ISSN 1870-4069

Selección de características y optimización de hiperparámetros para la mejora en la clasificación del cáncer de próstata

Andrea G. Plascencia-Rodríguez[1], Manuel A. Soto-Murillo[1], José M. Celaya-Padilla[1], Jorge I. Galván-Tejada[1], Carlos E. Galván-Tejada[1]

Universidad Autónoma de Zacatecas,
Unidad Académica de Ingeniería Eléctrica,
Zacatecas,
México

{andrea.plascencia, 28900587, jose.celaya ,gatejo, ericgalvan}@uaz.edu.mx

Resumen. La selección precisa de características en el análisis de expresión génica es crucial para comprender la biología subyacente y mejorar el diagnóstico y tratamiento del cáncer de próstata. En este estudio, SE aplica la técnica de suma de cuadrados entre grupos y dentro de grupos (BSS/WSS) para identificar genes relevantes en la clasificación de muestras tumorales y normales. Los resultados muestran que la selección de características mejoró la eficiencia de los modelos de Árboles de Decisión y Bosques Aleatorios, alcanzando una precisión prometedora en la clasificación de muestras de cáncer de próstata. La optimización de hiperparámetros, especialmente en los modelos de Bosques Aleatorios, demostró un rendimiento óptimo. Estos hallazgos resaltan la importancia de la selección de características en la investigación del cáncer de próstata y sugieren su relevancia clínica para la práctica médica y la salud pública.

Palabras clave: Selección de características, expresión génica, cáncer de próstata, suma de cuadrados entre grupos y dentro de grupos, regresión logística, árboles de decisión, bosques aleatorios.

Feature Selection and Hyperparameter Optimization for Improved Prostate Cancer Classification

Abstract. Precise feature selection in gene expression analysis is crucial for understanding underlying biology and improving the diagnosis and treatment of prostate cancer. In this study, we apply the between-group sum of squares and within-group sum of squares (BSS/WSS) technique to identify relevant genes in the classification of tumor and normal samples. The results show that feature selection enhanced the efficiency of decision trees and random forest models, achieving promising accuracy in prostate cancer sample classification. Hyperparameter optimization, especially in random forest models, demonstrated

optimal performance. These findings underscore the importance of feature selection in prostate cancer research and suggest its clinical relevance for medical practice and public health.

Keywords: feature selection, gene expresión, prostate cancer, between-group sum of squares and within-group sum of squares, logistic regre-ssion, decision tree, random forests.

1. Introducción

El cáncer es una enfermedad causada por alteraciones genómicas en el ADN, el ARN y las proteínas de una célula, que conducen a un crecimiento y desa-rrollo celular anormal. Comprender estas alteraciones genómicas es crucial para decodificar los mecanismos del desarrollo del cáncer y mejorar el diagnóstico y tratamiento de los diferentes tipos de cáncer en función de sus anomalías moleculares [1]. En el contexto del cáncer de próstata, la progresión hacia la malignidad de la próstata se caracteriza por una serie secuencial de pasos.

Estas etapas comienzan con el desarrollo de la neoplasia intraepitelial prostática (PIN), a la que sigue la aparición de un cáncer de próstata localizado. Posteriormente, aparece una forma avanzada de adenocarcinoma de próstata con invasión local que, en última instancia, culmina en un cáncer de próstata metastásico [12]. El cáncer de próstata, una neoplasia que afecta a la glándula prostática, se considera uno de los problemas de salud más importantes entre la población masculina.

Representa una carga importante a escala mundial, ya que se sitúa como la segunda causa principal de cáncer y la quinta causa principal de mortalidad relacionada con el cáncer en los hombres [6]. Esta dolencia presenta una amplia gama de perfiles moleculares y heterogeneidades genéticas, lo que complica su diagnóstico y la implementación de estrategias de tratamiento eficaces [11].

