# Proyecto – Analítica de datos

Anderson J. Alvarado[1]*

[1]Pontificia Universidad Javeriana

**ABSTRACT**

En este documento, se presenta un enfoque novedoso basado en Inteligencia Artificial (IA) para abordar los desafíos relacionados con la seguridad alimentaria y la sostenibilidad en Colombia. El objetivo principal es desarrollar un modelo predictivo integral que combine datos del Índice de Precios al Consumidor (IPC) y datos climáticos para anticipar fluctuaciones en los precios de los alimentos. Mediante la aplicación de algoritmos de clasificación y agrupamiento, se busca identificar patrones complejos y relaciones no lineales entre estas variables. El modelo propuesto tiene el potencial de proporcionar información valiosa sobre cómo las condiciones climáticas afectan los precios de los alimentos en diferentes regiones de Colombia. Estas predicciones permitirán a los tomadores de decisiones prepararse y mitigar los efectos adversos en la seguridad alimentaria y la estabilidad económica. Además, el análisis de agrupamiento revelará similitudes y diferencias entre las distintas ciudades o regiones en términos de patrones climáticos y variaciones de precios. Este enfoque integral aborda múltiples aspectos interrelacionados y podría adaptarse a diferentes contextos dentro de Colombia, generando beneficios económicos y sociales a través de una mejor planificación y toma de decisiones informada en el sector alimentario.

**Keywords:** IPC, Temperatura, Agrupamiento, Categorias.

## 1 INTRODUCCIÓN

La seguridad alimentaria y la sostenibilidad agrícola son temas cruciales en el contexto colombiano, donde la variabilidad climática y las fluctuaciones de precios pueden tener un impacto significativo en la producción y accesibilidad de los alimentos [1]. En las últimas décadas, el cambio climático ha exacerbado fenómenos meteorológicos extremos, como sequías e inundaciones, lo que ha afectado negativamente la productividad agrícola y ha ocasionado volatilidad en los precios de los alimentos [2]. Además, la falta de información oportuna y precisa sobre las tendencias climáticas y los patrones de precios dificulta la toma de decisiones informada para mitigar estos desafíos [3].

En este contexto, la inteligencia artificial emerge como una herramienta poderosa para abordar estos problemas complejos. Los avances en el aprendizaje automático y el análisis de grandes conjuntos de datos han demostrado su potencial para identificar patrones complejos y relaciones no lineales entre múltiples variables [4]. Específicamente, los algoritmos de clasificación, como los árboles de decisión, y los algoritmos de agrupamiento, como DBSCAN, han sido ampliamente utilizados en diversas áreas, incluyendo la predicción climática y la modelización agrícola [5, 6].

Varios estudios han explorado el uso de IA para abordar desafíos relacionados con la seguridad alimentaria y la agricultura sostenible, por ejemplo, Tran, et al. [7] analizan investigaciones recientes sobre algoritmos de aprendizaje automático para la predicción de precios agrícolas. Discuten la importancia de la agricultura en los países en desarrollo y los problemas asociados a las caídas en los precios de los cultivos.

Sin embargo, la mayoría de estos enfoques se han centrado en aspectos individuales, como la predicción de precios o la optimización de rendimientos. Nuestro enfoque propone un modelo predictivo integral que combina múltiples fuentes de datos, incluyendo el Índice de Precios al Consumidor (IPC) y datos climáticos, para abordar de manera holística los desafíos relacionados con la seguridad alimentaria y la producción agrícola en Colombia.

## 2 HIPÓTESIS

Se propone que las variaciones en las condiciones climáticas, específicamente la temperatura y la humedad, tienen un impacto significativo en el Índice de Precios al Consumidor (IPC) en las diferentes ciudades de Colombia. La hipótesis plantea que un incremento en la temperatura está asociado con un aumento en los precios de productos perecederos, debido a una mayor tasa de descomposición y a los costos adicionales de almacenamiento en frío necesarios para preservar estos productos. Por otro lado, se espera que altos niveles de humedad afecten negativamente la calidad y la disponibilidad de productos agrícolas, lo cual se reflejaría en un aumento de sus precios en el mercado.

En consecuencia, se plantea que las variables climáticas, como la temperatura y la humedad, pueden ser utilizadas para predecir el IPC en diferentes ciudades. Utilizamos Bogotá como referencia inicial para evaluar la viabilidad del enfoque, con el objetivo de extender la aplicabilidad de esta investigación a otras ciudades. Esto permitirá el desarrollo de modelos predictivos robustos basados en datos climáticos en un contexto más amplio.

## 3 EXPLORACIÓN DE LOS DATOS

En este proyecto, se han recopilado y utilizado varios conjuntos de datos de diferentes fuentes para construir un modelo predictivo integral basado en inteligencia artificial. A continuación, se describen los datos obtenidos:

### 3.1 IPC mensual por ciudad

El Índice de Precios al Consumidor (IPC) es un indicador económico que mide la variación promedio de los precios de un conjunto específico de bienes y servicios representativos del consumo habitual de los hogares.

En nuestro proyecto, se enfoca exclusivamente en el IPC mensual como indicador económico clave. Esta decisión se basa en la naturaleza del objetivo, que es implementar un algoritmo de IA que analice la relación entre el IPC mensual y los datos mensuales del clima de cada ciudad. Al centrarse en el IPC mensual, podemos capturar de manera más precisa las fluctuaciones de precios en períodos cortos, lo que nos permite correlacionar estos datos de manera más efectiva con las variables climáticas mensuales.

*e-mail: andersonjalvarado@javeriana.edu.co

Sin embargo, el dataset "IPC_Por_ciudad_IQY.xlsx" contiene todas las variantes del IPC en una sola hoja de Excel, estructurándose de la siguiente manera: tiene 2184 filas y 24 columnas. La primera columna está en formato de fecha (AAAA-MM), y las demás contienen los valores del IPC de varias ciudades de Colombia, incluyendo Bogotá, Neiva, Medellín, entre otras, en formato decimal con negativos. Estos datos se encuentran apilados, incluyendo las diferentes variantes del IPC.

| Fecha | Bogotá | Medellín | Cali |
|---|---|---|---|
| 2024-01 | 105.2 | 103.5 | 104.0 |
| 2024-02 | 106.0 | 104.0 | 105.0 |

Table 1: Estructura IPC mensual por ciudad

Los datos fueron tomados directamente de la pagina del DANE [11].

### 3.2 Datos Climáticos

Para los datos climáticos, primero se obtuvieron las coordenadas geográficas de longitud y latitud de Geodatos [9]. Posteriormente, se utilizó NASA Power [10] para obtener los datos meteorológicos de cada ciudad. Se recopilaron datos diarios desde el año 2010, hasta el 2024, con un total de 5205 filas.

