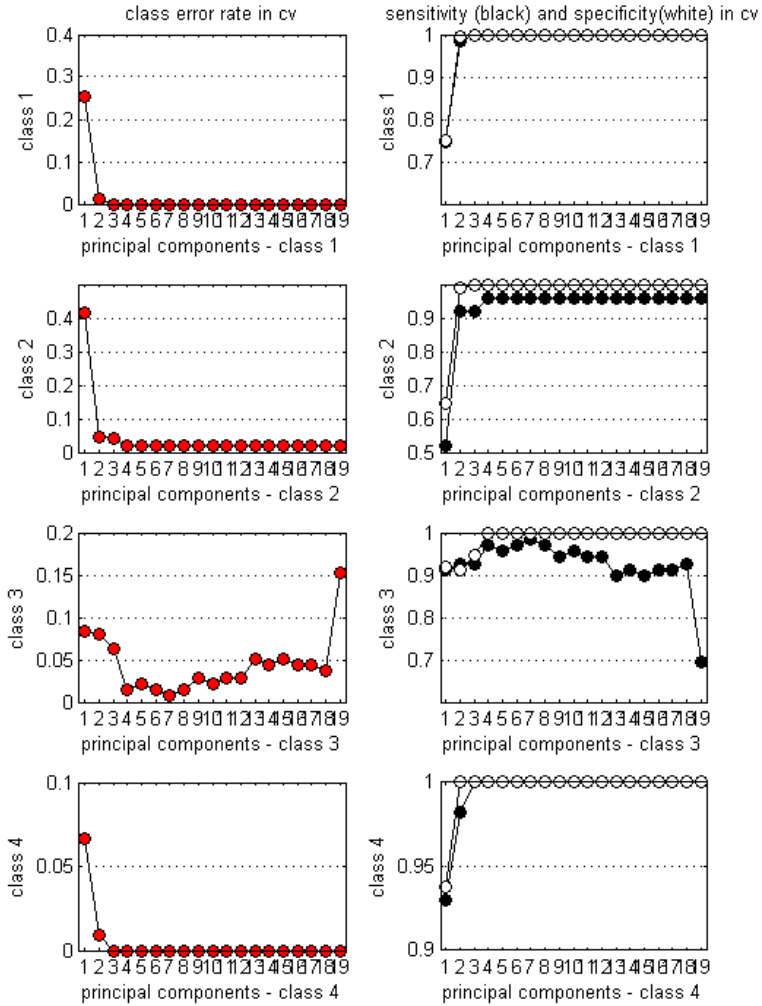



***Gráfica 40*. Estimación de grupo "y calc cv" vs. *leverages* para (a) TR y (b) TA mediante PLS-DA.**

6.11 Modelación Suave e Independiente por Analogías de Clase (SIMCA)

Se dejó esta técnica al final porque fue la que más complicaciones generó, puesto que fue necesario explorar el intervalo completo y el intervalo reducido en PCA (3000–1100 $cm^{-1}$), ambos con malos resultados. Para obtener la matriz de trabajo para esta técnica, se realizó un PCA con el objetivo de recuperar los *scores* del intervalo 3000–1100 $cm^{-1}$ con datos espectrales autoescalados de las 235 muestras. Esto es útil para reducir la dimensionalidad de los datos sin afectar la información de las muestras.

En la *gráfica 41*, se exhiben las tasas de error por clase, en función de los componentes principales, así como los parámetros de sensibilidad y especificidad del modelado en la validación cruzada. Vale la pena recordar que esta técnica permite decidir el número de componentes a utilizar por clase.

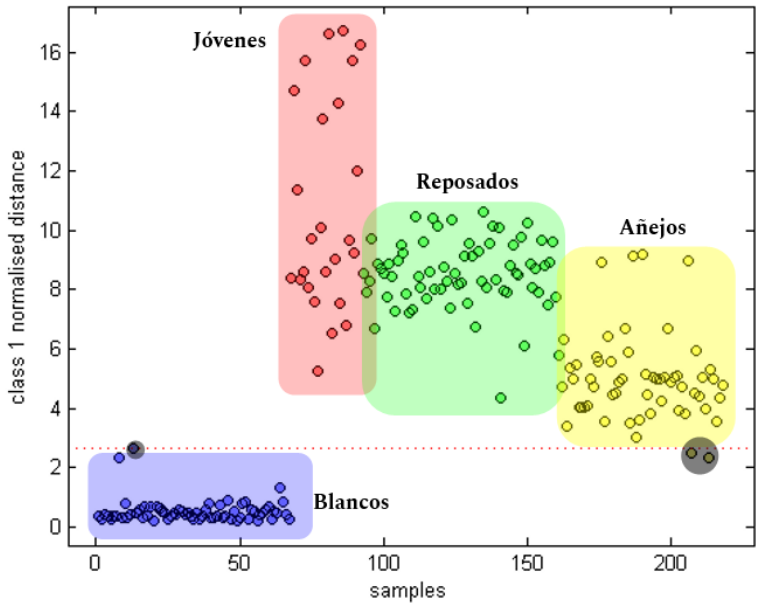



***Gráfica 41*. Tasa de error y valores de parámetros de desempeño para SIMCA con 235 *scores*.**

Tras explorar diferentes números de componentes por clase, se encontró que las mejores condiciones eran: dos PC's para TB, tres PC's para TJ y TA y cuatro para TR, además, se aplicó validación cruzada con diez grupos de cancelación y el criterio de modelación fue *class modelling*. Cabe mencionar, que no se escalaron los datos al ingresarse en *Matlab*©, puesto que los *scores* se recuperaron ya con autoescalado.

Las varianzas explicadas por clase fueron: **TB** (95.93%): PC1= 62.27% y PC2= 33.66%, **TJ** (98.55%): PC1= 76.33%, PC2= 21.04% y PC3= 1.18%, **TR** (98.57%): PC1= 87.95%, PC2= 6.94%, PC3= 2.3% y PC4= 1.38% y **TA** (97.73%): PC1= 86.23%, PC2= 10.07% y PC3= 1.43%.

