ReTEC

Revista Electrónica de Tecnología, Educación y Ciencia
ISSN: 2953-5654
http://retec.unsa.edu.ar
Universidad Nacional de Salta

Estimación de la Densidad de Columna de CO mediante Aprendizaje Automático en Regiones de Formación Estelar

Rocio D. Taboada[1,3], Cristina C. Mendez[2]

[1]Instituto de Investigaciones en Energía No Convencional, CONICET-UNSa, Argentina
[2] Departamento de Física, Facultad de Ciencias Exactas, UNSa, Argentina
[3] Departamento de Física, Facultad de Ciencias Exactas y Naturales, UBA, Argentina
rocio.taboada@exa.unsa.edu.ar , mendezcris.75@gmail.com

Revista Electrónica de Tecnología, Educación y Ciencia,

Volumen 1, Número 3, pág. 126-133, jun, 2026. ISSN: 2953-5654

Disponible en http://retec.unsa.edu.ar/

Estimación de la Densidad de Columna de CO mediante Aprendizaje Automático en Regiones de Formación Estelar

Rocio D. Taboada[1,3], Cristina C. Mendez[2]

[1]Instituto de Investigaciones en Energía No Convencional, CONICET-UNSa, Argentina
[2] Departamento de Física, Facultad de Ciencias Exactas, UNSa, Argentina
[3] Departamento de Física, Facultad de Ciencias Exactas y Naturales, UBA, Argentina
rocio.taboada@exa.unsa.edu.ar , mendezcris.75@gmail.com

Resumen: La determinación precisa de la distribución de densidades de columna moleculares en nubes interestelares es esencial para caracterizar la estructura química y física de las regiones de formación estelar. Sin embargo, los métodos tradicionales, como aquellos basados en el equilibrio termodinámico local (LTE) o en modelos de transferencia radiativa no-LTE como RADEX, presentan limitaciones asociadas a suposiciones simplificadas o a altos costos computacionales. En este trabajo, exploramos el uso de algoritmos de aprendizaje automático como herramienta alternativa para estimar la densidad de columna del 13CO (J=3–2), a partir de un cubo espectral observacional y su mapa asociado de temperatura de excitación. Utilizamos datos de la región G29.96–0.02 observados con el telescopio James Clerk Maxwell (JCMT) para entrenar y validar dos modelos supervisados: Random Forest y una red neuronal multicapa (MLP). Ambos modelos mostraron correlaciones razonables con las densidades derivadas mediante el método LTE, alcanzando coeficientes de determinación R 2 de 0.67 y 0.51, respectivamente. Si bien se evaluó la incorporación de parámetros físicos derivados como la profundidad óptica y la intensidad integrada, estos no mejoraron el desempeño del modelo. Los resultados sugieren que futuras mejoras podrían requerir la inclusión de información proveniente de otras moléculas o transiciones espectrales. Este estudio representa un primer paso hacia el uso de técnicas de aprendizaje automático para inferir propiedades físicas del medio interestelar de manera eficiente y escalable.

Palabras claves: Densidad de columna, Aprendizaje automático, regiones de formación estelar.

1. Introducción

En la teoría de la formación estelar, resulta fundamental comprender la estructura interna de las nubes moleculares, ya que es en sus regiones más densas donde ocurre el colapso gravitacional que da origen a las estrellas. Una de las herramientas estadísticas más utilizadas para caracterizar dichas estructuras es la función de distribución de probabilidad de la densidad de columna (PDF, por sus siglas en inglés). Este enfoque permite analizar el comportamiento global de la materia en una nube, identificar firmas de procesos físicos dominantes como turbulencia, gravedad o retroalimentación estelar, y comparar observaciones con simulaciones numéricas.

