Clasificación de Tumores con Datos de RNA-Seq del TCGA Pan-Cancer Atlas

INTRODUCCIÓN A CIENCIA DE DATOS

24 de Noviembre de 2025

Jessica Rubí Lara Rosales jessica.lara@cimat.mx

1. Introducción

El análisis de datos genómicos a gran escala ha transformado la comprensión de la biología del cáncer, permitiendo identificar patrones moleculares que distinguen distintos tipos de tumores. El proyecto *The Cancer Genome Atlas* (TCGA) `https://portal.gdc.cancer.gov/` constituye uno de los recursos más importantes en este ámbito, ya que reúne datos moleculares, clínicos y patológicos de más de 11,000 pacientes con cáncer a lo largo de 33 tipos tumorales. Estos datos incluyen información de expresión génica, metilación del ADN, mutaciones somáticas, variación en el número de copias y perfiles proteómicos, entre otros.

Dentro de este proyecto, el *TCGA Pan-Cancer Atlas* (PanCanAtlas) representa un subconjunto estandarizado de los datos de TCGA. Su objetivo principal es permitir comparaciones sistemáticas entre múltiples tipos de cáncer, integrando diversas capas de información molecular para identificar patrones compartidos entre tumores de distinta procedencia. Este programa posibilitó el estudio de temas comunes, como firmas moleculares compartidas, mutaciones recurrentes, vías biológicas conservadas y otros mecanismos fundamentales de la oncogénesis.

La colección completa de estudios del *TCGA Pan-Cancer Atlas*, publicada en 2018 en la revista *Cell*, se encuentra disponible en el siguiente enlace: `https://www.cell.com/pb-assets/consortium/pancanceratlas/pancani3/index.html`

2. Descripción de los datos

Los datos utilizados en este estudio provienen del proyecto *TCGA Pan-Cancer Atlas*, específicamente aquellos empleados en el artículo *"Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer"* publicado en la revista *Cell* en 2018 [1]. Este trabajo integra análisis moleculares de aproximadamente 10,000 tumores pertenecientes a 33 tipos de cáncer, con el objetivo de comprender cómo los patrones relacionados con el *cell-of-origin* (tipo celular de origen) influyen en la clasificación molecular de los tumores.

En nuestro caso, utilizamos únicamente el archivo que contiene la **matriz de expresión génica (RNA-Seq)** generada mediante la plataforma Illumina HiSeq. Esta matriz incluye 20,502 genes medidos en 11568 muestras tumorales, abarcando los 33 tipos de cáncer analizados en el Pan-Cancer Atlas. Los tipos de cáncer que se consideraran son los que se encuentran en el Cuadro 1

Código	Nombre corto	Tipo	Descripción	N
BRCA	Breast Invasive Carcinoma	Mama	Tumores invasivos de mama, ductales o lobulillares.	1188
STAD	Gastric Adenocarcinoma	Digestivo	Tumores malignos del estómago.	670
KIRC	Clear Cell RCC	Riñón	Tumor renal de células claras.	587
LUAD	Lung Adenocarcinoma	Pulmón	Adenocarcinoma pulmonar, frecuente en no fumadores.	569
OV	Ovarian Serous Adenocarcinoma	Ginecológico	Carcinoma seroso de ovario, de alto grado.	568
THCA	Thyroid Carcinoma	Endocrino	Carcinoma papilar de tiroides.	567
UCEC	Uterine Endometrioid Carcinoma	Ginecológico	Cáncer de endometrio, frecuente en mujeres posmenopáusicas.	564
HNSC	Head and Neck SCC	Epithelial	Carcinomas escamosos de cabeza y cuello.	560
PRAD	Prostate Adenocarcinoma	Reproductivo	Principal cáncer maligno de próstata.	545
LUSC	Lung Squamous Cell Carcinoma	Pulmón	Tumor pulmonar escamoso asociado al tabaquismo.	537
LGG	Lower Grade Glioma	SNC	Gliomas de bajo grado (II–III).	532
COAD	Colon Adenocarcinoma	Digestivo	Tumores malignos del colon.	499
SKCM	Skin Cutaneous Melanoma	Piel	Melanoma cutáneo derivado de melanocitos.	473
BLCA	Bladder Urothelial Carcinoma	Urinario	Cáncer del epitelio urotelial de vejiga.	423
LIHC	Hepatocellular Carcinoma	Hígado	Principal tipo de cáncer hepático primario.	421
LAML	Acute Myeloid Leukemia	Sangre	Leucemia agresiva de precursores mieloides.	340
KIRP	Papillary RCC	Riñón	Cáncer renal de tipo papilar.	318
CESC	Cervical Carcinoma	Ginecológico	Tumor cervical asociado con HPV.	306
SARC	Sarcoma	Conectivo	Tumores de tejidos blandos y óseos.	260
ESCA	Esophageal Carcinoma	Digestivo	Cáncer escamoso o adenocarcinoma de esófago.	193
PCPG	Pheochromocytoma/Paraganglioma	Endocrino	Tumores neuroendocrinos del sistema simpático.	186
GBM	Glioblastoma Multiforme	SNC	Tumor cerebral agresivo de alto grado.	174
READ	Rectum Adenocarcinoma	Digestivo	Adenocarcinoma rectal.	166
PAAD	Pancreatic Adenocarcinoma	Digestivo	Tumor agresivo del páncreas exocrino.	161
TGCT	Testicular Germ Cell Tumors	Reproductivo	Tumores germinales testiculares.	155
THYM	Thymoma	Mediastino	Tumor epitelial del timo.	122
KICH	Chromophobe RCC	Riñón	Tumor renal de células cromófobas.	89
MESO	Mesothelioma	Pleura	Tumor agresivo asociado al asbesto.	87
UVM	Uveal Melanoma	Ojo	Melanoma de la úvea (coróides, iris, cuerpo ciliar).	80
ACC	Adrenocortical Carcinoma	Endocrino	Tumor maligno de la glándula suprarrenal.	78
UCS	Uterine Carcinosarcoma	Ginecológico	Tumor uterino mixto epitelial/mesenquimal.	57
DLBC	Diffuse Large B-Cell Lymphoma	Sangre	Linfoma agresivo de células B grandes.	48
CHOL	Cholangiocarcinoma	Biliar	Tumor de vías biliares intra o extrahepáticas.	45