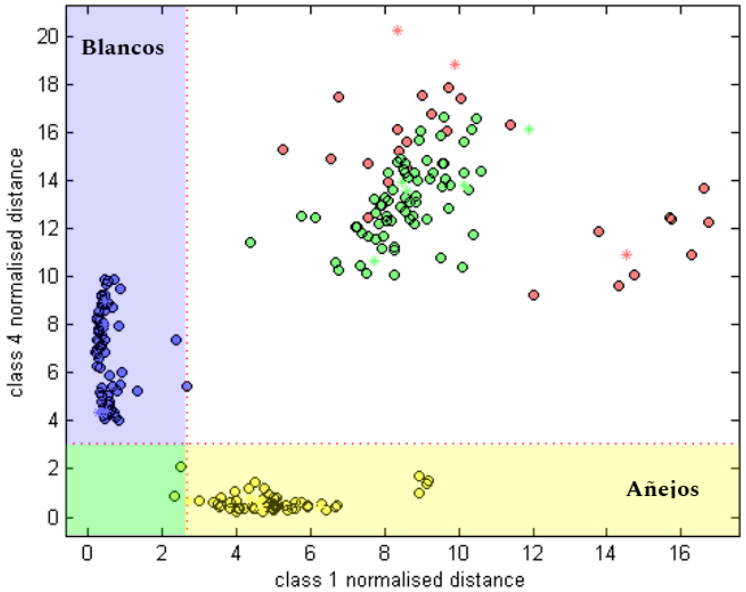
Se decidió comenzar por el análisis de los gráficos de distancia normalizada contra muestras, los cuales, se exhiben a continuación.




***Gráfica 42*. Distancia normalizada para los tequilas blancos vs. muestras a partir de los *scores*.**

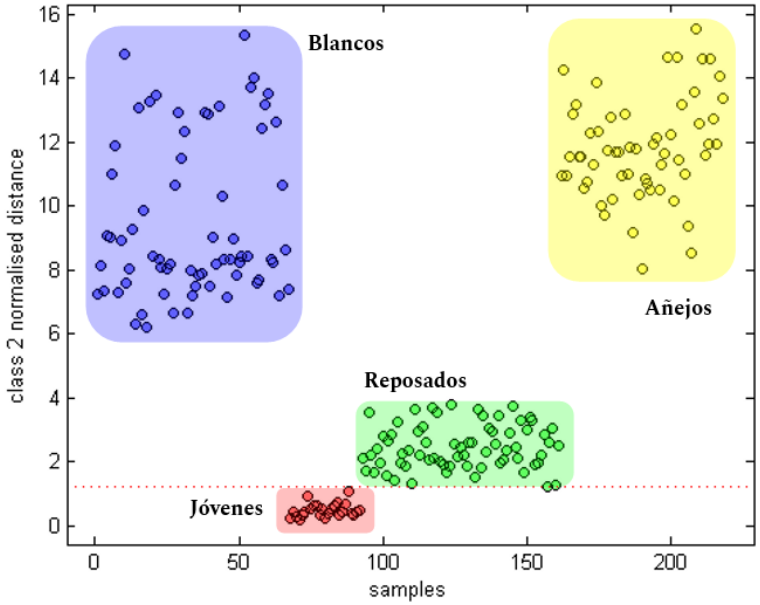
En la *gráfica 42*, se presenta el modelo para tequilas blancos con dos componentes, donde, se aprecia que la muestra 13 (TB4599) estaba en el límite de distancia entre su clase y el resto. Lo cual, podría deberse a alguna propiedad del tequila similar a los tequilas añejos, puesto que es el conjunto más cercano a su límite de distancia (apróx. 2.7). Por otro lado, se dice que las muestras 207 y 213 (TEA2847 y TEA5426, respectivamente) podrían tener características similares a las de los tequilas blancos, ya que invaden el área de los TB's. Con la finalidad de esclarecer las características de estas muestras, se procedió a graficar la distancia normalizada de los tequilas blancos vs. los tequilas añejos

(*gráfica 43*), donde se observó que el tequila TB4599 no tenía características similares al conjunto de los TA's, sin embargo, sí era distinto a su clase. Además, los tequilas TEA2847 y TEA5426 sí se comportaban similar al grupo de los TB's, puesto que entran en la región de color verde.




***Gráfica 43*. Distancias normalizadas para los tequilas blancos vs. tequilas añejos.**

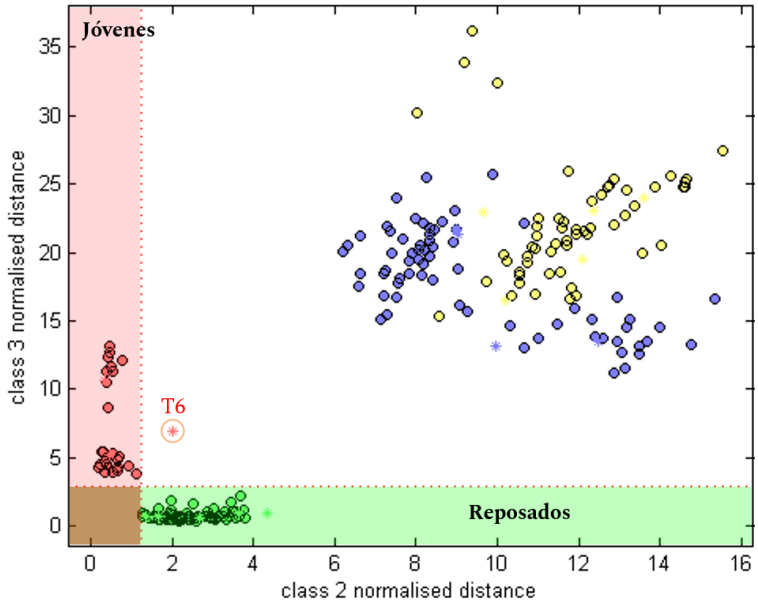
Profundizando en la interpretación de las *gráficas 42* y *43*, se dice que a una distancia normalizada de aproximadamente 2.7 (eje x), los TB's se diferencian del resto de clases; y que, a una distancia de 3.0 (eje y), los TA's se diferencian del resto. De esta forma, la intersección de ambas distancias representa el área de similitud entre TB's y TA's.




