UNIVERSIDADE DA CORUÑA

CLASIFICACIÓN DE TUMORES SOBRE DATOS DE EXPRESIÓN GÉNICA OBTENIDOS POR RNA-SEQ CON TÉCNICAS DE APRENDIZAJE MÁQUINA

Máster en Bioinformática para las Ciencias de la Salud

Daniel Iglesias
Antía Fraga
Elena Beade
Verónica Aranda

INTRODUCCIÓN

- Utilización de técnicas de aprendizaje máquina ("machine learning", en inglés) para la clasificación de tipos de tumores empleando datos de expresión génica (obtenidos mediante la técnica RNA-Seq).

- El Dataset empleado se ha obtenido del repositorio "UCI Machine Learning" (https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq#).

Características del Dataset	Número de instancias	Número de atributos
Multivariante	801	20531

- Los tipos de tumores estudiados son: BRCA, KIRC, COAD, LUAD y PRAD.

- Existe un desequilibrio entre el número de instancias que pertenecen a cada clase.

- El Dataset consta de dos ficheros:

  - "**data.csv**"
  - "**labels.csv**"

METODOLOGÍA

Para llevar a cabo el trabajo, es necesario realizar los siguientes pasos:

- LECTURA Y PREPROCESAMIENTO DE LOS DATOS.
- ENTRENAMIENTO DE LOS MODELOS.

METODOLOGÍA

LECTURA Y PREPROCESAMIENTO DE LOS DATOS

¿Por qué preprocesar los datos?:

- Número de atributos inmanejable de forma manual (20531), fichero de entrada (~ 200 MB).
- Mucha información redundante.
- **Necesidad de reducción de la dimensionalidad.**
- **Estandarización:** normalizar los datos facilita el aprendizaje de los modelos.

Posibilidades:

- *Features selection* o *features extraction*.
- Algoritmos para reducción de la dimensionalidad (PCA). **Más adecuado para este caso.**

Desarrollo:

- **¿Número de atributos idóneo para resolver el problema?**
- Desconocido a priori. Selección del número a base de prueba y error.
- Información manejable ahora útil para entrenamiento de modelos.

METODOLOGÍA

ENTRENAMIENTO DE LOS MODELOS

Experimentos a realizar:

- PCAs con diferente número de atributos.
- Distintos modelos de machine learning: algoritmo kNN, máquinas de soporte vectorial (SVM) y random forest.

Desarrollo:

- Splitting aleatorio en dos conjuntos: entrenamiento y test (60%-40%).
- Definir modelo y parámetros. Entrenar N veces (por elementos aleatorios).
- Obtener métricas de efectividad y computar estadísticos:
  - **Media:** bondad de los modelos en los N entrenamientos.
  - **Desviación típica:** robustez de los modelos en los N entrenamientos.

