MANUAL

CLASIFICACIÓN BINARIA PARA PREDICCIÓN DE ACTIVIDAD MOLECULAR.

DESCRIPCIÓN.

El modelo realiza clasificación binaria para identificar compuestos con una actividad biológica específica (clase objetivo) de un conjunto de compuestos no etiquetados. Utiliza aprendizaje automático de ensamble (Random Forest y Redes Neuronales) con descriptores moleculares para hacer predicciones. El modelo está refinado para funcionar de manera "exploratoria" con baja demanda computacional, y de una manera de "confirmación" que es mas demandante, pero entrega una predicción mas robusta, todo depende del numero de corridas que se le asigne al modelo.

Lo que hace el modelo:

- Entrena múltiples modelos de aprendizaje automático con tus datos etiquetados
- Elimina fugas de datos y asegura evaluación robusta
- Selecciona los descriptores moleculares más informativos
- Hace predicciones sobre compuestos nuevos no etiquetados
- Proporciona estimaciones de confianza para cada predicción
- Identifica qué características moleculares impulsan la clasificación
- Genera visualizaciones 3D interactivas del espacio PCA

Flujo de Trabajo

1. Prepara datos → Conjunto de entrenamiento + Conjunto nuevo
2. Configura parámetros → Clase objetivo, clase de comparación (opcional)
3. Ejecuta script → Espera finalización (tiempo de ejecución depende del tamaño de datos y numero de corridas)
4. Verifica diagnósticos → Verifica AUC, balance de clases, duplicados
5. Revisa visualización PCA → Examina separación de clases en 3D
6. Revisa importancia de características → Ayuda con el entendimiento/investigación
7. Examina predicciones → Comienza con candidatos de alta confianza

REQUERIMIENTOS

1. Datos de Entrenamiento (train_data) Un conjunto de datos con compuestos etiquetados que contiene:
   - Columna "ID": Identificador único para cada compuesto
   - Columna "Type": etiqueta de clase (ej., "AFP", "antiAChE", "other")
   - Columnas de descriptores: Descriptores moleculares numéricos

2. Datos Nuevos (new_data) Un conjunto de datos con compuestos no etiquetados para clasificar:
   - Columna "ID": Identificador único
   - SIN columna "Type"
   - Mismas columnas de descriptores que los datos de entrenamiento

NOTA: Los datos deben ser cargados desde la interfase del programa (Rstudio) y estas tablas deben ser llamadas "train_data" y "new_data" respectivamente, de lo contrario, el script no funcionará. Igualmente las columnas ID y Type deben tener esos nombres para que sean reconocidas por el script..

3. Configuración de Usuario Configura estas variables al inicio del script:

target_class ← "una de las clases en columna Type" (La actividad que se quiere predecir)

comparison_class ← "una de las clases en columna Type distinta a target_calss" o "NULL" (NULL para todos los demás, sin comparación)

Ejemplo:

target_class= "X"

comparison_class = "Y" → Entrenar solo X vs Y

comparison_class = NULL → Entrenar con X vs todo lo demás

n_runs ← Número de corridas (por defecto 20 para análisis robusto, usar 2-5 para exploración rápida)

prob_threshold ← Umbral de probabilidad (por defecto 0.90)

auc_lower ← Límite inferior AUC para selección de características (por defecto 0.60)
auc_upper ← Límite superior AUC para selección de características (por defecto 0.90)

4. Bibliotecas de R Requeridas

El script requiere las siguientes bibliotecas:

• readxl - Lectura de archivos Excel

• openxlsx - Escritura de archivos Excel

• caret - Entrenamiento de modelos de machine learning

• ranger - Implementación de Random Forest

• nnet - Redes neuronales

• pROC - Cálculo de curvas ROC y AUC

• recipes - Preprocesamiento de datos

• dplyr - Manipulación de datos

• plotly - Visualizaciones 3D interactivas

PASOS DEL SCRIPT

Paso 1: Carga y Limpieza de Datos

1. Los nombres de columnas se estandarizan a nombres válidos de R
2. Si "comparison_class" está definido, filtra solo a esas dos clases
3. Crea clasificación binaria: target_class vs "other"
4. Elimina valores de cadena vacía (reemplazados con NA)

Paso 2: Selección de Descriptores

1. Identifica todas las columnas numéricas (descriptores moleculares)
2. Elimina columnas no numéricas o de datos de identificación (ID, Type)

Paso 3: Eliminación de Duplicados

Elimina muestras exactamente duplicadas basándose en valores de descriptores para prevenir fuga de datos.