***Gráfica 44*. Distancia normalizada para los tequilas jóvenes vs. muestras a partir de los *scores*.**

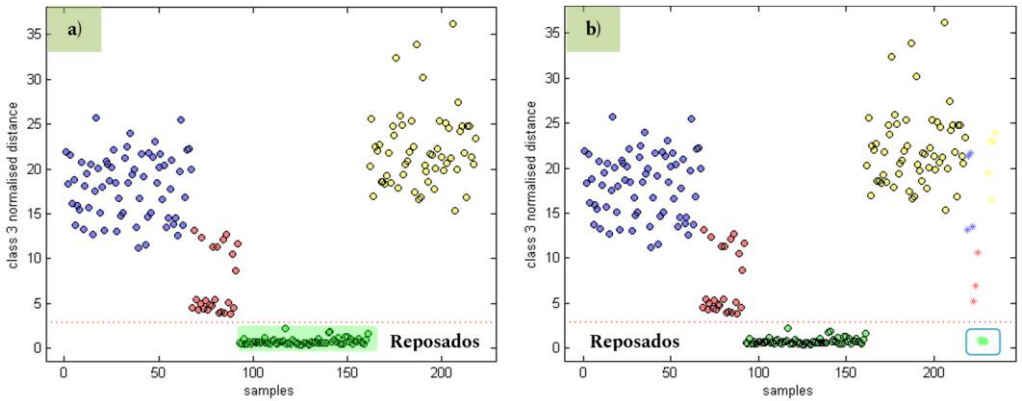
En la *gráfica 44*, se presenta el gráfico de distancia normalizada para el conjunto de los tequilas jóvenes, donde, fue posible determinar que ninguna muestra se encuentra fuera de la distancia límite de la clase (1.25). Sin embargo, existieron muestras como la 88, 157 y 160 (TJ6914, TR8641 y TR8688, respectivamente) que se ubicaron cerca de dicho límite.

De igual forma, se graficaron las distancias normalizadas que los TJ's y TR's con el objetivo de verificar similitudes (*gráfica 45*), donde, se encontró que ninguna muestra de tequilas jóvenes o reposados poseen características similares, puesto que no hay muestras en la intersección de distancias de cada clase (área marrón). No obstante, la muestra T6 (*test* 6), que corresponde a aquella etiquetada como TJ5448, salió de su clase, comportándose de forma similar a TB's y TA's.




***Gráfica 45*. Distancias normalizadas para los tequilas jóvenes vs. tequilas reposados.**

Al analizar el gráfico de distancia normalizada de los tequilas reposados vs. muestras, no se observó ningún tequila fuera del límite, en este caso su valor fue de aproximadamente 3 (*gráfica 46a*).

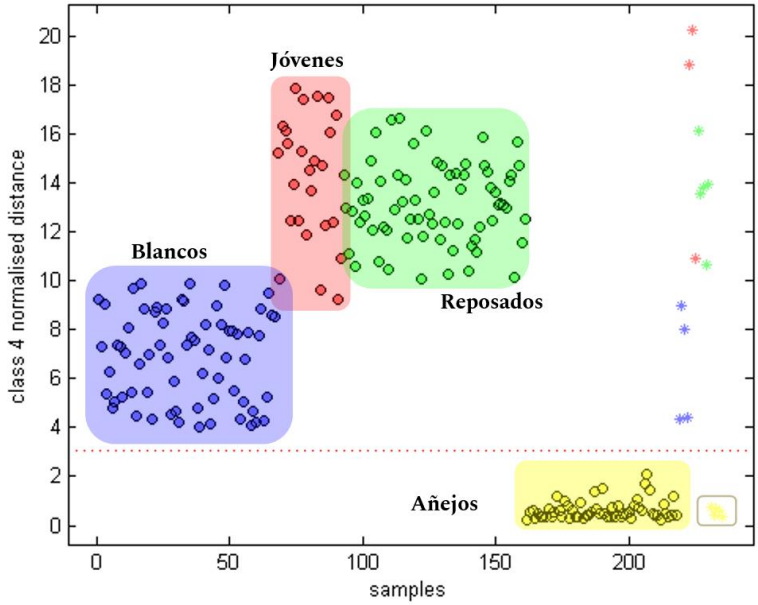



***Gráfica 46*. Distancia normalizada para los TR's vs. muestras a partir de los *scores*.**

Debido a estos resultados, no fue necesario realizar el gráfico de comparación de distancias entre TR's y cualquier otra clase. Con lo cual, fue posible aseverar que todas las muestras de predicción fueron correctamente asignadas (*gráfica 46b*).

Finalmente, se graficó la distancia normalizada para los tequilas añejos con los conjuntos de calibración-validación interna y predicción (*gráfica 47*). En ella, podemos ver que todas las muestras de predicción fueron asignadas correctamente, dado que se ubican debajo del límite de distancia (3.0).

Como se dijo al comenzar el análisis de modelos por SIMCA, otro resultado que ofrece esta técnica son los gráficos de residuales Q y *hotelling* $T^2$, que tienen como objetivo identificar muestras anómalas o *outliers*. Para facilitar su interpretación, en la *tabla 11*, se enlistan las muestras identificadas como *outliers* en cada técnica, donde, se puede ver que la única muestra en la que coinciden es la 140 (TR5577), además, en las *gráficas 42-47* no está registrada esta muestra como anómala, por lo que se decidió no eliminarla de ningún modelo.




***Gráfica 47*. Distancia normalizada para los TR's vs. muestras a partir de los *scores*.**

En cuanto a los tequilas blancos y jóvenes, las muestras consideradas *outliers* por el estadístico de residuales Q, fueron aquellas que se comportaban diferente a su conjunto en los modelos de las *gráficas 42-45*; no obstante, por *hotelling* $T^2$, no fueron consideradas anómalas, por lo que no se eliminaron tampoco de sus respectivos modelos.

Cabe mencionar que los estadísticos de residuales Q se refieren a los factores (o muestras) no modelados, mientras que, $T^2$ es la distancia al centro del grupo de clasificación, es decir, considera al resto de muestras del modelo.

