El análisis exploratorio realizado al comienzo del trabajo es en parte una repetición del primer trabajo práctico. Sin embargo, se pueden sacar algunas conclusiones útiles:

El gráfico por pares permite apreciar que la mayoría de las combinaciones de dos variables permite una discriminación entre las dos variedades de vino. La cantidad de alcohol, de ácido cítrico y la calidad no son muy buenas para clasificación debido al solapamiento de sus distribuciones para las dos variedades.

Ninguna de las variables numéricas tiene una distribución normal. Eso necesariamente implica que se rechazará la hipótesis de normalidad multivariada. Se rechazó tanto la homocedasticidad como la igualdad entre los vectores medias de los atributos de los dos tipos de vinos.

Al realizar un análisis por componentes principales, se utiliza el criterio de Kaiser por dos razones: la cantidad de componentes mínima recomendada es razonablemente pequeña y, al menos para este trabajo, funcionó razonablemente bien.

Elegir los tres primeros componentes llega a explicar dos tercios de la varianza (lo que puede parecer insuficiente), aunque como puede verse en el biplot, inclusive dos componentes son suficientes para poder generar dos clústeres distintivos correspondientes a cada variedad de vino. Componentes adicionales permitirían refinar el resultado. Uno puede concluir de este hecho que la varianza del base de datos condensa adecuadamente la información presente.

Una posible interpretación de las componentes (a partir de lo que puede observarse en el corrplot tanto como del biplot) es: PC1 es principalmente influenciada por los anhídridos sulfurosos y en menor medida por la acidez volátil, por lo que puede entenderse a PC1 como “anhídridos”, PC2 es principalmente influenciada por densidad y alcohol por lo que puede entenderse como “contenido alcohólico” y PC3 es principalmente influenciada por el ácido cítrico. Igualmente son resultados que deberían tomarse con cierto escepticismo debido a que muchas otras variables tienen peso también.

El análisis exploratorio muestra que no se cumplen los supuestos del análisis discriminante (normalidad multivariada) ni tampoco se cumple con la homocedasticidad. Sin embargo, es un muy buen clasificador en la práctica (el lineal funciona mejor que el cuadrático en este caso), y parece ser una buena idea considerarlo inclusive si se violan los supuestos.

En el análisis discriminante lineal (y el cuadrático), se obtiene una sola dimensión debido que se pueden obtener como máximo un número de dimensiones igual al número de clases menos uno (no se necesitan más funciones discriminantes para poder clasificar observaciones adecuadamente). Lo que se puede observar de la distribución de los resultados de la función discriminante es que hay una muy buena separación entre ambas variedades de vino.

Se me ocurren tres posibles explicaciones de por qué funcionan estos algoritmos a pesar de que no se cumplen los supuestos:

La base de datos es trivial de separar en el espacio, por lo que no importa mucho las desviaciones de los supuestos.

El análisis discriminante lineal funciona como una especie de perceptrón que resulta ser útil para este problema (este también genera un hiperplano).

(probablemente la interpretación más interesante, aunque muy especulativa) El capítulo de clasificadores lineales del libro The Elements of Statistical Learning, de Hastie, Tibshirani y

Friedman, afirma que una regresión que codifica las dos clases con diferentes valores puede utilizarse como un clasificador binario y que sus coeficientes son proporcionales a los del análisis discriminante lineal. Esto relajaría la suposición de normalidad multivariada (aunque la regresión tiene sus supuestos), porque indirectamente se estaría realizando esta regresión. También explicaría por qué el clasificador lineal funciona mejor que el cuadrático.

Las máquinas de soporte vectorial resultaron ser muy precisas. Se observa que cambiar el umbral (threshold) de predicción puede reducir la precisión de clasificación; en este caso, valores alrededor de 0,5 funcionan bien en la práctica. Algo curioso pero que es probable que sea accidental es que los modelos generados a veces predicen mejor el conjunto de prueba que el de entrenamiento.

El kernel lineal presentó resultados mejores que el sigmoide, pero el radial es el mejor. Aún así, todos los métodos dan resultados similares, lo que hace sospechar que el problema de clasificación es trivial.

De entre los métodos supervisados, el orden de peor a mejor (por accuracy) es SVM sigmoide, QDA, SVM lineal/LDA y SVM radial, aunque la diferencia es insignificante para los tres últimos.

Mediante el análisis de conglomerados se logró recuperar los grupos originales correspondientes a cada variedad de vino mediante la elección correcta de los hiperparámetros del modelo.

La elección de la distancia euclídea y el número de grupos es obvia en este caso (debido a que se usan atributos numéricos y porque ya se sabe de antemano cuántos grupos se deben encontrar, aunque quizás debería hacerse una evaluación honesta del número de clústeres en trabajos futuros).

La elección de la distancia Ward para evaluar la distancia entre clústeres es menos obvia. Esta distancia busca minimizar la varianza en el interior de un clúster. También se conocido que una interpretación alternativa del análisis discriminante lineal, la de Fisher, define la separación entre dos grupos de observaciones como el coeficiente entre la varianza entre grupos y dentro de los grupos y busca una transformación que maximice la separación entre grupos. El hecho que el discriminante lineal funciona razonablemente bien permite sospechar que la minimizar la varianza dentro de cada clúster es la mejor opción.

Para K-medias, el gráfico de suma de errores cuadrados muestra un decrecimiento monótono por lo que no es de mucha utilidad para estimar el número óptimo de clústeres. El ancho de silueta promedio muestra sus mejores resultados para cinco clústeres (recordar que valores cercanos a uno significan una separación adecuada, valores cercanos a cero son indiferentes y negativos indican una separación incorrecta).

K-medias con dos clústeres aproximadamente recuperó los grupos originales. K-medias con cinco clústeres genera dos clústeres correspondientes a vino blanco y tres a tinto (aunque uno es un clúster prácticamente hecho de outliers. Aunque en principio se mezclan más los vinos de las dos variedades en cada clúster comparado con dos clústeres, esto no significa que los clústeres sean peores. Podría haber algún patrón oculto (¿quizás bajo ciertos valores de algunas de las variables?) que tenga sentido.

De entre los métodos no supervisados, K-medias con dos clústeres es el que más se aproxima a recuperar los grupos originales, seguido por el método jerárquico. Aunque es más difícil elegir elegir los hiperparámetros del modelo jerárquico.

Algo interesante para probar en trabajos futuros sería remover aquellas variables cuyas distribuciones se solapan demasiado entre diferentes variedades por considerarse que su poder de discriminación es bajo.