La detección temprana del cáncer de próstata es imperativo para mejorar las tasas de supervivencia y la calidad de vida de los pacientes. Aunque se han establecido métodos convencionales como el antígeno prostático específico (PSA) y la biopsia, estos enfrentan desafíos significativos en términos de especificidad y sensibilidad, lo que subraya la necesidad apremiante de enfoques más precisos y no invasivos [7]. En este contexto, la aplicación de técnicas de aprendizaje automático ha surgido como un enfoque innovador y prometedor para abordar los desafíos asociados con la detección y clasificación del cáncer de próstata.

La capacidad del aprendizaje automático para analizar grandes conjuntos de datos ómicos y otros recursos proporciona una oportunidad única para identificar patrones moleculares y biomarcadores asociados con la enfermedad [10]. Estos elementos clave resultan fundamentales para esclarecer la complejidad de la expresión génica, un principio fundamental en la biología molecular. El principio fundamental de la biología molecular postula que el ácido deso-xirribonucleico (ADN) engendra ácido ribonucleico (ARN) y el ARN engendra proteínas. Este fenómeno se conoce como expresión génica, en el que la información genética se utiliza dentro de una entidad celular para generar las proteínas necesarias para la funcionalidad celular.

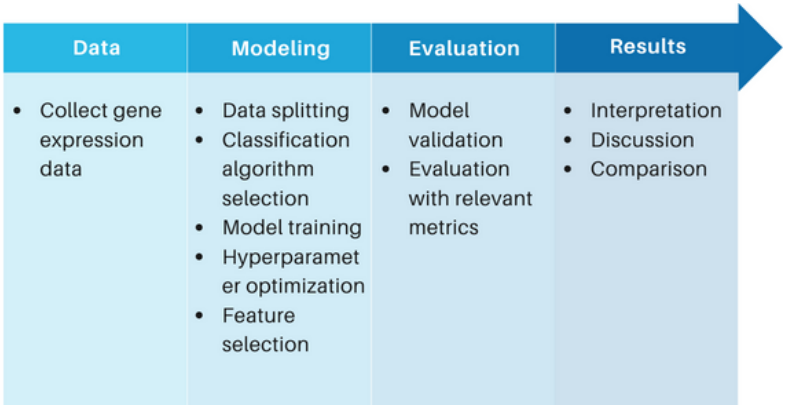

Fig. 1. Flujo del proceso.

Más específicamente, el modelo para la síntesis de proteínas está integrado en la secuencia de nucleótidos del ADN. La transformación de la información codificada por los genes en proteínas constituye un proceso celular que depende de los ácidos nucleicos. La expresión génica abarca dos procedimientos complejos: la transcripción y la traducción. La transcripción denota el acto de sintetizar ARN a partir del ADN, iniciando así la cascada de expresión génica y sirviendo como punto de control crítico en la síntesis de proteínas y la manifestación génica [9].

La investigación en el campo de la expresión génica ha revolucionado nuestra comprensión de las enfermedades. La abundancia de datos de expresión génica proporciona una ventana única hacia los complejos mecanismos moleculares subyacentes a esta enfermedad, pero también presenta desafíos significativos debido a la alta dimensionalidad y variabilidad inherentes a estos conjuntos de datos.

Dentro de este ámbito, la selección de características emerge como un componente esencial para destilar la información más relevante y biológicamente significativa de los datos de expresión génica del cáncer de próstata. La identificación de biomarcadores específicos y la comprensión de las firmas genéticas distintivas pueden desempeñar un papel crucial en el diagnóstico temprano, la estratificación de pacientes y el desarrollo de terapias más personalizadas.

En esta revisión se explora la selección de características aplicadas a datos de expresión génica, con un enfoque especial en su aplicación al estudio del cáncer de próstata. Se examina la evolución de estas técnicas, desde enfoques clásicos hasta enfoques más avanzados, destacando sus aplicaciones, limitaciones y contribuciones específicas a la comprensión de la biología subyacente al cáncer de próstata.

2. Materiales y métodos

La metodología empleada en este estudio sigue un enfoque sistemático y organizado para llevar a cabo la investigación como se muestra en la Fig. 1 Comienza con la recopilación de datos de expresión génica, esenciales para el análisis y la comprensión del problema en estudio.

