Universidad de San Andrés

Departamento de Economía

Maestría en Economía

Análisis y evolución de la clase media peruana 2004-2021

Luis Villazon Sanchez

118166510

Mentor: Maria EDO

Victoria

05 de Mayo, 2023

Tesis de Maestría en Economía de

Luis VILLAZON SANCHEZ

"*Análisis y evolución de la clase media peruana 2004-2021*"

Resumen

Este estudio examina la identificación y evolución de la clase media en el Perú, utilizando un enfoque multidimensional basado en una amplia gama de variables de bienestar obtenidas de encuestas de hogares realizadas entre 2004 y 2021. Al emplear el análisis de componentes principales (PCA) para construir un índice de bienestar multidimensional, se segmentó la población en grupos ricos, clase media y pobres. El análisis revela mejoras en todas las variables socioeconómicas y una reducción de la brecha en algunas áreas, como el acceso al agua y la electricidad. Sin embargo, persisten importantes desigualdades en ingresos, nivel educativo y acceso a ciertos bienes y servicios. El estudio también identifica a partir de métodos de machine learning, variables clave que replican la clasificación basada en PCA con un 80% de precisión. Estos hallazgos resaltan la necesidad de políticas públicas para mejorar las condiciones de vida y el bienestar en Perú, priorizando el acceso a servicios básicos y reduciendo las disparidades en otras áreas.

Palabras clave: pobreza, clase media, aprendizaje automatico, PCA, análisis de componentes principales

"Analysis and Evolution of the Peruvian Middle Class 2004-2021"

Abstract

This study examines the identification and evolution of the middle class in Peru, using a multidimensional approach based on a wide range of welfare variables obtained from household surveys conducted between 2004 and 2021. Using principal component analysis (PCA) to construct a multidimensional welfare index, the population was segmented into rich, middle class and poor groups. The analysis reveals improvements in all socioeconomic variables and a narrowing of the gap in some areas, such as access to water and electricity. However, significant inequalities persist in income, educational level and access to certain goods and services. The study also identifies from machine learning methods, key variables that replicate the PCA-based classification with 80% accuracy. These findings highlight the need for public policies to improve living conditions and well-being in Peru, prioritizing access to basic services and reducing disparities in other areas.

Keywords: poverty, middle class, machine learning, PCA, principal component analysis

Códigos JEL: [I30, I32, Z13]

1. Introducción:

La literatura económica sugiere que la clase media ayuda a fomentar el desarrollo económico a través de la inversión en capital humano, el consumo y el ahorro. Además, una mayor clase media representa un mayor porcentaje de familias que salen de la pobreza y mejoran sus condiciones de vida, favoreciendo también a la cohesión social y a la estabilidad social y política.

Si bien el rol positivo de la clase media no parece estar en discusión, la definición y caracterización de esta es menos clara y ha sido abordada por la literatura sociológica y económica desde diversos enfoques y criterios. Así, por ejemplo, desde el campo sociológico se han enfatizado definiciones en términos de estratificación del mercado laboral, acumulación de capital humano, así como visiones generales, valores y estilo de vida; mientras que, desde la literatura económica, por razones prácticas, se pone énfasis en el ingreso o gasto de las familias como variables relevantes, concentrándose en la caracterización de los umbrales de ingreso o gasto que delimitan los bordes de los estratos medios de la sociedad.

En el contexto de los países en desarrollo, la definición de la clase media y, por lo tanto, su medición, resulta más compleja debido a la existencia de segmentos sociales que si bien no son considerados pobres, se encuentran en condiciones de vulnerabilidad y alto riesgo de caer en pobreza. Incluso, si fuera posible definir el umbral más bajo que separa a los pobres del resto de la población, establecer un límite superior que separe a la clase media de los ricos puede resultar también complicado.

En las últimas décadas, el Perú se ha caracterizado por tener una economía con tasas de crecimiento relativamente constantes y por una reducción significativa de la pobreza monetaria, que decreció desde un máximo histórico de 59% en 2004 a un mínimo récord de 20% en el 2019. La mejora económica de los hogares ha contribuido a mejorar el bienestar de las familias, sin embargo, muchas otras siguen viviendo en condiciones de vulnerabilidad. Además, dependiendo del criterio con el que se analicen las mejoras en las condiciones de vida, estas adquieren mayor relatividad, lo que indica que otros aspectos relacionados al bienestar habrían registrado modestos avances en relación con los ingresos.

En el Perú, no existe una definición oficial sobre que significa ser clase media. La literatura que ha buscado aproximarse a esta, se ha basado únicamente en medidas alrededor de los ingresos y no ha incorporado otros factores que la literatura señala como relevantes en la definición de esta categoría social. Asimismo, la mayoría de los estudios al respecto, no han tenido como objetivo realizar un ejercicio de identificación riguroso, pues han tenido el foco en el análisis macro de los factores que habrían impulsado la reducción de la pobreza y por consiguiente de la clase media. Además, la literatura existente no ha abarcado el periodo reciente de menor crecimiento debido a la desaceleración económica y crisis/recuperación por el COVID 19.

En ese sentido, este documento representa el primer ejercicio de identificación multidimensional y caracterización de la clase media para el Perú, abarcando el periodo 2004-2021, de crecimiento económico y desaceleración reciente. Basándose en una definición multidimensional de la clase media que toma en cuenta aspectos más allá de los ingresos, y buscando conciliar los diversos enfoques que se han aproximado a esta, tiene como objetivo identificar la clase media en el contexto de un país en desarrollo como el Perú, utilizando una gran cantidad de datos de variables de bienestar provenientes de encuestas de hogares durante los años 2004-2021.

Este documento contribuye a la literatura y se diferencia de estudios anteriores a partir de tres innovaciones. En primer lugar, integra el enfoque de vulnerabilidad a la pobreza (relevante para países en desarrollo) junto con una identificación multidimensional multivariada de la clase media. En segundo lugar, es la única investigación que analiza la clase media peruana de manera multidimensional abarcando un periodo tan largo y reciente (2004-2021) a partir de microdatos de encuestas de hogares. En tercer lugar, se presenta un ejercicio de selección de variables utilizando métodos de big data, los cuales no han sido utilizados anteriormente en la literatura sobre clase media. Esto permite facilitar la operatividad de este ejercicio y adaptarlo a contextos donde puede haber escasez de datos.

2. Revisión de literatura:

2.1 Definición e identificación de la clase media:

La definición e identificación de la clase media ha sido abordada por la literatura sociológica y económica a partir diversos enfoques y criterios. Desde el campo sociológico, se han enfatizado definiciones en términos de estratificación del mercado laboral, acumulación de capital humano, valores y estilo de vida (Weber, 1946, 1978; Goldthorpe y McKnight, 2006), mientras que, desde el campo económico, se ha puesto énfasis en el ingreso o gasto de las familias.

Debido a la mayor disponibilidad de datos de hogares y mejor procesamiento, los economistas han trasladado la definición de la clase media al dominio de los datos. Así, la literatura económica rara vez ha pretendido definir la clase media en términos de su composición profesional, características educativas o el sistema de valores y creencias. La mayoría de los estudios han optado por una definición basada en los ingresos, utilizando principalmente definiciones relativas que abordan un estrato de la distribución del ingreso en lugar de realizar un análisis más comprehensivo de clase.

En un sentido práctico, utilizar el ingreso como variable de identificación de la clase media proporciona un indicador natural en una sola dimensión, facilitando la localización de un “grupo medio”. A grandes rasgos, este es el procedimiento que ha seguido la literatura económica que se basa en definiciones relativas de la clase media, aunque difiriendo en definir cuáles son los umbrales escogidos que definen este segmento medio. Un primer grupo de estudios define los umbrales en relación con el ingreso medio de la distribución. Así, por ejemplo, el umbral inferior es definido por una porción del ingreso medio, mientras que el umbral superior por un múltiplo del ingreso (Blackburn y Bloom, 1985; Davis y Huston, 1992; Birdsall, Graham y Pettinato,

2000; Castellani y Parent, 2011; Rasch, 2017). Otro conjunto de estudios sitúa los umbrales no en el espacio de los ingresos sino en el de los rangos o posiciones en su distribución (Alesina y Perotti, 1996; Easterly, 2001; Solimano, 2008; Nissanov y Pitau, 2016). En este enfoque, el tamaño de la clase media (en términos de población) está fijado naturalmente por la definición y se cuantifica el porcentaje del total de los ingresos apropiado por este grupo.

Una limitación de las definiciones relativas es que, al diferir las distribuciones del ingreso entre países o regiones, se comparan diferentes clases medias de un lugar a otro. En contraste, un enfoque absoluto identifica a la clase media utilizando un rango específico y comparable de ingresos o consumos al usar una gama específica de dólares estandarizados. La pregunta fundamental es cómo definir ese nivel absoluto, pues hasta ahora, estos parecen haber sido fijados de manera algo arbitraria.

Así, por ejemplo, Milanovich y Yitzhaki (2002) dividieron a la población mundial en tres grupos y utilizaron encuestas de hogares para definir a la clase media como aquellos hogares con ingresos per cápita entre los ingresos per cápita promedio de Brasil e Italia. Banerjee y Duflo (2008) definen a la clase media como aquellos hogares que viven con un gasto per cápita de US$2–US$10 al día y analizan los patrones de consumo y empleo de este grupo en 11 países en vías de desarrollo. Estudios más recientes (para Estados Unidos y Rusia, respectivamente), como el de Rose (2020) y Slobodenyuk y Mareeva (2020), utilizan también líneas federales de pobreza, para definir estos umbrales. De manera similar, Ravallion (2010) propuso el concepto de "clase media del mundo en vías de desarrollo", definido como una gama entre el umbral medio de la pobreza de los países en vías de desarrollo y el umbral de la pobreza en Estados Unidos.

2.2 **Definición e identificación de la clase media en países en desarrollo:**

La definición e identificación de la clase media resulta más compleja si se considera que, en el contexto de los países en desarrollo, esta puede ser muy heterogénea entre sí al incluir hogares con diversas trayectorias. En ese sentido, parte de la literatura ha precisado las dificultades de incluir en este grupo a individuos que pueden ser vulnerables a shocks inesperados y que, en ausencia de protección social, pueden caer fácilmente en pobreza. Este enfoque de "vulnerabilidad a la pobreza" argumenta que es importante identificar y distinguir a este grupo de la clase media, en tanto tienen necesidades particulares que no son las mismas de los pobres ni tampoco las de la clase media o alta, especialmente en términos de política social.

