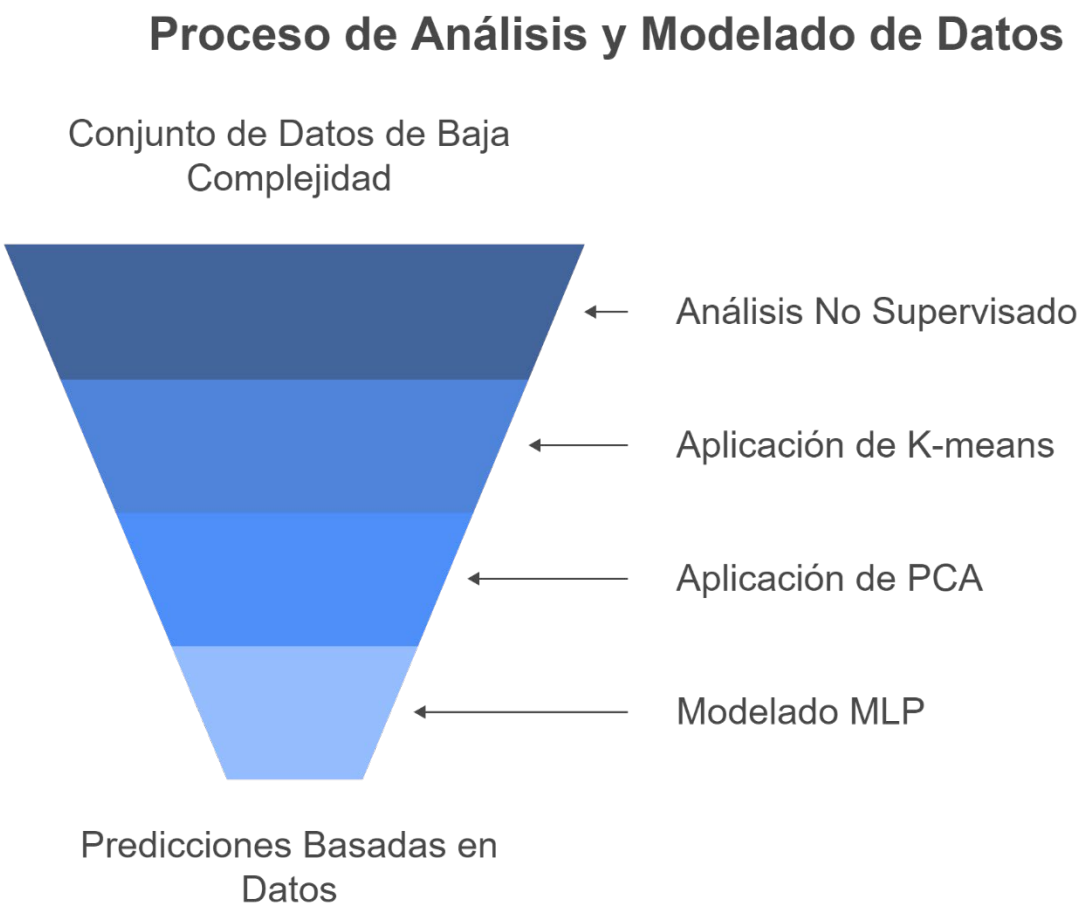
**Examen de cinturón AML – Opción A – Customer Personality Analysis**

**Introducción**

El presente informe documenta un análisis detallado del dataset "Customer Personality Analysis", con el propósito de:

Aplicar técnicas de clustering o reducción de dimensionalidad para analizar los datos, seguido de la implementación de un modelo MLP para realizar predicciones.



**Planteamiento del problema**

El análisis de la personalidad del cliente es un análisis detallado de los clientes ideales de una empresa. Ayuda a la empresa a comprender mejor a sus clientes y les permite modificar los productos de acuerdo con las necesidades, los comportamientos y las inquietudes específicas de los diferentes tipos de clientes.

