Guía de Interpretación

Visión General del Archivo

El archivo Excel contiene 10 hojas, las cuales se mencionan a continuación:

Estructura del Archivo

{clase objetivo}_Complete_Analysis.xlsx

1_Summary	→	Resumen
2_Simple_Predictions	→	Predicciones SÍ/NO
3_Detailed_Predictions	→	Información completa
4_High_Confidence	→	Candidatos prioritarios
5_All_Model_Stability	→	Detalles técnicos de modelos
6_Consensus	→	Acuerdo entre modelos
7_Model_Evaluation	→	Métricas de rendimiento
8_Feature_Importance	→	Qué descriptores importan
9_Top_Features_Stats	→	Estadísticas de mejores características
10_Diagnostics	→	Información del conjunto de datos
11_PCA_coordinates	→	Coordenadas en espacio PCA

Hoja 1: Summary

Contenido

Seccion	Qué Significa
Dataset Info	
Target Class	La actividad que estás buscando
Comparison	Contra qué se compara
Model Performance	
Best Model AUC	Qué tan bien predice el modelo (0.5-1.0)
Number of runs	Cuántas veces se entrenó para estabilidad
Predictions Summary	
Total predictions	Cuántos compuestos nuevos se clasificaron
Predicted as {target_class} (50%)	Cuántos predichos como {target_class}
High confidence candidates	Cuántos con alta confianza
Models agree (90%)	Cuántos donde todos los modelos concuerdan

Interpretación

Best Model AUC:

- **0.95-1.0**: Excelente (pero hay que verificar si no hay fuga de datos)
- **0.85-0.94**: Muy bueno
- **0.75-0.84**: Bueno
- **0.65-0.74**: Aceptable
- **<0.65**: Pobre – se tienen que revisar los datos/descriptores

High confidence candidates:

- Este número dice cuántos compuestos se deben priorizar para la validación experimental
- Estos tienen la mayor probabilidad de ser verdaderos positivos

Models agree (90%):

- Cuando los 3 modelos concuerdan fuertemente, la predicción es más confiable
- Estos son los candidatos MÁS seguros

Hoja 2: Simple_Predictions

Columnas

Columna	Significado
ID	Identificador del compuesto
Is_{target_class}	Predicción binaria simple
AFP_Probability_%	Probabilidad como porcentaje (0-100%)
Confidence	Nivel de confianza en la predicción

Niveles de Confianza

Very High (Muy Alta)

- Probabilidad >90% o <10%
- Los 3 modelos están muy seguros
- **Acción:** Máxima prioridad para validación

High (Alta)

- Probabilidad >70% o <30%
- Al menos 2 modelos concuerdan
- **Acción:** Alta prioridad para validación

Medium (Media)

- Probabilidad >60% o <40%
- Acuerdo mixto de modelos
- **Acción:** Baja prioridad para validación

Low (Baja)

- Probabilidad 40-60%
- Modelos no concuerdan
- **Acción:** No priorizar, predicción incierta

Hoja 3: Detailed_Predictions

Columnas Principales

Columna	Qué Te Dice
ID	Identificador
Ensemble_Probability	Probabilidad promedio de los 3 modelos (0-1)
Prediction_50	Predicción con umbral estándar (50%)
Prediction_70	Predicción conservadora (70%)
Prediction_90	Predicción muy conservadora (90%)
Confidence	Nivel de confianza general
Models_Agree_90	Cuántos modelos están >90% seguros (0-3)
RF_Prediction	Predicción del Random Forest
NN_PCA_Prediction	Predicción de la Red Neuronal con PCA
NN_DESC_Prediction	Predicción de la Red Neuronal con ANOVA
Interpretation	Resumen legible

Entendiendo los Umbrales

Umbral 50% (Prediction_50)

- **Cuándo usar:** Baja seguridad, resultado exploratorio
- **Ventaja:** Captura más candidatos potenciales
- **Desventaja:** Más falsos positivos

Umbral 70% (Prediction_70)

- **Cuándo usar:** Para reducir falsos positivos
- **Ventaja:** Mejor balance
- **Desventaja:** Se pueden perder algunos verdaderos positivos

Umbral 90% (Prediction_90)

- **Cuándo usar:** Umbral estricto, para asegurar solo positivos mas robustos
- **Ventaja:** Mínimos falsos positivos
- **Desventaja:** Lista pequeña de candidatos

Models_Agree_90 - Interpretación

Valor	Significado	Acción
3	Los 3 modelos muy confiados (>90%)	Máxima prioridad - casi seguro
2	2 modelos muy confiados	Alta prioridad - muy probable
1	Solo 1 modelo muy confiado	Prioridad media - revisar con cuidado
0	Ningún modelo muy confiado	Baja prioridad - alta incertidumbre