Tabla 1. Impacto de la profundidad máxima en el rendimiento del árbol de decisión.

Profundidad máxima	AUC
1	0.735
2	0.619
3	0.629
10	0.705
20	0.705
30	0.685
40	0.655
50	0.695
100	0.690
None	0.705

Posteriormente, se procede con el modelado de los datos, donde se dividen en conjuntos de entrenamiento, prueba, y validación, se seleccionan los algoritmo de clasificación adecuados, se entrenan los modelos con los datos de entrenamiento y se optimizan los hiperparámetros para mejorar su rendimiento. Luego, se evalúan los modelos resultantes utilizando métricas pertinentes para validar su precisión y efectividad. En la etapa de resultados, se interpreta el significado de los hallazgos.

Posteriormente, se procede con el proceso de selección de características, donde se identifica un subconjunto óptimo de características a partir del conjunto original. Este proceso de selección de características es fundamental para destilar la información más relevante y biológicamente significativa de los datos de expresión génica del cáncer de próstata. Este nuevo subconjunto se utiliza para comenzar de nuevo desde el modelado. En conjunto, esta metodología proporciona una guía clara y estructurada para el proceso de investigación, desde la recolección de datos hasta la interpretación y discusión de los resultados, asegurando la coherencia y la rigurosidad en cada etapa del estudio.

3. Selección de características

El objetivo de la selección de características es identificar un subconjunto óptimo, representado por M **características**, a partir del conjunto original de N **dimensiones** (donde $M \leq N$), con el fin de maximizar la función objetivo. Dado un conjunto de características $X = \{x_i, i = 1, \ldots, N\}$, se busca un subconjunto $Y_M = \{x_{1i}, x_{2i}, \ldots, x_{iM}\}$ con $M \leq N$, que optimice la función objetivo $J(Y)$, la cual está relacionada con la probabilidad de clasificación correcta de alguna manera. La función objetivo, que evalúa la calidad del subconjunto de características, puede vincularse con la precisión predictiva, en el caso del enfoque del wrapper, o calcularse a partir del contenido de información del propio subconjunto (por ejemplo, distancia entre clases, correlación o medidas teóricas de información), siguiendo el enfoque de los filtros. La selección de características permite que las características seleccionadas mantengan su interpretación física original, lo cual facilita la comprensión del proceso físico subyacente en la generación de patrones.

Tabla 2. Evaluación de parámetros del árbol de decisión.

AUC			
Profundidad máxima	**1**	**10**	**50**
Número mínimo de muestras por hoja			
1	0.735	0.705	0.650
5	0.735	0.831	0.856
10	0.735	0.826	0.826
15	0.735	0.818	0.819
20	0.735	0.670	0.692
100	0.830	0.790	0.680
200	0.500	0.500	0.500
500	0.500	0.500	0.500

Además, puede resultar en una reducción de los costos de medición y/o computacionales, ya que solo se necesitará calcular las características seleccio-nadas. Sin embargo, es importante tener en cuenta que el proceso de encontrar el mejor subconjunto de características puede ser computacionalmente exigente, y en algunos casos, podría ser necesario conformarse con una solución subóptima [2].

3.1. Suma de cuadrados entre grupos y dentro de grupos

El método de selección de características basada en la suma de cuadrados entre grupos y dentro de grupos (BSS/WSS) clasifica las características según una proporción tal que las características con una gran variación entre clases y pequeñas variaciones dentro de las clases reciben calificaciones más altas. Este algoritmo de selección de características univariado determina las características que tienen un mayor poder de discriminación entre clases [3]. Para la característica k, $x_{i,k}$ denota el valor de la característica k para el ejemplo de entrenamiento i, $\overline{x_{z,k}}$ el valor promedio de la característica k en los ejemplos de clase z, y $\overline{x_k}$ el valor promedio de la característica k en todos los ejemplos. La relación BSS/WSS del gen k la proporciona la ecuación (1), donde $\delta_{i,z}$ es igual a 1 si el ejemplo i pertenece a la clase z y 0 en caso contrario:

$$\frac{\mathrm{BSS}(k)}{\mathrm{WSS}(k)} = \frac{\sum_i \sum_z \delta_{i,z} (\overline{x_{z,k}} - \overline{x_k})^2}{\sum_i \sum_z \delta_{i,z} (x_{i,k} - \overline{x_{z,k}})^2}. \tag{1}$$