- **PRECTOTCORR**: Media corregida del sesgo de la precipitación total en la superficie de la tierra en masa de agua (incluye el contenido de agua en la nieve).
- **T2M**: Temperatura media del aire (bulbo seco) a 2 metros por encima de la superficie terrestre.
- **T2M_MAX**: Temperatura máxima horaria del aire (bulbo seco) a 2 metros por encima de la superficie de la tierra en el periodo de interés.
- **T2M_MIN**: Temperatura mínima horaria del aire (bulbo seco) a 2 metros sobre la superficie de la tierra en el periodo de interés.
- **RH2M**: Relación entre la presión parcial real de vapor de agua y la presión parcial de saturación, expresada en porcentaje.
- **WS10M**: Media de la velocidad del viento a 10 metros por encima de la superficie de la tierra.

| Fecha | T2M_MAX | T2M_MIN | RH2M | PRECTOTCORR |
|---|---|---|---|---|
| 2024-01-01 | 30.3 | 25.58 | 79.31 | 0.079677 |
| 2024-01-02 | 29.93 | 25.76 | 79.62 | 0.720357 |

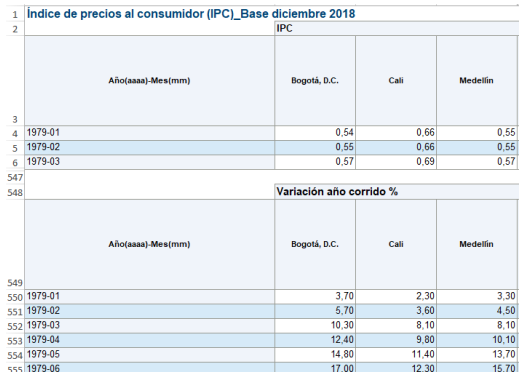
Table 2: Estructura índices climáticos

## 4 Tratamiento de Datos

En este apartado se busca explicar como, después de obtener los datos, se tuvieron que reestructurar los datasets para la aplicación de los modelos.

### 4.1 IPC

En el contexto de los datos asociados al Índice de Precios al Consumidor, fue necesario llevar a cabo un proceso de mapeo para identificar las filas correspondientes a los diferentes tipos de IPC, Puesto que estos se encontraban apilados dentro de un mismo archivo Excel, como se mencionó anteriormente.



| | | | |
|---|---|---|---|
| 1 | Índice de precios al consumidor (IPC)_Base diciembre 2018 | | |
| 2 | | IPC | |
| 3 | Año(aaaa)-Mes(mm) | Bogotá, D.C. | Cali | Medellín |
| 4 | 1979-01 | 0,54 | 0,66 | 0,55 |
| 5 | 1979-02 | 0,55 | 0,66 | 0,55 |
| 6 | 1979-03 | 0,57 | 0,69 | 0,57 |
| 547 | | | | |
| 548 | | Variación año corrido % | | |
| 549 | Año(aaaa)-Mes(mm) | Bogotá, D.C. | Cali | Medellín |
| 550 | 1979-01 | 3,70 | 2,30 | 3,30 |
| 551 | 1979-02 | 5,70 | 3,60 | 4,50 |
| 552 | 1979-03 | 10,30 | 8,10 | 8,10 |
| 553 | 1979-04 | 12,40 | 9,80 | 10,10 |
| 554 | 1979-05 | 14,80 | 11,40 | 13,70 |
| 555 | 1979-06 | 17,00 | 12,30 | 15,70 |

Figure 1: Excel del IPC

En la figura 1 se muestra inicialmente el IPC general, y al ocultar ciertas filas, se revela que más abajo se encuentran los datos correspondientes al IPC del año corrido. Por consiguiente, la obtención de estos datos requirió la selección manual de las filas pertinentes. Considerando que el propósito fundamental de este proyecto consiste en analizar exclusivamente la evolución del IPC mensual, se procedió a seleccionar las filas correspondientes a los registros desde la posición 1638 hasta la 2181 del dataset, con el fin de obtener los datos pertinentes para dicho análisis.

### 4.2 Datos meteorológicos

En este caso, los datos obtenidos de NASA Power [10] estaban bien estructurados, pero consistían en índices climáticos diarios, es decir, se calculaban para cada día. Para relacionarlos con los datos del IPC, que se generan mensualmente, se decidió calcular el promedio de cada columna agrupando por mes y año. Además, fue necesario ajustar el formato de la fecha para que coincidiera con el formato del IPC, que es (AAAA-MM). Cada conjunto de datos generado hacia referencia a información una ciudad en especifico.

### 4.3 Relacionar el clima con el IPC

Después de recopilar los datos del IPC y los datos climáticos de cada ciudad, se procedió a cruzar la información por fecha, con el objetivo de identificar posibles relaciones entre los índices climáticos y el IPC en las 23 ciudades de Colombia.

| Fecha | IPC_Neiva | IPC_Año_corr_Neiva | IPC_Año_Neiva | IPC_Mes_Neiva | T2M_MAX_Neiva | T2M_MIN_Neiva | RH2M_Neiva | PRECTOTCORR_Neiva | T2M_Neiva | WS10M_Neiva |
|---|---|---|---|---|---|---|---|---|---|---|
| 2010-01 | 73.17 | 0.58 | 2.69 | 0.58 | 29.160645 | 16.627419 | 66.133871 | 0.308065 | 21.875161 | 2.750000 |
| 2010-02 | 73.73 | 1.36 | 3.1 | 0.77 | 29.632500 | 18.466429 | 67.959643 | 1.274643 | 23.156429 | 2.861786 |
| 2010-03 | 73.79 | 1.44 | 2.58 | 0.08 | 28.434839 | 18.032903 | 73.184839 | 2.984839 | 22.306452 | 2.820968 |
| 2010-04 | 74.08 | 1.84 | 2.68 | 0.39 | 25.628667 | 17.539333 | 82.152667 | 8.171667 | 20.886333 | 2.468667 |
| 2010-05 | 74.36 | 2.22 | 2.89 | 0.37 | 24.181613 | 16.968065 | 86.241613 | 9.566774 | 20.122903 | 2.273226 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 2023-11 | 137.24 | 7.88 | 9.4 | 0.23 | 25.591000 | 17.066000 | 81.062667 | 3.494000 | 20.664667 | 2.322667 |
| 2023-12 | 137.53 | 8.11 | 8.11 | 0.21 | 26.244194 | 17.478710 | 80.604194 | 4.420000 | 21.184839 | 2.511290 |
| 2024-01 | 138.92 | 1.01 | 7.48 | 1.01 | 28.046452 | 17.181935 | 74.010968 | 0.810645 | 21.698387 | 2.687742 |
| 2024-02 | 139.93 | 1.75 | 6.73 | 0.73 | 25.958276 | 17.847586 | 81.676552 | 8.476897 | 21.246897 | 2.631724 |
| 2024-03 | 140.59 | 2.23 | 6.32 | 0.47 | 25.916774 | 18.080000 | 82.194839 | 9.477097 | 21.426129 | 2.784516 |

Figure 2: Dataframe de IPC unido con los datos del clima.

## 5 Limitaciones

Este estudio se centra exclusivamente en el IPC mensual. Esta decisión se tomó con el objetivo de mantener la coherencia temporal y permitir un análisis detallado de las variaciones mensuales de los precios. El enfoque en el IPC mensual ofrece una perspectiva más precisa y puntual de las fluctuaciones económicas a corto plazo. Se trabajara solo con el IPC Mensual.