***Tabla 16*. Comparación de *outliers* con residuales Q y *hotelling* T² con categoría y %v/v de etanol.**

| | **Residuales Q** | | **Categoría** | **%v/v Etanol** | ***Hotelling* T²** | | **Categoría** | **%v/v Etanol** |
|---|---|---|---|---|---|---|---|---|
| | **No. de tequila** | **Etiqueta del tequila** | | | **No. de tequila** | **Etiqueta del tequila** | | |
| **TB** | 8 | TB3138 | S/D | S/D | - | - | - | - |
| | 13 | TB4599 | 100% agave | 40.16 | | | | |
| **TJ** | 88 | TJ6914 | 100% agave | 40.30 | - | - | - | - |
| | T6 (*test* 6) | TJ5448 | Mixto | 40.25 | | | | |
| **TR** | 117 | TR4371 | 100% agave | 35.42 | **140** | **TR5577** | **100% agave** | **40.18** |
| | 135 | TR5436 | 100% agave | 35.32 | 141 | TR5590 | Mixto | 54.74 |
| | **140** | **TR5577** | **100% agave** | **40.18** | 149 | TR6912 | 100% agave | 49.59 |
| | 145 | TR6859 | 100% agave | 40.02 | 161 | TR8719 | 100% agave | 53.64 |
| | 150 | TR6915 | 100% agave | 38.20 | - | - | - | - |
| **TA** | 203 | TEA2493 | S/D | S/D | 173 | TA2848 | S/D | S/D |
| | 207 | TEA2847 | S/D | S/D | 176 | TA3043 | S/D | S/D |
| | 208 | TEA2912 | S/D | S/D | 187 | TA5288 | Mixto | 52.41 |
| | 217 | TEA8643 | 100% agave | 40.12 | 190 | TA5427 | 100% agave | 54.01 |
| | - | - | - | - | 206 | TEA2591 | S/D | S/D |

En el caso de *hotelling* T² para tequilas reposados, omitiendo la muestra TR5577 (40.18%), se trató de tequilas con más de 49 grados de alcohol, lo que podría suponer un comportamiento atípico en el conjunto, puesto que el resto tiene entre 35.15 y 41.09 grados (*tabla 16*). Al coincidir los análisis de Residuales Q y *hotelling* T² en la identificación de la muestra TR5577 como *outlier*, podría suponerse que esta muestra es de mala calidad.

En el caso de los tequilas añejos, por falta de información, y dado que en los modelos ninguno se ubica en otro conjunto, se decidió no eliminar ninguna muestra.

Con base en los parámetros de desempeño, tasas de error y muestras no asignadas (*tabla 14*), se dice que los modelos obtenidos por SIMCA fueron satisfactorios, aunque, comparados con otras técnicas, no son los más útiles en la clasificación de tequilas de acuerdo con su tiempo de añejamiento salvo que

se establezcan nuevas condiciones de modelado (en muestras contenidas y tratamiento de datos).

### 6.12 Muestras no asignadas en los modelos

Como comparación final de los modelos, en la *tabla 17* se enlistan los tequilas que fueron identificados como anómalos, a partir de la cual podemos decir que las muestras con mayor incidencia en comportamiento distinto al de su clase son: TEA2847 (4), TB4599 (3), TR5590 (3), TEA5426 (3), TB8710 (2), TJ4757(2), TR5452 (2), TR6912 (2), TR8719 (2), las cuales podrían ser realmente *outliers*; no obstante, esta información debe ser confirmada por el Consejo Regulador del Tequila.

***Tabla 17.*** **Muestras identificadas como *outliers* en diferentes técnicas supervisadas.**

| Tequila | Técnica | | | | |
|---|---|---|---|---|---|
| | PC | PF | k–NN | PCA – DA | SIMCA |
| TB4599 | ✓ | ✓ | | | Tendencia a TJ y TR. |
| TB5428 | | ✓ | | | |
| TB6848 | | ✓ | | | |
| TB8710 | ✓ | ✓ | | | |
| TJ2656 | | | | ✓ | |
| TJ4757 | | ✓ | ✓ | | |
| TJ5448 | | | | | Similitud con TB y TA. |

***Tabla 17.*** **(Continuación)**

| Tequila | Técnica | | | | |
|---|---|---|---|---|---|
| | PC | PF | k–NN | PCA – DA | SIMCA |
| TJ6914 | | | | | Tendencia a TJ. |
| TJ8733 | | ✓ | | | |
| TR5452 | ✓ | ✓ | | | |
| TR5577 | | ✓ | | | |
| TR5590 | ✓ | ✓ | ✓ | | |
| TR6912 | ✓ | ✓ | | | |
| TR8594 | | | ✓ | | |
| TR8641 | | | | | Tendencia a TJ. |
| TR8688 | | | | | Tendencia a TJ. |
| TR8719 | ✓ | ✓ | | | |
| TA3043 | | ✓ | | | |
| TA5427 | ✓ | | | | |
| TEA 2483 | | ✓ | | | |
| TEA2591 | ✓ | | | | |
| TEA 2847 | ✓ | ✓ | | ✓ | Similitud con TB. |
| TEA3094 | | | ✓ | | |

| | | | | | |
|---|---|---|---|---|---|
| **TEA5426** | | ✓ | | ✓ | Similitud con TB. |

**NOTA:** Las muestras en color **azul**, son aquellas que se identificaron como *outliers* en el conjunto de predicción.

**Resumen Capítulo III** – PF, k–NN, PCA–DA, PLS–DA y SIMCA

Para Funciones de Potencia (PF), se ocupó la matriz de 235 tequilas, datos autoescalados, intervalo 3000–1100 $cm^{-1}$, percentil de 95%, 2 componentes principales y validación cruzada mediante persianas venecianas con 10 grupos de cancelación. Además, los factores de suavizado fueron de 0.5 para tequilas blancos, 0.7 para tequilas jóvenes y de 0.6 para tequilas reposados y añejos. En este análisis, hubo trece muestras no asignadas durante la etapa de calibración (6%), veinte durante la validación (10%) y dos durante la predicción (12%). No obstante, tuvo excelentes parámetros de desempeño y tasa de error del 0%, por lo cual, se consideró un modelo adecuado de clasificación, aunque no el mejor con respecto a los obtenidos mediante otras técnicas.

El método del vecino más cercano (k-NN), se erigió con los espectros de 235 tequilas, todas las variables (4000-450 $cm^{-1}$), autoescalado, distancia euclídea y se realizó una validación cruzada mediante persianas venecianas con diez grupos de cancelación. El valor óptimo de k fue de 1 vecino, que tuvo excelentes parámetros de desempeño (1.0) y una tasa de error del 0%. Aquí, también se observó que sin escalado y k=1, el modelo no era tan malo (tasa de error del 1% y exactitud del 99% en la calibración y 98% en la validación interna -*cross validation*-, con 4 tequilas no asignados o asignados a otra clase).