La formación estelar es un proceso fundamental en la astrofísica que ocurre en las regiones más densas y frías del medio interestelar (MIE), conocidas como nubes moleculares [1]. El hidrógeno molecular (H2), el componente principal de estas nubes, es difícil de observar directamente. Por lo tanto, se utilizan trazadores moleculares como el monóxido de carbono (CO) para inferir la distribución y propiedades del gas [2]. El isótopo más abundante, 12CO, a menudo se encuentra ópticamente grueso (saturado) en las regiones densas, lo que limita su capacidad para sondear el interior de las nubes [3]. En contraste, el isótopo menos abundante 13CO, es

significativamente menos ópticamente grueso, lo que lo convierte en un trazador más fiable de la masa de gas molecular total y de las condiciones físicas en los núcleos densos donde se forman las estrellas [4]. La generación de mapas de densidad de columna de 13CO (N(13CO)) es, por tanto, crucial para entender la estructura de las nubes moleculares, los procesos de formación estelar y la evolución del gas en galaxias [5]. Este isótopo también es valioso para investigar fenómenos como el fraccionamiento químico y el agotamiento en regiones de alta densidad.

Los métodos tradicionales para derivar la densidad de columna N presentan limitaciones importantes. Aquellos basados en suposiciones de equilibrio termodinámico local (LTE) pueden no ser válidos en condiciones reales, mientras que los métodos no-LTE, como los implementados en RADEX, requieren múltiples líneas y son computacionalmente costosos. Además, son sensibles al ruido y a las condiciones iniciales. Esto dificulta el análisis de grandes muestras de nubes moleculares.

En este contexto, el aprendizaje automático (ML) emerge como una alternativa prometedora. Su capacidad para modelar relaciones no lineales complejas, procesar grandes volúmenes de datos y automatizar tareas lo convierte en una herramienta adecuada para inferir propiedades físicas a partir de observaciones espectrales. En particular, podría permitir una estimación eficiente del grado de fotodisociación mediante el análisis indirecto de líneas moleculares.

En este trabajo, exploramos la aplicación de modelos de aprendizaje supervisado específicamente Random Forest y redes neuronales multicapa (MLP) para predecir la densidad de columna del 13CO a partir de un cubo espectral y su mapa de temperatura de excitación.

2. Datos y preprocesamiento

La región analizada corresponde a G29.96–0.02, una conocida zona de formación estelar masiva, observada con el telescopio James Clerk Maxwell (JCMT). Se utilizó un cubo espectral del 13CO (J=3–2), compuesto por 399 canales espectrales, junto con un mapa de temperatura de excitación de dimensiones 117×97 píxeles. La densidad de columna de 13CO, N(13CO), fue estimada píxel a píxel a partir de la emisión (J=3–2) bajo la suposición de equilibrio térmico local (LTE), siguiendo el procedimiento detallado en [6]. Por su parte, la temperatura de excitación (Tex) fue derivada a partir de la emisión del 12CO (J=3–2), asumiendo que esta línea es ópticamente gruesa.

Obtener un mapa de densidad de columna y temperatura en una región de formación estelar es comparable a cartografiar un bosque denso. El enfoque LTE equivale a estimar la densidad de árboles y la temperatura ambiente basándose en la sombra proyectada (es decir, la emisión más brillante), asumiendo que la luz es absorbida completamente. En contraste, un análisis no-LTE, por ejemplo utilizando herramientas como RADEX, se asemeja a emplear múltiples imágenes obtenidas desde distintos ángulos y sensores para construir un modelo 3D más detallado del bosque. Este enfoque permite capturar la compleja interacción de la luz con las copas de los árboles, revelando variaciones locales en densidad y temperatura que el método LTE no detecta.

Para preparar los datos, se reordenaron los cubos espectrales, pasando del formato ZYX a YXZ. Se descartaron los píxeles con valores inválidos (NaNs) o físicamente inconsistentes (densidad de columna o temperatura de excitación ≤ 0). Los espectros fueron suavizados mediante un filtro de Savitzky-Golay, utilizando una ventana de 11 canales y un polinomio de

orden 3, con el objetivo de reducir el ruido preservando las características de las líneas espectrales.

Cada píxel fue representado por un vector de entrada de 400 dimensiones: los 399 valores del espectro de 13CO más el valor de temperatura de excitación correspondiente. La variable objetivo fue la densidad de columna asociada a ese píxel. El conjunto completo se dividió en subconjuntos de entrenamiento (aproximadamente 72–80 %), validación (aproximadamente 8–10 %) y prueba (10–20 %), adaptando la proporción según el modelo empleado (Random Forest o MLP). Tanto las variables de entrada como las etiquetas fueron normalizadas utilizando StandardScaler, asegurando media cero y desviación estándar unitaria para cada característica.