Sin embargo, esta elección presenta limitaciones significativas; se pierde la perspectiva a largo plazo de la inflación, se limita la comparabilidad entre períodos, y se pierde la oportunidad de integrar estos datos con otros indicadores económicos medidos

en diferentes frecuencias. Además, la disponibilidad de datos mensuales puede no ser consistente para todos los períodos o regiones, lo que puede afectar la continuidad y la validez del análisis.

## 6 Análisis Exploratorio

Antes de comenzar con esta sección, cabe mencionar que los datos anteriormente mencionados fueron seleccionados a partir del año 2010. Se ha verificado que ninguno de ellos contiene valores repetidos o nulos.

### 6.1 Análisis de la distribución de Bogotá

Para simplificar el análisis, inicialmente el enfoque de los datos fue en la ciudad de Bogotá. Sin embargo, se puede modificar para analizar datos de cualquiera de las 22 ciudades de Colombia que cuentan con estos datos. En esta sección y la siguiente, nos centramos en realizar un análisis detallado exclusivamente sobre los datos relacionados con Bogotá.

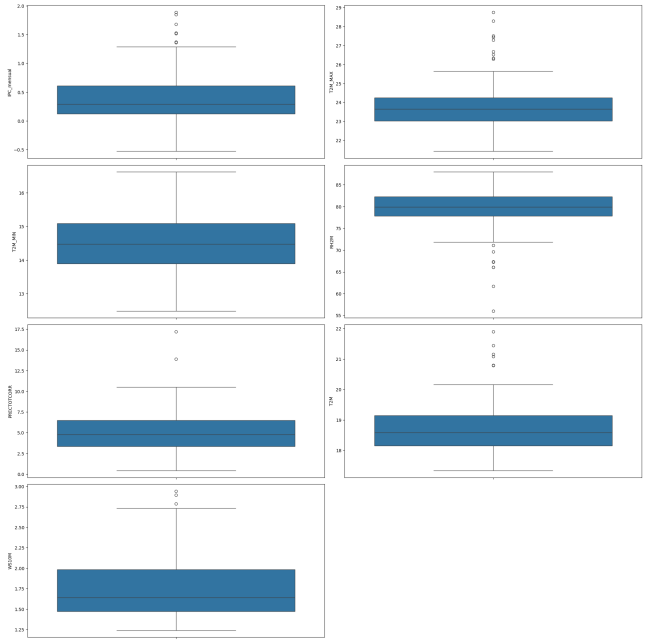


Figure 3: Gráfico BoxPlot de las variables de Bogotá

El gráfico 3 ilustra la distribución de las variables contenidas en el dataset seleccionado para el modelo analítico. Cada variable se representa en un diagrama de caja y bigotes, también conocido como diagrama de caja. Este tipo de gráfico es útil para visualizar la dispersión y la forma de la distribución de los datos, mostrando la mediana, los cuartiles y los valores atípicos de cada variable. En el caso de este estudio, el gráfico proporciona una visión general de la variabilidad de las variables, lo que puede ayudar a identificar patrones o anomalías en los datos.

- **IPC Mensual**: El rango intercuartílico es de alrededor de 1, lo que significa que la mitad central de los meses tuvo un IPC dentro de 1 unidad alrededor de la mediana. La distribución muestra una ligera asimetría hacia la derecha, sugiriendo que hay más meses con IPC por encima de la mediana que por debajo. Se identifican algunos valores atipicos, pocos a comparación de las demás variables.

- **T2M_MAX**: La distribución de la variable T2M_MAX muestra varios picos distintos de temperatura, lo cual es coherente dada la variabilidad de las temperaturas en Bogotá, donde ocasionalmente se registran valores elevados.

- **T2M_MIN**: En esta instancia, la distribución de la variable ronda entre 14 y 15. No se evidencian observaciones atípicas dentro de este intervalo, lo cual se alinea con el comportamiento esperado para las temperaturas mínimas en Bogotá, que generalmente no experimentan descensos abruptos.

- **RH2M**: Esta instancia evidencia una alta cantidad de datos atípicos y dispersos, ya que se encuentran significativamente alejados de los valores centrales representados en la caja.

- **PRECTOTCORR**: En este caso tiene una distribución de datos normal, con unos pocos valores atípicos, acorde con la realidad de precipitaciones en Colombia.

- **T2M**: En esta distribución se puede evidenciar la presencia de valores atípicos, lo cual se considera normal debido a su estrecha relación con las temperaturas mínimas y máximas. Además, estas temperaturas son representativas de los valores usuales en la ciudad especificada.

- **WS10M**: Por último, esta distribución muestra muy pocos valores atípicos, además de una caja considerablemente inclinada hacia abajo, lo que indica que en Bogotá la velocidad del viento suele ser baja.

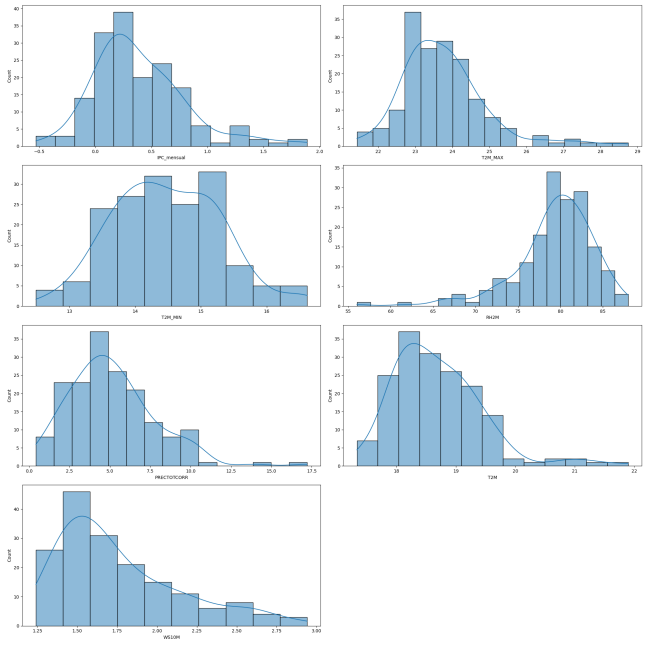


Figure 4: Distribución de las variables de Bogotá

En la figura 4 se presenta una agrupación de histogramas de cada una de las variables del conjunto de datos. Algunas de estas variables muestran similitudes con la distribución Borel-Tanner hacia la izquierda, como es el caso de WS10M, PRECTOTCRR, T2M y T2M_MAX, o hacia la derecha, como RH2M. Sin embargo, en su mayoría, las variables presentan un cumplimiento general de una distribución normal.

## 6.2 Análisis de Correlación

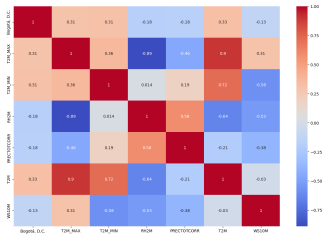


Figure 5: Correlación índices del clima

Un diagrama de correlación es una herramienta visual que muestra la relación lineal entre diferentes variables mediante una matriz de correlación. En este diagrama, cada celda representa el coeficiente de correlación entre dos variables, que puede variar entre -1 y 1. Un valor de 1 indica una correlación positiva perfecta, -1 una correlación negativa perfecta y 0 ninguna correlación.