En esta línea, resalta el estudio de López-Calva y Ortiz-Juárez (2013), quienes sostienen que la vulnerabilidad a la pobreza facilita identificar a la clase media fijando el umbral inferior a un nivel absoluto. Así, en lugar de elegir un umbral de pobreza específico como el límite inferior de la clase media, los autores buscan un valor de ingresos que corresponde a un requisito mínimo para los funcionamientos que definen el ser de clase media. Según los autores, se puede considerar a la capacidad de tener una alimentación adecuada o de participar significativamente en un conjunto mínimo de actividades sociales como funcionamientos que definen la pobreza, y un umbral de la pobreza como una demarcación en el margen de ingresos de lo que se requiere para alcanzar esos funcionamientos mínimos y salir de la pobreza. De manera análoga, se

podría buscar el conjunto de funcionamientos asociados con la pertenencia a una clase media, y luego intentar cuantificar un nivel de ingreso que permita esto.

Una ventaja de este enfoque es que se acerca al concepto sociológico de "estilo de vida". Aunque López-Calva y Ortiz-Juárez (2013) no definen un vector de bienes de consumo asociados con el estatus de clase media, eligen un "funcionamiento" concreto, la seguridad económica, como la característica que define la clase media (opuesto a la vulnerabilidad de caer en la pobreza).

El anclaje de una definición de clase media a la seguridad económica es conceptualmente atractivo, como es el hecho de que estos autores han aplicado su propuesta a tres países de América Latina específicamente. Así, concretamente, calculan la probabilidad de caer en la pobreza en Chile, México y Perú, según un conjunto de covariantes observadas (indicadores demográficos y del mercado laboral). Los autores sugieren adoptar una probabilidad de 10% de caer en la pobreza como línea divisoria "operativa" entre la seguridad económica y la vulnerabilidad. Toman como referencia, además, a Cruces et al. (2015), quienes encuentran que el 10 % de las personas en América Latina cayó en la pobreza cada año en un período de 15 años. La opción de una probabilidad de 10% de caer en la pobreza en un intervalo de cinco años, además, arroja diferentes umbrales de ingresos para los tres países analizados.

Si bien los enfoques económicos se han basado principalmente en el ingreso o consumo para definir a la clase media, algunos autores han explorado la relación de otras variables con la pertenencia a la clase media, aunque sin incluir estas en la definición de esta. Así, por ejemplo, Davis y Huston (1992) propusieron un análisis multivariado de la clase media investigando las causas de la pertenencia a la clase baja y alta a través de una regresión sobre muchos atributos socioeconómicos. Eisenhauer (2011), propone un umbral de riqueza para distinguir a los ricos de la clase media y un índice de intensidad para medir el grado de riqueza dentro en Italia durante el periodo 2002 – 2004, encontrando que la pertenencia a cierta clase social está vinculada estadísticamente a la edad, género, estado civil, tamaño del hogar, educación, empleo y geografía.

Recientemente, diversos estudios han explorado la relación entre la pertenencia a la clase media utilizado con la tenencia de activos (como proxy de riqueza de los hogares) (Thurlow et al., 2015; Shimeles y Ncube, 2015; Johnston y Abreu; 2016). Si bien estas variables toman en cuenta los activos con los que el hogar cuenta, estas se incluyen en su valor monetario. Asimismo, aunque los documentos mencionados valoran la relevancia de diversos atributos en la definición de la clase media, la identificación de este grupo sigue estando fuertemente vinculada a variables predominantemente económicas, como el ingreso o la riqueza, dejando de lado otros aspectos relevantes como el nivel educativo, la calidad de la vivienda, la calidad del empleo, etc. Es decir, la inclusión de un análisis multidimensional se realiza a posteriori de su definición inicial y clasificación.

Estudios como el de Gigliriano y Mosler (2009), dan un paso más allá al proponer una exploración más completa de cómo debería definirse la clase media y de cómo medir su potencial declive, así como la de otros segmentos cruciales de la sociedad en términos de estatus socioeconómico, cuando se considera más de un atributo relevante. Los autores visualizan la clase media como un porcentaje específicamente

establecido de individuos cuyas dotaciones se hallan en un segmento central de la distribución conjunta de las características socioeconómicas. Si bien este enfoque incorpora diversas dimensiones en la definición de la clase media, metodológicamente no se garantiza que el grupo identificado pertenece a la región central de la distribución multivariada y tampoco proporciona una clasificación que identifique a la clase rica y pobre.

Edo et al. (2021) buscan superar estos problemas proporcionando una definición de la clase media para Argentina (2004-2014) que sea también compatible con la existencia de los pobres o ricos en sentido multidimensional. Para ello, presentan un nuevo enfoque para identificar la clase media basado en cuantiles multivariados. Los autores definen un índice de bienestar unidimensional partiendo de datos multivariados y proyectando los datos del espacio original en una dirección de crecimiento de bienestar creciente a partir de la primera componente principal.

Los cuantiles se definen sobre este índice, estableciendo un umbral inferior y otro superior. Los autores muestran que los cuantiles identifican verdaderamente a la clase media multidimensionalmente y además presentan un nuevo enfoque para reducir la dimensionalidad de bienestar a través de la selección de variables. El documento innova al adoptar un enfoque multidimensional para definir la clase media, incorporando un extenso conjunto de datos que cubre un amplio rango de variables y periodos de tiempo. Metodológicamente, mejora los enfoques anteriores al garantizar que la clasificación de clase media corresponda realmente al núcleo central de la distribución multivariada, redefiniendo así la comprensión de esta segmentación y estableciendo un nuevo estándar en su análisis.

La literatura sobre clase media en el Perú ha seguido principalmente la misma tendencia de la literatura internacional. En los países en desarrollo, incluido el Perú, no existe una definición oficial de lo que es la clase media. La literatura que se ha aproximado al estudio de esta y su evolución ha utilizado principalmente el criterio del ingreso monetario y no se han explorado enfoques multidimensionales.

En el caso específico del Perú, por ejemplo, Jaramillo y Zambrano (2013) han analizado el cambio ocurrido en la porción de la población definida como clase media, durante el periodo 2005 y 2011. Los autores consideran diversos criterios (enfoque relativo, enfoque de seguridad económica, enfoque de estratos socioeconómicos y enfoque de no-pobres no-ricos) para definir distintos umbrales en la distribución del ingreso que identifiquen a la clase media, y encuentran que este segmento abarca al 2011, al 40% - 50% de la población aproximadamente (70% si se incluyen a los vulnerables).

Similares hallazgos obtienen Canavire-Bacarreza et al. (2018) y Castilleja-Vargas y Enciso (2019) quienes utilizan datos a nivel estatal y encuestas de hogares para investigar cómo el crecimiento en diferentes sectores afecta las tasas de pobreza y la clase media. Canavire-Bacarreza et al. (2018) categoriza como clase media a aquellos cuyo nivel de ingresos está por debajo de los cuantiles medios de la distribución, mientras que Castilleja-Vargas y Enciso (2019) categorizan como clase media a aquellos cuyo gasto per cápita del hogar supera la línea de pobreza del ámbito geográfico correspondiente y que está por debajo de diez veces el valor de esta. Los resultados de ambos estudios muestran conclusiones similares, resaltando la importancia del crecimiento del sector manufacturero para el alivio de la pobreza y el crecimiento en el

sector de servicios para fortalecer a la clase media (Canavire-Bacarreza et al., 2019), a través del aumento en los ingresos provenientes de actividades en este sector con empleo preponderantemente informal y de cuenta propia, mientras que el incremento de los ingresos del empleo formal jugó un papel acotado (Castilleja-Vargas y Enciso, 2019).

Winkelried y Torres (2019), desde enfoques macroeconómicos han explorado los beneficios del crecimiento en la evolución de pobreza y la clase media. Los autores definen la clase media a partir del gasto per cápita y encuentran un comportamiento pro-cíclico de la movilidad económica y de las transiciones de pobreza. La movilidad es descendente (mayor caída en pobreza) durante la fase de recesión entre 1997 y 2004 y ascendente (mayor salida de pobreza) durante la fase de expansión entre 2004 y 2016.

Estudios para otros países en desarrollo, también se han concentrado en variables como el ingreso, el gasto o los activos (cuantificados), encontrando resultados similares en la evolución positiva y crecimiento de la clase media específicamente en las últimas décadas. Así, por ejemplo, el estudio mencionado anteriormente de López-Calva y Ortiz-Juárez (2013) para Chile, México y Perú, utilizó el ingreso para delimitar a la clase media. Documentos más recientes, como el de Schotte et al. (2018) y Burger et al. (2015) para Sudáfrica, y el de Birdsall et al. (2014) y Stampini et al. (2017) para Latinoamérica, se han basado también en variables monetarias para la identificación de este segmento. Sin embargo, estos estudios reconocen la complejidad de la identificación de la clase media en los países en desarrollo la alta movilidad de los sectores pobres a los medios (y viceversa), y a su alta heterogeneidad, ya que hay grupos más vulnerables que se encuentran en riesgo de caer en la pobreza, así como otros más consolidados que gozan de una mayor estabilidad económica. La heterogeneidad está asociada a diferentes factores, como el nivel de educación, el tipo de ocupación, el grado de calificación y la formalidad laboral. Además del empleo y la ocupación, una mayor tenencia de activos (incluyendo la vivienda) y bienes también parece ser un factor relevante en la clase media en los países en desarrollo.

En resumen, la literatura económica ha utilizado enfoques tanto relativos como absolutos para definir la clase media, considerando variables como el ingreso/gasto o la valorización/monetización de activos y bienes. Sin embargo, no ha incorporado variables importantes como la educación, la calidad de la vivienda o el tipo de empleo y ocupación, a pesar de reconocer su relevancia en la pertenencia a un determinado segmento social. En los países en desarrollo, estos factores adquieren aún mayor importancia debido a la alta movilidad entre la clase media y los segmentos pobres, así como a la elevada heterogeneidad presente dentro de la clase media, donde coexisten tanto un sector vulnerable como otro más consolidado.

En ese sentido, buscando aportar a la literatura, como se mencionó anteriormente, este documento aporta tres innovaciones: fusiona el enfoque de vulnerabilidad a la pobreza con la identificación multidimensional de la clase media, analiza la clase media peruana en un periodo largo (2004-2021) usando microdatos de encuestas, y emplea métodos de big data para seleccionar variables, mejorando la operatividad del estudio en posibles contextos de escasez de datos.