Paso 4: Selección de Características Basada en AUC (área bajo la curva)

1. Calcula qué tan bien cada descriptor individual separa las clases (basado en AUC)
2. Muestra la distribución de calidad de descriptores
3. Selecciona características dentro de un rango (AUC 0.60-0.90 por defecto) para evitar fuga de datos
4. Elimina predictores perfectos que hacen el problema demasiado fácil (fuga de datos u overfitting)
5. Muestra los 10 mejores descriptores por AUC

Paso 5: División Entrenamiento/Prueba

1. Divide los datos en 80% entrenamiento, 20% prueba
2. Mantiene el balance de clases en ambos conjuntos
3. Verifica duplicados restantes entre conjuntos
4. Genera tabla de diagnóstico con tamaños de conjuntos y proporciones de clases

Paso 6: Preparación de Recetas (Recipes)

Crea dos recetas de preprocesamiento:

1. rec_rf: Para Random Forest
   - Elimina predictores de varianza cero
   - Imputa valores faltantes con la mediana
2. rec_nn_pca: Para Red Neuronal con PCA
   - Elimina predictores de varianza cero
   - Imputa valores faltantes
   - Normaliza todas las características
   - Aplica PCA para capturar 90% de la varianza

Paso 7: VISUALIZACIÓN 3D PCA

Genera visualizaciones interactivas en 3D del espacio de componentes principales:

1. Crea receta PCA específica con exactamente 3 componentes
2. Transforma datos de entrenamiento y nuevos al espacio PCA
3. Calcula y muestra varianza explicada por los 3 primeros componentes principales
4. Genera dos gráficos 3D interactivos HTML:

a) Gráfico de Datos de Entrenamiento → Muestra solo datos de entrenamiento

b) Gráfico Combinado → Muestra datos de entrenamiento + nuevas muestras

5. Exporta coordenadas PCA a hoja Excel para análisis adicional

Interpretación de Gráficos PCA 3D:

- Separación clara entre clases = buena clasificación esperada
- Clases mezcladas = clasificación más difícil
- Nuevas muestras cerca de target_class = alta probabilidad de clasificación positiva
- Los ejes muestran el % de varianza explicada

Paso 8: Configuración de Validación Cruzada

Configura validación cruzada repetida:

- 5 folds (divisiones)
- 3 repeticiones
- Guarda probabilidades de clase y predicciones

Paso 9: Entrenamiento de Modelos (Multi-Run)

Entrena tres modelos diferentes, cada uno ejecutado múltiples veces (definido por n_runs) para asegurar estabilidad:

Modelo 1: Random Forest (RF)

- Método de ensamble basado en árboles de decisión
- Maneja bien relaciones no lineales
- Usa todos los descriptores seleccionados
- Proporciona puntuaciones de importancia de características

Modelo 2: Red Neuronal con componentes principales (NN_PCA)

- Reduce dimensionalidad usando Análisis de Componentes Principales
- Captura 90% de varianza en menos características
- Bueno para datos de alta dimensión
- Parámetros: size = 2-4 neuronas, decay = 0.1-0.3

Modelo 3: Red Neuronal con selección ANOVA (NN_DESC)

- Selecciona las 30 mejores características usando pruebas estadísticas (ANOVA)
- Se enfoca en características con diferencias de clase más fuertes
- Más interpretable que PCA
- Parámetros: size = 2-4 neuronas, decay = 0.1-0.3

Cada modelo se evalúa en el conjunto de prueba usando:

- AUC (Área Bajo la Curva ROC): Capacidad de discriminación general (0.5 = aleatorio, 1.0 = perfecto)
- Matriz de confusión: Precisión de clasificación real
- Rango de predicción: Dispersión de puntuaciones de probabilidad

Para cada corrida:

- Se entrena el modelo con validación cruzada
- Se hacen predicciones en datos nuevos
- Se evalúa rendimiento en conjunto de prueba
- Se almacenan probabilidades para análisis de estabilidad

Paso 10: Análisis de Estabilidad

Combina resultados de todas las corridas para calcular: • Probabilidad media de cada modelo • Desviación estándar (estabilidad) • Frecuencia de predicciones por encima del umbral • Estadísticas de AUC (media, SD, min, max) para cada modelo

Paso 11: Predicciones Finales y Consenso

Genera predicciones de ensamble combinando los tres modelos: • Probabilidad de ensamble (promedio de los 3 modelos) • Predicciones a diferentes umbrales (50%, 70%, 90%) • Niveles de confianza basados en consenso entre modelos • Acuerdo de modelos al 90%