3. Metodología

En esta sección describimos los modelos de aprendizaje automático utilizados para la predicción de la densidad de columna, junto con los criterios empleados para su evaluación.

3.1. Modelos de Machine Learning

a) Random Forest (RF)

El modelo *Random Forest* es un método de aprendizaje conjunto (*ensemble learning*) basado en árboles de decisión [7]. Combina la técnica de *bagging* (*Bootstrap Aggregating*) con la aleatoriedad en la selección de características en cada nodo para construir múltiples árboles de decisión independientes. Las predicciones finales se obtienen promediando las salidas de todos los árboles, lo cual reduce el sobreajuste y mejora la robustez del modelo [8].

Los hiperparámetros del modelo fueron optimizados mediante búsqueda en cuadrícula (*Grid SearchCV*). Se evaluaron 30 combinaciones de parámetros utilizando validación cruzada con 5 *folds*. La combinación óptima fue:

- **n_estimators**: 478
- **max_depth**: 15
- **min_samples_leaf**: 3
- **min_samples_split**: 6
- **max_features**: 0.7

El modelo fue entrenado con el conjunto de entrenamiento (~80 %) y evaluado tanto en los subconjuntos de validación como de prueba (~10 % cada uno), definidos manualmente después del preprocesamiento.

También exploramos la posibilidad de mejorar el desempeño del modelo RF mediante la inclusión de variables adicionales como la profundidad óptica y la intensidad integrada. Sin embargo, esta expansión del conjunto de atributos no resultó en un aumento significativo del coeficiente de determinación R^2, lo cual sugiere que, en este caso, la incorporación de nuevas variables no aportaba información sustancialmente independiente o relevante para el modelo. Esto podría deberse a que las nuevas features estaban altamente correlacionadas con las ya existentes, o a que el modelo alcanzaba una capacidad predictiva saturada con el conjunto inicial de variables.

b) Red Neuronal Multicapa (MLP)

Las MLP son un tipo fundamental de red neuronal *feedforward*, compuestas por una capa de entrada, varias capas ocultas y una capa de salida [9]. Los pesos de las conexiones se ajustan durante el entrenamiento mediante el algoritmo de retropropagación (*backpropagation*) [10].

La arquitectura de la MLP utilizada fue la siguiente:

Capa de entrada: 400 neuronas (399 canales espectrales + temperatura de excitación).

Primera capa oculta: 64 neuronas con activación ReLU.

Segunda capa oculta: 32 neuronas con activación ReLU.

Capa de salida: 1 neurona con activación lineal (regresión).

El modelo fue compilado con la función de pérdida mean_squared_error y el optimizador Adam con una tasa de aprendizaje (*learning rate*) de 0.001. Se aplicó regularización L2 (kernel_regularizer) con un coeficiente de 0.05 para evitar el sobreajuste. El entrenamiento se realizó durante 100 épocas, con validación temprana (*early stopping*) en base al rendimiento del conjunto de validación.

3.2. Métricas de evalución

Para evaluar el rendimiento predictivo de ambos modelos, se utilizaron las siguientes métricas sobre el conjunto de prueba:

Coeficiente de Determinación (R^2): Mide la proporción de la varianza explicada por el modelo. Un valor cercano a 1 indica un buen ajuste, mientras que valores negativos sugieren un desempeño peor que una predicción constante.

Error Cuadrático Medio (MSE): Representa el promedio de los cuadrados de los errores entre las predicciones y los valores reales. Valores más bajos indican mejor rendimiento.

4. Resultados

4.1. Rendimiento del modelo Random Forest

El modelo *RF*, optimizado mediante GridSearchCV, alcanzó un coeficiente de determinación R^2 = 0,9054 y un error cuadrático medio (MSE) de 0.0976 en el conjunto de entrenamiento. En el conjunto de validación, el modelo obtuvo R^2 = 0,6147 y MSE = 0.3543, mientras que en el conjunto de prueba el rendimiento fue de R^2 = 0,6750 y MSE = 0.3023.