3. Enfoque metodológico

3.1 Bienestar multidimensional e identificación de la clase media peruana

En el presente documento, adoptaremos una noción relativa de la clase media y se desarrolla un ejercicio de identificación multidimensional.

Como se mencionó previamente, en un contexto basado únicamente en ingresos, esto implica trazar un límite inferior y superior de ingresos que contiene a la clase media con una probabilidad determinada. Los enfoques absolutos requieren definir umbrales de manera exógena, como es el caso de las líneas de pobreza basadas en el valor monetario de una canasta de bienes cuyo consumo separa a los pobres de los no pobres. Esto, sin embargo, se complica en un escenario multidimensional, donde factores más allá de los ingresos deben ser considerados para determinar las fronteras de la clase media.

La noción de la clase media multidimensional relativa implica identificar, dentro de un marco multivariante, aquellas observaciones que se ubican entre un límite inferior y un límite superior. El límite inferior diferencia a la clase media de la población en situación de pobreza, mientras que el límite superior establece la frontera entre la clase media y la población acaudalada. Asimismo, siguiendo el enfoque de "vulnerabilidad a la pobreza", se introduce un tercer límite que divide a la clase media en "vulnerables" y "seguros". El desafío consiste entonces en segmentar el espacio de bienestar en cuatro grupos ordenados: pobres, clase media vulnerable, clase media segura y ricos.

Antes de definir los umbrales que separan las diferentes categorías, se precisa construir un indicador que permita proyectar los datos de las diferentes dimensiones relevantes. Esta proyección, además, debe tener una dirección de crecimiento, es decir, debe poseer una dirección especifica en la cual el bienestar aumenta de manera inequívoca. Así, considerando el enfoque propuesto por Edo et al. (2021), se puede definir cada variable aleatoriamente en el espacio multidimensional de bienestar de tal forma que se establezca un orden creciente, es decir, niveles más altos se corresponden con mayores niveles de bienestar. De esta manera, se pueden proyectar los datos en una dirección de crecimiento que puede representar el promedio de las dimensiones de bienestar. Esta proyección unidireccional podría realizarse a través del primer componente principal.

El primer componente principal en el Análisis de Componentes Principales (PCA) sirve como un indicador eficiente para proyectar datos multidimensionales en un indicador unidimensional debido a su habilidad para maximizar la varianza, es decir, captura la mayor cantidad de variabilidad en los datos. Al mismo tiempo, reduce la dimensionalidad de los datos sin perder mucha información, lo cual es útil tanto para la visualización de datos de alta dimensión como para la reducción del ruido y redundancia antes de aplicar otros análisis. Además, este primer componente puede desacoplar correlaciones entre variables, facilitando la interpretación de los resultados. Por último, el PCA es un método eficiente y escalable, capaz de manejar grandes conjuntos de datos, lo que significa que proyectar los datos en el primer componente principal puede ofrecer una comprensión útil de los datos con un costo computacional relativamente bajo.

Este enfoque supera las limitaciones de los métodos anteriores. En primer lugar, ofrece una identificación verdaderamente multidimensional de la clase media. Como se describió en la revisión de literatura, algunos autores afirman estudiar la clase

media en varias dimensiones, pero la definen en términos de ingresos/gastos y utilizan varias dimensiones para el análisis posterior. Este método, sin embargo, proporciona una identificación multidimensional verdadera basada en la proyección de las variables en el espacio original multidimensional.

En segundo lugar, garantiza que el grupo identificado como la clase media realmente pertenezca a la región central de la distribución multivariante, dado que las variables representan el bienestar y tienen un orden natural creciente. Esto no es un resultado obvio. A diferencia de otros enfoques como el de Gigliriano y Mosler (2009), que no pueden garantizar la identificación de un subconjunto en la región central de la distribución, el método propuesto define que la clase media pertenecerá a la región central de la distribución multivariante, considerando el orden natural que implican las variables de bienestar.

Cuando se aplica el PCA a un conjunto de datos que contienen varias medidas de bienestar (por ejemplo, ingresos, educación, vivienda, etc.), el primer componente principal identificará la dirección en la que estas medidas están más correlacionadas. En este caso, las variables representan el bienestar y tienen un orden natural creciente. Esto significa que, a medida que una variable de bienestar aumenta, las demás también tienden a hacerlo. Por lo tanto, si se usa el primer componente principal para clasificar a los individuos en la clase media, estos se ubicarán en la región central de la distribución multivariante, porque el primer componente principal detecta las direcciones donde las variables presentan correlaciones positivas y captura la mayor cantidad de varianza en el conjunto de datos (James et al., 2010).

Finalmente, la propuesta genera una clasificación exhaustiva de los individuos a lo largo de un índice de bienestar creciente. Es decir, no solo se define a la clase media, sino también a los pobres y a la clase alta. Esto permite hacer comparaciones entre nuestro grupo de interés y otros grupos identificados multidimensionalmente.

Es relevante aclarar que el enfoque adoptado en nuestra investigación se aleja de los métodos tradicionales formulados por Alkire et al. (2015) para medir la pobreza multidimensional. Específicamente, nuestra metodología se caracteriza por su naturaleza relativa. Esto implica que designamos un segmento específico dentro de la distribución de bienestar para definirlo como la clase media.

Este método contrasta con la postura de carácter más absoluta de Alkire y Foster (2011), donde la pobreza se define mediante el establecimiento de umbrales en cada dimensión, muchos de los cuales se fundamentan en estándares internacionales consensuados.

Además, el proceso de agregación de las dimensiones en nuestro enfoque difiere del método de Alkire y Foster. Mientras que la literatura de pobreza multidimensional se enfoca en cuantificar "privaciones" en cada dimensión (donde la clasificación final de pobreza se basa en la cantidad de privaciones y los criterios establecidos para marcar un hogar como pobre), en nuestra investigación empleamos la primera componente principal como un indicador de bienestar. Este uso nos impide establecer un criterio absoluto para definir la "privación", tal como se interpreta en el enfoque de pobreza multidimensional.

Adicionalmente, nuestro objetivo es delimitar la clase media, lo cual requiere establecer un criterio que separe a aquellos en el medio de aquellos en la parte superior de la distribución de bienestar. En este contexto, la metodología de Alkire y Foster no resulta apropiada, ya que no existen convenciones internacionales que definan un umbral que delimite la transición de la clase media a la riqueza, ni un equivalente teórico para la noción de privación en este caso.

Para obtener el primer componente principal, se debe realizar un análisis de componentes principales (PCA). El objetivo de esta metodología será proyectar la data en una dirección de crecimiento que capture la mayor variabilidad posible en un índice de bienestar multidimensional. Para ello, se utilizan diversas variables relacionadas al bienestar multidimensional, como ingresos, educación, acceso a servicios básicos, entre otros.

El análisis de componentes principales es un método utilizado para reducir la dimensionalidad de los datos, lo que permite una mejor comprensión de estos. En el proceso de PCA, se calculan los componentes principales y luego se utilizan para explicar la variabilidad de los datos. Es importante destacar que el PCA es un enfoque no supervisado, ya que implica representar n observaciones con un conjunto de características $p (X_1, X_2, \dots, X_P)$ y no una respuesta asociada Y.

La idea subyacente es que cada una de las n observaciones habita en un espacio con p dimensiones; sin embargo, no todas estas dimensiones resultan igualmente relevantes. El Análisis de Componentes Principales (PCA) intenta encontrar un número reducido de dimensiones que sean lo más significativas posible, tomando en cuenta la variación de las observaciones en cada dimensión como indicador de interés.

Cada dimensión identificada por PCA es una combinación lineal de las p características. Las dimensiones (o componentes principales) se determinan de la siguiente manera. El primer componente principal de un conjunto de características $X_1, X_2, \dots, X_P$ corresponde a la combinación lineal normalizada de dichas características que tiene la más larga varianza.

$$Z_1 = \emptyset_{11}X_1 + \emptyset_{21}X_2 + \cdots + \emptyset_{p1}X_p$$

La normalización implica que $\sum_{j=1}^{p} \emptyset_{j1}^2 = 1$. Estos elementos $(\emptyset_{11}, \emptyset_{21}, \dots, \emptyset_{p1})$ se denominan cargas del primer componente; y en conjunto componen el vector de cargas del componente principal, $\emptyset_1 = (\emptyset_{11}X_1 + \emptyset_{21}X_2 + \cdots + \emptyset_{p1}X_p)^T$. Las cargas se encuentran restringidas de tal manera que la suma de sus cuadrados es 1, de otra manera, valores arbitrarios podrían resultar en una alta varianza arbitraria.

Considerando un conjunto de datos (X_{nxp}), el primer componente principal se calcula de la siguiente manera. Dado que interesa la varianza, se asume que cada una de las variables en X ha sido centrada para tener una media de cero (es decir, las medias de las columnas de X son cero). Luego, buscamos la combinación lineal de los valores de muestra de las características de la forma:

$$Z_{i1} = \emptyset_{11}x_{i1} + \emptyset_{21}x_{i2} + \cdots + \emptyset_{p1}x_{ip}$$

que tiene la más larga varianza, sujeta a la restricción $\sum_{j=1}^{p} \emptyset_{j1}^2 = 1$. Es decir, el vector de cargas del primer componente principal resuelve el problema de optimización:

$$\underset{\emptyset_{11},\ldots,\emptyset_{p1}}{maxim} \left\{ \frac{1}{n} \sum_{i=1}^{n} (\sum_{j=1}^{p} \emptyset_{j1}x_{ij})^2 \right\} \; s.a. \sum_{j=1}^{p} \emptyset_{j1}^2 = 1$$

La función objetivo puede ser reescrita como $\frac{1}{n}\sum_{i=1}^{n} z_{i1}^2$. Dado que $\frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0$, el promedio del $z_{11}, \ldots, z_{n1}$ será cero también. Por lo tanto, el objetivo que estamos maximizando es solo la varianza muestral de los valores de z_{i1}. Nos referimos a $z_{11}, \ldots, z_{n1}$ como los scores del primer componente principal. El problema puede ser resuelto por descomposición en valores propios.

Una interpretación geométrica interesante para la primera componente principal considera que el vector de cargas $\emptyset_1$ con elementos $(\emptyset_{11}, \emptyset_{21}, \ldots, \emptyset_{p1})$, define una dirección en el espacio de las características a lo largo de la cual los datos varían más. Si proyectamos las n observaciones de $x_1, \ldots, x_n$ sobre esta direccion, los valores proyectados serán los scores del primer componente principal $z_{11}, \ldots, z_{n1}$.