Hoja 4: High_Confidence

Qué Contiene

Solo compuestos que cumplen **ambos criterios**:

1. Predicción al 70% = clase objetivo
2. Confianza = "High" o "Very High"

Columnas Clave

Mismas columnas que Hoja 3 (Detailed_Predictions), pero solo los mejores candidatos

Hoja 5: All_Model_Stability

Columnas

Columna	Qué Significa
ID	Identificador
RF_Mean	Probabilidad promedio del Random Forest
RF_SD	Desviación estándar del RF (qué tan estable)
RF_Freq	Frecuencia de veces que RF predijo >90%
NN_PCA_Mean	Probabilidad promedio de Red Neuronal PCA
NN_PCA_SD	Desviación estándar de NN_PCA
NN_PCA_Freq	Frecuencia >90% para NN_PCA
NN_Desc_Mean	Probabilidad promedio de Red Neuronal ANOVA
NN_Desc_SD	Desviación estándar de NN_DESC
NN_Desc_Freq	Frecuencia >90% para NN_DESC

Interpretando la Desviación Estándar (SD)

SD < 0.02 (Muy estable)

- El modelo da predicciones muy consistentes
- Alta confiabilidad

SD 0.02-0.05 (Estable)

- Variación aceptable
- Confiabilidad buena

SD > 0.05 (Inestable)

- El modelo varía mucho entre ejecuciones
- Predicción poco confiable para este compuesto

Interpretando la Frecuencia (Freq)

Freq > 0.9 (Muy consistente)

- En >90% de las ejecuciones, el modelo predijo probabilidad >90%
- Predicción muy robusta

Freq 0.7-0.9 (Consistente)

- La mayoría de las veces predice alto
- Buena confiabilidad

Freq < 0.7 (Inconsistente)

- Predicción variable entre ejecuciones
- Menor confianza

Hoja 6: Consensus

Las columnas que se encuentran en la hoja 5 mas unas columnas adicionales, estas son:

Columna	Qué Significa
Consensus_Mean	Promedio de acuerdo entre los 3 modelos
Models_80pct	Cuántos modelos están >80% seguros (0-3)
Confidence_Tier	Nivel de confianza basado en consenso

Confidence_Tier - Cómo se Calcula

Very High:

- Consensus_Mean ≥ 0.9 **Y** Models_80pct = 3
- Los 3 modelos muy de acuerdo y muy confiados
- **Máxima prioridad**

High:

- Consensus_Mean ≥ 0.8 **Y** Models_80pct ≥ 2
- Al menos 2 modelos confiados
- **Alta prioridad**

Low:

- Todo lo demás
- Modelos no concuerdan bien
- **Baja prioridad**

Hoja 7: Model_Evaluation

Columnas

Columna	Qué Mide
Model	Nombre del modelo
Mean_AUC	AUC promedio en las 20 ejecuciones
SD_AUC	Variabilidad del AUC entre ejecuciones
Min_AUC	Peor AUC obtenido
Max_AUC	Mejor AUC obtenido

Entendiendo el AUC (Area Under Curve)

El AUC mide qué tan bien el modelo separa las clases:

AUC = 1.0 (Perfecto)

- El modelo clasifica TODO correctamente
- **ADVERTENCIA:** Es muy posible que esto sea un indicativo de fuga de datos
- En ocasiones muy especificas es este valor aceptable, siempre y cuando de justifique

AUC = 0.95-0.99 (Excelente)

- Rendimiento muy alto
- Se tiene que verificar posible fuga de datos
- Probablemente utilizable

AUC = 0.85-0.94 (Muy Bueno)

- Rendimiento sólido
- Modelos confiables
- **Ideal para uso práctico**

AUC = 0.75-0.84 (Bueno)

- Rendimiento aceptable
- Útil para selección inicial
- Considerar mejoras si es posible

AUC = 0.65-0.74 (Regular)

- Rendimiento marginal
- Usa con precaución
- Intentar mejorar descriptores/datos

AUC = 0.50-0.64 (Malo)

- Apenas mejor que azar (AUC=0.5)
- No recomendado para uso
- Revisar datos y descriptores

Interpretando SD_AUC

SD_AUC < 0.01 (Muy estable)

- Rendimiento muy consistente
- Alta confiabilidad del modelo

SD_AUC 0.01-0.03 (Estable)

- Variación aceptable
- Confiabilidad buena

SD_AUC > 0.03 (Inestable)

- Rendimiento inconsistente
- Problema: datos o tamaño de muestra

Hoja 8: Feature_Importance

Columnas

Columna	Significado
Feature	Nombre del descriptor molecular
Importance	Puntuación de importancia

Cómo Leer las Puntuaciones

Las puntuaciones son **relativas** - compara características entre sí:

Top 5 características:

- Las MÁS importantes para la clasificación
- Si eliminas estas, el rendimiento cae mucho
- Investiga qué representan químicamente

Características medias:

- Contribuyen moderadamente
- Útiles en conjunto

Características bajas (cerca de 0):

- Casi no contribuyen
- Podrías eliminarlas sin afectar rendimiento

Hoja 9: Top_Features_Stats

Columnas

Columna	Significado
Feature	Nombre del descriptor
Target_Mean	Valor promedio en la clase objetivo (ej. AFP)
Other_Mean	Valor promedio en la otra clase
Difference	Diferencia entre las clases
Importance	Puntuación de importancia

Interpretando la Diferencia

Diferencia Grande (>0.1)

- Las clases se separan bien en esta característica
- Descriptor muy discriminativo
- Probablemente clave para la clasificación

Diferencia Moderada (0.05-0.1)

- Separación moderada
- Útil en combinación con otros descriptores

Diferencia Pequeña (<0.05)

- Poca separación individual
- Importante en conjunto con otras características
- Efecto sutil pero relevante

Hoja 10: Diagnostics

Métricas

Metric	Qué Significa
Total_Train_Size	Muestras usadas para entrenar
Total_Test_Size	Muestras usadas para evaluar
Train_AFP	Compuestos {target_class} en entrenamiento
Train_other	Compuestos "other" en entrenamiento
Test_AFP	Compuestos {target_class} en prueba
Test_other	Compuestos "other" en prueba
Test_AFP_Proportion	Proporción de {target_class} en prueba (49.5%)
Test_other_Proportion	Proporción de "other" en prueba (50.5%)
Features_Used	Número de descriptores seleccionados

Hoja 11: PCA Coordinates

Esta hoja del Excel contiene las coordenadas exactas de cada muestra en el espacio PCA:

Columnas:

- ID: Identificador de la muestra
- PC1, PC2, PC3: Valores de componentes principales
- Type: Clase (clase objetivo, other, o New Sample)
- Dataset: Training o New

Usos de las Coordenadas:

- Análisis cuantitativo de distancias entre muestras
- Crear visualizaciones personalizadas en otros programas
- Identificar muestras outliers por valores extremos
- Calcular distancias euclidianas para análisis de similitud
- Exportar a software de estadística para análisis adicional

Consejos Prácticos

1. Siempre revisar ambos gráficos:

- Gráfico de entrenamiento: Evaluar calidad del modelo
- Gráfico combinado: Validar predicciones individuales

2. Combinar visualización con métricas:

- No confiar solo en inspección visual
- Verificar probabilidades en hojas de predicción
- Confirmar con AUC en hoja Model_Evaluation

3. Identificar patrones sospechosos:

- Nueva muestra muy alejada de todos: Posible error de datos
- Separación perfecta: Verificar fuga de datos
- Distribución inesperada: Revisar preprocesamiento

Adicional: Visualizaciones 3D PCA

El análisis ademas genera dos archivos HTML con visualizaciones 3D de los primeros 3 PCAs. Estos gráficos permiten ver la separación de clases y la posición de nuevas muestras.

Archivos Generados

1. {target_class}_PCA_3D_Training.html

- Visualización 3D de solo los datos de entrenamiento
- Muestra cómo se separan las clases en el espacio PCA
- Colores: Azul = clase objetivo, Naranja = otras clases
- Útil para evaluar la calidad de separación de clases

2. {target_class}_PCA_3D_Combined.html

- Visualización 3D con datos de entrenamiento + nuevas muestras
- Permite ver dónde caen las nuevas muestras respecto a clases conocidas
- Colores:
  - Azul = clase objetivo (entrenamiento)
  - Naranja = otras clases (entrenamiento)
  - Verde (diamantes) = nuevas muestras
- Útil para validación visual de predicciones

Componentes Principales (PC)

Los gráficos muestran los primeros 3 componentes principales:

- PC1: Primera componente principal (mayor varianza explicada)
- PC2: Segunda componente principal
- PC3: Tercera componente principal

Cada eje muestra el porcentaje de varianza explicada. Por ejemplo:

- PC1 (45.3%): Este componente explica el 45.3% de la varianza total
- PC2 (23.1%): Explica el 23.1% adicional
- PC3 (12.8%): Explica el 12.8% adicional
- Total: 81.2% de varianza capturada en 3D

Cómo Usar los Gráficos 3D

Características Interactivas:

- Rotar: Click y arrastrar para rotar el gráfico en 3D
- Zoom: Usar scroll del mouse para acercar/alejar
- Pan: Click derecho y arrastrar para mover
- Hover: Pasar el mouse sobre puntos para ver información
- Leyenda: Click en elementos de la leyenda para mostrar/ocultar grupos

Interpretación de los Gráficos

Separación de Clases:

Clases Bien Separadas:

- Los puntos azules y naranjas forman clusters distintos
- Hay espacio visible entre los grupos
- Interpretación: Alta confiabilidad esperada en clasificación
- Acción: Confiar en las predicciones del modelo

Clases Parcialmente Superpuestas:

- Los clusters se tocan pero mantienen centros distintos
- Algunos puntos de ambas clases se mezclan en la frontera
- Interpretación: Clasificación moderadamente confiable
- Acción: Prestar atención a niveles de confianza en predicciones

Clases Completamente Mezcladas:

- No se distinguen clusters separados
- Puntos azules y naranjas distribuidos aleatoriamente
- Interpretación: Clasificación difícil o imposible
- Acción: Considerar usar descriptores diferentes o más datos

Posición de Nuevas Muestras

Nuevas Muestras en Gráfico Combinado:

Muestra Verde Cerca de Cluster Azul:

- Interpretación: Alta probabilidad de ser clase objetivo
- Candidato fuerte para validación

Muestra Verde Cerca de Cluster Naranja:

- Interpretación: Baja probabilidad de ser clase objetivo
- Probablemente no es clase objetivo

Muestra Verde en Zona Intermedia:

- Interpretación: Predicción incierta
- Revisar con detalle, requiere evidencia adicional

Muestra Verde como Outlier (Alejada):

- Interpretación: Muestra muy diferente a ambas clases
- Verificar calidad de datos, revisar predicción con precaución

Varianza Explicada

La varianza total explicada por PC1+PC2+PC3 indica qué tan bien la visualización 3D representa los datos completos:

>70% Varianza Explicada:

- Excelente representación 3D
- La visualización captura la mayoría de la información
- Confiar en interpretación visual

50-70% Varianza Explicada:

- Buena representación 3D
- Visualización útil pero no completa
- Combinar con métricas AUC para decisiones

<50% Varianza Explicada:

- Representación 3D limitada
- Los datos son muy complejos/multidimensionales
- Usar visualización solo como guía aproximada
- Confiar más en métricas de AUC y probabilidades

Casos Especiales

Separación Perfecta (Clases Totalmente Separadas):

- Señal: No hay ningún punto azul cerca de naranjas (o viceversa)
- Posibles causas:
- 1. Las clases son genuinamente muy diferentes (válido)
- 2. Fuga de datos (problema)
- Verificar: Revisar hoja Diagnostics para duplicados
- Verificar: AUC = 1.0 podría indicar fuga de datos

Múltiples Subclusters:

- Señal: Cada clase forma varios grupos separados
- Interpretación: Subtipos o variantes dentro de cada clase
- Normal en datos biológicos complejos
- El modelo puede manejar esta complejidad

Distribución Lineal:

- Señal: Puntos forman una línea o plano en lugar de clusters
- Interpretación: Transición gradual entre clases
- Puede indicar que la clasificación binaria es artificial
- Considerar si regresión sería más apropiada

Verificaciones de Calidad

1. Tamaño de Conjuntos

Bueno:

Total_Train_Size: >200
Total_Test_Size: >50
Suficientes datos para aprendizaje y evaluación confiable

Marginal:

Total_Train_Size: 100-200
Total_Test_Size: 20-50
Puede funcionar, pero modelos menos robustos

Insuficiente:

Total_Train_Size: <100
Total_Test_Size: <20
No hay suficientes datos, recolecta más muestras

2. Balance de Clases

Bien Balanceado:

Train_{target_class}: 216
Train_other: 219
Proporción ~ 1:1
Ideal para aprendizaje

Moderadamente Desbalanceado:

Train_{target_class}: 150
Train_other: 285
Proporción ~ 1:2
Aceptable, pero menos óptimo

Muy Desbalanceado:

Train_{target_class}: 50
Train_other: 385
Proporción > 1:5
Problema: el modelo puede sesgarse hacia clase mayoritaria

3. Consistencia Train/Test

Consistente:

Train_{target_class}_Proportion: 0.497
Test_{target_class}_Proportion: 0.495
Diferencia < 0.05 → La división mantuvo las proporciones

Inconsistente:

Train_{target_class}_Proportion: 0.497
Test_{target_class}_Proportion: 0.350
Diferencia > 0.10
Cuidado: evaluación podría no ser representativa

4. Número de Características (basado en 1600+ iniciales)

Buena Cantidad:

Features_Used: 200-500
Suficiente información sin sobrecargar modelos

Pocas:

Features_Used: <50
Puede limitar capacidad predictiva

Demasiadas:

Features_Used: >800
Riesgo de sobreajuste, considera más filtrado

Versión: 2.0

Fecha: Enero 2026

Autor: Javier Badilla

Actualización: Incluye visualizaciones 3D PCA interactivas