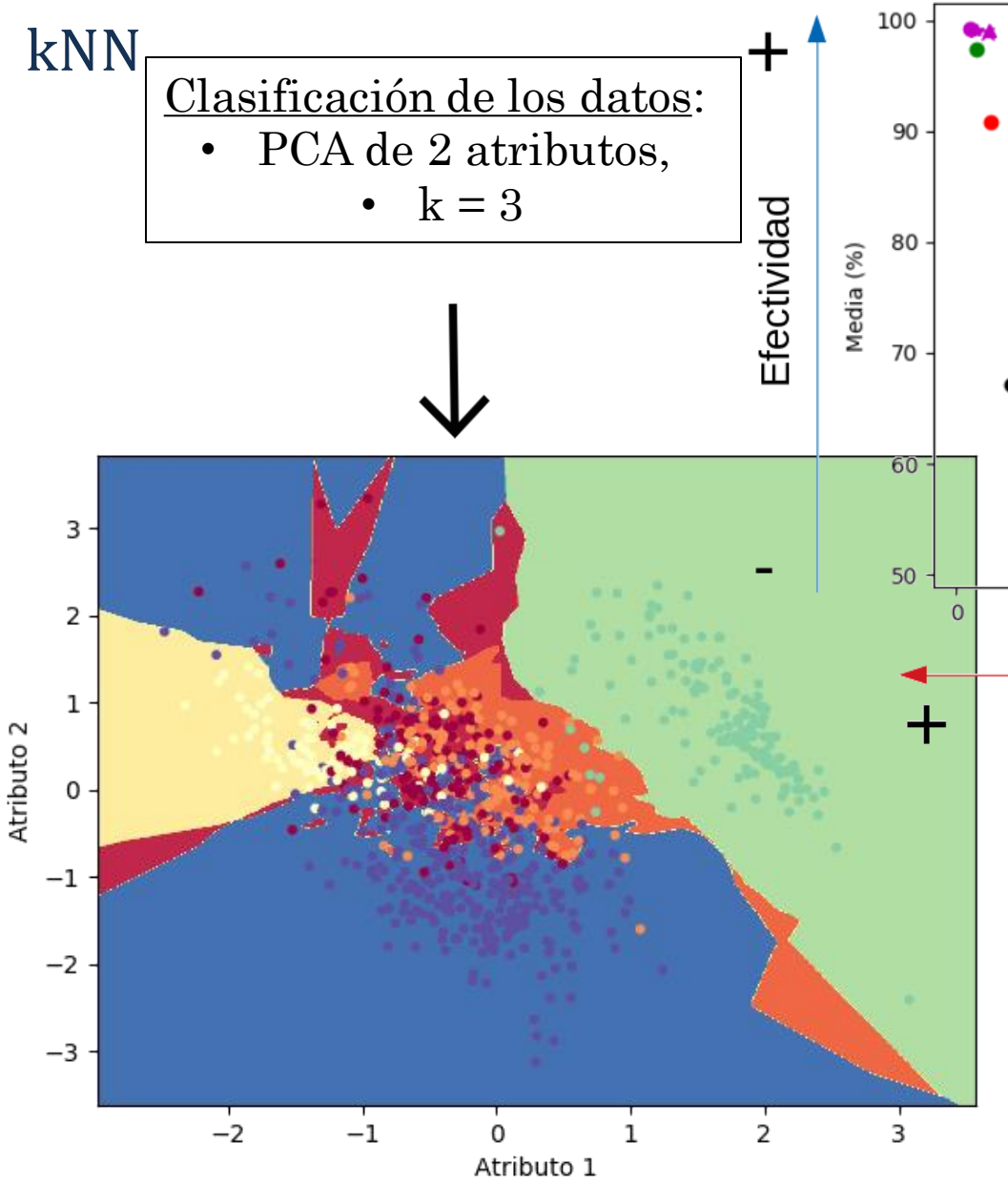
Resultados:

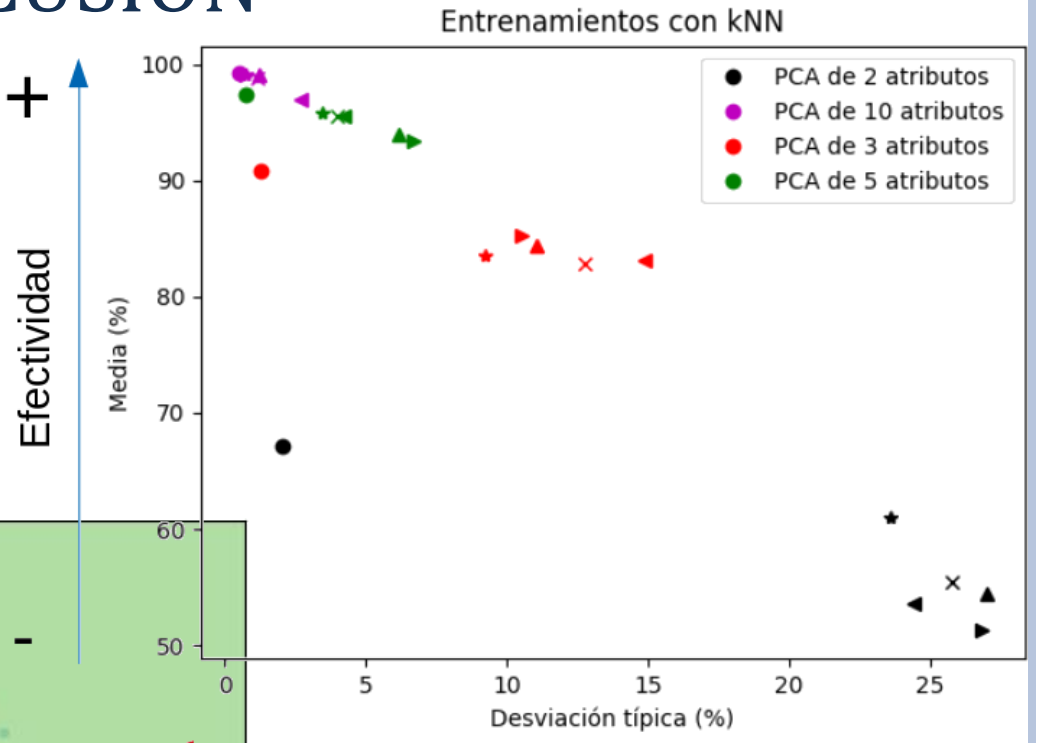
- Métrica de precisión para un punto de vista global.
- Métrica de AUC-PR (curva de *precision-recall*) para evaluar en cada clase.

RESULTADOS Y DISCUSIÓN

kNN

Clasificación de los datos:
- PCA de 2 atributos,
- k = 3





Efectividad: + / -

Robustez: + / -

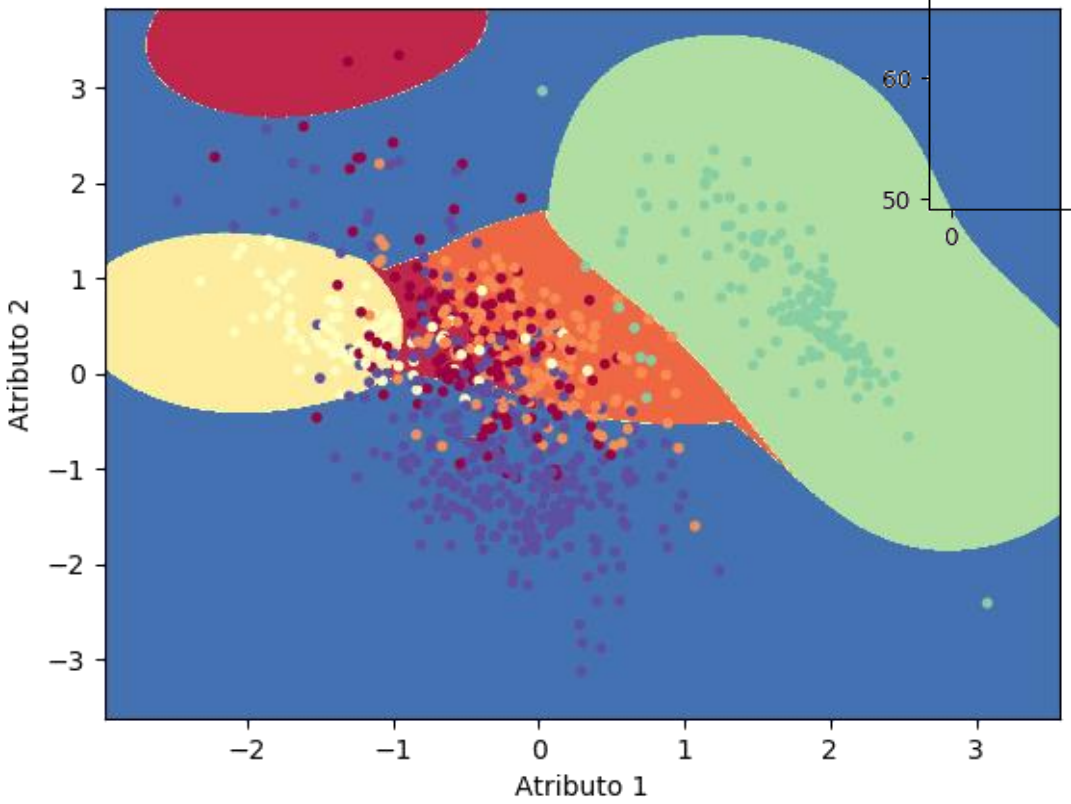
Comparación entre entrenamientos con PCAs diferentes