Esta interpretación resulta particularmente útil para los objetivos de esta investigación. La interpretación geométrica del primer componente principal (PC1) en relación con el objetivo de construir un índice de bienestar para identificar a la clase media puede ser abordada desde la perspectiva de las cargas y la proyección geométrica.

Las cargas de la primera componente principal representan los pesos asignados a cada variable original en la combinación lineal que conforma el PC1. Estas cargas indican la importancia y el efecto de cada variable en el índice de bienestar. Variables con cargas altas (ya sean positivas o negativas) tienen un impacto más significativo en el índice de bienestar, mientras que variables con cargas bajas influyen menos en el índice.

Geométricamente, la proyección de los puntos de datos sobre el eje del primer componente principal representa la posición de cada individuo en el índice de bienestar. La proyección permite observar cómo se distribuyen los datos en relación con el PC1, resaltando las diferencias en bienestar entre individuos. Al proyectar los datos sobre este eje, se obtiene un valor para cada individuo que refleja su posición en el índice de bienestar.

Para identificar a la clase media utilizando este índice de bienestar basado en el primer componente principal, se puede establecer un rango de valores que represente a esta clase social. Los individuos cuyas proyecciones en el eje del PC1 se encuentren dentro de este rango serán considerados parte de la clase media. Este enfoque permite una clasificación más precisa de la población y facilita la identificación de aquellos que se encuentran en el rango de bienestar correspondiente a la clase media.

3.2 Reducción de dimensionalidad vía selección de variables

Se realizará un ejercicio adicional para identificar un conjunto de variables que puedan replicar los resultados obtenidos a partir de todas las variables del espacio original. Por razones de interpretación y restricciones operativas, es conveniente usar menos variables para monitorear la performance de la clase media, ya que esto facilita la obtención de información a partir de muestras y encuestas específicas y simplifica la interpretación y comunicación de resultados.

Para lograr este objetivo, se emplearán modelos estadísticos o de aprendizaje automático que seleccionen variables según su importancia en la predicción del índice de bienestar multidimensional. Estos métodos eliminan variables redundantes o irrelevantes, mejorando la precisión y capacidad de generalización de los modelos predictivos asociados al índice socioeconómico. Además, estos métodos pueden manejar relaciones no lineales entre variables, a diferencia del análisis basado en valores absolutos de pesos del primer componente principal o análisis de clúster jerárquico.

La flexibilidad de estos métodos permite probar diferentes enfoques y elegir el que mejor preserve la similitud con el índice original. Entre los métodos adecuados para la selección de variables en este contexto, se encuentran la Regresión LASSO, que agrega una penalización a los coeficientes de regresión, resultando en la reducción de algunos coeficientes a cero; la Regresión Ridge, que utiliza una penalización cuadrática en lugar de una penalización absoluta en los coeficientes de regresión, ayudando a identificar variables importantes y reducir el impacto de variables menos relevantes en el modelo; y Elastic Net, que combina aspectos de LASSO y Ridge, permitiendo la selección de variables y la regularización de los coeficientes de regresión al mismo tiempo, proporcionando un equilibrio entre la selección de variables y la estabilidad del modelo, especialmente cuando hay correlaciones entre las variables.

Enfoques basados en árboles de decisión, también son útiles en la selección de variables, ya que pueden manejar relaciones no lineales y complejas entre variables y proporcionar una medida de la importancia de las variables en el modelo, además de manejar eficientemente variables categóricas y datos faltantes. Dentro de la familia de árboles, Random Forest y los modelos Gradient Boosting (GBM), incluido XGBoost, son técnicas populares y efectivas para la selección de variables en el contexto del aprendizaje automático (James et al., 2010).

3.2.1 Lasso, Ridge y Elastic Net

La regresión LASSO (“Least Absolute Shrinkage and Selection Operator”) es un método de regresión lineal que aplica una penalización basada en la norma L1 a la suma de los valores absolutos de los coeficientes. La norma L1 es una medida de la magnitud de un vector que se calcula sumando los valores absolutos de sus componentes. Esta penalización provoca que algunos coeficientes se reduzcan a cero, lo que lleva a la exclusión de esas variables del modelo final y facilita la identificación de un subconjunto más pequeño de variables relevantes. De esta manera, se simplifica el modelo y mejora la interpretación de los resultados. A diferencia de la regularización basada en la norma L2, como en la regresión Ridge, la regresión LASSO produce modelos dispersos y permite una interpretación más sencilla. La norma L2, en cambio, es la suma de los cuadrados de los componentes del vector, y su aplicación en la regresión Ridge no lleva a la eliminación de variables del modelo.

Las soluciones Lasso son problemas de programación cuadrática. Considerando N observaciones, cada uno con p variable y una sola variable dependiente. Sea y_i la variable dependiente y $x_i = (x_1, x_2, \dots, x_p)^T$ el vector con variables para el caso j y λ es el parámetro de penalización que controla la intensidad de la penalización en los coeficientes. El objetivo del algoritmo es minimizar la siguiente expresión:

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p}|\beta_j|$$

El primer término de la ecuación $\sum_{i=1}^{n}(y_i - \sum_j x_{ij}\beta_j)^2$, corresponde al error cuadrático de la regresión lineal ordinaria. Este término busca minimizar la diferencia entre los valores observados (y) y los valores predichos por el modelo $(\beta_0 + \sum_j x_{ij}\beta_j)$. El segundo término $\lambda \sum_{i=1}^{p}|\beta_j|$ es la penalización proporcional al valor absoluto de los coeficientes de las variables predictoras. Esta penalización tiene el efecto de encoger algunos de los coeficientes hacia cero, lo que resulta en la exclusión de esas variables del modelo final. El parámetro λ controla la intensidad de la penalización: valores más altos de λ conducen a una mayor penalización y, por lo tanto, a modelos más simples con menos variables.

En el caso de la regresión Ridge, la función de costo se modifica agregando un término de penalización equivalente al cuadrado de la magnitud de los coeficientes. El objetivo es minimizar el mismo término de la expresión anterior, correspondiente al error cuadrático de la regresión lineal ordinaria. En el caso del segundo término, es la penalización L2, que agrega una penalización proporcional al cuadrado de los coeficientes de las variables predictoras. A diferencia de LASSO, la penalización L2 no reduce directamente los coeficientes a cero, pero sí reduce su magnitud, lo que disminuye la influencia de las variables menos relevantes en el modelo y ayuda a mitigar el problema de multicolinealidad en los datos.

Al igual que la regresión Lasso, λ controla la intensidad de la penalización. La Regresión Ridge busca minimizar la suma de ambos términos, lo que implica encontrar un equilibrio entre ajustar el modelo a los datos (minimizando el error cuadrático) y mantener un modelo estable y menos sensible a la multicolinealidad (minimizando la suma de los cuadrados de los coeficientes) aunque la Regresión Ridge no reduce directamente los coeficientes a cero y no produce modelos dispersos como LASSO.

La regresión Elastic Net es un algoritmo de clasificación que combina las penalizaciones de regularización de los métodos Lasso y Ridge, superando sus limitaciones individuales. Este enfoque híbrido permite seleccionar un subconjunto de variables relevantes, similar a LASSO, mientras mantiene la estabilidad y robustez de Ridge.

En el contexto de selección de variables, Elastic Net es especialmente útil en casos donde hay correlación entre las variables, ya que tiende a seleccionar grupos de variables correlacionadas. Así, puede ser una herramienta valiosa al reducir la cantidad de variables en un índice socioeconómico y obtener resultados lo más parecidos posible al índice original.

La implementación de la regresión Lasso y Elastic Net se llevó a cabo para seleccionar cinco variables relevantes año tras año. Se utilizaron variables estandarizadas y validación cruzada k-fold.

Tanto en el caso de Lasso como de Elastic Net, los valores de λ desempeñan un papel crucial. Para seleccionar el mejor valor de λ en el modelo, se emplea la técnica de validación cruzada k-fold (Hastie et al., 2009). Los datos se dividen en k subconjuntos de tamaño similar, y uno de estos subconjuntos se utiliza como conjunto de validación. Los $k-1$ subconjuntos restantes se emplean como datos de entrenamiento. Este proceso se repite k veces, cada vez con un conjunto de validación diferente, y el valor óptimo de λ se determina de tal manera que se maximice la función de log-verosimilitud con validación cruzada (Goeman, 2010).

Una vez obtenido el mejor valor de λ para cada año, se ajustó el modelo Lasso y se seleccionaron las variables relevantes, es decir, aquellas cuyos coeficientes no eran iguales a cero. En el caso de la regresión Elastic Net, se sigue un enfoque similar, pero el modelo equilibra las ventajas de Lasso y Ridge al combinar elementos de ambos métodos de regularización, permitiendo una selección de variables más precisa y un mejor manejo de la multicolinealidad.

Con los mejores valores de λ obtenidos mediante la validación cruzada, se ajusta el modelo Elastic Net y se seleccionan las variables relevantes. El proceso de ajuste y selección de variables en Elastic Net es similar al de Lasso, pero con la ventaja adicional de combinar las características de Lasso, que promueve la dispersión y selección de variables, y Ridge, que ayuda a mitigar la multicolinealidad al encoger los coeficientes de variables correlacionadas. Después de ajustar el modelo Elastic Net con los mejores parámetros encontrados, se seleccionan las variables más relevantes, es decir, aquellas cuyos coeficientes no son iguales a cero y se almacenan en un diccionario, junto con los resultados de Lasso, para su análisis y comparación.

3.2.2 Decision Trees

Dentro de la familia de árboles, Random Forest y los modelos Gradient Boosting (GBM), incluido XGBoost, son técnicas populares y efectivas para la selección de variables en el contexto del aprendizaje automático.

Random Forest es un método de conjunto que utiliza múltiples árboles de decisión, cada uno de los cuales se entrena con una muestra distinta generada mediante bootstrapping. Al introducir aleatoriedad en las muestras utilizadas, se disminuye la correlación entre los árboles individuales, lo que resulta en un modelo más robusto. Random Forest proporciona una medida de la importancia de las variables, lo que ayuda a identificar las características más relevantes para el problema en cuestión.

Por otro lado, los modelos Gradient Boosting (GBM) y su variante XGBoost también ofrecen selección de variables. Estos modelos están compuestos por un conjunto de árboles individuales entrenados secuencialmente, donde cada nuevo árbol mejora los errores de los árboles anteriores. Para predecir, se agregan las predicciones de todos los árboles individuales que forman el modelo. XGBoost es una implementación específica de GBM que utiliza aproximaciones más precisas para encontrar el mejor modelo de árbol y evita de mejor manera el sobreajuste.