Cuadro 1: Resumen de las cohortes TCGA con código, tipo de cáncer, descripción y número de muestras.

3. Objetivo

El objetivo principal de este trabajo es evaluar la capacidad de diferentes métodos de clasificación supervisada para identificar correctamente el tipo de cáncer a partir de datos de expresión génica. Además, nos interesa determinar qué genes tienen un mayor peso en la clasificación, es decir, cuáles contribuyen de manera más significativa a la separación entre los distintos tipos de tumores. Para ello utilizamos técnicas de reducción de dimensionalidad, métodos clásicos de clasificación y análisis de importancia de variables.

4. Resultados

Para comenzar, aplicamos Análisis de Componentes Principales (PCA) con el fin de reducir la alta dimensionalidad del conjunto de datos y conservar la mayor proporción posible de variabilidad. En la Figura 1 se muestra la varianza explicada acumulada por los componentes principales.

A partir de la gráfica observamos que la varianza explicada deja de incrementarse significativamente después de aproximadamente 30–40 componentes. Por esta razón, decidimos conservar las primeras 35 componentes principales para los modelos de clasificación posteriores. En la Figura 2, donde se muestran las dos primeras componentes,

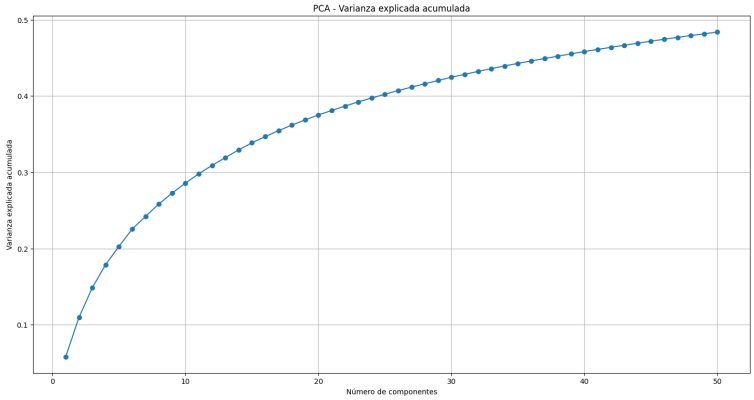

Figura 1: Varianza explicada por los componentes principales.

se aprecia que los datos se encuentran todavía bastante mezclados, lo que sugiere que los tipos de tumores no son fácilmente separables en un espacio lineal de baja dimensión.