El análisis de la personalidad de los clientes ayuda a una empresa a modificar su producto en función de sus clientes objetivo de distintos tipos de segmentos de clientes. Por ejemplo, en lugar de gastar dinero para comercializar un nuevo producto a todos los clientes de la base de datos de la empresa, una empresa puede analizar qué segmento de clientes tiene más probabilidades de comprar el producto y luego comercializarlo solo en ese segmento en particular.

**Fuente de Datos**

**Dataset:** Customer Personality Analysis

**Origen:** Kaggle

**Exploración y Preprocesamiento de Datos**

**Carga y Exploración de los Datos:**

El conjunto de datos fue cargado y analizado para identificar las variables clave en la clasificación de tipos de compras. Este conjunto de datos incluye nivel de educación, estado civil, nivel de ingresos, tipos de compañas aceptadas y preferencias de compras.

**Manejo de Valores Nulos y Normalización:**

Se verificaron si existen valores faltantes en el conjunto de datos, los resultanos arrojaron algunos valores, esos fueron imputados por un modelo de regresión simple capaz de entrenar con filas donde los valores no fueran nulos. Posteriormente, las variables fueron estandarizadas mediante StandardScaler. Este paso es fundamental

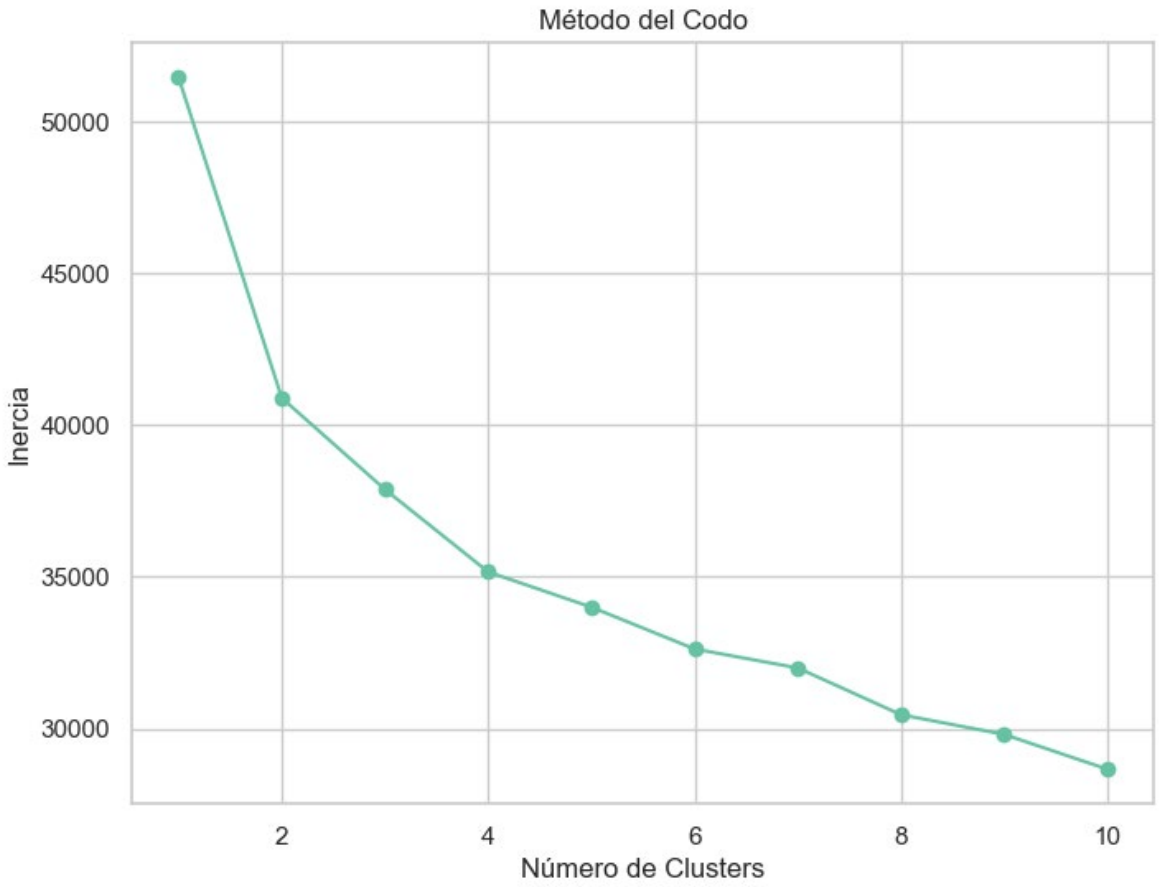
para garantizar que todas las variables tengan una escala uniforme, evitando que algunas influyan más que otras debido a discrepancias en sus magnitudes.