Una ventaja clave de GBM y XGBoost en comparación con otros métodos es que permiten manejar datos con valores faltantes. Esto significa que se pueden incluir más variables y observaciones en el análisis, lo que puede mejorar la selección de variables y, en última instancia, el rendimiento del modelo.

En resumen, tanto Random Forest como GBM y XGBoost son métodos de aprendizaje automático basados en árboles que pueden ayudar en la selección de variables. Estos modelos proporcionan medidas de importancia de variables y pueden manejar valores faltantes de manera eficiente, lo que los hace adecuados para la selección de variables en contextos donde se busca reducir la cantidad de variables y mantener un resultado lo más parecido posible al índice original.

En este documento, se implementaron tres algoritmos de aprendizaje automático basados en árboles de decisión, Random Forest, Gradient Boosting y XGBoost, para llevar a cabo la selección de variables en un conjunto de datos con información anual. El objetivo fue identificar las cinco variables más importantes que contribuyen a la predicción de la variable dependiente en cada año específico.

Los tres algoritmos utilizados combinan múltiples árboles de decisión para construir un modelo más robusto y preciso. Estos algoritmos evalúan la importancia de las características en función de su capacidad para mejorar la precisión de las predicciones y reducir el error en la estimación.

Para cada algoritmo, se llevó a cabo el siguiente proceso: se iteró sobre cada año en la lista, filtrando el conjunto de datos para incluir solo las observaciones correspondientes a ese año. Luego, se extrajeron las variables independientes y dependientes y se entrenó el modelo con ese subconjunto de datos.

Una vez ajustado el modelo, se calculó la importancia de las características y se clasificaron en orden descendente según su relevancia. A continuación, se seleccionaron las cinco características más importantes para cada año y se almacenaron en un diccionario de resultados.

3.3 Datos

La información utilizada en este estudio proviene de la Encuesta Nacional de Hogares (ENAHO) y consta de muestras de corte transversal anuales desde 2004 hasta 2021. Se incluye un amplio conjunto de variables que abarcan información sobre aspectos demográficos, educación, empleo e ingresos familiares, así como características de las viviendas para hogares en todo el país. Este amplio conjunto de variables permite identificar multidimensionalmente la clase media peruana.

Como se observa en la Tabla 1, se consideraron 14 variables. El primer grupo de variables incluye a aquellas vinculados al ingreso. Al ser el Perú una economía de alto empleo informal, los ingresos dependientes representan un porcentaje importante del ingreso de los hogares. En ese sentido, se incluyó el Ingreso bruto total y el Ingreso proveniente del trabajo independiente (como porcentaje del Ingreso bruto total).

El segundo grupo de variables incluye variables relacionadas a la ocupación y el nivel educativo. Así, se incluye una variable relacionada al empleo del jefe de hogar, donde el nivel más alto representa a ocupaciones en cargos directivos y profesionales y el nivel más bajo a ocupaciones elementales. Asimismo, se ha incluido el número de años de educación del jefe de hogar como indicador del nivel educativo.

El tercer conjunto de variables incluye las características de las viviendas, como el material de construcción, el tipo de piso y el acceso a servicios básicos (agua, saneamiento y electricidad). Esto resulta relevante en un país como Perú, ya que gran parte de la expansión urbana en las últimas décadas ha ocurrido de manera no planificada y mediante la autoconstrucción. En este contexto, aunque la propiedad de la vivienda generalmente estaba garantizada por las autoridades a través de títulos de propiedad otorgados después de invasiones, muchas viviendas fueron adquiriendo los servicios públicos a lo largo de los años. Por lo general, las viviendas ubicadas en áreas periféricas y pobres accedieron más tardíamente a estos servicios esenciales.

Finalmente, se incluye otras variables de acceso a determinados bienes y etnicidad. Así, por ejemplo, se considera la tenencia de celular, internet, auto y refrigeradora. Si bien en los últimos anos, la telefonía celular se ha masificado, existen aún brechas en el acceso a internet. Estos elementos reflejan un cierto nivel socioeconómico y capacidad adquisitiva, facilitan el acceso a la educación, el empleo y la comunicación, y mejoran la calidad de vida. Durante la pandemia (y post), la importancia de algunos de estos elementos se hizo aún más evidente, ya que la refrigeradora y el automóvil permitieron enfrentar las restricciones de movilidad, mientras que el acceso a internet y el uso de celulares fueron cruciales para el teletrabajo y la educación a distancia.

Tabla 1 Variables relacionadas al bienestar (a nivel de hogar)

Variable	**Descripción**
Ingresos	
Ingreso Bruto Total	Monto recibido anualmente
Ingreso Independiente	Como porcentaje del Ingreso Bruto Total
Empleo y Educación	
Ocupación	Tipo de ocupación del jefe de hogar
Nivel Educativo	Anos de educación del jefe de hogar
Características de la vivienda	
Material	Vivienda es de concreto o cemento
Piso	Piso de la vivienda es de concreto, madera, etc. (no tierra)
Acceso a agua	Accede a agua por red publica
Acceso a saneamiento	Accede a saneamiento por red publica
Electricidad/Combustible	Usa electricidad o gas para cocinar
Otras	
Celular	Jefe de hogar tiene celular
Internet	Cuenta con internet
Refrigeradora	Cuenta con refrigeradora
Auto	Cuenta con automóvil

Etnicidad	Si lengua materna es español

El periodo cubierto es 2004-2014. El análisis se lleva a cabo para cada ano. Cada muestra anual cuenta con aproximadamente 25 mil observaciones y se incluye un total de aproximadamente 500 mil observaciones para todo el periodo de estudio.

4 Resultados

4.1 Índice de bienestar multidimensional

Para cada año de análisis, se estandarizaron las variables seleccionadas utilizando la media y la desviación estándar de la muestra, de modo que todas las variables tuvieran la misma escala y pudieran ser comparables. Luego, se aplicó el PCA a las variables estandarizadas para cada año, lo que permitió reducir la dimensionalidad de los datos y obtener un índice de bienestar que capta la mayor variabilidad posible en las variables originales.

Los resultados sugieren que las catorce variables durante el período de análisis pueden ser resumidas en 5 factores ortogonales, que abarcan el 91% de la variabilidad total.

La dirección de crecimiento adoptada es el módulo del primer componente principal. Dos resultados del análisis de componentes principales respaldan esta decisión. En primer lugar, el primer componente principal representa, en promedio, el 34.8% de la variabilidad a lo largo de los años, lo cual es alto en relación con la magnitud de nuestro espacio original. En segundo lugar, los pesos asignados a cada variable en el primer componente principal son, en promedio, los mismos en todos los años, lo que garantiza una comparación coherente entre períodos.

El producto final de este ejercicio es un índice de bienestar que permitirá ordenar a los individuos en un ranking a partir de la información utilizada. Con este índice de bienestar, se pudo ordenar a los individuos según su nivel de bienestar. La definición de clase media utilizada en este análisis es relativa, y se basa en la posición de los individuos según el índice de bienestar. Se definió la clase media como aquel grupo de individuos cuyo índice de bienestar se encuentra entre los cuantiles 0.25 y 0.90. La elección de estos percentiles es arbitraria sin embargo se sustenta en la literatura previa que encuentra que este límite superior funciona bien en mediciones unidimensionales para países en desarrollo (Edo et al. 2021) y en el porcentaje de población en situación de pobreza durante los últimos anos (entre 20% y 25%) para el límite inferior.

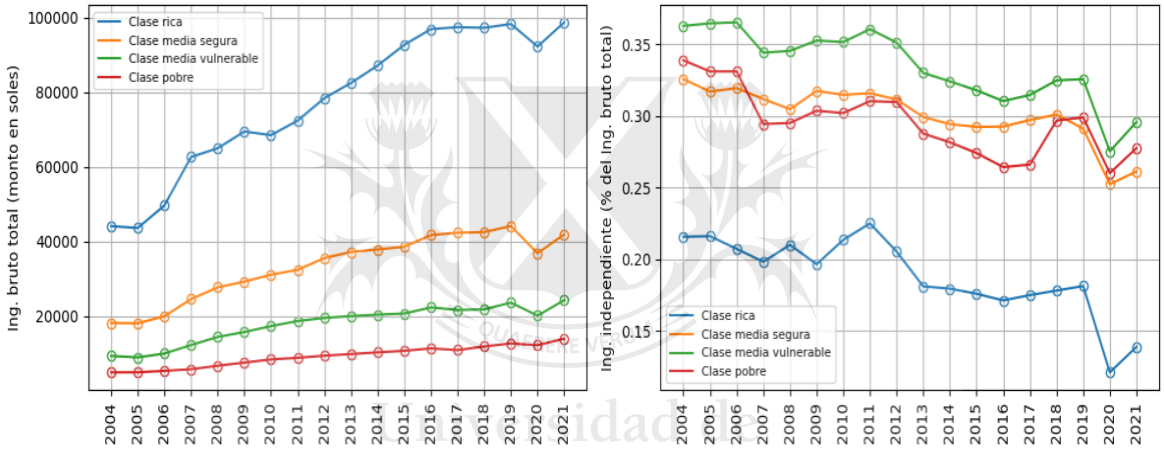
Cabe mencionar que la clase media puede ser heterogénea, y dentro de ella se pueden encontrar distintos niveles de vulnerabilidad. Hay un sector vulnerable que, si bien no es pobre, enfrenta grandes riesgos de volver a serlo y está expuesto a shocks económicos y sociales. Por ese motivo, el análisis distingue un grupo de clase media vulnerable y otro de clase media segura, siendo el límite que divide a estos dos grupos el cuantil 0.60 en el índice de bienestar. El límite se define arbitrariamente, sin embargo, está sustentado en literatura que encuentra que aproximadamente entre 25% y 35% de hogares se encuentra en situación de vulnerabilidad (fuera de la pobreza) (Herrera y Cozzubo, 2016). De esta manera, se busca analizar la evolución de

la clase media en el país y examinar su relación con otros factores socioeconómicos, considerando también la heterogeneidad y la vulnerabilidad dentro de este grupo social.

El enfoque planteado en esta investigación implica que la clase media esta fija. Así, en los siguientes gráficos se analiza la evolución de las variables relacionadas al bienestar en este periodo en los diferentes segmentos sociales a lo largo del periodo de estudio (2004-2021). Además de ser un objetivo de esta investigación, el análisis de las variables asociadas al bienestar permite validar el cálculo del índice de bienestar y la clasificación realizada por esta.