Por lo tanto, es posible ordenar las características de mayor a menor en función de la proporción BSS/WSS. Este ratio puede funcionar como un peso vinculado a una característica, ya que a mayor sea, más relevante será su capacidad para discriminar. Un interrogante relevante en este tipo de selección de características, es determinar qué grupo de características elegir. El enfoque comúnmente empleado para resolver esta cuestión es analizar la curva de precisión en función del número de características y

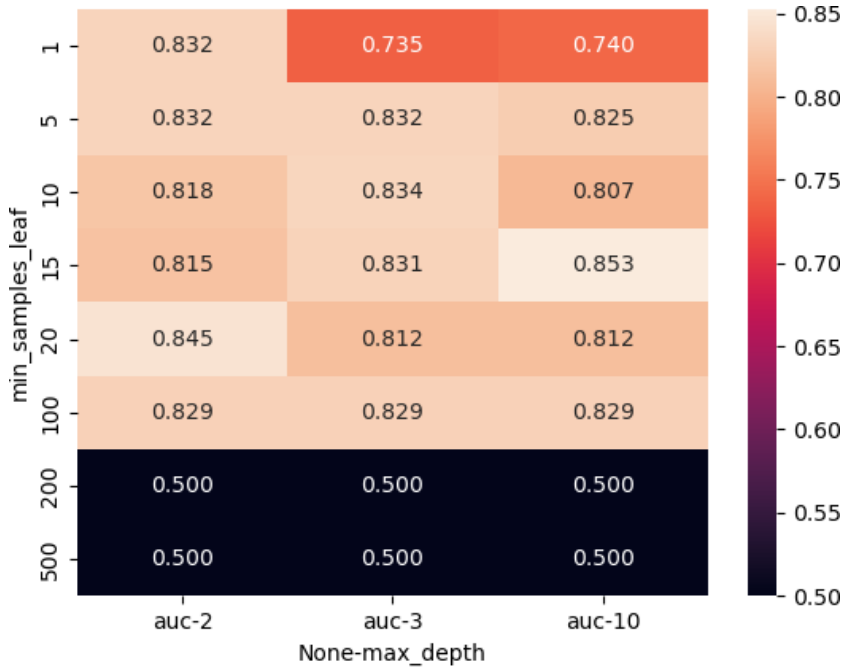

Fig. 2. Mapa de calor de puntuaciones de área bajo la curva (AUC) para árboles de decisión.

hallar un máximo local, o idealmente, un máximo global después del cual la precisión disminuye a medida que se incorporan más características. De manera más sencilla, también es factible seleccionar un número determinado de características, o limitar el conjunto de características seleccionadas en un punto de corte natural en la lista completa. Este es el enfoque utilizado en este caso.

4. Datos

La tecnología de secuenciación (RNA-seq) examina la cantidad y las secuencias de ARN en una muestra utilizando la secuenciación de próxima generación (NGS) [8]. Esta técnica analiza el transcriptoma, es decir, los patrones de expresión génica codificados dentro de nuestro ARN. En otras palabras, RNA-seq nos permite investigar y descubrir el contenido celular total de ARN, incluyendo ARNm, ARNr y ARNt. Al comprender el transcriptoma, podemos conectar la información de nuestro genoma con su expresión proteica funcional. El RNA-seq revela qué genes se activan en una célula, cuál es su nivel de expresión y cuándo se activan o desactivan [4, 5].

Por otro lado, el conjunto de datos sobre cáncer de próstata se descarga de Firehose[1]. Estos datos son el resultado de secuenciación de próxima generación, que ya ha sido normalizada. El conjunto de datos está organizado con genes en filas y pacientes en columnas.