Como se observa en la Figura 5, las relaciones fuertes y negativas: Hay una fuerte correlación positiva entre T2M_MAX y T2M (0.9) y entre T2M_MIN y T2M (0.72), lo que indica una estrecha relación entre las temperaturas máximas, mínimas y medias. Además, T2M_MAX y WS10M tienen una relación positiva moderada (0.31). Por otro lado, existe una fuerte correlación negativa entre T2M_MAX y RH2M (-0.89), y entre RH2M y T2M (-0.64), lo que sugiere que a mayores temperaturas, la humedad relativa tiende a ser menor.

## 6.3 Serie temporal IPC

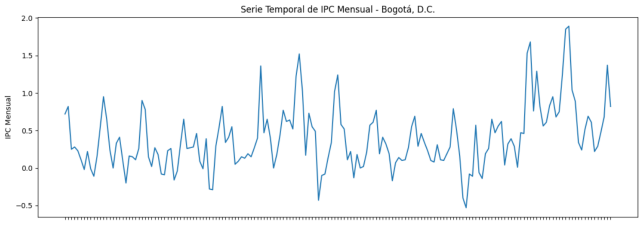



Figure 6: Serie temporal del IPC desde 2010 hasta 2023

El gráfico de la Figura 6 es una serie temporal que muestra la evolución del IPC mensual para Bogotá, a lo largo del tiempo. La serie temporal es una representación visual de los datos de IPC mensual, que puede ayudar a identificar patrones, tendencias y posibles estacionalidades en el comportamiento del IPC. La serie temporal muestra fluctuaciones frecuentes, indicando que el IPC mensual varía considerablemente de un mes a otro. La serie temporal muestra fluctuaciones frecuentes, indicando que el IPC mensual varía considerablemente de un mes a otro. Además, se identifican picos sobresalientes se observan alrededor de la mitad y hacia el final del período analizado, alcanzando valores cercanos a 1.5 y 2.0.

## 6.4 Análisis de Agrupamiento con DBSCAN

Para analizar si los datos siguen alguna agrupación en especial, se optó por utilizar el algoritmo DBSCAN (Density-Based Spatial Clustering of Applications with Noise) para el análisis de los datos debido a su capacidad para identificar grupos. DBSCAN es especialmente útil para descubrir clusters de formas arbitrarias en datasets con ruido [13], lo que lo hace ideal para este tipo de análisis. En este caso, se aplicó DBSCAN con el fin de encontrar relaciones entre las variables climáticas (como la temperatura y la precipitación) y el IPC en la ciudad de Bogotá. La capacidad de DBSCAN para manejar ruido y detectar patrones complejos sin requerir un número predeterminado de clusters es particularmente ventajosa para analizar estos datos multifacéticos .

### 6.4.1 Encontrar los hiper-parámetros

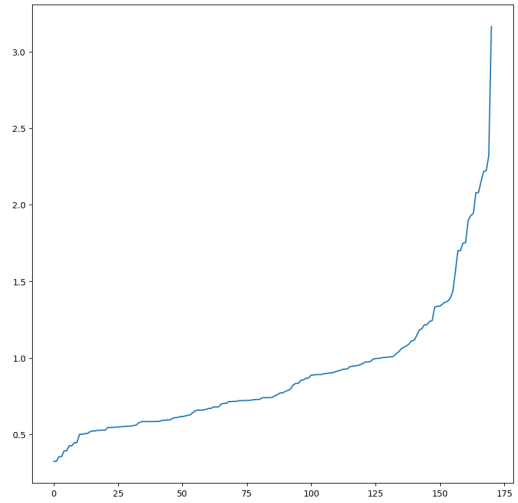


Figure 7: Distancia ordenadas de todos los puntos

Se usó el algoritmo de Vecinos Más Cercanos para calcular las distancias a los dos vecinos más cercanos para cada punto en el conjunto de datos previamente escalado. Posteriormente, se ordenaron estas distancias de manera ascendente y selecciona las distancias al segundo vecino más cercano para cada punto. En general, las distancias tienden a aumentar a medida que avanzamos en el conjunto de datos, indicando que los puntos de datos son menos similares entre sí en regiones más alejadas. La pendiente inicial pronunciada sugiere una mayor variabilidad entre los puntos de datos iniciales, lo que puede deberse a una mayor heterogeneidad en esos puntos. Además, la distribución relativamente uniforme de las distancias implica que los puntos de datos están bien distribuidos en el espacio de características.

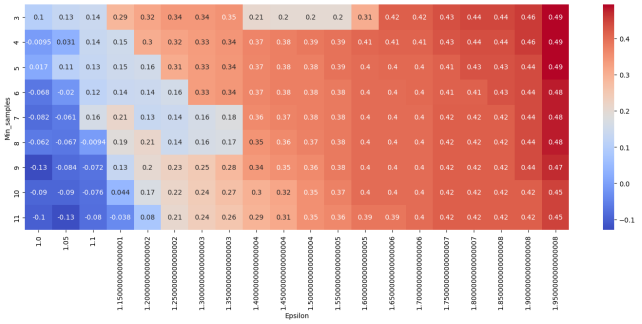



Figure 8: Score de silueta

Se buscó explorar de manera sistemática diferentes configuraciones de dos parámetros clave, epsilon y min_samples, para el algoritmo de clustering DBSCAN. Estos parámetros son cruciales para determinar cómo se agrupan los datos en clusters. Para cada configuración de parámetros evaluada, se ajusta el modelo DBSCAN al conjunto de datos de interés y se calcula el número de clusters resultantes, así como el score de silueta, que es una medida

de la coherencia de los clusters. Gracias a la imagen 8 se puede identificar fácilmente qué combinaciones de épsilon y min_samples, los cuales son 1.65 para epsilon y min_samples de 3.

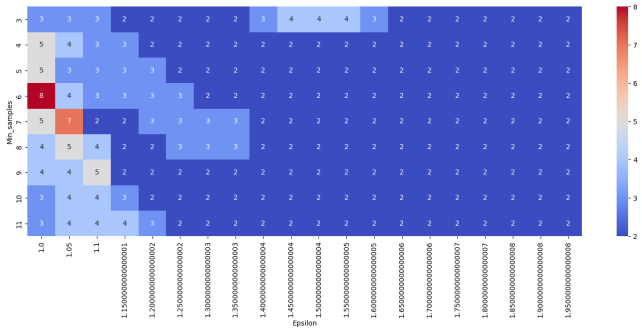



Figure 9: Clusters encontrados DBSCAN

En la imagen 9 presenta una visualización gráfica de la cantidad de clusters obtenidos por el algoritmo DBSCAN para distintas configuraciones de los parámetros épsilon y min_samples. Se utiliza un mapa de calor, para representar esta información de manera intuitiva.

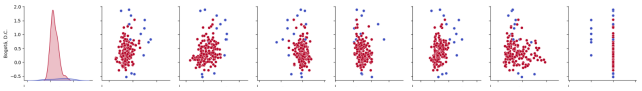


Figure 10: Variables con Clusters DBSCAN en Bogotá

Se procede a implementar el algoritmo DBSCAN con los parámetros óptimos determinados previamente, siendo estos un valor de épsilon de 1.65 y un mínimo de min_samples de 3. Una vez ajustado el modelo al conjunto de datos normalizado de Bogotá, se obtienen las etiquetas de cluster para con respecto a cada punto, en este caso se ve esta relación de los índices del clima con el IPC.