En los siguientes gráficos, se observa la evolución del ingreso bruto total y el porcentaje que representa el ingreso independiente del ingreso bruto total.

Gráfico 1. Evolución del ing. bruto total e ing. independiente (como % del ing. bruto total)



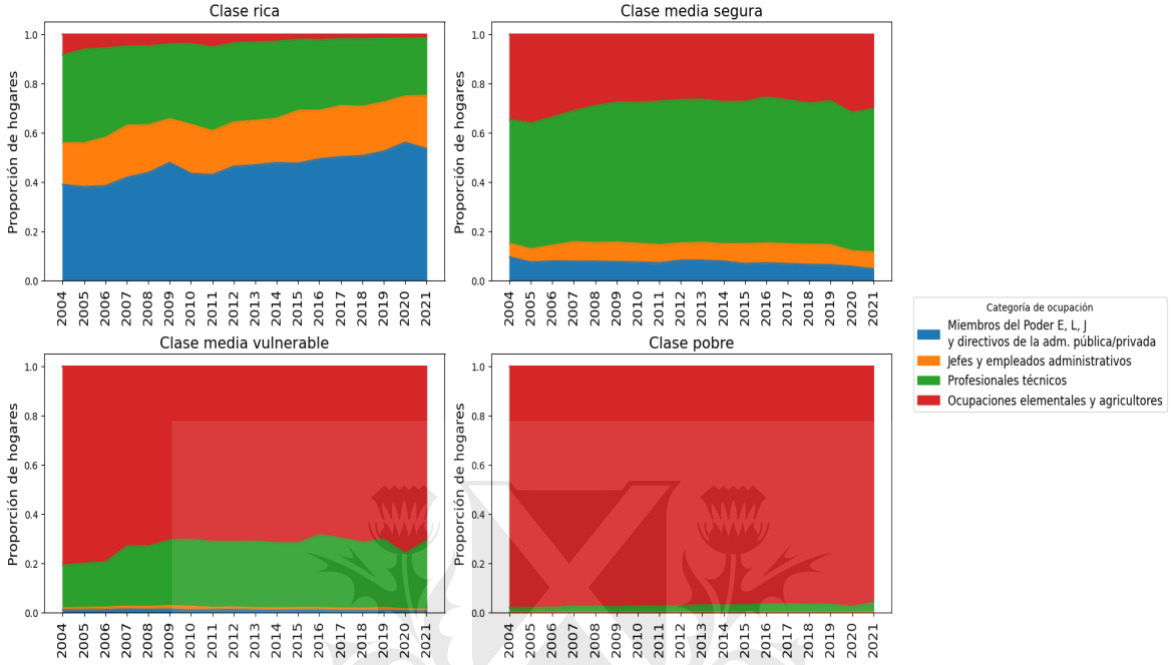
Se puede apreciar que, en todos los segmentos, el ingreso bruto del hogar aumenta considerablemente más del 100%. Se observa también un deterioro en el año 2020 debido a la pandemia del COVID 19, seguido de recuperación durante el 2021 y 2022.

En el caso del ingreso independiente, se observa una redacción de la importancia de este en el ingreso total en todos los segmentos socioeconómicos. En el 2020, se observa una caída pronunciada relacionada a posibles pérdidas de ingreso debido a las restricciones impuestas durante la pandemia. La diferencia en la importancia del ingreso independiente entre los más ricos y el resto de los segmentos indica que los hogares con mayor bienestar suelen basar sus fuentes económicas en trabajo dependiente formal, mientras que el resto de los sectores aún depende de manera importante de fuentes alternativas e informales al trabajo dependiente asalariado.

Lo observado al analizar el ingreso independiente puede asociarse a lo que se observa en el Gráfico 2 con respecto a las categorías ocupacionales. Por ejemplo, se observa que un alto porcentaje de ocupaciones directivas y profesionales de ramas administrativas en el segmento más rico. En la clase media segura, los profesionales técnicos son mayoría, mientras que, en los segmentos vulnerable y pobre, las ocupaciones elementales (por ejemplo, vendedores ambulantes, recolectores de desechos, peones, etc.) y los agricultores suelen representar el mayor porcentaje de

ocupaciones. Este tipo de actividades suelen ser actividades independientes de carácter precario e informal.

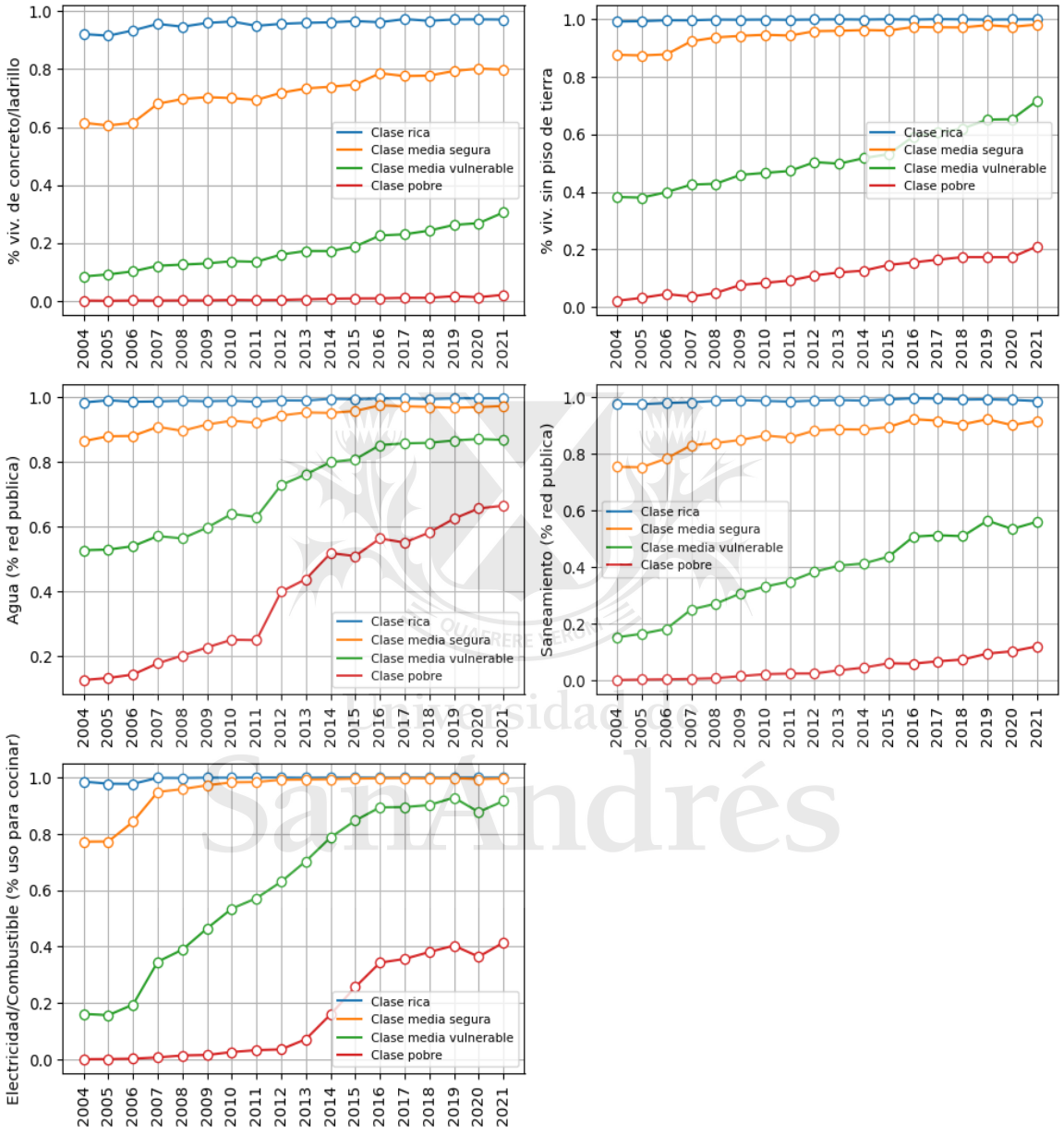
Gráfico 2. Evolución de la proporción de hogares por categoría de ocupación



En el caso de características de la vivienda con respecto a materiales con los que está construida la vivienda y el acceso a servicios, se observa una diferencia importante entre clases en el material de la vivienda (es decir si fue construida con cemento o concreto) y en si las viviendas tienen o no piso de tierra. A pesar de una evolución positiva en el segmento vulnerable y pobre, las diferencias aún son resaltantes. Al 2021, el 100% de la clase rica (y más del 80% de la clase media vulnerable) dispone de viviendas hechas de cemento/concreto, mientras que, en el segmento vulnerable, este grupo representa poco más del 30% y no llega ni al 5% en los pobres. En el caso del tipo de piso, casi el 100% de los hogares ricos y de la clase media segura tienen piso que no es de tierra (parquet, madera, cerámica u otro material). En contraste, alrededor del 80% de hogares pobres tienen piso de tierra.

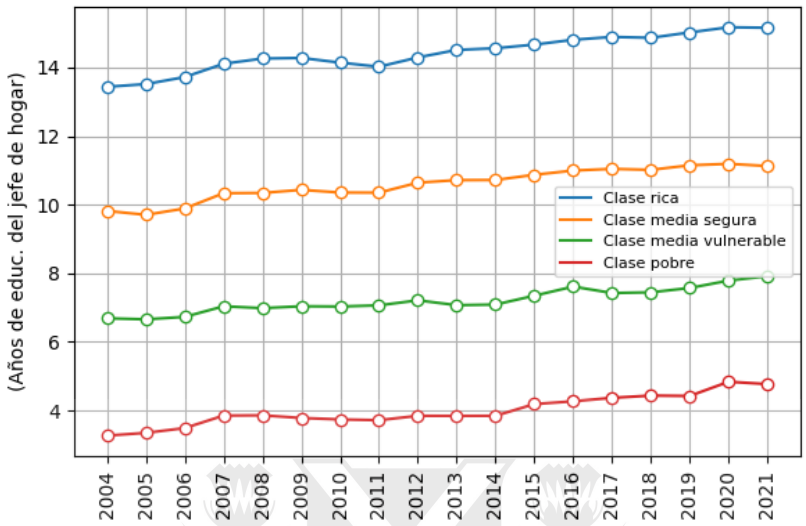
En el caso del acceso de servicios, las diferencias son también resaltantes, aunque se observa cierta convergencia de los sectores vulnerable y pobre especialmente en el acceso a agua y electricidad en la vivienda. En el caso de acceso a saneamiento, sin embargo, las diferencias se mantienen de manera importante durante todo el periodo de análisis (ver Gráfico 3).