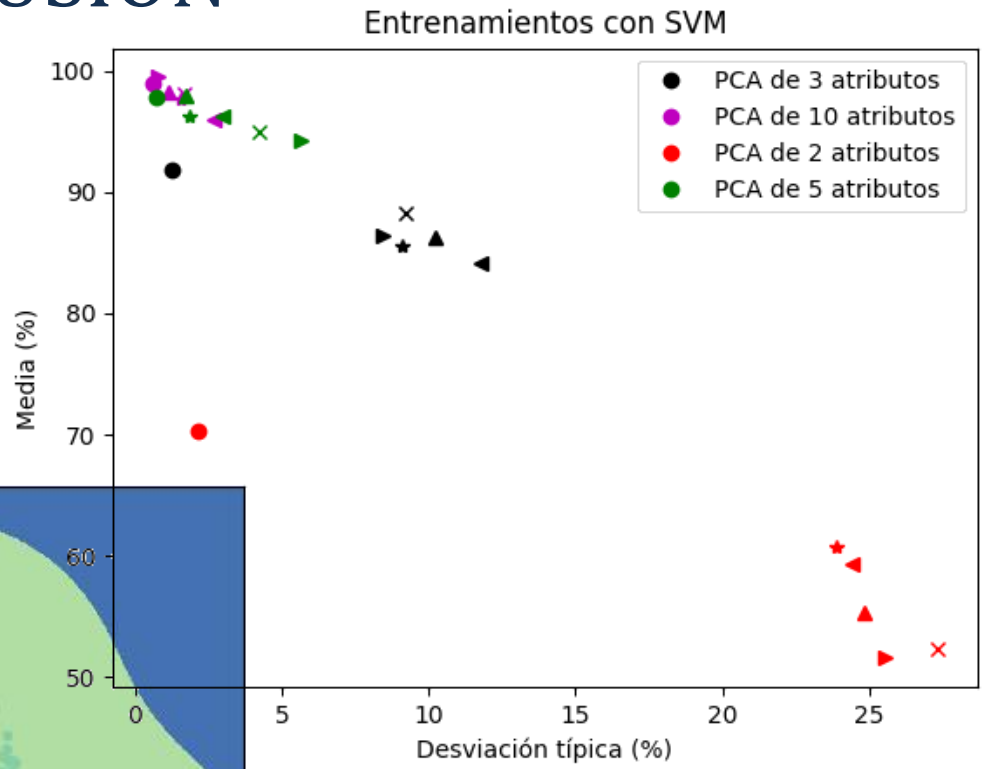
RESULTADOS Y DISCUSIÓN

SVM

Clasificación de los datos:

- PCA de 2 atributos,
- kernel de función de base radial con $C = 1.0$





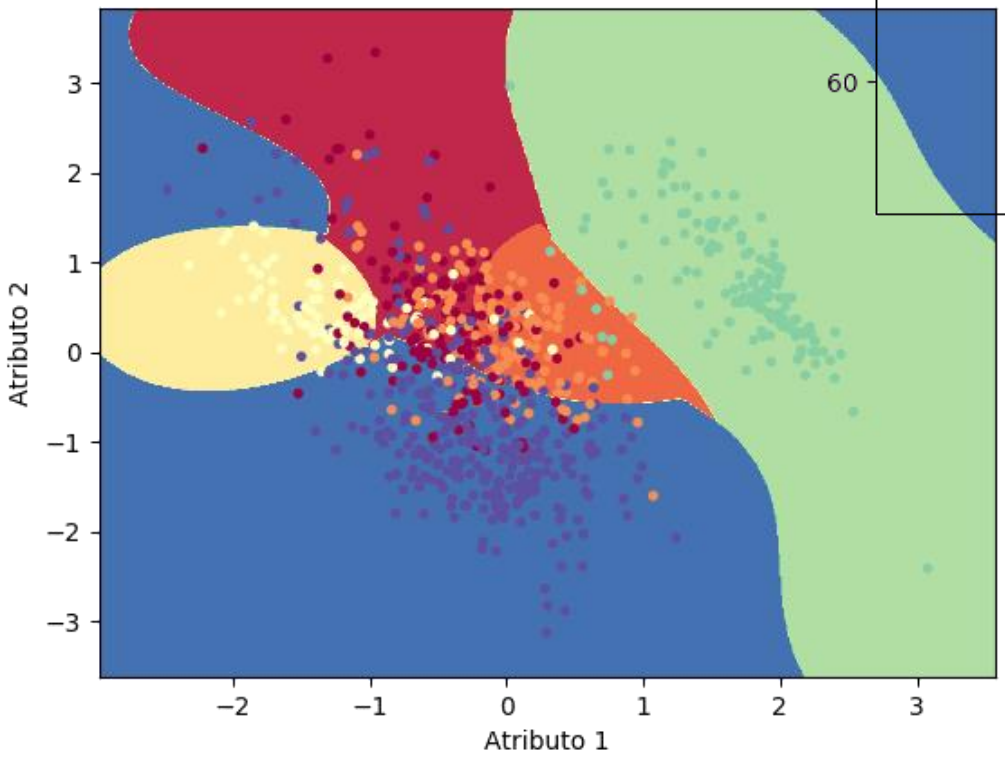
Comparación entre entrenamientos con PCAs diferentes

