Existen dos razones principales por las cuales se estima : i) inferencia y ii) predicción. Hasta este punto, en este documento se ha discutido el primer punto (la inferencia). Ahora, se trabajará sobre la predicción de los ingresos a partir de métodos de remuestreo con distintas especificaciones.

En primer lugar, es importante motivar la necesidad de un modelo de predicción de ingresos en Bogotá. En el 2022, se estimó para Colombia un índice Gini de 0,4609 (DANE, 2022) lo cual expone la importante diferencia de ingresos entre los ciudadanos del país, lo cual no es heterogéneo en la región latinoamericana. Recientemente, se han realizado esfuerzos de política pública para contribuir en una mejor distribución de los ingresos en el país, siendo la política fiscal una de las que tienen mayor impacto distributivo. Un estudio realizado por la CEPAL en 2014, mostró que los impuestos directos y las transferencias públicas monetarias tenían un impacto de hasta tres puntos porcentuales en el coeficiente Gini en Latinoamérica y 17 puntos porcentuales para países de la OCDE (CPELA/IEF, 2014). El mayor impacto se genera a través de impuestos a la renta personal, los cuales permiten corregir las dinámicas de concentración del ingreso. Sin embargo, esta diferencia entre los países de la OCDE y los países latinoamericanos radica en los bajos ingresos tributarios y la baja capacidad estatal para recolectar los tributos. La evasión de impuestos y los beneficios para los contribuyentes de rentas más altas, generan que incluso los impuestos sean regresivos, es decir, que tengan mayor impacto sobre los individuos en los deciles de menores ingresos (CEPAL, 2016). En Colombia, por ejemplo, la evasión se estima entre 50 y 80 billones de pesos.

De acuerdo con los expertos, aquellos impuestos que tienen un mayor impacto distributivo son los impuestos directos, dentro de los cuales resalta el impuesto sobre la renta (Gómez Sabaíni, Jiménez y Podestá, 2010). Sin embargo, como se mencionó anteriormente, los altos niveles de evasión generan que este impuesto sea uno de los más difíciles de implementar y recolectar en los países de la región. Según Fedesarrollo (2022), la evasión del impuesto sobre la renta de personas naturales en Colombia lleva a la pérdida de ingresos que constituyen 0,7% del PIB. Así las cosas, se han buscado soluciones para solventar el problema de la evasión.

Dentro de las propuestas más comunes para corregir esta falta de contribución, es implementar modelos de predicción de ingresos para evitar los ingresos no declarados de los individuos. En particular, las instituciones podrían adoptar metodologías basadas en el uso de técnicas de Machine Learning para complementar los modelos estadísticos. En este caso, los modelos se pueden entrenar con una muestra y una serie de variables explicativas y probar su poder predictivo con otra parte de la muestra. Estas técnicas han demostrado un poder predictivo mayor a los modelos estadísticos comunes (Shin et al., 2021). En esta sección, se evaluará el poder predictivo de las especificaciones trabajadas hasta el momento y se introducirán nuevas especificaciones.

El algoritmo consiste de dividir la muestra en dos: una muestra de entrenamiento y una muestra de evaluación. En este caso, se realizó la muestra de *training* con 70% de los datos y la de *testing* con 30% de la muestra original. A partir de estas muestras, se utilizaron los modelos propuestos para comparar sus errores de raíz cuadrada media, como indicador del error de predicción al ser la desviación estándar de los valores residuales.

Los modelos utilizados fueron:

A continuación se muestra la tabla con los RMSE:

|  |  |
| --- | --- |
| **Modelo** | **RMSE** |
| Modelo 1 | 0,68287347 |
| Modelo 2 | 0,69597081 |
| Modelo 3 | 0,68291471 |
| Modelo 4 | 0,66057742 |
| Modelo 5 | 0,5677023 |
| Modelo 6 | 0,50933346 |
| Modelo 7 | 0,49864982 |
| Modelo 8 | 0,48748093 |
| Modelo 9 | 0,52230999 |
| Modelo 10 | 0,49651658 |
| Modelo 11 | 0,4979176 |

Como se puede observar, la complejidad de los modelos en algunos casos mejora la capacidad predictiva y en otros disminuye su rendimiento. Al utilizar polinomios