El Análisis Discriminante por Componentes Principales (PCA–DA), se llevó a cabo con 235 muestras, de las cuales 218 fueron del conjunto de calibración-validación interna y 17 del conjunto de predicción, con autoescalado, intervalo completo (4000–450 $cm^{-1}$), tres factores discriminantes (al ser 4 clases), discriminación lineal y validación cruzada mediante persianas venecianas con diez grupos de cancelación. Tuvo una varianza explicada del 88.14% (PC1: 46.39%, PC2: 24.41% y PC3: 17.34%). El modelo más adecuado se construyó con el intervalo 3000–1100 $cm^{-1}$, el cual, tuvo una varianza explicada del 94.67% (PC1= 44.81, PC2= 32.23 y PC3= 17.63).

Posteriormente, para desarrollar el Análisis Discriminante por el Método de Mínimos Cuadrados Parciales (PLS–DA), se introdujeron los datos autoescalados del intervalo 3000–1100 $cm^{-1}$ de 235 muestras y se llevó a cabo el análisis con tres factores discriminantes, criterio de asignación bayes,

validación cruzada mediante persianas venecianas y diez grupos de cancelación. El modelo más adecuado fue construido con cuatro variables latentes (Varianza explicada del 96.27%: 44.78% para LV1, igual a 32.03% para LV2, 15.61% para LV3 y de 3.85% para LV4).

Finalmente, para la Modelación Suave e Independiente por Analogías de Clase (SIMCA), se ocupó la matriz de *scores* de 235 tequilas en el intervalo 3000–1100 $cm^{-1}$ con autoescalado (que implica la reducción se dimensionalidad de los datos), a partir de la cual, se encontró que las mejores condiciones para desarrollar un modelo satisfactorio fueron: dos PC's para TB, tres PC's para TJ y TA y cuatro para TR. Se aplicó validación cruzada con diez grupos de cancelación y el criterio de asignación fue *class modelling* con varianzas explicadas del **TB** (95.93%): PC1= 62.27% y PC2= 33.66%, **TJ** (98.55%): PC1= 76.33%, PC2= 21.04% y PC3= 1.18%, **TR** (98.57%): PC1= 87.95%, PC2= 6.94%, PC3= 2.3% y PC4= 1.38% y **TA** (97.73%): PC1= 86.23%, PC2= 10.07% y PC3= 1.43%.

# CAPÍTULO IV
# Productos alcohólicos (prueba de modelos)

Existen muestras formuladas física y químicamente en forma muy similar a un tequila auténtico y aunque estas son usualmente detectadas por las técnicas empleadas en el Consejo Regulador del Tequila, ha habido algunas que no fue posible identificar. Por ello, en el presente capítulo se exhibe el modelo de PCA-DA que permitió identificar como apócrifas a dos muestras con estas características. Cabe destacar que este modelo no fue diseñado para tal fin, no obstante, fue capaz de diferenciarlas de los tequilas auténticos.

Las muestras que se emplearon en la etapa de predicción en los *capítulos II y III* pasaron al conjunto de calibración y, de esta forma, las nuevas muestras (no autentificadas por el Consejo Regulador del Tequila), fueron las de predicción.

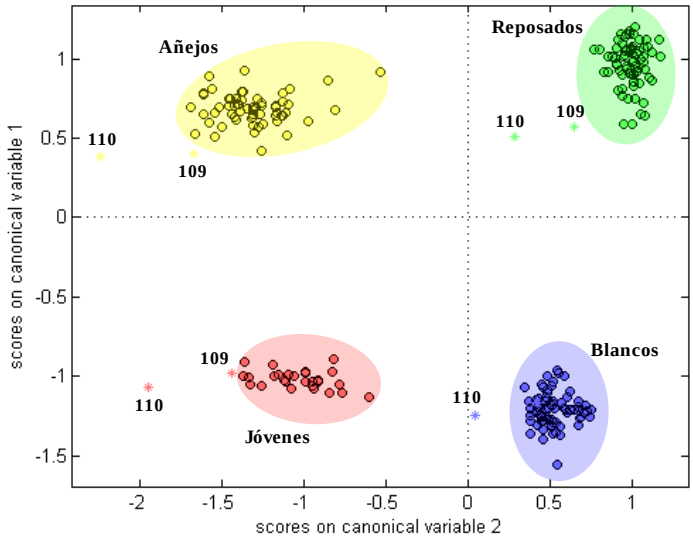
Para empezar, es necesario aclarar que las muestras empleadas en los siguientes modelos son formulaciones químicas, es decir, una persona con conocimiento en química y de las normas que regulan la bebida, las elaboró artificialmente, otorgándole a estas el sabor, color y aroma característicos.

El objetivo del presente capítulo es crear un primer acercamiento (no definitivo) a la aplicación de los modelos con muestras no auténticas. Para lo cual, se aplicaron los modelos con las técnicas expresadas en los *capítulos I, II y III*; sin embargo, la única que permitió la identificación de estos productos como *outliers* (sabiendo que se trataba de muestras ofertadas como TB 100% agave y con 55 grados de alcohol) fue el Análisis Discriminante por Componentes Principales (PCA-DA), abordado a continuación.

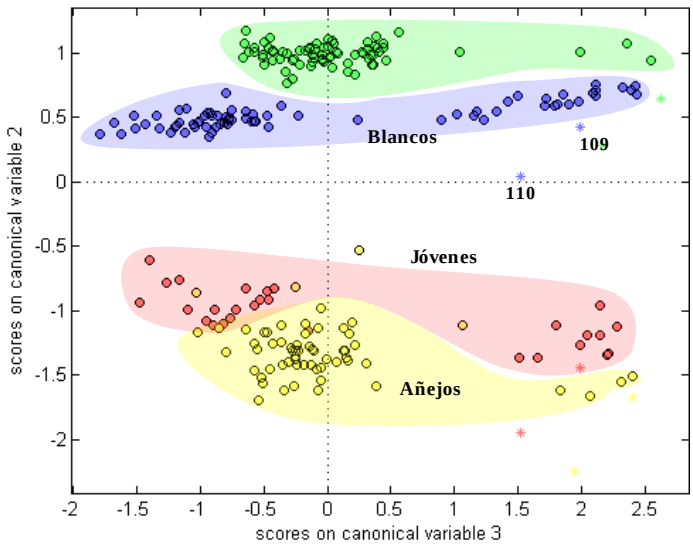
**Análisis Discriminante por Componentes Principales (PCA-DA)**

Las condiciones más adecuadas de modelado fueron: 235 tequilas, intervalo completo (4000–450 $cm^{-1}$, datos autoescalados, tres factores discriminantes, discriminante lineal y validación cruzada por persianas venecianas con diez grupos de cancelación. La varianza explicada por el modelo fue del 88.07% (PC1: 46.43%, PC2: 24.49% y PC3: 17.15%). El cual, tuvo excelentes parámetros de desempeño en todas sus etapas (calibración, validación y predicción), una tasa de error del 0% y una exactitud igual a la unidad.