### 6.4.2 Con PCA

El Análisis de Componentes Principales (PCA) es una técnica estadística utilizada para reducir la dimensionalidad de conjuntos de datos complejos, conservando al mismo tiempo la mayor cantidad posible de información. PCA logra esto transformando las variables originales en un nuevo conjunto de variables no correlacionadas llamadas componentes principales. Estos componentes principales están ordenados de tal manera que el primero captura la mayor variabilidad presente en los datos, seguido por el segundo componente, y así sucesivamente.

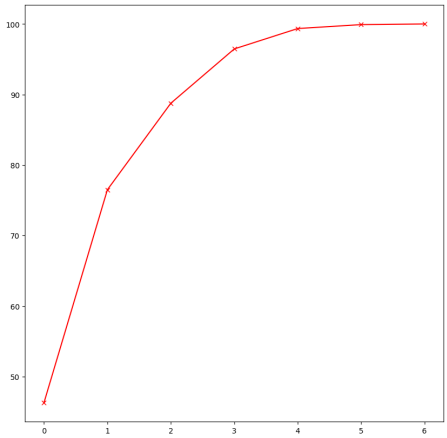


Figure 11: varianza explicada acumulada

El cálculo de la varianza explicada por cada componente principal, expresada como un porcentaje de la varianza total, es crucial en la evaluación de la efectividad de la reducción de la dimensionalidad realizada por PCA. Esta información es especialmente relevante en la determinación del número óptimo de componentes principales a retener, ya que permite identificar el punto en el que se conserva una cantidad significativa de la estructura original de los datos. Gracias al gráfico se puede evidenciar que los componentes principales necesarios para llegar a una varianza de 85% es de 2. Luego de eliminar los componentes menos importantes dio estos resultados:

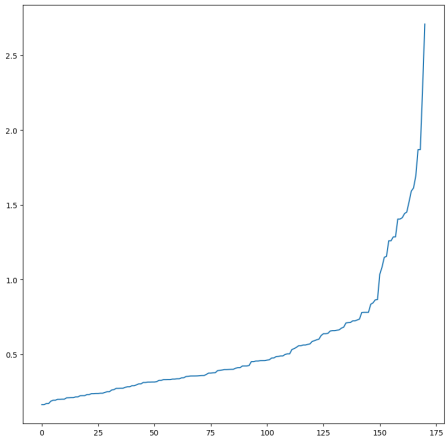


Figure 12: Distancia ordenadas de todos los puntos (PCA)

En el contexto del cálculo de los vecinos más cercanos, se observó que los resultados fueron comparables entre el método de vecinos más cercanos y el algoritmo DBSCAN sin la reducción de dimensionalidad mediante PCA.

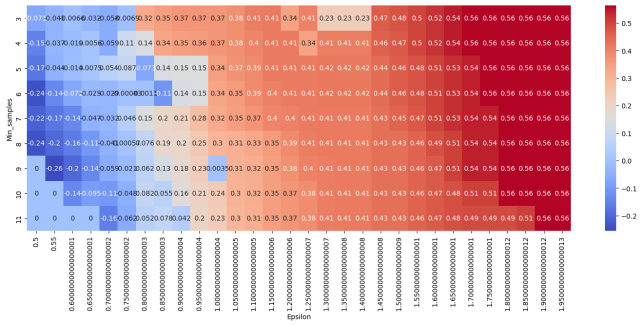



Figure 13: Score de silueta (PCA)

En el presente gráfico se observa que los parámetros del clúster, con la aplicación de PCA, han sido modificados a un valor de epsilon de 1.7 y un número mínimo de muestras de 3. Este resultado sugiere que no se ha producido un cambio drástico en los hiperparámetros.

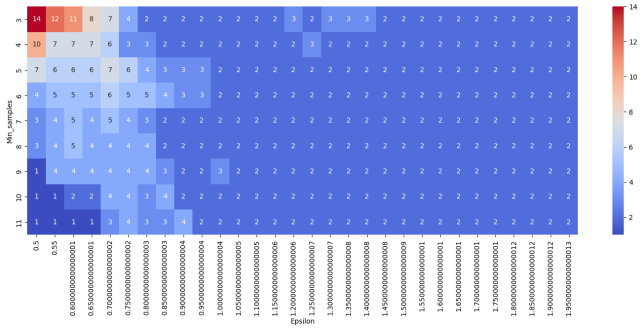



Figure 14: Clusters encontrados DBSCAN (PCA)

En comparación con la figura 9 se repitió el patrón de encontrar numerosos clusters de 2 agrupaciones, incluso se generó uno de 14.

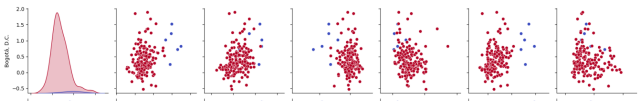


Figure 15: Variables con Clusters DBSCAN en Bogotá

Tras un análisis exhaustivo de las variaciones en la agrupación realizada por el algoritmo DBSCAN, se ha llegado a la conclusión de que, para el contexto del proyecto en cuestión, no resulta idóneo para la clasificación del IPC en relación con los índices climáticos en Bogotá. A pesar de presentar un coeficiente de silueta elevado, no aporta un valor significativo a la estructura de agrupación de los datos. No obstante, su implementación podría resultar útil en la detección de valores atípicos (outliers) en este ejercicio.

# 7 Modelo Predictivo

## 7.1 Random Forest (Todas las características)

La implementación de Random Forest para el análisis de datos climáticos e IPC implica varios pasos clave. En primer lugar, se definen las características del modelo, que incluyen las variables climáticas definidas en la sección 3.2. La variable objetivo se establece como la categorización del IPC para Bogotá, que se logra mediante la cuantificación de los valores de IPC en tres grupos Bajo, Medio y Alto (0, 1 y 2) utilizando cuartiles, buscando que haya una muestra balanceada de cada clase. Luego, teniendo en cuenta que el conjunto de datos se encuentra balanceado, los datos se dividen en conjuntos de entrenamiento y prueba.

### 7.1.1 Encontrar los mejores hiperparámetros con validación cruzada

En la búsqueda de hiperparámetros para ajustar el modelo de Random Forest, se hicieron los siguientes pasos: En primer lugar, se define un conjunto de hiper parámetros que se van a ajustar, como el criterio de división de los árboles, el número de árboles en el bosque, la profundidad máxima de los árboles, etc. Luego, se utiliza la función *GridSearchCV* de la biblioteca scikit-learn [12] para realizar una búsqueda exhaustiva de los mejores hiper parámetros dentro del conjunto definido, utilizando validación cruzada para evaluar el rendimiento del modelo en diferentes subconjuntos de datos.