Gráfico 3. Evolución del material/piso de la vivienda y acceso a servicios



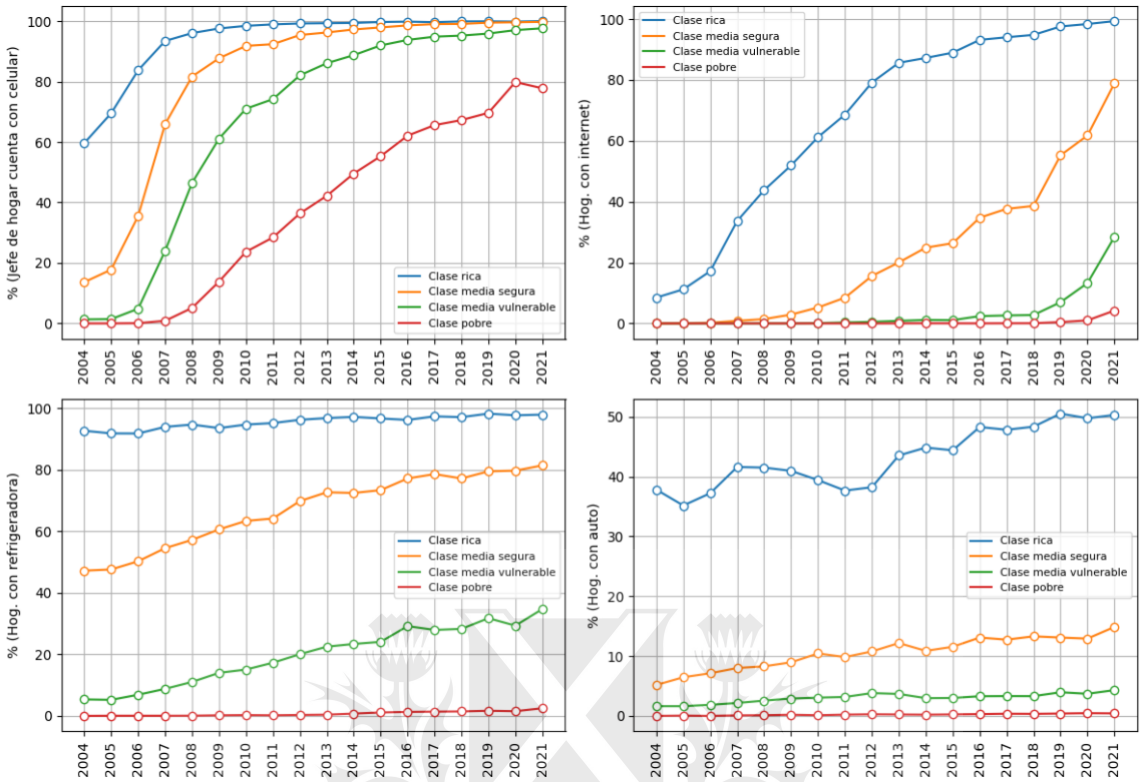
En el caso del nivel educativo, si se consideran los años de educación del jefe de hogar, si bien se observa una ligera evolución positiva de este indicador en todos los segmentos, las diferencias no parecen acortarse significativamente. Considerando que el año escolar es de 11 años, llama la atención los promedios menores a este valor en los niveles de los segmentos vulnerables y pobre. Esto probablemente indica un mayor porcentaje de personas sin haber culminado la educación básica regular.

Gráfico 4. Evolución del nivel educativo



Con respecto a la evolución del acceso a otros bienes como el celular, internet, refrigeradora o auto, se observa que, en el caso del celular, este no parece ser un bien característico de la clase media en los últimos años pues casi todos los segmentos cuentan en más de 90% con este bien. En el caso del internet, las diferencias suelen ser más notorias principalmente entre el segmento más rico y la clase media segura versus el segmento vulnerable y pobre, aunque se observa una tendencia creciente en todos los niveles socioeconómicos especialmente durante los últimos anos. En el caso de la refrigeradora, también se observan diferencias entre segmentos, aunque con una tendencia creciente menos pronunciada. Esta diferencia es aún más pronunciada si se analiza la tenencia de automóvil especialmente entre el segmento más rico versus el resto.

Gráfico 5. Evolución del acceso a otros bienes



4.3 Selección de variables

En este estudio, se exploraron diversos métodos para seleccionar variables como Lasso, Elastic Net, Random Forest, GBM y XGBoost.

En el caso de Lasso y Elastic Net, para poder elegir el mejor λ del modelo se utiliza el error cuadrático medio obtenido por validación cruzada para cada valor del logaritmo de λ junto con la barra de error correspondiente. En el caso de Lasso, el λ con el que consigue el menor error es 0,0136; mientras que en el caso de Elastic Net es 0,0116. Ambos métodos resultaron en la misma selección de variables. Aunque estos métodos son diferentes en términos de cómo aplican la regularización, en este caso específico, han convergido en la misma selección de variables.

Este resultado sugiere que las variables seleccionadas por Lasso y Elastic Net tienen un impacto significativo en el modelo y son esenciales para la predicción. La selección idéntica de variables a partir de dos métodos diferentes aumenta nuestra confianza en la relevancia de estas variables y nos permite proceder con un análisis más profundo basado en esta selección. La selección de variables se muestra en detalle en la siguiente tabla:

Tabla 2. Selección de variables por Lasso y Elastic Net

Variables	Selección por Lasso y Elastic Net

Material de la vivienda	**18**
Nivel Educativo	**18**
Acceso a saneamiento	**14**
Refrigeradora	**12**
Electricidad/Combustible	**10**
Ocupación	**8**
Piso de la vivienda	6
Internet	3
Ingreso bruto total	1

Esta tabla muestra la cantidad de veces durante el periodo de análisis que las variables fueron seleccionadas por el modelo para cada año. A partir de esta tabla, se pueden seleccionar las siguientes 6 variables más importantes: Material de la vivienda, nivel educativo, acceso a saneamiento, posesión de refrigeradora, acceso a electricidad/combustible y ocupación.

En el caso del modelo Random Forest, se identificaron a las seis variables más importantes en base a sus promedios de importancia a lo largo de los años. Asimismo, para validar esta selección, se utilizó la técnica OOB (out of bag) error estimate, que se basa en el hecho de que, en promedio, cada árbol del bosque se ajusta solo al 63% de las observaciones de entrenamiento, lo que deja alrededor del 37% de las observaciones (las OOB) para estimar la precisión del modelo. Aunque no es un "test" per se, ofrece una medida de qué tan bien está funcionando tu modelo con datos no vistos.

Las variables más importantes según los promedios de importancia se listan a continuación:

Tabla 3. Selección de variables por Random Forest

Variables	Promedio de importancia
Electricidad/Combustible	0.3463
Acceso a saneamiento	0.2720
Material de la vivienda	0.0777
Nivel Educativo	0.0677
Refrigeradora	0.0357
Ocupación	0.0354
OOB-error	5%

Asimismo, el OOB error es de 5%, lo que indica que, en promedio, la diferencia entre las predicciones del modelo y los valores reales es del 5%. Dicho de otra manera, el modelo acierta en sus predicciones el 95% de las veces, lo cual es bastante bueno.

Llama la atención que las 6 variables con mayor importancia según el modelo Random Forest sean las mismas que las seleccionadas por el modelo Lasso. Tanto LASSO como Random Forest pueden identificar el mismo conjunto de variables importantes cuando las relaciones lineales entre características y la variable objetivo son fuertes y consistentes. Aunque Random Forest captura interacciones no lineales, las relaciones lineales dominantes pueden ser suficientes para que ambos modelos seleccionen un conjunto similar de características importantes. Estas variables influyentes podrían ser robustas en diferentes técnicas de modelado. Sin embargo, las importancias relativas de las variables pueden variar entre los dos modelos, ya que capturan las relaciones de manera diferente.

Los resultados del modelo Gradient Boosting Machines (GBM) en la selección de variables muestran similitudes y diferencias con el modelo Random Forest. En comparación con Random Forest, GBM otorga más importancia al material de la vivienda y menos importancia a Electricidad/Combustible.

Esto puede deberse a que, aunque ambos modelos son basados en árboles y capturan interacciones no lineales y complejas entre las características, GBM construye árboles de manera secuencial, enfocándose en corregir los errores del árbol anterior. Esta capacidad de enfocarse en los errores residuales podría permitir a GBM identificar patrones y relaciones en los datos que podrían no ser tan evidentes para Random Forest.

Tabla 4. Selección de variables por GBM

Variables	Promedio de importancia
Material de la vivienda	0.3395
Acceso a saneamiento	0.2697
Refrigeradora	0.1231
Ocupación	0.1131
Nivel educativo	0.0943
Electricidad/Combustible	0.0602

Al observar la selección de variables hechas a partir del modelo XG Boost, las variables "Internet" y "Piso de la vivienda" aparecen como más importantes en lugar de "Ocupación" y "Nivel educativo". Esto podría deberse a diferencias en la forma en que los modelos manejan las variables y las interacciones entre ellas.

XGBoost es un algoritmo de aprendizaje automático basado en árboles de decisión que utiliza gradient boosting. A diferencia del GBM, XGBoost incluye una regularización L1 y L2 en su función de pérdida, lo que puede mejorar la precisión y evitar el sobreajuste

en ciertos casos. Además, XGBoost tiene mejor manejo de las variables categóricas y suele ser más eficiente computacionalmente que el GBM.

Estas diferencias entre los modelos podrían hacer que el modelo XGBoost considere que "Internet" y "Piso de la vivienda" son más importantes que "Ocupación" y "Nivel educativo" en el contexto de la selección de variables. "Internet" podría ser considerado relevante porque puede reflejar el acceso a la información y la adopción de tecnología en los hogares, lo cual está relacionado con el nivel socioeconómico. Por otro lado, "Piso de la vivienda" podría ser un indicador de la calidad y el nivel de inversión en el hogar, lo cual también puede estar asociado con el índice socioeconómico.