[1] gdac.broadinstitute.org/

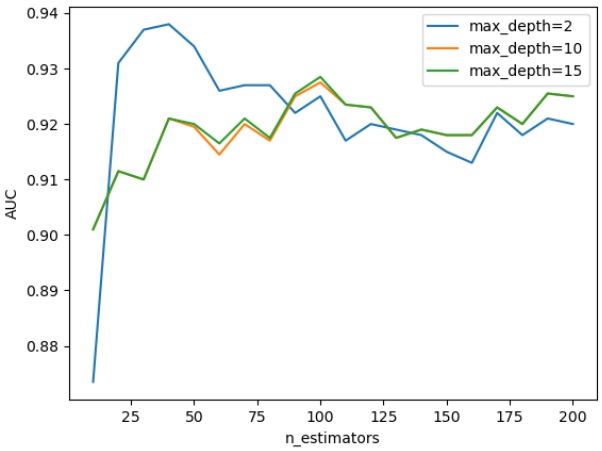

Fig. 3. Variación del AUC con respecto al número de estimadores para distintas profundidades máximas.

En total, consta de 20,531 filas que representan los genes y sus niveles de expresión, junto con 551 columnas (550 sujetos y una columna adicional para los nombres de los genes). De estas columnas, 498 corresponden a muestras tumorales y 52 a muestras normales.

5. Modelado y evaluación

Para la construcción de los modelos predictivos, primeramente se utilizó el conjunto de datos completos (550 muestras y 20,531 genes), se procedió a utilizar regresión logística, árboles de decisión y bosque aleatorio. Se dividieron los datos en conjuntos de entrenamiento (60 %), validación (20 %), y prueba (20 %) utilizando la función train_test_split de scikit-learn. Se evaluó el rendimiento de los modelos utilizando métricas relevantes como precisión y AUC (área bajo la curva) la cual resume la información de la curva ROC en una sola métrica.

Para la regresión logística, se ajustó un modelo utilizando el conjunto de entrenamiento y se evaluó su rendimiento en el conjunto de validación. Los resultados mostraron una precisión promedio del 90 % en la clasificación de muestras tumorales y normales. La matriz de confusión reveló una sensibilidad del 95 % y una especificidad del 90 %. Para los árboles de decisión, se llevó a cabo un análisis exhaustivo para determinar el impacto de la profundidad máxima del árbol de decisión en la capacidad predictiva del modelo.

Se exploraron varias profundidades, incluyendo valores específicos y sin límite de profundidad (None), y se evaluó el rendimiento del modelo utilizando el AUC. A continuación se presentan los resultados obtenidos en la Tabla 1. Los resultados muestran que la profundidad máxima del árbol de decisión influye significativamente en su capacidad para generalizar patrones en los datos de validación.

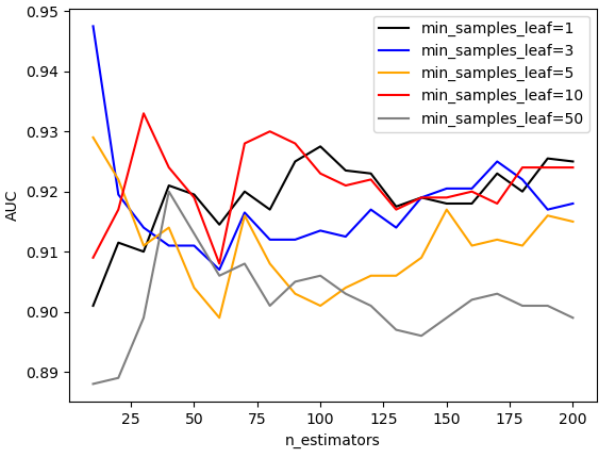

Fig. 4. Relación entre el número de estimadores y el AUC para distintos valores de min_samples_leaf.

Se observa un aumento en el AUC hasta una profundidad máxima de 10, seguido de una estabilización y, en algunos casos, un ligero descenso a profundidades mayores. Se decidió seleccionar una profundidad máxima de 10 para un equilibrio entre capacidad predictiva y complejidad del modelo.