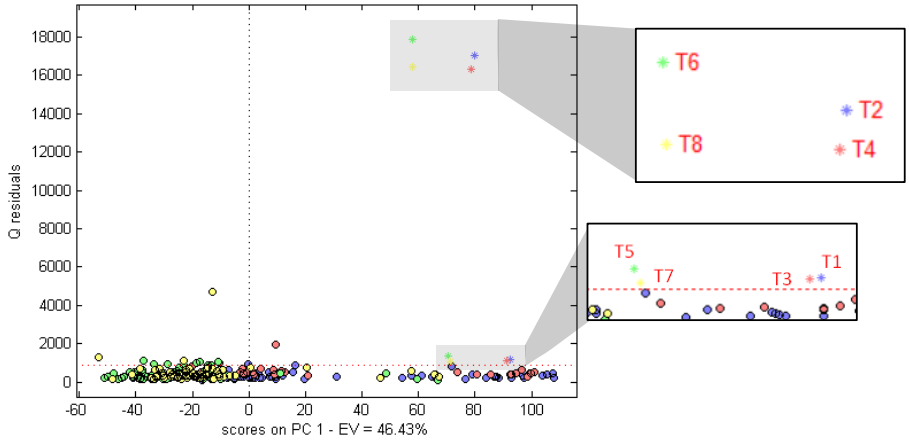
En la *gráfica 48*, se muestra el subespacio CV1-CV2 construido a partir de PCA-DA con 243 muestras (235 tequilas y 2 productos alcohólicos que no son tequilas con las CILB's de las cuatro clases), donde, fue posible observar que ningún producto alcohólico (109 y 110) entró en el conjunto de las clases jóvenes, reposados y añejos, no obstante, el PA109 si entró en el conjunto de los tequilas blancos.




***Gráfica 48.*** **Diagrama de *scores* de CV1-CV2 con 235 muestras de tequila del CRT y dos muestras identificadas como no auténticas.**




***Gráfica 49.*** **Diagrama de *scores* de CV1-CV3 con 235 muestras de tequila del CRT y dos muestras identificadas como no auténticas.**

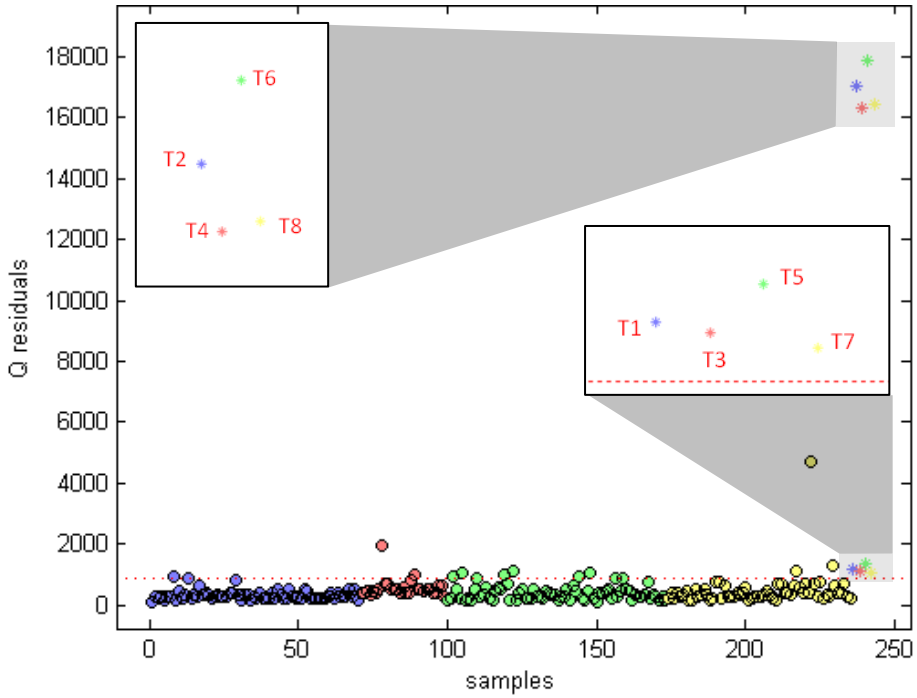



***Gráfica 50*. Diagrama de scores de PC1 vs. Residuales Q con PCA-DA.**

Por otro lado, en la *gráfica 50*, se exhibe un diagrama de *scores* (correspondientes al componente principal PC1) vs. el valor de residuales Q del análisis. En ella, se observó que todas las muestras de productos alcohólicos quedaron por arriba el límite trazado por el algoritmo, lo cual implica que estas muestras son *outliers*. En este caso, como se trata de muestras ofertadas como TB, fueron de interés únicamente la T1 (PA109) y T2 (PA110).

Finalmente, en la *gráfica 51*, se presentan los residuales Q contra todas las muestras, donde fue posible determinar que estas muestras (T1 y T2, específicamente), son *outliers*; especialmente la T2 que corresponde al TA110.

Después de obtener los presentes resultados, se plantearon diferentes ideas; una de ellas fue la posibilidad de realizar modelos bajo técnicas supervisadas y no supervisadas con muestras de tequilas blancos de aproximadamente 55% v/v de etanol y 100% agave, ya que los modelos trazados en la presente tesis fueron con el objetivo de discriminar entre clases y no si una muestra es apócrifa o no. Otra idea, fue utilizar un espectro de un tequila blanco y 100% agave bien comportado para restarle el espectro del producto alcohólico y así conocer la región del espectro que permita diferenciar un tequila de una muestra no auténtica.