Este mejor modelo se utiliza entonces para hacer predicciones sobre un conjunto de datos de prueba (‘X_test‘) y se evalúa su rendimiento utilizando métricas como la precisión (‘accuracy‘) y el informe de clasificación (‘classification_report‘), definiendo los siguientes hiperparametros:

- **Criterion**: Criterio para medir la calidad de una división - ”entropy”.
- **n_estimators**: Número de árboles en el bosque - 100.
- **max_depth**: Profundidad máxima de los árboles - 10.
- **max_samples**: Porcentaje de muestras a considerar - 0.5.
- **min_samples_split**: Número mínimo de muestras requeridas para dividir un nodo - 2

### 7.1.2 Evaluación del modelo

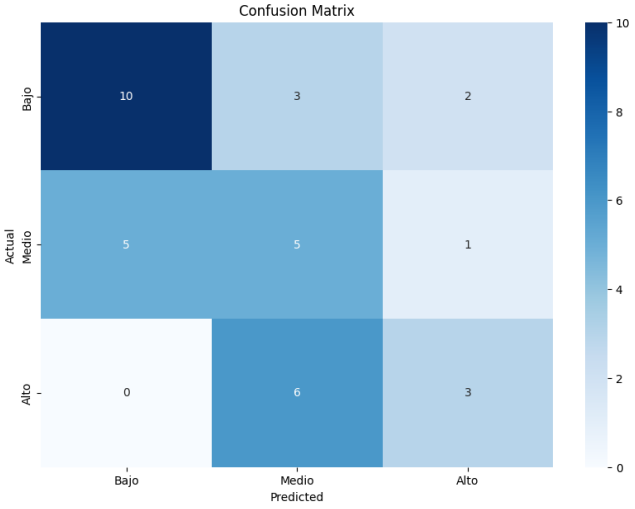



Figure 16: Matriz de confusión

El modelo exhibe un rendimiento relativamente sólido en la predicción de la clase ’Bajo’, con diez instancias correctamente clasificadas. No obstante, también se observaron cinco instancias que fueron erróneamente clasificadas como ’Medio’ cuando en realidad pertenecían a la clase ’Bajo’. Este fenómeno sugiere que el modelo podría estar experimentando cierta confusión entre las clases ’Bajo’ y ’Medio’. En lo que respecta a la clase ’Medio’,

el modelo logró predecir correctamente cinco instancias, aunque también clasificó incorrectamente tres instancia como 'Bajo' y seis instancias como 'Alto'. Estos resultados apuntan a posibles dificultades del modelo para distinguir entre las clases 'Medio' y 'Alto', así como entre 'Medio' y 'Bajo'. Finalmente, en el caso de la clase 'Alto', el modelo acertó en la predicción de seis instancias, pero también clasificó de manera incorrecta tres instancias como 'Medio'. Esta situación indica una posible confusión entre las clases 'Alto' y 'Medio' por parte del modelo.

Si bien el modelo muestra un rendimiento aceptable en la predicción de la clase 'Bajo', con una cantidad significativa de instancias correctamente clasificadas, presenta dificultades para distinguir entre las clases 'Medio' y 'Alto', así como entre 'Bajo' y 'Medio'. Esto sugiere que la matriz de confusión refleja un desempeño mixto del modelo, con áreas de mejora en la precisión de la clasificación, especialmente entre las clases mencionadas.

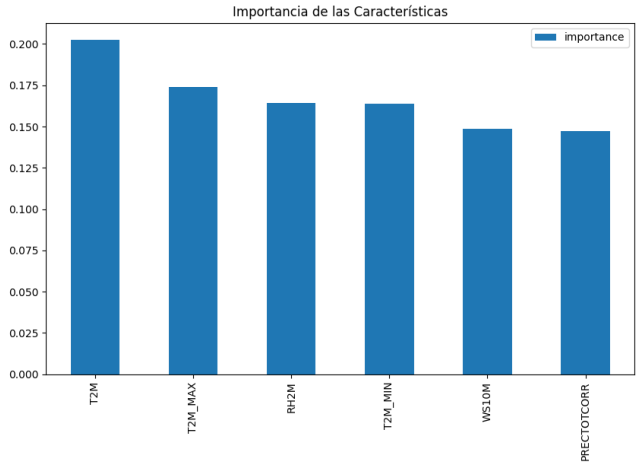



Figure 17: Características importantes

El análisis de características revela que la temperatura media diaria (T2M) es el atributo más influyente en el modelo, con una ponderación de 0.202574. Le siguen en importancia la temperatura máxima diaria (T2M_MAX) con 0.173787 y la humedad relativa (RH2M) con 0.164068. Estos resultados sugieren que las variables relacionadas con la temperatura y la humedad son determinantes en la predicción realizada del IPC. Otros factores evaluados incluyen la temperatura mínima diaria (T2M_MIN), la velocidad del viento a 10 metros sobre la superficie del suelo (WS10M), y la precipitación total corregida (PRECTOTCORR), con ponderaciones de 0.163801, 0.148402, y 0.147367 respectivamente. Estos hallazgos indican la relevancia de las condiciones climáticas en el proceso predictivo del modelo, subrayando la importancia de la temperatura y la humedad en particular.

En ultima instancia, el modelo de árbol de decisión implementado para analizar si las condiciones climáticas influyen en el IPC en Bogotá ha mostrado una precisión del 54.07%. Este nivel de precisión sugiere que, aunque existe cierta relación entre las variables climáticas y el IPC, el modelo actual tiene una capacidad limitada para realizar predicciones exactas. Estos resultados iniciales indican la necesidad de realizar mejoras en el modelo, tales como la optimización de los parámetros, como seleccionar otros índices climáticos, o la exploración de otros algoritmos de aprendizaje automático. Continuar refinando este enfoque podría proporcionar una comprensión más profunda y precisa de cómo las variaciones climáticas afectan el IPC en Bogotá.

## 7.2 Random Forest (Sin todas las características)

Como se observa en la correlación de las características en le Figura 5, las características de temperatura T2M_MIN, T2M_MIN están con altas correlaciones positivamente con T2M, por ende se decide dejar solamente la variable T2M como representante de la temperatura. y analizar nuevamente los resultados obtenidos.

### 7.2.1 Evaluación del modelo

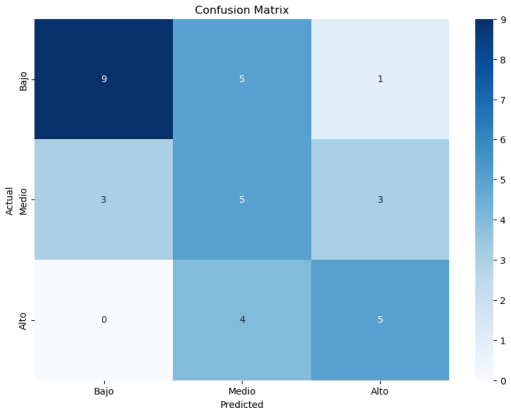



Figure 18: Matriz de Confusión

Como se observa en la Figura 18, al disminuir los datos siguen habiendo datos que no se predicen bien, pero la mayoría de los datos en las tres categorías se están prediciendo bien.