Además, se llevaron a cabo pruebas adicionales variando tanto la profundidad máxima como el número mínimo de muestras por hoja para explorar la sensibilidad del modelo a estos parámetros. Los resultados se presentan en la Tabla 2. Estos resultados proporcionan información valiosa sobre la configuración óptima de los hiperparámetros del árbol de decisión para este conjunto de datos específico. Se realizó una evaluación comparativa entre dos modelos de árbol de Decisión para predecir la clasificación de muestras de cáncer de próstata basadas en la expresión génica.

Los modelos fueron entrenados con diferentes configuraciones de parámetros, lo que permitió comparar su rendimiento utilizando el AUC como métrica de evaluación. El primer modelo se caracterizó con una profundidad máxima de 10 y un número mínimo de muestras por hoja de 5, mientras que el segundo modelo se ajustó con una profundidad máxima de 1 y un número mínimo de muestras por hoja de 100.

Los resultados obtenidos revelan una diferencia notable en el rendimiento de los modelos. El modelo con una profundidad máxima de 1 y un número mínimo de muestras por hoja de 100 alcanzó un AUC de 0.79 en el conjunto de validación, mientras que el modelo con una profundidad máxima de 10 y un número mínimo de muestras por hoja de 5 obtuvo un AUC ligeramente inferior, con un valor de 0.76.

Este hallazgo sugiere que, a pesar de su mayor complejidad, el modelo con una profundidad máxima de 10 no logró superar significativamente al modelo más simple con una profundidad máxima de 1. La precisión comparable del modelo más simple destaca su eficacia en la tarea de clasificación de muestras de cáncer de próstata basadas en la expresión génica.

Con bosque aleatorio, de igual manera se realizó un proceso de optimización de hiperparámetros para mejorar el rendimiento en la clasificación. Se exploraron los efectos de variar los siguientes parámetros: el número de estimadores (n_estimators), la profundidad máxima del árbol (max_depth), y el número mínimo de muestras por hoja (min_samples_leaf).

Inicialmente, se evaluó el impacto del número de estimadores en el rendimiento del modelo. Se construyeron múltiples modelos de bosques aleatorios, variando el número de estimadores de 10 a 200 en incrementos de 10. Se continuó explorando el efecto de la profundidad máxima del árbol en el rendimiento del modelo. Se ajustaron modelos de bosques aleatorios con diferentes profundidades máximas (2, 10, y 15). Finalmente, se examinó la influencia del número mínimo de muestras por hoja en el desempeño del modelo. Se entrenaron múltiples modelos con diferentes valo-res de este parámetro (1, 3, 5, 10, y 50), manteniendo constante la profundidad máxima.

Los resultados finales muestran que el modelo de bosque aleatorio con 200 estimadores, una profundidad máxima de 10 y un número mínimo de muestras por hoja de 1 logró alcanzar un AUC de 0.9785 en el conjunto de validación. Esta configuración óptima demuestra la importancia de ajustar cuidadosamente los hiperparámetros para obtener un rendimiento óptimo del modelo en la clasificación de muestras de cáncer de próstata.

5.1. Selección de características utilizando BSS/WSS

La selección de características utilizando BSS/WSS permite reducir la dimensionalidad del conjunto de datos y mejorar la eficiencia computacional del mode-lo, al tiempo que conserva la información más relevante para la tarea de clasificación de muestras de cáncer de próstata. Al reducir la redundancia en los datos, esta técnica puede ayudar a mejorar la precisión y la generalización del modelo.

El cálculo de la relación entre la suma de cuadrados entre clases (BSS) y la suma de cuadrados dentro de las clases (WSS) para cada gen en el conjunto de datos permitió evaluar la capacidad de cada gen para discriminar entre clases de muestras. Utilizando los resultados de BSS/WSS, se seleccionaron los 100 mejores genes que contribuyen significativamente a la variabilidad entre las clases. Estos genes seleccionados forman un conjunto de datos reducido que conserva las características más importantes para la clasificación de muestras de cáncer de próstata.