Paso 12: Análisis de Importancia de Características

Identifica qué descriptores moleculares son más importantes para la clasificación:

1. Extrae puntuaciones de importancia del Random Forest
2. Muestra los 20 descriptores más importantes
3. Calcula concentración de importancia: • >70% en top 5 = clasificación basada en pocos descriptores • 50-70% = diversidad moderada de características • <50% = información distribuida en muchas características
4. Estadísticas detalladas de los 5 mejores descriptores: • Medias por clase • Diferencias entre clases • Rangos de valores • Correlación con importancia

Paso 13: Generación de Predicciones Claras

Crea predicciones finales interpretables: • Clasificación binaria (YES/NO) • Probabilidad en porcentaje • Nivel de confianza • Interpretación en lenguaje natural

Identifica candidatos de alta confianza: • Umbral 70% con confianza Alta o Muy Alta • Ordenados por probabilidad de ensamble • Lista top 10 candidatos

Paso 14: Exportación de Resultados

Crea archivo Excel con 10 hojas (ver sección ARCHIVO DE SALIDA)

ARCHIVO DE SALIDA

El modelo genera los siguientes archivos de salida:

ARCHIVO EXCEL: (target_class)_Complete_Analysis.xlsx

Este archivo contiene 10 hojas:

1. Summary

Resumen ejecutivo del análisis:

- Información del dataset (clase objetivo, comparación)
- Rendimiento de modelos (mejor AUC)
- Resumen de predicciones (totales, alta confianza)
- Acuerdo entre modelos

2. Simple_Predictions

- Formato simple (tipo si/no) de la predicción:
- ID de la muestra
- Is_(target_class): YES/NO
- (target_class)*Probability*%: 0-100%
- Confidence: Very High, High, Medium, Low

3. Detailed_Predictions

- La predicción más detallada:
- Probabilidad de ensamble
- Predicciones a 3 umbrales (50%, 70%, 90%)
- Nivel de confianza
- Número de modelos en acuerdo al 90%
- Predicción individual de cada modelo
- Interpretación en texto

4. High_Confidence

- Los mejores candidatos para análisis posteriores:
- Solo muestras con predicción positiva al 70%
- Confianza Alta o Muy Alta
- Ordenados por probabilidad descendente

5. Consensus

- La combinación de los resultados de los modelos:
- Frecuencia de predicciones por modelo
- Consenso promedio entre modelos
- Número de modelos con acuerdo al 80%
- Tier de confianza (Very High, High, Low)

- Ordenado por confianza

6. Model_Evaluation

- AUC y valores de rendimiento:
- Mean_AUC: rendimiento promedio
- SD_AUC: estabilidad (menor = más estable)
- Min_AUC y Max_AUC: rango de rendimiento
- Comparación entre RF, NN_PCA, NN_DESC

7. Feature_Importance

- Valoración de todos los descriptores:
- Nombre del descriptor
- Puntuación de importancia (Random Forest)
- Ordenado de mayor a menor importancia
- Útil para entender qué características moleculares son clave

8. Top_Features_Stats

- Estadística de los 5 mejores descriptores:
- Medias por clase (target_class vs other)
- Diferencia entre medias
- Puntuación de importancia
- Ayuda a interpretar biológicamente

9. Diagnostics

- Información de la división del set de datos:
- Tamaños de conjuntos (train/test)
- Distribución de clases
- Proporciones
- Número de características usadas
- Verificación de balance

10. PCA_Coordinates

Coordenadas en espacio PCA:

- ID de muestra
- PC1, PC2, PC3 (valores de componentes principales)
- Type (clase o "New Sample")
- Dataset (Training o New)
- Útil para análisis exploratorio adicional

ARCHIVOS HTML INTERACTIVOS:

1. (target_class)_PCA_3D_Training.html
   - Visualización 3D interactiva de datos de entrenamiento
   - Separación de clases en espacio PCA
   - Rotable, con zoom, y explorable
   - Muestra varianza explicada por cada eje
2. (target_class)_PCA_3D_Combined.html
   - Visualización 3D con entrenamiento + nuevas muestras
   - Permite ver dónde caen nuevas muestras respecto a clases conocidas -
   - Códigos de color: azul (target), naranja (other), verde (new)
   - Interactivo y exportable

Versión: 2.0
Fecha: Enero 2026
Autor: Javier Badilla
Actualización: Incluye visualizaciones 3D PCA interactivas