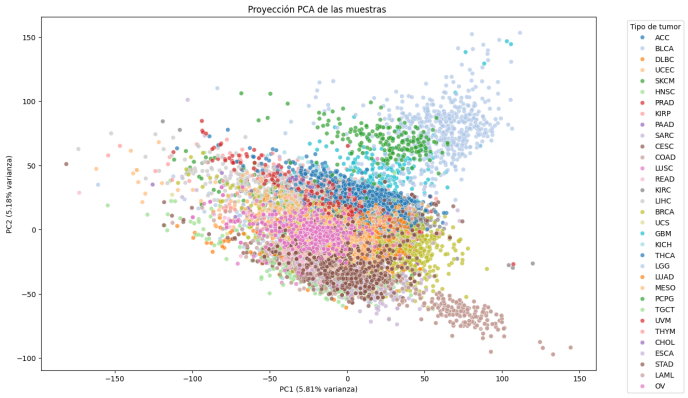

Figura 2: Proyección de los datos sobre las dos primeras componentes principales.

Posteriormente, ajustamos distintos clasificadores supervisados utilizando los 35 componentes principales. Los resultados se resumen en el Cuadro 2. Entre los modelos evaluados, la red neuronal multicapa obtuvo el mejor desempeño global.

Modelo	Accuracy	Precision	Recall	F1
Random Forest	0.9032	0.8987	0.9032	0.8963
Naive Bayes	0.8456	0.8630	0.8456	0.8490
k-NN (n=5)	0.8879	0.8874	0.8879	0.8859
Regresión Logística	0.9015	0.9118	0.9015	0.9040
Red Neuronal	0.9133	0.9128	0.9133	0.9122

Cuadro 2: Resultados comparativos de varios métodos usando PCA.

La matriz de confusión correspondiente a la red neuronal (el mejor clasificador) se muestra en la Figura 3, donde se observa un desempeño notablemente equilibrado entre todas las clases evaluadas.

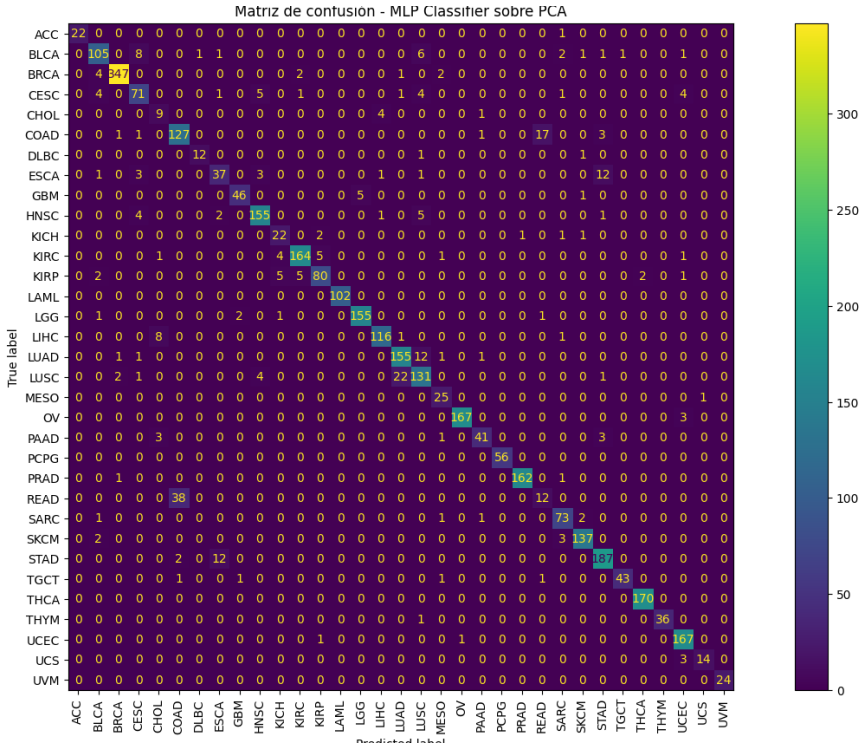

Figura 3: Matriz de confusión del mejor modelo (red neuronal).