Como anteriormente se realizó con el conjunto de datos completo, se aplicaron los algoritmos de aprendizaje automático, incluidos la regresión logística, árboles de decisión y bosques aleatorios, para realizar la clasificación de muestras de cáncer de próstata únicamente con las características más relevantes. Esta vez el modelo de regresión logística obtuvo una precisión del 91.82 %, una sensibilidad del 94 % y una especificidad del 70 %. Para el árbol de decisión, se exploraron diferentes profundidades máximas y diferentes números mínimos de muestras por hoja. Se calculó el AUC para cada configuración de hiperparámetros, obteniendo los resultados de la Fig. 2.

Se ajustó un modelo de árbol de decisión con una profundidad máxima de 10 y un número mínimo de muestras por hoja de 15 utilizando el conjunto de entrenamiento. Luego, se realizaron predicciones sobre el conjunto de validación y se calculó el AUC, obteniendo un valor de 0.8454.

Para el bosque aleatorio, se observó un aumento en el AUC con el incremento del número de árboles y la profundidad máxima como se puede observar en la Fig. 3. Además, se evaluaron diferentes valores del hiperparámetro de número mínimo de muestras por hoja (Fig. 4) para el bosque aleatorio con una profundidad máxima de 10. Se encontró que la AUC era más alta para el valor igual a 1 de dicho hiperparámetro.

En general, el modelo de bosque aleatorio con 200 estimadores, una profundidad máxima de 10 y un número mínimo de muestras por hoja igual a 1 tuvo el mejor rendimiento con un AUC de 0.925 en el conjunto de validación.

6. Conclusión

La investigación analizó la aplicación de técnicas de selección de características, específicamente basadas en la relación de suma de cuadrados entre grupos y dentro de grupos (BSS/WSS), en el análisis de expresión génica para clasificar muestras de cáncer de próstata. Los resultados revelaron la eficacia de esta técnica para identificar genes relevantes en la clasificación de muestras tumorales y normales.

La selección de características basada en BSS/WSS identificó un conjunto óptimo de genes que contribuyen significativamente a la variabilidad entre las clases de muestras de cáncer de próstata. Esta reducción de la dimensionalidad facilitó el modelado y mejoró la eficiencia computacional de los algoritmos de aprendizaje automático utilizados.

Los modelos de regresión logística, árboles de decisión y bosques aleatorios, luego de la selección de características, mostraron un rendimiento prome-tedor en la clasificación de muestras tumorales y normales, evidenciando la informatividad y relevancia de los genes seleccionados. Además, al ajustar cuidadosamente los hiperparámetros, se observó una mejora en el rendimiento de los modelos de bosques aleatorios. A pesar de reducir el número de genes de más de 20,500 a solo 100, el AUC no disminuyó significativamente.

Esto resalta la importancia de la optimización para lograr un rendimiento óptimo en la clasificación de muestras de cáncer de próstata. Para futuras investigaciones en el campo de la clasificación del cáncer de próstata, se pueden explorar diversas áreas. En primer lugar, sería interesante investigar técnicas avanzadas para seleccionar el número óptimo de características (genes) que realmente contribuyen a la clasificación precisa. Además, se podría comparar el enfoque utilizado en este estudio con otros métodos que emplean cómputo evolutivo, identificando posibles áreas de mejora y validando la efectividad de las características seleccionadas.

También se recomienda probar los modelos y las características en conjuntos de datos externos e independientes, asegurándose de que estén equilibrados para evaluar la generalización del enfoque. Por último, sería valioso crear modelos utilizando los hiperparámetros optimizados en este estudio para verificar su validez en un entorno clínico real, lo que podría tener implicaciones significativas para el diagnóstico y tratamiento del cáncer de próstata.

Agradecimientos. El autor principal agradece el apoyo recibido por el Consejo Nacional de Humanidades, Ciencias y Tecnologías (CONAHCYT) a través del Programa de Becas para Estudios de Posgrado en México.