De igual forma la característica relacionada con la temperatura media T2M fue la característica que más importancia tuvo en el modelo, como se ilustra en la Figura 19:

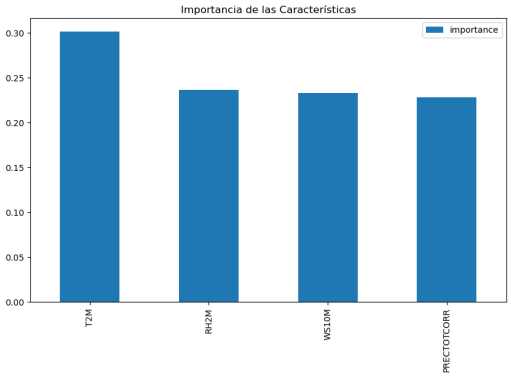



Figure 19: Importancia de Características

EL reporte de las diferentes métricas fue el siguiente:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| **0** | 0.75 | 0.60 | 0.67 | 15 |
| **1** | 0.36 | 0.45 | 0.40 | 11 |
| **2** | 0.56 | 0.56 | 0.56 | 9 |
| **accuracy** | | | 0.54 | 35 |
| **macro avg** | 0.55 | 0.54 | 0.54 | 35 |
| **weighted avg** | 0.58 | 0.54 | 0.55 | 35 |

Table 3: Reporte de métricas de clasificación

1. **Precision (Precisión)**:
    - La precisión mide la proporción de predicciones positivas correctas del modelo en cada clase.
    - Para la clase 0, la precisión es de 0.75, lo que significa que el 75% de las muestras que el modelo predijo como clase 0 son realmente de la clase 0.
    - Para la clase 1, la precisión es de 0.36, lo que indica que solo el 36% de las muestras que el modelo predijo como clase 1 son realmente de la clase 1.
    - Para la clase 2, la precisión es de 0.56, lo que significa que el 56% de las muestras que el modelo predijo como clase 2 son realmente de la clase 2.
2. **Recall (Exhaustividad)**:
    - El recall mide la proporción de muestras positivas reales que el modelo identificó correctamente en cada clase.
    - Para la clase 0, el recall es de 0.60, lo que significa que el modelo identificó correctamente el 60% de las muestras reales de la clase 0.
    - Para la clase 1, el recall es de 0.45, lo que indica que el modelo identificó correctamente el 45% de las muestras reales de la clase 1.
    - Para la clase 2, el recall es de 0.56, lo que significa que el modelo identificó correctamente el 56% de las muestras reales de la clase 2.
3. **F1-score (Puntaje F1)**:
    - El F1-score es la media armónica de la precisión y el recall, y proporciona un equilibrio entre ambas métricas.
    - Para la clase 0, el F1-score es de 0.67, lo que indica un buen equilibrio entre precisión y recall para esta clase.
    - Para la clase 1, el F1-score es de 0.40, lo que sugiere un rendimiento relativamente bajo en términos de precisión y recall para esta clase.
    - Para la clase 2, el F1-score es de 0.56, lo que indica un rendimiento moderado en términos de precisión y recall para esta clase.
4. **Support (Soporte)**:
    - El soporte indica el número de muestras reales en cada clase.
    - Hay 15 muestras de la clase 0, 11 muestras de la clase 1 y 9 muestras de la clase 2 en el conjunto de datos de prueba.
5. **Accuracy (Exactitud)**:
    - La exactitud mide la proporción de predicciones correctas del modelo en general.
    - La exactitud general del modelo es de 0.54, lo que significa que el modelo predijo correctamente el 54% de las muestras en el conjunto de datos de prueba.
6. **Macro Average (Promedio macro)**:
    - El promedio macro calcula el promedio no ponderado de las métricas (precisión, recall, F1-score) para todas las clases.
    - El promedio macro de la precisión es de 0.55, el promedio macro del recall es de 0.54 y el promedio macro del F1-score es de 0.54.
    - Estas métricas proporcionan una visión general del rendimiento del modelo en todas las clases, sin tener en cuenta el desequilibrio de clases.
7. Weighted Average (Promedio ponderado):
    - El promedio ponderado calcula el promedio ponderado de las métricas (precisión, recall, F1-score) para todas las clases, teniendo en cuenta el n´umero de muestras en cada clase.
    - El promedio ponderado de la precisión es de 0.58, el promedio ponderado del recall es de 0.54 y el promedio ponderado del F1-score es de 0.55.
    - Estas métricas proporcionan una visión general del rendimiento del modelo, considerando el desequilibrio de clases.

## 7.3 Regresión Logística

De la misma forma sin tomar todas las características se entreno el modelo de clasificación Regresión Logística.

### 7.3.1 Evaluación del modelo

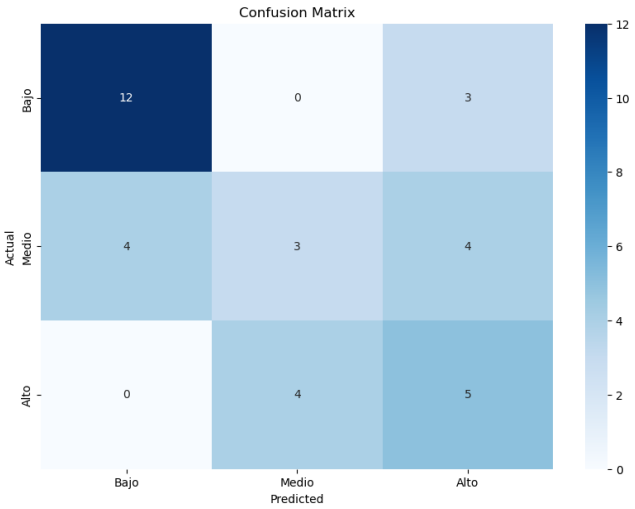



Figure 4: Matriz de Confusión

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.75 | 0.80 | 0.77 | 15 |
| 1 | 0.43 | 0.27 | 0.33 | 11 |
| 2 | 0.42 | 0.56 | 0.48 | 9 |
| accuracy | | | 0.57 | 35 |
| macro avg | 0.53 | 0.54 | 0.53 | 35 |
| weighted avg | 0.56 | 0.57 | 0.56 | 35 |

Table 4: Métricas de clasificación

1. **Precision (Precisión)**:
    - Para la clase 0, la precisión es de 0.75, lo que significa que el 75% de las muestras que el modelo predijo como clase 0 son realmente de la clase 0.
    - Para la clase 1, la precisión es de 0.43, lo que indica que el 43% de las muestras que el modelo predijo como clase 1 son realmente de la clase 1.

- Para la clase 2, la precisión es de 0.42, lo que significa que el 42% de las muestras que el modelo predijo como clase 2 son realmente de la clase 2.

2. **Recall (Exhaustividad)**:

   - Para la clase 0, el recall es de 0.80, lo que significa que el modelo identificó correctamente el 80% de las muestras reales de la clase 0.
   - Para la clase 1, el recall es de 0.27, lo que indica que el modelo identificó correctamente el 27% de las muestras reales de la clase 1.
   - Para la clase 2, el recall es de 0.56, lo que significa que el modelo identificó correctamente el 56% de las muestras reales de la clase 2.