***Gráfica 51.*** **Residuales Q vs. muestras (235 teq. y 2 pa. por clase) por PCA-DA.**

# 7. Conclusiones

En la presente tesis, se construyeron nueve modelos satisfactorios de clasificación de tequilas de acuerdo con su tiempo de añejamiento en cuatro clases (blancos, jóvenes, reposados y añejos/extra añejos), a partir de datos de espectrofotometría de FT-MIR de 236 muestras auténticas del Consejo Regulador del Tequila, mediante las técnicas no supervisadas: Análisis de Componentes Principales (PCA), Análisis Cluster (CA); y las técnicas supervisadas: *Support Vector Machines* (SVM), Curvas de Potencia (PC), Funciones de Potencia (PF), Vecino más cercano (k–NN), Análisis Discriminante por Componentes Principales (PCA–DA), Análisis Discriminante por el Método de Mínimos Cuadrados Parciales (PLS–DA) y por Modelación Suave e Independiente por Analogías de Clase (SIMCA).

El fundamento de cada técnica fue descrito en el marco teórico y constituye un material de estudio valioso para quienes deseen familiarizarse con dichas técnicas, ya que, al día de hoy, no son tópicos considerados todavía en los planes de estudio de las carreras de licenciatura en el área de química.

El conjunto de muestras en estudio consideró características muy variables, tal como diferentes regiones de la Denominación de Origen (DOT), procesos de elaboración distintos, diferentes clases, ambas categorías (100% agave y mixtos), diferente grado de abocamiento y contenido alcohólico (aproximadamente de 35 a 55 %), provenientes de diferentes casas tequileras y, por tanto, distintos lotes. Así, esta diversidad química, aunado al número de muestras analizadas, favoreció la representatividad de los modelos.

Se efectuaron etapas de desarrollo (entrenamiento) y de validación de los modelos, cuyos procesos son independientes. Es importante hacer notar que la utilización de estos modelos requerirá de su actualización mediante un conjunto de muestras de tequila de clase desconocida.

Por otra parte, se proyectaron y predijeron dos muestras externas no auténticas en los modelos óptimos construidos, donde se observó que, a pesar de que

estos no fueron construidos con el fin de diferenciar muestras apócrifas de auténticas, la técnica PCA–DA (capítulo IV) logró llevarlo a cabo.

En resumen, se logró establecer una estrategia química innovadora para la diferenciación de tequilas en función de su clase, a través de la construcción de modelos quimiométricos (supervisados y no supervisados) con uso de la técnica FT-MIR. La aplicación de dichos procedimientos de clasificación establece una opción viable de carácter químico y no físico (actualmente empleado por el CRT) para la identificación de muestras apócrifas. Los requisitos de partida para dichos métodos son la confiabilidad (veracidad y precisión), eficacia, rapidez y bajo costo, con lo cual, pueden ser utilizadas por las diversas instancias involucradas en la comercialización de tequila.

Finalmente, se logró coadyuvar a una estrategia química de carácter innovador para la mejora en el control de calidad del tequila, ofreciendo una alternativa eficiente y económica de análisis químico para el Consejo Regulador del Tequila.

## 8. Prospectivas

- Introducir en los modelos construidos en la presente tesis, muestras de clase desconocida químicamente, para verificar la funcionalidad de éstos.
- Elaborar modelos de clasificación, de acuerdo con la categoría de los tequilas descrita en la NOM-006, para diferenciar los 100% agave de los tequilas mixtos.
- Elaborar modelos de clasificación entre tequilas y destilados de agave.
- Desarrollar modelos de clasificación para tequilas auténticos y bebidas adulteradas.

# 9. Referencias


Libros

[1] Aldás, J., Uriel, E. (2017). Análisis Multivariante Aplicado con R. Madrid, España: Alfacentauro.

[2] Mongay, C. (2005). Quimiometría. España: Universitat de Valencia.

[3] Martínez, W.L. (2003). Estadística Descriptiva. Bolivia: La Hoguera.

[4] Ramis, G. et.al. (2010). Quimiometría. Madrid, España: Síntesis.

[5] Tuya, J. et.al. (2007). Técnicas cuantitativas para la gestión de la ingeniería del software. La Coruña, España: Netbiblo.

[6] Véliz, C. (2016). Análisis Multivariante: Métodos estadísticos multivariantes para la investigación. Buenos Aires, Argentina: Cengage Learning.

Tesis

[7] Arriagada, M. (2015). Comparación de métricas de distancia en el algoritmo k-vecinos más cercanos para el problema de reconocimiento automático de dígitos manuscritos. Valparaíso, Chile: Facultad de Ingeniería: Pontificia Universidad Católica de Valparaíso.

[8] García-Montiel, R. (2015). Espectrometría Infrarroja Media (MIR) y Reconocimiento de Pautas para la Diferenciación de Tequilas. México: FES Cuautitlán, UNAM.

[9] Mateos García, D. (2013). Ponderación local evolutiva de la regla KNN. España: Universidad de Sevilla.

[10] Salgado, I. (2000). Suavización no Paramétrica para Análisis de Datos. México: FES Zaragoza, UNAM.

[11] Fundamentos Teóricos de la Tesis Doctoral https://documentslides.org/the-philosophy-of-money.html?utm_source=teoria-quimiometria-regression-analysis-physics-mathematics

Artículos

[12] Ballabio, D. et.al. (año desconocido). Classification tools in chemistry. Part 1: Linear Models. PLS-DA. Analythical Methods. 5, 3790-3798.

[13] Bautista-Justo, M. et.al. (2001). El Agave tequilana Weber y la Producción de Tequila. Acta Universitaria, Vol. 11, No. 2. Guanajuato, México.

[14] Brereton, R.G. et.al. (2010). Support Vector Machines for classification and regression, Analyst, 135, 230-267.

[15] Comesaña, Y. et.al. (2009). Comparación de dos métodos supervisados de reconocimiento de patrones para la clasificación de destilados medios de petróleo mediante espectroscopia infrarroja. Revista CENIC Ciencias Químicas. Vol. 40, No. 2.
[16] Ferreira, M. et.al. (1999). Quimiometria I: calibração multivariada, um tutorial. Quím. Nova. Vol. 22 n.5 São Paulo, Brasil.
[17] Forina, M. et.al. (1991). A class-modelling technique based on potential functions. Journal of Chemometrics Vol. 5 Issue 5. 435-453
[18] Liang, Y. et.al. (2016). Support vector machines and their application in chemistry and biotechnology. CRC Press.
[19] Moreno, A. et.al. (año desconocido). El análisis de correlación canónica como instrumento para la evaluación de la eficiencia. España.
[20] Pérez, G. et.al. (2017). Authentication of tequilas using pattern recognition and supervised classification. Trends in Analytical Chemistry 94 (2017) 117-129.
[21] Zayas, E. et.al. (2014). Clasificación multivariante de rones añejos cubanos. Revista Cubana de Ingeniería, 5(2), 62-67.