Referencias

1. Bar, Y., Keenan, J. C., Ryan, L., Juric, D., Shin, J., Wander, S. A., Spring, L. M., Moy, B., Ellisen, L., Isakoff, S. J., Bardia, A., Vidula, N.: Abstract P4-01-14: Changes in the Genomic spectrum of actionable alterations in HER2 negative metastatic breast cancer in serial cell free DNA (cfDNA) analysis. Cancer Research, vol. 83, no. 5_Supplement (2023) doi: 10.1158/1538-7445.SABCS22-P4-01-14
2. Barrué, C.: The i-walker: An intelligent pedestrian mobility aid. Computational Intelligence in Healthcare 4 Studies in Computational Intelligence, vol. 309 (2010) doi: 10.1007/978-3-642-14464-6
3. Bichindaritz, I.: Comparison of reuse strategies for case-based classification in bioinformatics. In: Case-Based Reasoning Research and Development, pp. 393–407 (2011) doi: 10.1007/978-3-642-23291-6_29
4. Deshpande, D., Chhugani, K., Chang, Y., Karlsberg, A., Loeffler, C., Zhang, J., Muszyńska, A., Munteanu, V., Yang, H., Rotman, J., Tao, L., Balliu, B., Tseng, E., Eskin, E., Zhao, F., Mohammadi, P., P. Łabaj, P., Mangul, S.: RNA-seq data science: From raw data to effective interpretation. Frontiers in Genetics, vol. 14 (2023) doi: 10.3389/fgene.2023.997383
5. Farrell, R. E.: Chapter 24 - RNA-seq: The premier transcriptomics tool. RNA Methodologies (Sixth Edition), pp. 697–721 (2023) doi: 10.1016/B978-0-323-90221-2.00045-X
6. Ferlay, J., Colombet, M., Soerjomataram, I., Mathers, C., Parkin, D., Piñeros, M., Znaor, A., Bray, F.: Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. International Journal of Cancer, vol. 144, no. 8, pp. 1941–1953 (2019) doi: 10.1002/ijc.31937
7. Loeb, S., Vellekoop, A., Ahmed, H. U., Catto, J., Emberton, M., Nam, R., Rosario, D. J., Scattoni, V., Lotan, Y.: Systematic review of complications of prostate biopsy. European Urology, vol. 64, no. 6, pp. 876–892 (2013) doi: 10.1016/j.eururo.2013.05.049
8. Million, M., Feyissa, T.: RNA-Seq as an effective tool for modern transcriptomics, a review-based study. Journal of Applied Research in Plant Sciences, vol. 3, no. 02, pp. 236–241 (2022) doi: 10.38211/joarps.2022.3.2.29
9. Shen, C. H.: Chapter 3 - Gene expression: Transcription of the genetic code. Diagnostic Molecular Biology (Second Edition), pp. 57–89 (2023) doi: 10.1016/B978-0-323-91788-9.00003-X
10. Subramanian, I., Verma, S., Kumar, S., Jere, A., Anamika, K.: Multi-omics data integration, interpretation, and its application. Bioinformatics and biology insights, vol. 14 (2020) doi: 10.1177/1177932219899051
11. Taylor, B. S., Schultz, N., Hieronymus, H., Gopalan, A., Xiao, Y., Carver, B. S., Arora, V. K., Kaushik, P., Cerami, E., Reva, B., Antipin, Y., Mitsiades, N., Landers, T., Dolgalev, I., Major, J. E., Wilson, M., Socci, N. D., Lash, A. E., Heguy, A., Eastham, J. A., et al.: Integrative Genomic Profiling of Human Prostate Cancer. Cancer Cell, vol. 18, no. 1, pp. 11–22 (2010) doi: 10.1016/j.ccr.2010.05.026
12. Wang, G., Zhao, D., Spring, D. J., DePinho, R. A.: Genetics and biology of prostate cancer. Genes & development, vol. 32, no. 17-18, pp. 1105–1140 (2018) doi: 10.1101/gad.315739.118