3. **F1-score (Puntaje F1)**:

   - Para la clase 0, el F1-score es de 0.77, lo que indica un buen equilibrio entre precisión y recall para esta clase.
   - Para la clase 1, el F1-score es de 0.33, lo que sugiere un rendimiento relativamente bajo en t´erminos de precisión y recall para esta clase.
   - Para la clase 2, el F1-score es de 0.48, lo que indica un rendimiento moderado en términos de precisión y recall para esta clase.

4. **Support (Soporte)**:

   - Hay 15 muestras de la clase 0, 11 muestras de la clase 1 y 9 muestras de la clase 2 en el conjunto de datos de prueba.

5. **Accuracy (Exactitud)**:

   - La exactitud general del modelo es de 0.57, lo que significa que el modelo predijo correctamente el 57% de las muestras en el conjunto de datos de prueba.

6. **Macro Average (Promedio macro)**:

   - El promedio macro de la precisión es de 0.53, el promedio macro del recall es de 0.54 y el promedio macro del F1-score es de 0.53.

7. **Weighted Average (Promedio ponderado)**:

   - El promedio ponderado de la precisión es de 0.56, el promedio ponderado del recall es de 0.57 y el promedio ponderado del F1-score es de 0.56.
   - Estas métricas proporcionan una visión general del rendimiento del modelo, considerando el desequilibrio de clases.

## 8 Conclusión

Tras analizar los resultados del reporte de clasificación de los modelos Random Forest y Regresión Logística, se pueden extraer varias conclusiones sobre la capacidad de estos enfoques para analizar la relación entre las condiciones climáticas y el IPC en Bogotá.

1. En cuanto a la Hipótesis:

   - La hipótesis propone que las variaciones en las condiciones climáticas, específicamente la temperatura y la humedad, tienen un impacto significativo en el Índice de Precios al Consumidor (IPC) en las diferentes ciudades de Colombia. Se esperaba que un incremento en la temperatura estuviera asociado con un aumento en los precios de productos perecederos, debido a una mayor tasa de descomposición y a los costos adicionales de almacenamiento en frío necesarios para preservar estos productos. Además, se esperaba que altos niveles de humedad afectaran negativamente la calidad y la disponibilidad de productos agrícolas, reflejándose en un aumento de sus precios en el mercado.
   - Sin embargo, los resultados obtenidos de los modelos Random Forest y Regresión Logística sugieren que, si bien existe cierta relación entre las variables climáticas y el IPC en Bogotá, la capacidad de estos modelos para realizar predicciones precisas es limitada. El modelo Random Forest tiene una exactitud general del 54%, mientras que el modelo de Regresión Logística alcanza una exactitud del 57%. Aunque estos resultados indican que las variables climáticas tienen cierta influencia en el IPC, no se puede afirmar de manera concluyente que sean predictores robustos y confiables.
   - Además, el análisis de los reportes de clasificación de ambos modelos revela que el rendimiento varía según las diferentes clases del IPC. En general, la clase 0 tiene el mejor rendimiento en términos de precisión y F1-score, mientras que las clases 1 y 2 tienen un rendimiento inferior. Esto sugiere que la relación entre las variables climáticas y el IPC puede ser más compleja y no lineal, y que otros factores no considerados en el análisis también podrían influir en el comportamiento del IPC. Por otro lado, el algoritmo DBSCAN, a pesar de su alto coeficiente de silueta, no contribuye significativamente a la estructura de agrupación de los datos en este contexto específico.

2. En cuanto a los modelos Random Forest y Regresión Logística:

   - Ambos modelos muestran resultados similares en términos de exactitud general, con una precisión del 54% para Random Forest y del 57% para Regresión Logística.
   - Los reportes de clasificación de ambos modelos proporcionan información adicional sobre el rendimiento en cada clase. La clase 0 tiene el mejor rendimiento en términos de precisión y F1-score en ambos modelos. Las clases 1 y 2 tienen un rendimiento inferior, con precisiones y recalls más bajos.
   - Estos resultados sugieren que tanto el modelo Random Forest como el modelo de Regresión Logística pueden capturar mejor la relación entre las variables climáticas y el IPC para ciertas clases, pero aún presentan limitaciones en otras.

3. En cuanto al algoritmo DBSCAN:

   - A pesar de su alto coeficiente de silueta, no contribuye significativamente a la estructura de agrupación de los datos en este contexto específico.
   - Sin embargo, su implementación podría ser útil en la detección de valores atípicos en el conjunto de datos.

Teniendo en cuenta estos resultados, se podría:

- Continuar refinando tanto el modelo de árbol de decisión, el modelo Random Forest y el modelo de Regresión Logística.

- Para mejorar el rendimiento, se pueden explorar técnicas como el ajuste de hiperparámetros, la selección de características más relevantes y el equilibrio de clases.

- Considerar el uso de otros algoritmos de clasificación o enfoques de ensemble para obtener resultados más precisos.

- En paralelo, el uso selectivo de DBSCAN para la identificación de outliers puede complementar el proceso de análisis y ayudar a detectar posibles anomalías en los datos que puedan afectar el rendimiento de los modelos.

## REFERENCES


[1] FAO. (2021). El estado de la seguridad alimentaria y la nutrición en el mundo 2021. Organización de las Naciones Unidas para la Alimentación y la Agricultura.

[2] IDEAM. (2013). Efectos del cambio climático en la producción y rendimiento de cultivos por sectores: Evaluación del riesgo agroclimático por sectores. Marzo de 2013. Páginas 29-32. Autora: Mery Esperanza Fernández.

[3] Costa, C. (2007). Adaptación al cambio climático en Colombia: retos y oportunidades. Revista de Ingeniería.

[4] Ebtehaj, I. (2024). Application of artificial intelligence in environmental, agriculture and earth sciences. Front. Earth Sci., 12:1382457. doi: 10.3389/feart.2024.1382457.

[5] Kulyal, M., Saxena, P. (2024). Machine Learning approaches for Crop Yield Prediction: A Review. Department of Computer Science, Soban Singh Jeena University, Almora.

[6] Chen, L., Han, B., Wang, X., Zhao, J., Yang, W., Yang, Z. (2023). Machine Learning Methods in Climate Prediction: A Survey. Preprints, 202309.1764.v1. doi: 10.20944/preprints202309.1764.v1

[7] Tran, N.-Q., Nguyen, T. N. N., Tran, Q., Felipe, A., Huynh, T., Tang, A., Nguyen, T. (2023). Predicting Agricultural Commodities Prices with Machine Learning: A Review of Current Research. School of Science, Engineering, and Technology, RMIT University Vietnam.

[8] Banco de la República de Colombia. Índice de Precios al Consumidor (IPC). Recuperado de https://www.banrep.gov.co/es/estadisticas/indice-precios-consumidor-ipc

[9] Geodatos. Recuperado de https://www.geodatos.net/coordenadas/colombia/

[10] NASA Power. Recuperado de https://power.larc.nasa.gov/docs/

[11] DANE. Recuperado de https://www.dane.gov.co/index.php/estadisticas-por-tema/precios-y-costos/indice-de-precios-al-consumidor-ipc/ipc-informacion-tecnica

[12] Scikit-learn. Recuperado de https://scikit-learn.org/stable/index.html

[13] Ester, M., Kriegel, H. P., Sander, J., Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96).