Tabla 5. Selección de variables por XGBoost

Variables	Promedio de importancia
Electricidad/Combustible	0.3478
Material de la vivienda	0.2088
Acceso a saneamiento	0.1119
Internet	0.1131
Refrigeradora	0.0943
Piso de la vivienda	0.0602

Para modelos de Gradient Boosting y XGBoost, no existe un equivalente directo del Out-of-bag (OOB) estimate como en Random Forest. Esto se debe a la forma en que funcionan estos algoritmos. Sin embargo, eso no significa que no se pueda evaluar la robustez de estos modelos. Así, se puede dividir la muestra en dos grupos: un conjunto de entrenamiento para entrenar el modelo (train), y un conjunto de prueba (test), para evaluarlo. Si bien no es el objetivo predecir la pertenencia a la clase media, es relevante ver la robustez de estos modelos con los datos utilizados en esta investigación. Asi, como se observa en la tabla 6, a partir de una matriz de confusión, que ambos modelos predicen bien fuera y dentro de la muestra pues muestran una precisión de alrededor del 90%.

Tabla 6. Performance de modelos GBM y XG Boost

	Accuracy - train	Accuracy - test
Modelo GBM	92%	90%
Modelo XGBoost	93%	89%

Una vez seleccionado las variables por los diferentes modelos, se puede recalcular el índice de bienestar multidimensional a partir del primer componente principal. Esta nueva versión del índice sirve para reclasificar a los hogares en los diversos segmentos socioeconómicos y se puede comparar con la clasificación anterior basada en las 14

variables. Un alto porcentaje de coincidencias indicaría que las 14 variables iniciales pueden ser resumidas por las 6 seleccionadas por los modelos, lo que validaría la efectividad de cualquiera de estos métodos en la selección de variables.

La Tabla 7 muestra el porcentaje de coincidencias de hogares según cada segmento socioeconómico considerando las variables seleccionadas por los modelos Lasso/Elastic Net/Random Forest/GBM y XG Boost.

Tabla 7. Coincidencias a partir de selección de variables de diversos modelos

Variable	Porcentaje de coincidencias Lasso/Elastic Net/Random Forest/GBM	Porcentaje de coincidencias XGBoost
Clase rica	72%	16%
Clase media segura	83%	47%
Clase media vulnerable	77%	65%
Clase pobre	89%	16%

La tabla muestra que la selección de variables varía según el modelo utilizado para clasificar las clases socioeconómicas. En general, los modelos Lasso, Elastic Net, Random Forest y GBM parecen coincidir en mayor medida en la selección de variables en comparación con XGBoost.

En cuanto al índice de bienestar multidimensional basado en PCA, los resultados sugieren que las siguientes seis variables pueden ser suficientes para capturar una buena cantidad de información contenida en las 14 variables originales:

1.Electricidad/Combustible

2. Acceso a saneamiento

3. Material de la vivienda

4. Nivel Educativo

5. Refrigeradora

6. Ocupación

Al reducir la cantidad de variables, se simplifica el índice y se facilita su interpretación sin sacrificar significativamente su capacidad para describir el bienestar socioeconómico. Sin embargo, es importante tener en cuenta que, aunque estas seis variables pueden capturar gran parte de la variabilidad en los datos, siempre existe el riesgo de perder información valiosa al reducir el número de variables.

5 Conclusión

El presente documento, basándose en una definición multidimensional de la clase media, realizó un ejercicio de identificación de la clase media en el contexto de un país en desarrollo como el Perú, utilizando una gran cantidad de datos de variables de bienestar provenientes de encuestas de hogares durante los años 2004-2021.

La inclusión de un amplio conjunto de variables, como ingresos, educación, empleo, características de las viviendas y acceso a ciertos bienes y servicios, proporciona una visión integral de la evolución y características de la clase media en el país. Se realizó una construcción de un índice multidimensional de bienestar a través de análisis de componentes principales (PCA), específicamente utilizando la primera componente principal. Este índice permitió dividir a la población en segmentos (ricos, clase media y pobres) y tener una visión general del bienestar de la clase media (dividida en clase media segura y vulnerable) y analizar su evolución a lo largo del tiempo.

En cuanto a los resultados, se observan mejoras en todas las variables socioeconómicas, y en algunas de estas variables, las brechas parecen acortarse (por ejemplo, el acceso a agua y electricidad. Sin embargo, persisten brechas importantes en el ingreso, nivel educativo y el acceso a ciertos bienes y servicios, como el acceso a saneamiento, el material y piso de la vivienda o conectividad a internet.

Se llevó a cabo, además, un análisis exhaustivo en la selección de variables mediante diferentes técnicas, como Lasso, Ridge, Elastic Net, Random Forest, Gradient Boosting y XGBoost, identificando un conjunto de variables clave que pueden intentar replicar la clasificación a partir del índice obtenido por PCA. Las variables identificadas (acceso a electricidad/combustible, acceso a saneamiento, material de la vivienda, nivel educativo, refrigeradora, ocupación), permitieron replicar en alrededor del 80 % la clasificación realizada a partir de PCA.

En base a estos hallazgos, se sugiere la necesidad de diseñar e implementar políticas públicas enfocadas en mejorar las condiciones de vida y el bienestar de la población peruana, priorizando el acceso a servicios básicos como saneamiento y la reducción de brechas en el acceso a otros bienes, educación, etc.

La información proporcionada por este estudio también puede ser útil para investigaciones futuras que busquen profundizar en el análisis y comprensión de la evolución de la clase media en otros contextos nacionales y regionales.

6 Referencias Bibliográficas:

Alesina, A., & Perotti, R. (1996). Income distribution, political instability, and investment. *European economic review*, *40*(6), 1203-1228.

Alkire, S., & Foster, J. (2011). Understandings and misunderstandings of multidimensional poverty measurement. *The Journal of Economic Inequality*, *9*, 289-314.

Alkire, S., Roche, J. M., Ballon, P., Foster, J., Santos, M. E., & Seth, S. (2015). *Multidimensional poverty measurement and analysis*. Oxford University Press, USA.

Banerjee, A. V., & Duflo, E. (2008). What is middle class about the middle classes around the world?. *Journal of economic perspectives*, *22*(2), 3-28.

Blackburn, M., & Bloom, D. (1985). What is happening to the middle class. *American Demographics*, *7*(1), 18-25.

Bourguignon, F., & Chakravarty, S. R. (2019). Multidimensional poverty orderings: theory and applications. In *Poverty, Social Exclusion and Stochastic Dominance* (pp. 143-166). Springer, Singapore.

Birdsall, N., Graham, C., & Pettinato, S. (2000). Stuck in tunnel: Is globalization muddling the middle?.

Birdsall, N., Lustig, N., & Meyer, C. J. (2014). The strugglers: The new poor in Latin America?. *World development*, *60*, 132-146.

Burger, R., Steenekamp, C. L., van der Berg, S., & Zoch, A. (2015). The emergent middle class in contemporary South Africa: Examining and comparing rival approaches. *Development Southern Africa*, *32*(1), 25-40.

Castellani, F., & Parent, G. (2011). Being "middle-class" in Latin America.

Castilleja Vargas, L., & Enciso, S. (2019). The Pattern of Growth and the Expansion of the Middle Class in Peru.

Cruces, G., Lanjouw, P., Lucchetti, L., Perova, E., Vakis, R., & Viollaz, M. (2015). Estimating poverty transitions using repeated cross-sections: a three-country validation exercise. *The Journal of Economic Inequality*, *13*(2), 161-179.

Davis, J. C., & Huston, J. H. (1992). The shrinking middle-income class: A multivariate analysis. *Eastern Economic Journal*, *18*(3), 277-285.

Easterly, W. (2001). The middle class consensus and economic development. *Journal of economic growth*, *6*(4), 317-335.

Eisenhauer, J. G. (2011). The rich, the poor, and the middle class: Thresholds and intensity indices. *Research in Economics*, *65*(4), 294-304.

Goeman, J. J. (2010). L1 penalized estimation in the Cox proportional hazards model. Biometrical journal, 52(1), 70-84.

Goldthorpe, J. H., & McKnight, A. (2006). The economic basis of social class. *Mobility and inequality: Frontiers of research in sociology and economics*, 109-136.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.

Herrera Zúñiga, J., & Cozzubo Chaparro, A. (2016). La Vulnerabilidad de los hogares a la pobreza en el Perú, 2004–2014.

Jaramillo, F., & Zambrano, O. (2013). La clase media en Perú: cuantificación y evolución reciente. *Banco Interamericano de Desarrollo*.

James, G., Witten, D., Hastie, T., Tibshirani, R., James, G., Witten, D., ... & Tibshirani, R. (2021). Statistical learning. *An introduction to statistical learning: with applications in R*, 15-57.

López-Calva, L. F., & Ortiz-Juarez, E. (2014). A vulnerability approach to thedefinition of the middle class. *The Journal of Economic Inequality*, *12*(1), 23-47.

Milanovic, B., & Yitzhaki, S. (2002). Decomposing world income distribution: Does the world have a middle class?. *Review of income and wealth*, *48*(2), 155-178.

Nissanov, Z., & Pittau, M. G. (2016). Measuring changes in the Russian middle class between 1992 and 2008: a nonparametric distributional analysis. *Empirical Economics*, *50*, 503-530.

Rasch, Rebecca. "Measuring the middle class in middle-income countries." *Forum for Social Economics*. Vol. 46. No. 4. Routledge, 2017.

Ravallion, M. (2010). The developing world's bulging (but vulnerable) middle class. *World development*, *38*(4), 445-454.

Rose, S. (2020). Squeezing the middle class: Income trajectories from 1967 to 2016. *Economic Studies at Brookings Institution, August*.

Schotte, S., Zizzamia, R., & Leibbrandt, M. (2018). A poverty dynamics approach to social stratification: The South African case. *World Development*, *110*, 88-103.

Shimeles, A., & Ncube, M. (2015). The making of the middle-class in Africa: Evidence from DHS data. *The Journal of Development Studies*, *51*(2), 178-193.

Slobodenyuk, E. D., & Mareeva, S. V. (2020). Relative poverty in Russia: Evidence from different thresholds. *Social Indicators Research*, *151*(1), 135-153.

Solimano, A. (2008). *The middle class and the development process*. ECLAC.

Stampini, M., Robles, M., Sáenz, M., Ibarrarán, P., & Medellín, N. (2016). Poverty, vulnerability, and the middle class in Latin America. *Latin American Economic Review*, *25*(1), 1-44.

Torche, F. (2010). Social status and public cultural consumption: Chile in comparative perspective. *Social Status and Cultural Consumption*, 109-138.

Weber, M. (1946). "Class, Status, Party." En From Max Weber: Essays in Sociology, ed. Hans H. Gerth y C. Wright Mills. Nueva York: Oxford University Press

Weber, M. (1978). *Economy and society: An outline of interpretive sociology*. University of California press.