Sitios WEB

[22] Diario Oficial de la Federación. (2005). NOM-006-SCFI-2005. 04/Marzo/2018. Sitio web: https://www.crt.org.mx/images/Documentos/NOM-006-SCFI-2005.pdf
[23] Consejo Regulador del Tequila A.C. 04/Marzo/2018. Sitio web: https://www.crt.org.mx/
[24] Historia del CRT 04/Marzo/2018 Sitio web: https://www.crt.org.mx/index.php?option=com_content&view=article&id=29
[25] Reglamento Interior del Consejo Regulador del Tequila A.C. 04/Marzo/2018 Sitio web: https://www.crt.org.mx/index.php?option=com_content&view=article&id=32
[26] Gómez, R., Murillo, R. (2005). Espectroscopia Infrarroja. 04/Marzo/2018, de Facultad de Ciencias, UNAM Sitio web: http://sistemas.fciencias.unam.mx/~fam/Infrarroja.pdf
[27] Arriortua, M.I. et. al. (2006). Espectroscopia Infrarroja. 04/Marzo/2018, de Facultad de Ciencia y Tecnología Sitio web: http://www.ehu.eus/imacris/PIE06/web/IR.htm#IR

[28] Alzate, E.J. (2011). Espectroscopia Infrarroja. 16/Marzo/2018, de Universidad de Pereira Sitio web: http://webdelprofesor.ula.ve/ ingenieria/antonioc/IR

[29] SICE, OAS. (2015). Espectrofotometría Infrarroja. 16/Marzo/2018.Sitio web: http://www.sice.oas.org/Trade/MRCSRS/Resolutions/RES_011_2015_s.pdf

[30] Sancho, F. (2017). Clasificación Supervisada y No Supervisada. 17/Marzo/2018, de Fernando Sancho Caparrini. Sitio web: http://www.cs.us.es/~fsancho/?e=77

[31] Todeschini, R. (2016). Classification Toolbox for Matlab - versión 4.2. 17/Marzo/2018, de Milano Chemometrics and QSAR Research Group Sitio web: http://michem.disat.unimib.it/chm/download/softwares/help_classification/index.htm

[32] GenEx Enterprise. Potential Curves. 19/Marzo/2018, de Multid Analysis Sitio web: https://www.multid.se/genex/onlinehelp/hs510.htm

[33] Portal Web de Información Estadística del CRT. 28/Agosto/2018. Sitio web: https://www.crt.org.mx/EstadisticasCRTweb/

[34] Demey, J. et. al. (2011) Capítulo V. Medidas de Distancia y Similitud. 28/Agosto/2018. Sitio web: https://www.researchgate.net/profile/Fernando_Casanoves/publication/260137073_MEDIDAS_DE_DISTANCIA_Y_SIMILITUD/links/5669c17b08ae430ab4f73d91/MEDIDAS-DE-DISTAN CIA-Y-SIMILITUD.pdf

[35] Módulo I. Método de Técnicas Estadísticas para el Desarrollo Territorial y Local. Máster en Ordenación y Gestión del Desarrollo Territorial y Local. Universidad de Sevilla, España. Impartido por el Dr. Ángel Luis Lucendo Monedero. 28/Agosto/2018 Sitio Web: http://titulaciongeografia-sevilla.es/master/archivos/recursos/ACluster_Alumnos_R.pdf

[36] Saavedra, P., Ibarra, V.H. El método Monte-Carlo y su aplicación a finanzas. UAM-Iztapalapa/ESFM-IPN. México. 29/Agosto/2018. Sitio Web: http://educommons.anahuac.mx:8080/eduCommons/matematica-aplicada/simulacion-seguros-y-finanzas/TEM_A01_Lectura%20de%20Numeros%20aleatorios.pdf

[37] Gómez, C. (2017) Validación de los cálculos de incertidumbre en química analítica con el método Monte Carlo. Parte I. 29/Agosto/2018. Sitio Web: https://www.analytical.cl/post/validacion-calculos-incertidumbre-quimica-analitica-metodo-monte-carlo/

[38] T-squared, Q-residuals and Contributions. 29/Agosto/2018. Sitio Web: http://wiki.eigenvector.com/index.php?title=T-Squared_Q_residuals_and_Contributions# Relative_Q_Contributions

[39] Capítulo 6. Análisis Cluster. 7/Septiembre/2018. Sitio web: http://halweb.uc3m.es/esp/Personal/personas/imolina/MiDocencia/TecnicasInvestigacion/SlidesAClusterEstudio809.pdf

[40] Estarrón-Espinosa, M. (2004). Calidad del tequila, composición volátil. Ciencia y Desarrollo, CONACYT. México. 7/Septiembre/2018. Sitio web: http://cienciaydesarrollo.mx/?p=articulo&id=288

[41] Prasad, D. (2015). What is leave-one-out cross validation? How can I use it for image fusion? ResearchGate. 18/Septiembre/2018. Sitio web: https://www.researchgate.net/post/What_is_leave-one-out_cross_validation_How_can_I_use_it_for_image_fusion

[42] Gorostizaga, J.C. Fundamentos matemáticos: 6. Matrices y determinantes. Escuela técnica superior de naútica y máquinas navales. 19/Septiembre/2018. Sitio web: http://www.ehu.eus/juancarlos.gorostizaga/mateI15/T_matrdeter/MatrDeter

[43] Calculadora de valores propios (eigenvalores). 19/Septiembre/2018. Sitio web: https://es.symbolab.com/solver/matrix-eigenvalues-calculator/eigenvalores%20%5Cbegin%7Bpmatrix%7D1%262%261%5C%5C6%26-1%260%5C%5C-1%26-2%26-1%5Cend%7Bpmatrix%7D

[44] Artículo en web de autor desconocido. (2010). Descomposición en valores singulares. 19/Septiembre/2018. Sitio web: http://www.mate.unlp.edu.ar/practicas/70_18_0911201012951

Cursos

[45] Extracción de Información a Partir de Datos Analíticos Multivariables. Impartido por el Dr. José Manuel Andrade Garda del 25 al 29 de Junio del 2018 en la Facultad de Estudios Superiores Cuautitlán, UNAM. Estado de México, México.