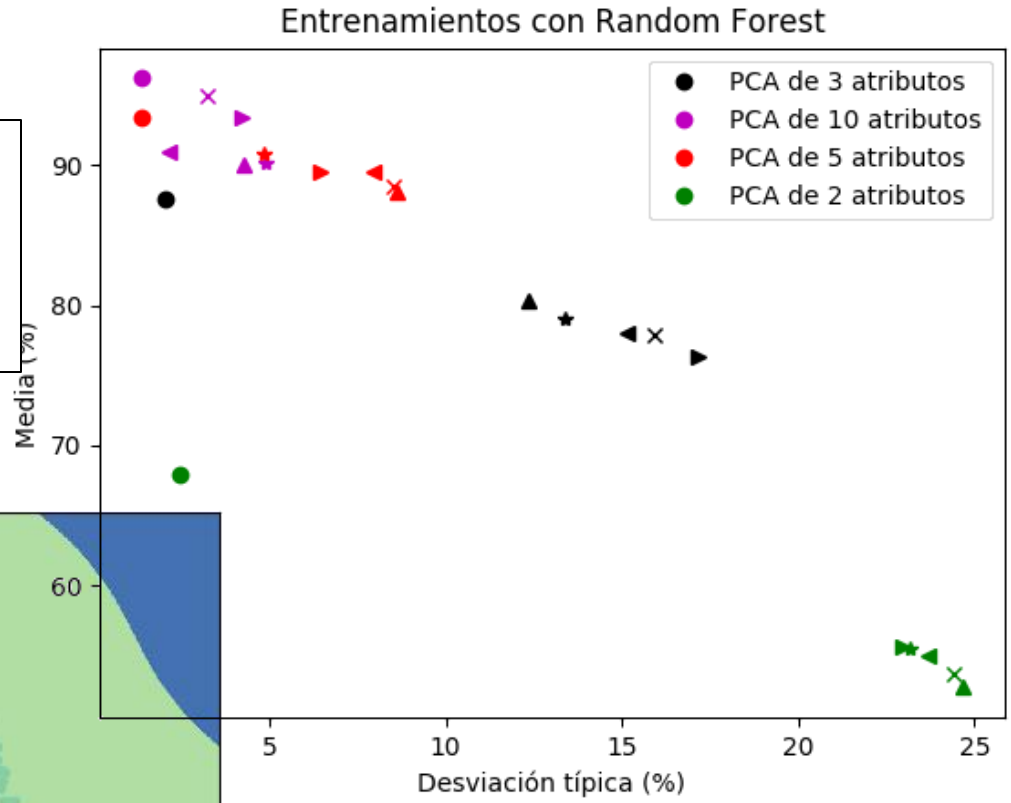
RESULTADOS Y DISCUSIÓN

Random Forest

Clasificación de los datos:

- PCA de 2 atributos,
- Nº arbóles = 100,
- max depth = 4



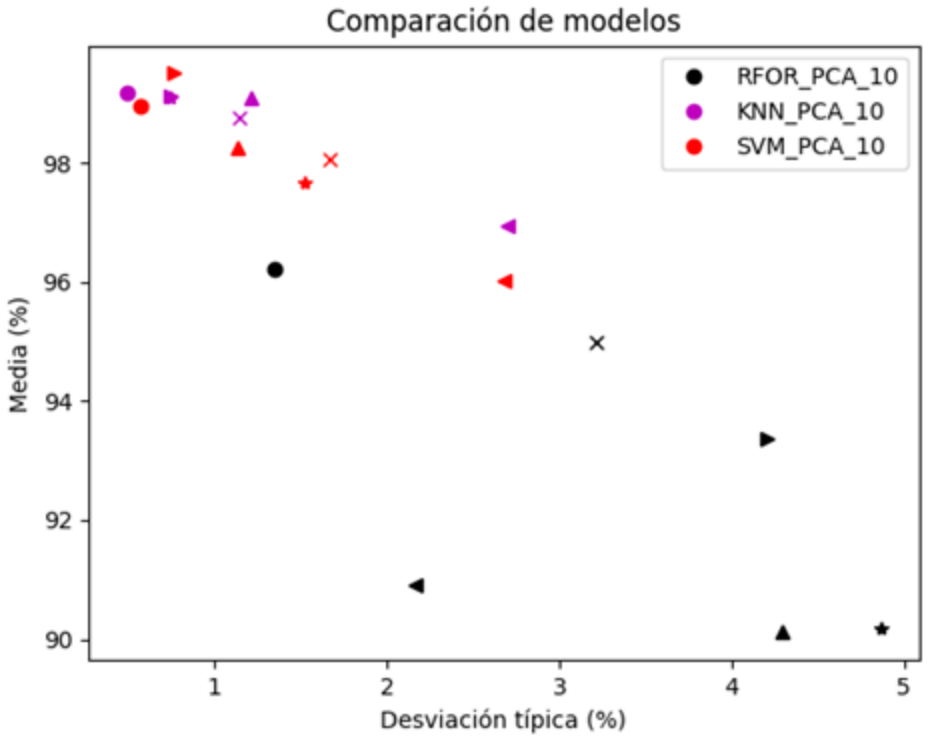


Comparación entre entrenamientos con PCAs diferentes

Resultados y discusión

Comparación de los modelos

AUC-PR	Clase 1 (BRCA)	Clase 2 (KIRC)	Clase 3 (COAD)	Clase 4 (LUAD)	Clase 5 (PRAD)
Algoritmo kNN	99.1000% ± 1.2127%	99.0921% ± 0.7439%	96.9458% ± 2.6949%	99.1105% ± 0.7452%	98.7622% ± 1.1424%
SVM	98.2397% ± 1.1324%	97.6723% ± 1.5262%	96.0261% ± 2.6794%	99.4992% ± 0.7597%	98.0714% ± 1.6709%
Random forest	90.1146% ± 4.2933%	90.1796% ± 4.8721%	90.9204% ± 2.1594%	93.3720% ± 4.2046%	94.9747% ± 3.2151%