En la Figura 1, se muestra el mapa de densidad de columna predicho por el modelo RF (panel central), el cual reproduce de manera general las estructuras principales observadas en el mapa de densidad real (panel izquierdo). Se aprecia una buena correspondencia en la ubicación de las regiones de mayor y menor densidad. Sin embargo, el mapa de residuos (panel derecho), que representa la diferencia entre los valores reales y los predichos (Real - Predicho), revela errores localizados. Estos errores tienden a ser más pronunciados en las regiones de mayor densidad de columna, donde el modelo podría estar subestimando o sobrestimando los valores en ciertos píxeles. Las zonas con residuos cercanos a cero (color gris claro) indican una buena

concordancia entre la predicción y el valor real, mientras que los colores más intensos (rojo y azul) señalan las áreas donde el modelo presenta las mayores discrepancias. Esta visualización de los residuos es crucial para identificar las limitaciones del modelo en diferentes regímenes de densidad.

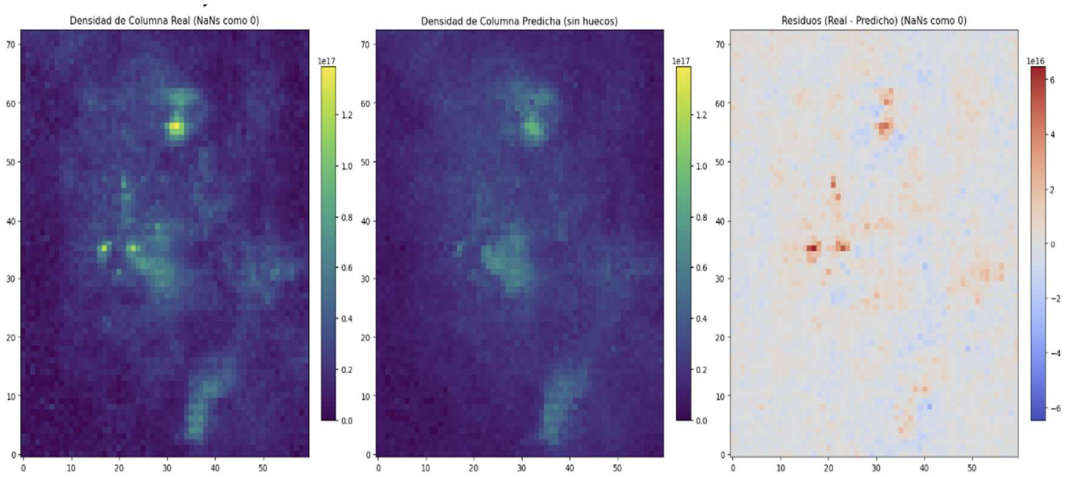

Figura 1: Izquierda: mapa de densidad de columna real. Centro: mapa de densidad predicho por RF. Derecha: mapa de residuos (Real - Predicho).

4.2. Rendimiento del modelo MLP

El modelo de Red Neuronal Multicapa (MLP) mostró un rendimiento consistente durante el entrenamiento y la validación, aunque con métricas inferiores a las del Random Forest. El MLP alcanzó un R^2 = 0,9890 y un MSE = 0.0113 en el conjunto de entrenamiento. En el conjunto de validación, los valores fueron R^2 = 0,5048 y MSE = 0.4554. En el conjunto de prueba, el rendimiento fue de R^2 = 0,5504 y MSE = 0.4183.