**División de Datos:**

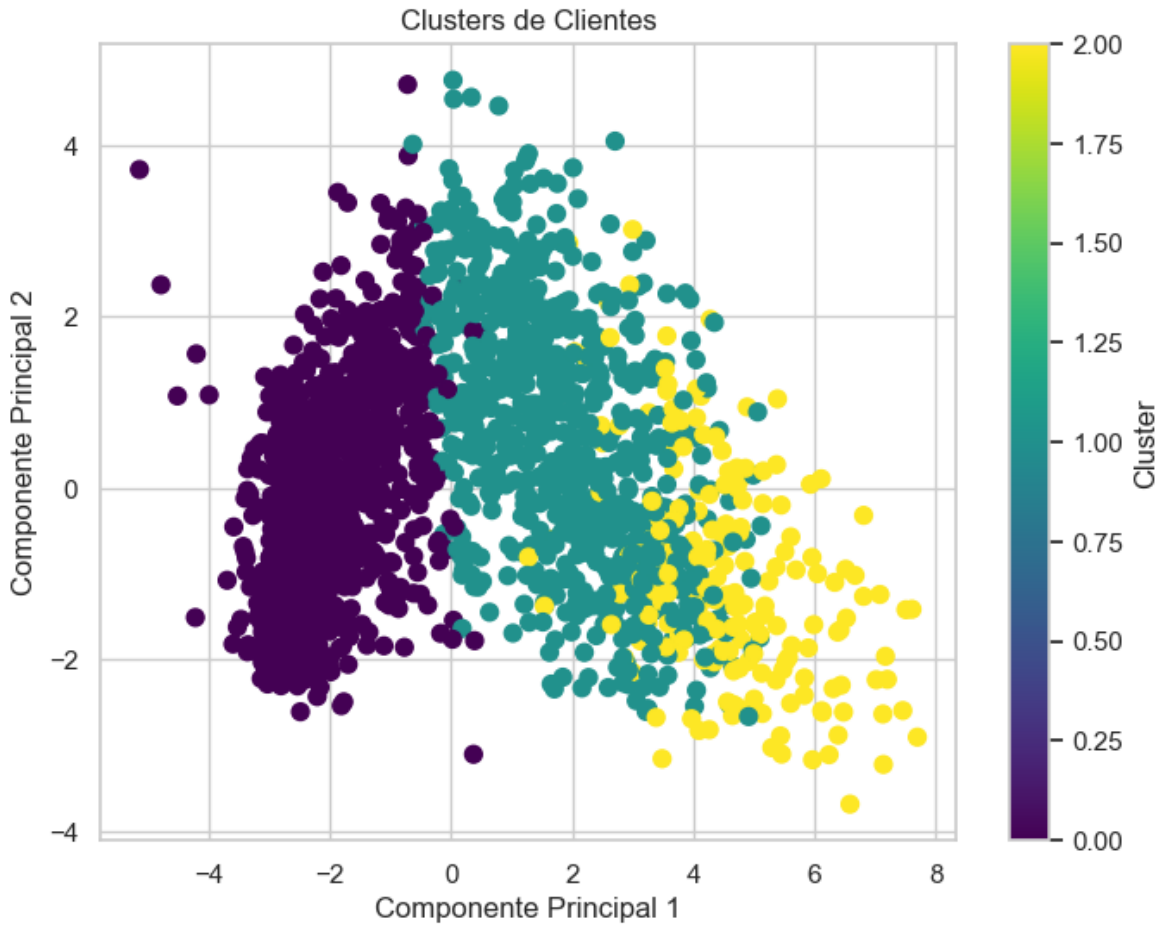
El conjunto de datos se dividió en subconjuntos de entrenamiento y prueba. La partición se llevó a cabo de manera que todas las variables estuvieran presentes en ambos conjuntos, asegurando que los modelos se entrenaran con una muestra representativa de las clases y pudieran evaluar su desempeño en datos no observados.