Clase	Precisión	Recall	F1-score	Soporte
ACC	1.0000	0.9565	0.9778	23
BLCA	0.8750	0.8268	0.8502	127
BRCA	0.9858	0.9747	0.9802	356
CESC	0.7978	0.7717	0.7845	92
CHOL	0.4286	0.6429	0.5143	14
COAD	0.7560	0.8467	0.7987	150
DLBC	0.9231	0.8571	0.8889	14
ESCA	0.6981	0.6379	0.6667	58
GBM	0.9388	0.8846	0.9109	52
HNSC	0.9281	0.9226	0.9254	168
KICH	0.6875	0.8148	0.7458	27
KIRC	0.9535	0.9318	0.9425	176
KIRP	0.9091	0.8421	0.8743	95
LAML	1.0000	1.0000	1.0000	102
LGG	0.9688	0.9688	0.9688	160
LIHC	0.9508	0.9206	0.9355	126
LUAD	0.8611	0.9064	0.8832	171
LUSC	0.8137	0.8137	0.8137	161
MESO	0.7812	0.9615	0.8621	26
OV	0.9940	0.9824	0.9882	170
PAAD	0.9111	0.8542	0.8817	48
PCPG	1.0000	1.0000	1.0000	56
PRAD	0.9939	0.9878	0.9908	164
READ	0.3871	0.2400	0.2963	50
SARC	0.8795	0.9359	0.9068	78
SKCM	0.9580	0.9648	0.9614	142
STAD	0.8990	0.9303	0.9144	201
TGCT	0.9773	0.9149	0.9451	47
THCA	0.9884	1.0000	0.9942	170
THYM	1.0000	0.9730	0.9863	37
UCEC	0.9278	0.9882	0.9570	169
UCS	0.9333	0.8235	0.8750	17
UVM	1.0000	1.0000	1.0000	24

Cuadro 3: Resultados de clasificación con Red Neuronal por clase.

Un caso particular donde la clasificación no resulta satisfactoria es el del tipo de cáncer **READ** (rectal adenocarcinoma) ver Cuadro 3. A partir de la matriz de confusión observamos que la mayoría de los errores se producen al confundirlo con el tipo **COAD** (colon adenocarcinoma). Este comportamiento es esperable, pues ambos corresponden a localizaciones anatómicas contiguas dentro del tracto gastrointestinal y comparten perfiles moleculares muy similares. Por lo tanto, es razonable que los modelos tengan dificultades para separar claramente ambos grupos, ya que las diferencias genéticas entre ellos pueden ser sutiles.

Otro caso con desempeño limitado es el de la clase **CHOL** (cancer biliar), cuya precisión es relativamente baja. Este bajo rendimiento puede explicarse en gran medida por el reducido tamaño de la muestra, lo cual dificulta que los modelos aprendan patrones robustos. En problemas con alta dimensionalidad, como el presente, las clases con pocos ejemplos tienden a ser clasificadas incorrectamente debido a la escasa representación estadística. Además de que nuevamente la confusion ocurre con otro tipo de cancer que tiene alta similitud en sus genes el cual es **LIHC** un cancer de higado.

Importancia de genes con Random Forest

Además del análisis mediante PCA, es de especial interés identificar qué genes aportan mayor información para discriminar entre los distintos tipos de tumores. Para ello, ajustamos un modelo de Random Forest utilizando todos los genes, ya que este método permite cuantificar la importancia de cada variable mediante el criterio de impureza de Gini.

Al entrenar el modelo completo, obtenemos los siguientes indicadores de desempeño:

Accuracy: 0.9461, **Precision:** 0.9500, **Recall:** 0.9461, **F1:** 0.9397.

La Figura 4 muestra los 30 genes más relevantes según este modelo.

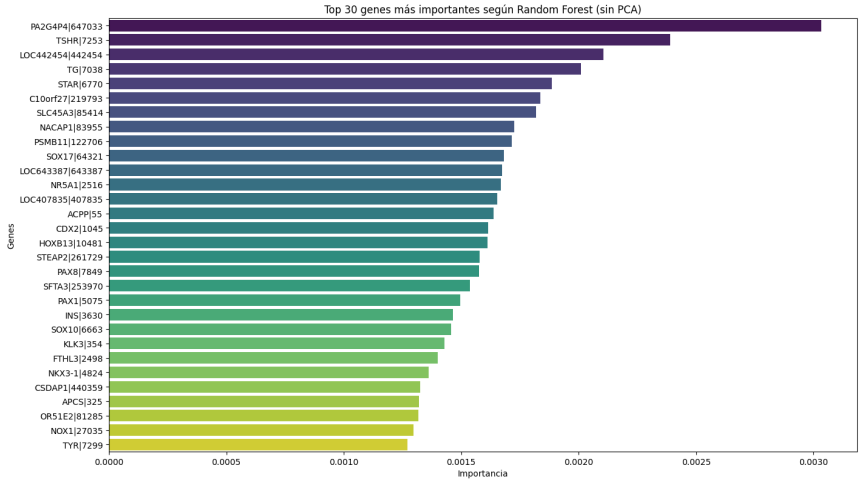

Figura 4: Treinta genes más importantes según Random Forest.