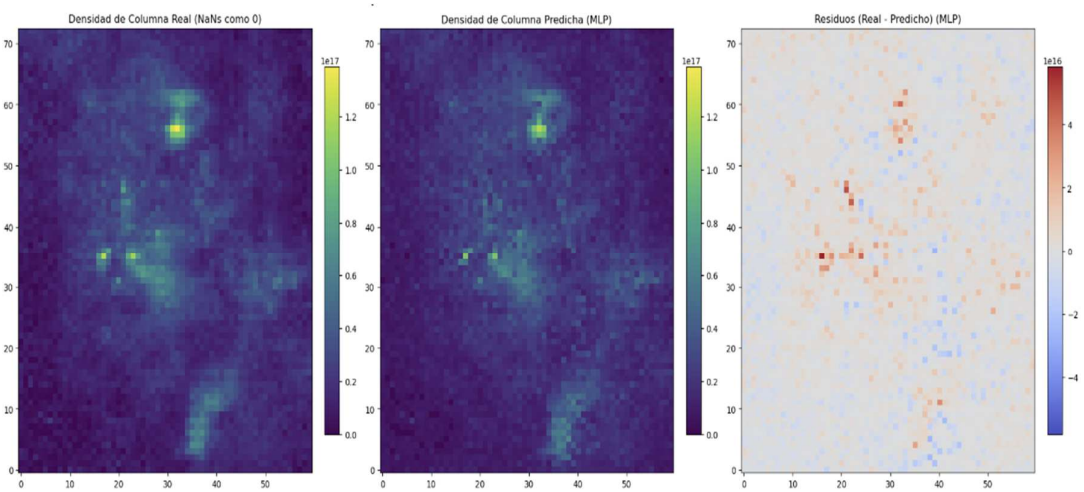

Figura 2: Izquierda: mapa de densidad de columna real. Centro: predicción con MLP. Derecha: mapa de residuos.

En la Figura 2, se observa el mapa de densidad de columna predicho por el modelo MLP (panel central), que también reproduce las estructuras principales del mapa real (panel izquierdo), de forma similar al modelo RF. No obstante, al examinar el mapa de residuos (panel derecho), se observa una mayor dispersión y errores más prominentes en comparación con el modelo RF, especialmente en las regiones de mayor densidad, esto sugiere que el MLP tiene más dificultad para capturar la variabilidad de los datos o puede estar más propenso a errores en los rangos extremos de la densidad de columna.

4.3. Comparación entre Modelos

El modelo RF demostró un mejor equilibrio entre precisión y robustez en comparación con el MLP. Si bien el MLP obtuvo resultados aceptables, se mostró más sensible tanto a la configuración de sus parámetros como al preprocesamiento de los datos.

En la Figura 3, se observa esta diferencia de rendimiento: en el panel izquierdo, los puntos azules se agrupan con mayor densidad en torno a la línea diagonal (que representa una predicción perfecta, donde los valores reales coinciden con los predichos). Aunque existe cierta dispersión, especialmente en los valores más altos, la tendencia general indica un ajuste más preciso.

Por el contrario, en el panel derecho, los puntos rojos presentan una mayor dispersión alrededor de la diagonal. Esto sugiere que el MLP tiene más dificultades para realizar predicciones precisas, mostrando una variabilidad más significativa en sus errores.