**Análisis No Supervisado Clusters y (PCA)**

Primer se hizo una agrupación por clusters, se determino el numero máximo mediante el método el codo.



Luego se utilizó el Análisis de Componentes Principales (PCA) para disminuir la dimensionalidad del conjunto de datos y observar las relaciones entre las distintas actividades físicas. PCA facilita la identificación de las componentes principales que explican la mayor parte de la variabilidad de los datos.

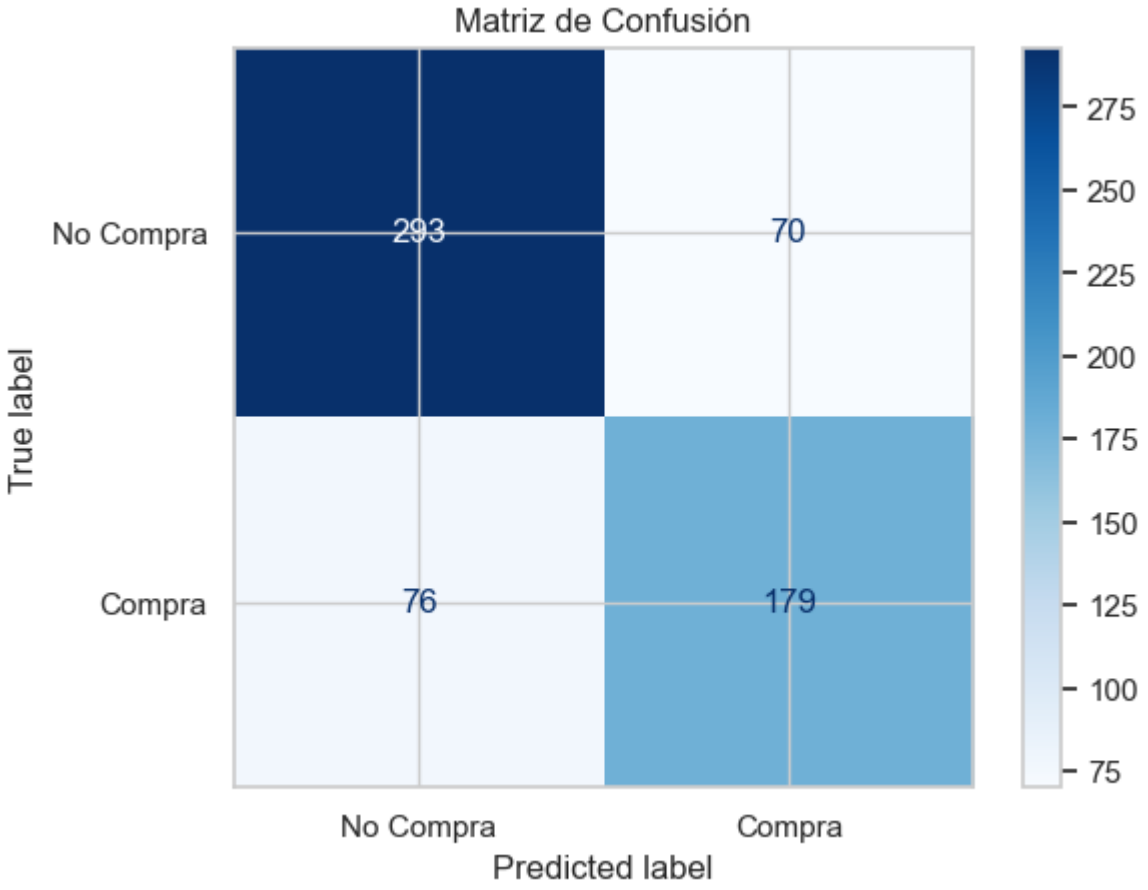


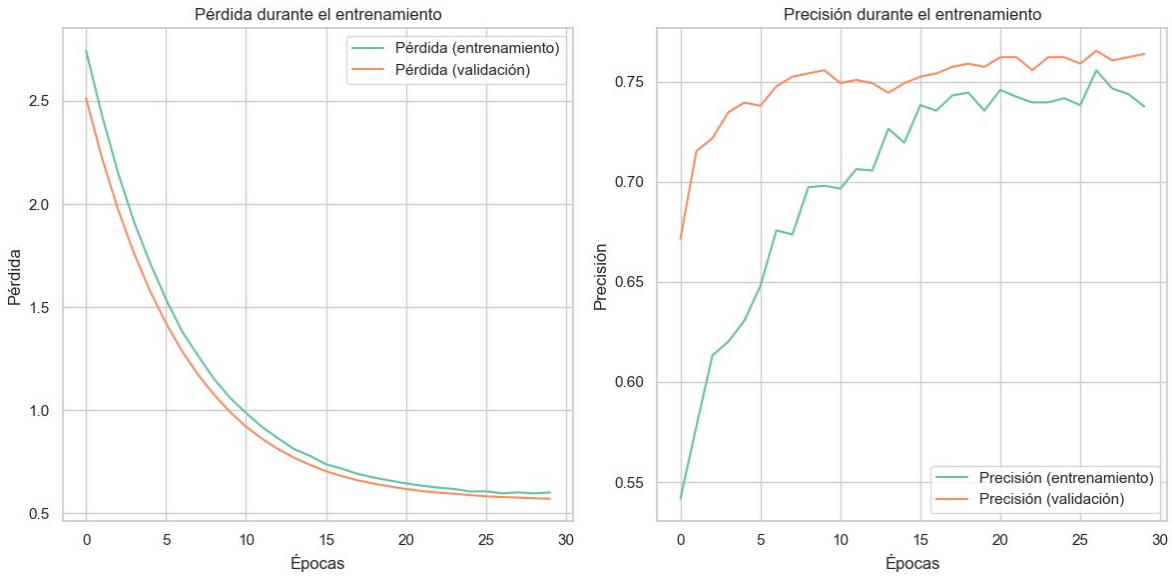

El modelo fue capaz de separar los componentes y dividirlos, aun así es difícil explicar porque PCA hace que se pierda la interpretabilidad.

**Modelado con MLP (Perceptrón Multicapa)**

El modelo MLP fue desarrollado para predecir las probabilidades de compra a partir de los datos de los clientes. Se diseñaron dos versiones del modelo:

La predicción se baso en la probabilidad de compra en MntWines

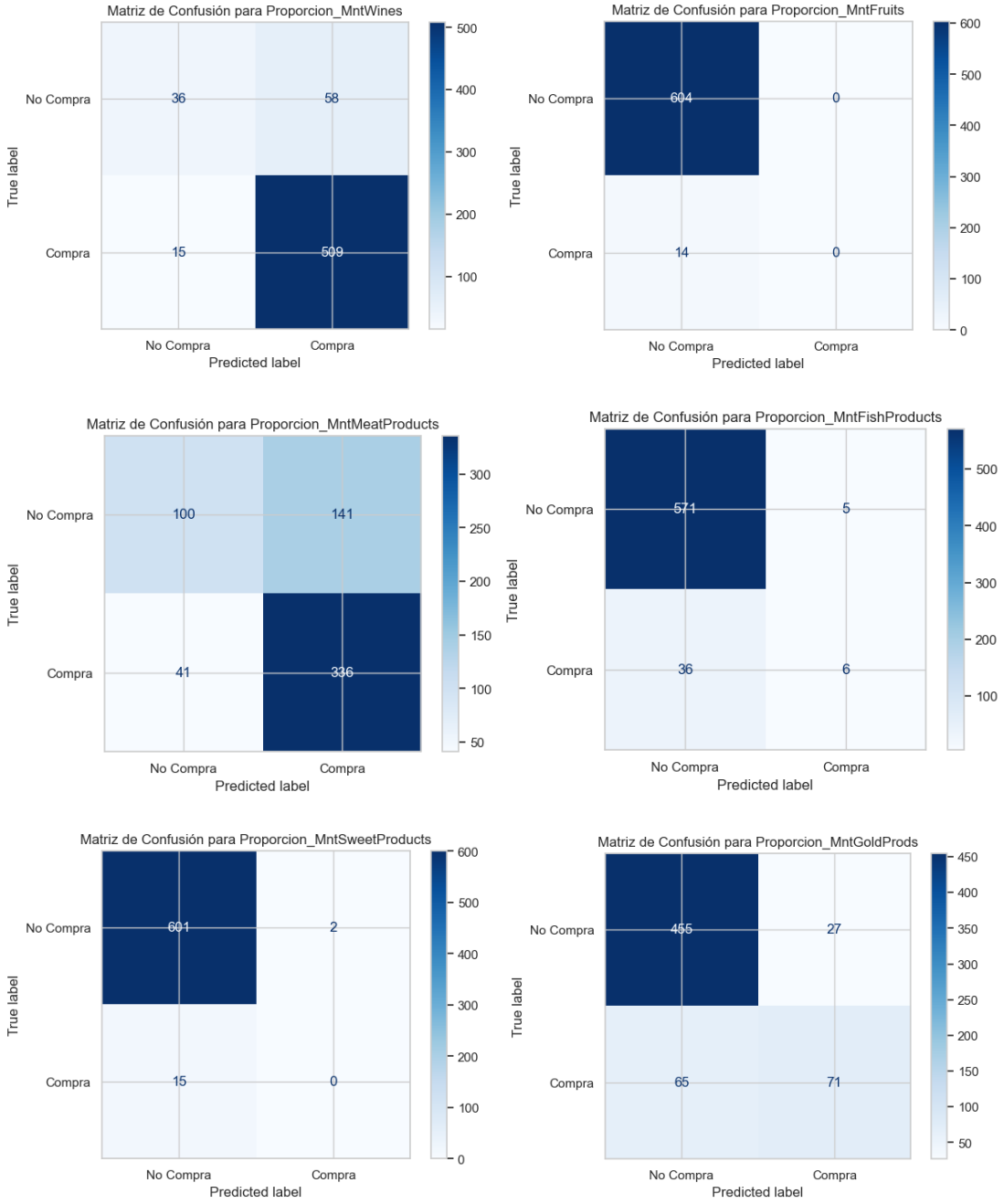





El modelo fue capaz de predecir si un cliente realiza compras en la categoría de vinos con una precisión del 75% en el conjunto de prueba. La matriz de confusión muestra que el modelo tuvo 77 falsos negativos y 72 falsos positivos.