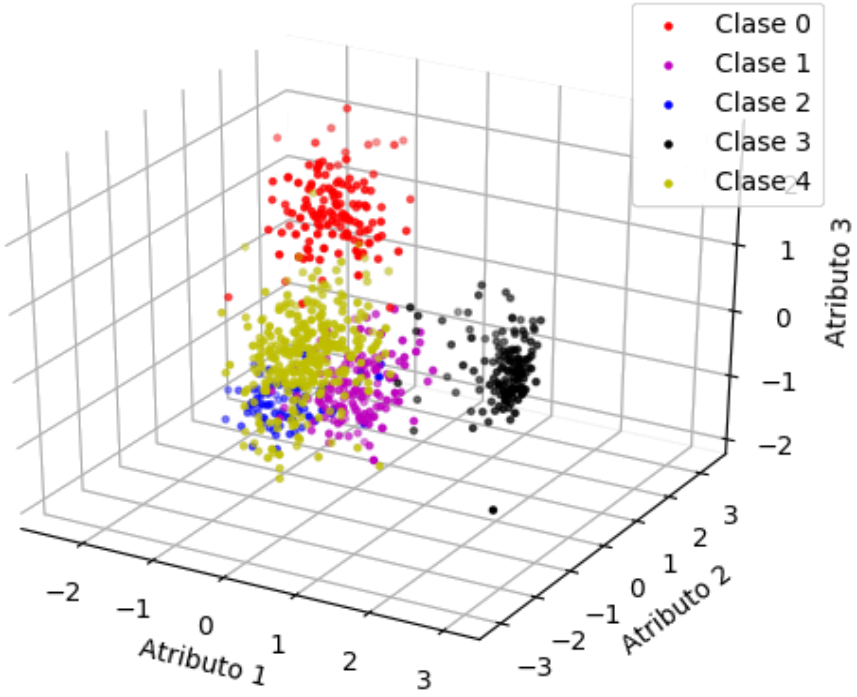
	Precisión
Algoritmo kNN	99.1651% ± 0.4864%
SVM	98.9533% ± 0.5702%
Random forest	96.2150% ± 1.3489%

Conclusiones

- Se ha empleado el algoritmo de PCA para reducir la cantidad de atributos.
- Se ha llevado a cabo un número considerable de experimentos (tres modelos y cuatro PCAs para cada modelo).
- Para la efectividad de los modelos se ha tomado como referencia la precisión global y el AUC-PR para cada clase.
- El número de atributos de entrada es relevante para la efectividad de los modelos, siendo 10 los que mejor ajustan el modelo.
- kNN y SVM tienen una efectividad similar, mientras que random forest mostró peores resultados.

TRABAJO FUTURO

- Realizar PCAs con un mayor número de atributos.
- Plantear un fine-tuning de los parámetros de los modelos.
- Probar otros modelos, por ejemplo, redes neuronales artificiales.
- Se podría realizar una fase de "oversampling" para equilibrar el número de patrones en función de las clases.



UNIVERSIDADE DA CORUÑA

CLASIFICACIÓN DE TUMORES SOBRE DATOS DE EXPRESIÓN GÉNICA OBTENIDOS POR RNA-SEQ CON TÉCNICAS DE APRENDIZAJE MÁQUINA

Máster en Bioinformática para las Ciencias de la Salud

Daniel Iglesias
Antía Fraga
Elena Beade
Verónica Aranda