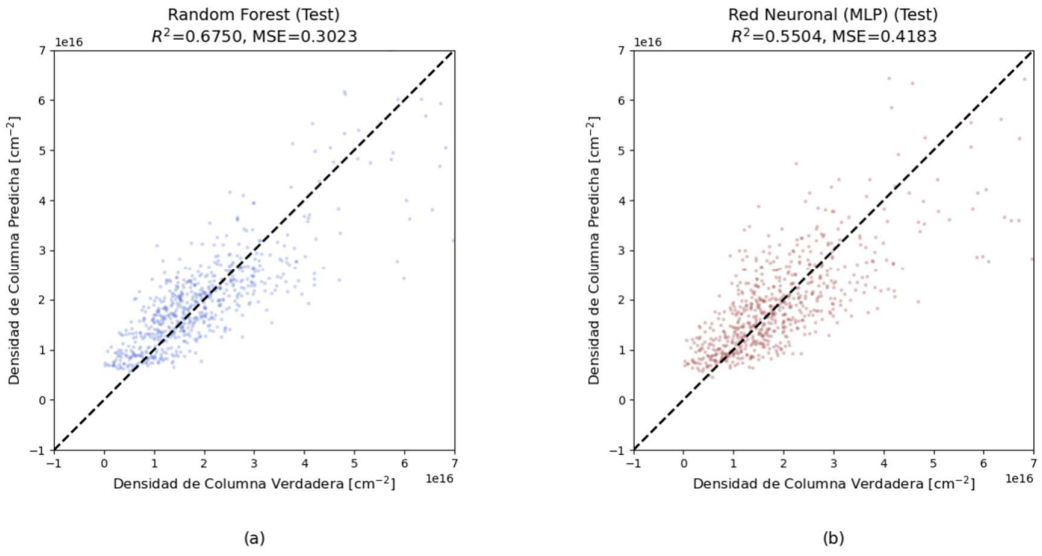

Figura 3: (a) Diagrama de dispersión para Random Forest. (b) Diagrama de dispersión para MLP.

Desde una perspectiva astrofísica, un valor de R^2 = 0,67 indica que el modelo logra capturar una parte significativa de la variabilidad en la densidad de columna, aunque con limitaciones en su precisión. Este grado de correlación es útil para identificar estructuras globales en la nube molecular, pero podría no ser suficiente para estudios detallados en zonas críticas, como los núcleos de formación estelar. Las mayores discrepancias observadas en las regiones de alta densidad sugieren que el modelo enfrenta dificultades para generalizar en rangos extremos. Esto

podría deberse a una mayor complejidad física en dichas zonas o a un desequilibrio en la representación de los datos. En el caso del modelo MLP, el contraste entre el alto rendimiento en el entrenamiento y el bajo desempeño en validación y prueba indica la presencia de sobreajuste. Este fenómeno sugiere que la red ha memorizado los patrones específicos del conjunto de entrenamiento, sin lograr una generalización robusta. Para mitigar este problema, se podrían explorar arquitecturas más simples, técnicas de regularización más fuertes o estrategias como el dropout.

5. Conclusiones

En este trabajo demostramos que los modelos de aprendizaje automático, como Random Forest y redes neuronales multicapa, pueden emplearse para estimar la densidad de columna de ^{13}CO a partir de datos espectrales y mapas auxiliares, como el de temperatura de excitación. Si bien el desempeño no alcanza aún la precisión de los métodos clásicos en todos los regímenes, los modelos logran capturar las estructuras globales y ofrecen tiempos de cómputo considerablemente menores. Las limitaciones observadas, particularmente en la predicción de valores extremos, sugieren que la incorporación de variables adicionales (como la emisión de otras transiciones o moléculas) podría mejorar significativamente el rendimiento. Como trabajo futuro, se propone extender este enfoque a cubos multitransición o multitrazador (por ejemplo, ^{13}CO y ^{18}CO), así como explorar arquitecturas más profundas con el objetivo de optimizar la inferencia de propiedades físicas en el medio interestelar a gran escala.

Referencias

[1] Lyman Spitzer. *Physical Processes in the Interstellar Medium*. Wiley, New York, 1978.
[2] Paul F. Goldsmith. Molecular depletion and thermal balance in dark cloud cores. *The Astrophysical Journal*, 680(1):428–445, 2008.
[3] Jorge E. Pineda, Paola Caselli, and Alyssa A. Goodman. Co isotopologues in the per seus molecular cloud complex: The x-factor and regional variations. *The Astrophysical Journal*, 679(1):481–496, 2008.
[4] Dario Colombo, Adam Ginsburg, Serena Viti, and et al. Physical properties of mole cular clouds in the galactic centre. *Monthly Notices of the Royal Astronomical Society*, 475(4):4504–4526, 2018.
[5] Yoshito Shimajiri, Philippe André, Jonathan Braine, and et al. Physical properties of the molecular gas in the orion b cloud as seen by the herschel gould belt survey and iram 30m. *Astronomy & Astrophysics*, 564:A68, 2014.
[6] S. Paron, M. B. Areal, and M. E. Ortega. Mapping the ^{13}co/C^{18}O abundance ratio in the massive star-forming region g29.96–0.02. *Astronomy and Astrophysics*, 617:A14, 2018.
[7] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
[8] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.
[9] Simon Haykin. *Neural networks and learning machines*. Pearson, 2009.
[10] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.