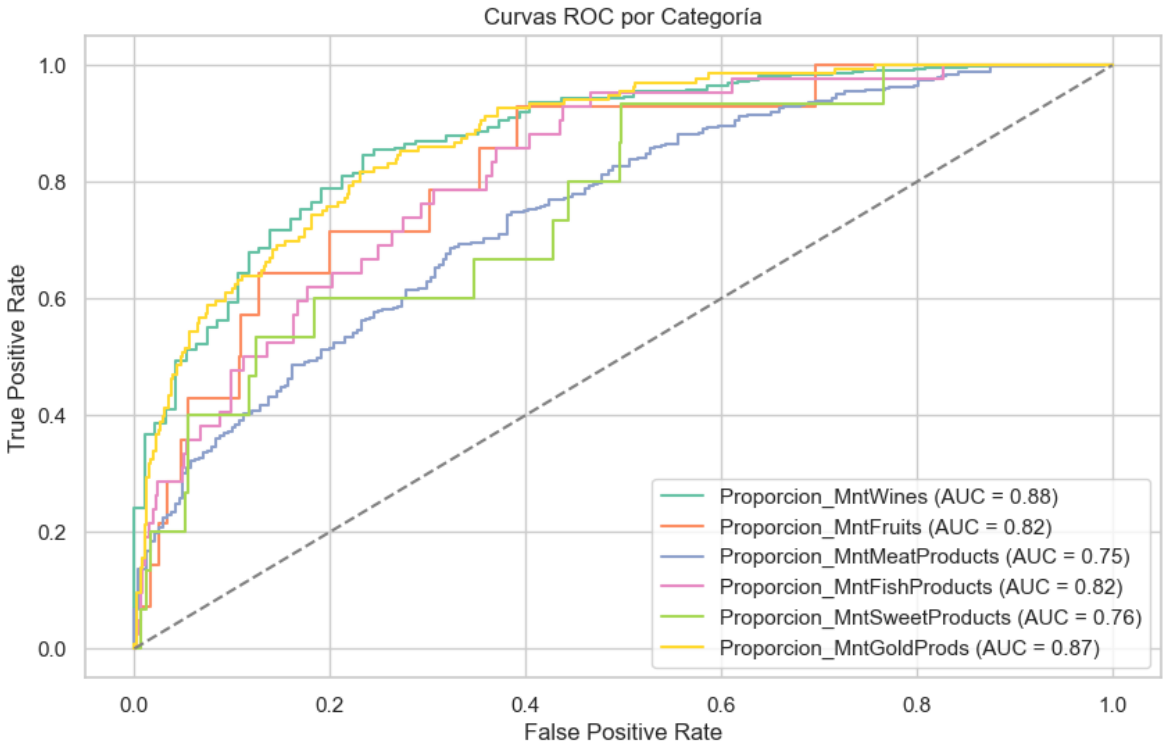
**Ramdon Forest Regressor**

El segundo modelo fue un random forest para poder comparar con el MLP, este modelo arrojo resultados aceptables, no muy buenos por la complejidad de lo que pide el examen.




Los resultados en las matrices arrojaron que hay productos que son mejores para predecir si una persona lo va a comprar o no, también hay otros que pudieron tener

resultados pesimos a la hora de predecir si realmente compra, dando muchos falsos positivos.




La curva roc muestra que MntMeatProducts es difícil de predecir, esto se debe a la falta de datos en muchos clientes, no podemos asumir que fueron errores a la hora de cargar dichos datos.

El proyecto aborto probabilidades de compra en distintas categorias, los resultados fueron aceptables, pero hay muchos falsos positivos a la hora de predecir.

**Conclusiones**

Se realizo un modelo de MLP con capas muy simples y parametros especificos porque el modelo siempre se sobreajusta. Aun asi, en las epocas mayores a 6 ya podemos notar que el modelo tuvo overfiting. Se puede solucionar pero requiere mas tiempo de analisis.

Cabe aclarar que los datos no estan preparados para ese tipo de analis, sin embargo se opto por hacer alguns calculos para adaptar los modelos a lo que pide la prueba.

**Recomendaciones**

Se podria probar con mas feactures engineering, tambien se puede optar por otros tipos de modelos como XgBoost Clasifier para predecir categorias, o regreseros en caso de agrupar por nivel de ingresos.