Entre los genes más influyentes destacan:

- **PA2G4P4:** pseudogén cuya posible asociación con procesos de tumorigenesis, particularmente en cáncer de vejiga, ha sido sugerida en estudios recientes.
- **TSHR:** receptor de la hormona estimulante de la tiroides, implicado en la regulación del metabolismo y procesos de diferenciación celular.
- **LOC442454:** gen cuya función aún no ha sido completamente caracterizada, pero que emerge como relevante en el modelo.

Modelos ajustados con los 30 genes más importantes

Finalmente, entrenamos nuevamente los modelos previos, pero ahora utilizando únicamente los 30 genes con mayor importancia. Los resultados se muestran en la Tabla 4.

En este caso observamos que el desempeño de los modelos disminuye notablemente respecto al escenario anterior. Esto sugiere que, aunque existen genes con fuerte capacidad discriminante, la clasificación correcta entre 33 tipos

Modelo	Accuracy	Precision	Recall	F1
Random Forest	0.9067	0.9028	0.9067	0.8980
Naive Bayes	0.6825	0.7276	0.6825	0.6721
k-NN (n =5)	0.7560	0.7547	0.7560	0.7492
Regresión Logística	0.7952	0.8263	0.7952	0.8051
Redes Neuronales	0.7822	0.7712	0.7822	0.7682

Cuadro 4: Resultados usando sólo los 30 genes más importantes del Random Forest.

de tumores requiere un conjunto de características mucho más amplio. Es decir, el problema es altamente complejo y depende de múltiples señales distribuidas en gran parte del genoma.

5. Conclusiones

En este trabajo se analizaron datos de expresión génica correspondientes a 33 tipos de cáncer del proyecto TCGA Pan-Cancer Atlas, con el objetivo de evaluar distintos métodos de clasificación y determinar qué genes aportan mayor información para distinguir entre los diferentes tumores. La reducción de dimensionalidad mediante PCA permitió conservar una parte importante de la variabilidad, sin embargo, la proyección en pocas dimensiones no logró separar claramente los tipos de cáncer, lo cual evidencia la complejidad biológica del problema.

Entre los modelos supervisados evaluados, la red neuronal multicapa obtuvo el mejor desempeño cuando se utilizaron las 35 componentes principales, alcanzando métricas superiores al 90 %. Al analizar la importancia de las variables por medio de Random Forest, se identificaron genes potencialmente relevantes en la diferenciación entre tumores, destacando algunos pseudogenes y genes con funciones parcialmente caracterizadas que podrían ser de interés para estudios posteriores.

No obstante, al restringir los modelos únicamente a los 30 genes más importantes, el desempeño disminuyó de manera notable. Esto sugiere que la discriminación entre los 33 tipos de cáncer no depende únicamente de un conjunto reducido de genes, sino que requiere información distribuida en un número mucho mayor de variables. Asimismo, la clasificación mostró dificultades en tipos de cáncer anatómicamente o molecularmente cercanos, como COAD y READ, y en clases con muy pocas muestras, como CHOL, donde los modelos no pudieron aprender patrones suficientemente representativos.

En conjunto, los resultados muestran que la clasificación de tumores a partir de expresión génica es viable y puede alcanzar altos niveles de precisión, pero también destacan los retos inherentes al problema, especialmente la alta dimensionalidad, la similitud biológica entre ciertos tumores y el desequilibrio en el tamaño de las clases.

Posibles propuestas incluyen realizar un análisis más exhaustivo del desbalance de clases, incorporando información adicional disponible en el proyecto TCGA, así como abordar el problema de la alta dimensionalidad mediante métodos de reducción no lineales, como t-SNE o UMAP, que podrían capturar estructuras complejas en los datos.

Referencias

[1] K. A. Hoadley et al., *Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer*, Cell, Vol. 173, Issue 2, pp. 291–304.e6, 2018. doi:10.1016/j.cell.2018.03.022

[2] National Cancer Institute. (s.f.). *The Cancer Genome Atlas Program*. National Institutes of Health. Recuperado 25 de noviembre de 2025, de `https://www.cancer.gov/ccg/research/genome-sequencing/tcga`

[3] Cell Press. (s.f.). *Pan-Cancer Atlas*. Cell. Recuperado 25 de noviembre de 2025, de `https://www.cell.com/pb-assets/consortium/pancanceratlas/pancani3/index.html`

[4] Li, Y., Kang, K., Krahn, J. M., Croutwater, N., Lee, K., Umbach, D. M. & Li, L. (2017). A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data. *BMC Genomics*, *18*(1), 508. https://doi.org/10.1186/s12864-017-3906-0