PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

FACULTAD DE CIENCIAS SOCIALES

Ni una menos en el mercado laboral: La vulnerabilidad del empleo de las mujeres en el Perú urbano del 2019

Tesis para obtener el título profesional de Licenciada en Economía presentado por:

Riega Escalante, Stephy Rosario

Asesor:

Cozzubo Chaparro, Angelo

Lima, 2025

Informe de Similitud

Yo, *Cozzubo Chaparro, Angelo*, docente de la Facultad de Ciencias Sociales de la Pontificia Universidad Católica del Perú, asesor(a) de la tesis/el trabajo de investigación titulado *Ni una menos en el mercado laboral: La vulnerabilidad del empleo de las mujeres en el Perú urbano del 2019* del/de la autor (a)/ de los(as) autores(as) *Riega Escalante, Stephy Rosario* dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 21%. Así lo consigna el reporte de similitud emitido por el software *Turnitin* el 20/02/2025.
- He revisado con detalle dicho reporte y la Tesis o Trabajo de Suficiencia Profesional, y no se advierte indicios de plagio.
- Las citas a otros autores y sus respectivas referencias cumplen con las pautas académicas.

Lugar y fecha: Lima, 21 de febrero del 2025

Apellidos y nombres del asesor / de la asesora: *Cozzubo Chaparro, Angelo*	
DNI: 47901197	Firma
ORCID: 0000-0002-7838-0256	

Agradecimientos

A mis padres Juan y Rosario, mi hermano Christopher y mis abuelos y abuelas, por acompañarme en todos los pasos que doy y motivarme a seguir adelante. Ustedes son la luz que ilumina mi camino. A mis amigos, por quedarse, por ser mi soporte y alentarme a no rendirme.

A mi asesor, Angelo, por su comprensión infinita y sus valiosos aportes para mantener esta investigación en la frontera del conocimiento. A mi segundo asesor en el bachiller, Javier, por sus investigaciones que inspiraron esta tesis y sus comentarios constructivos. Al profesor Alexander, por la introducción y enseñanza de esta fascinante rama de la Economía. Asimismo, a mi segunda lectora, Giannina, por sus valiosos comentarios que me permitieron mejorara este trabajo.

A todos los autores que he citado, sin sus trabajos nada de esto hubiera sido posible. A la comunidad de programadores y a aquellos que defienden el libre acceso a la información, sin sus trabajos nada de esto hubiera sido posible.

Y a las mujeres, por ser símbolo de resistencia, por persistir dentro y fuera del mercado laboral, y por luchar por un mundo más justo para ellas mismas y para las que les seguimos.

Resumen

La investigación analiza la pérdida de empleo entre mujeres en áreas urbanas del Perú en el 2020, utilizando datos de la ENAHO Panel 2016 - 2018 y un modelo logit penalizado con undersampling confirmando su eficiencia mediante el área bajo la curva ROC. El estudio se enfoca en la vulnerabilidad al desempleo, destacando cómo el mercado laboral precario afecta desproporcionadamente a las mujeres, influenciado por la segregación laboral y las normas de género en las tareas domésticas. El objetivo principal fue identificar a las mujeres que en el 2019 tenían el riesgo de perder su empleo en 2020, proponiendo que más del 18% estaría en esta situación basado en tendencias históricas. Los resultados revelaron que el 34.68% de las mujeres empleadas en 2019 estaban en riesgo. El análisis revela que las variables más influyentes en la predicción de la pérdida de empleo de las mujeres están relacionadas con sus lugares de trabajo y las características de sus hogares. Las pruebas adicionales de robustez del modelo mantuvieron resultados consistentes, sugiriendo una alta vulnerabilidad. La investigación subraya la necesidad de profundizar en el estudio del empleo femenino y explorar políticas públicas efectivas para mitigar impactos en futuras crisis. Se recomienda emplear metodologías avanzadas para una comprensión más completa de las dinámicas laborales y su impacto en políticas públicas.

Palabras clave: Vulnerabilidad del empleo, mujeres, Perú urbano, participación laboral femenina, mercado laboral peruano.

Abstract

The research analyzes job loss among women in urban areas of Peru in 2020, using data from the ENAHO Panel 2016 - 2018 and a penalized logit model with undersampling confirming its efficiency through the area under the ROC curve. The study focuses on vulnerability to unemployment, highlighting how the precarious labor market disproportionately affects women, influenced by labor segregation and gender norms in household chores. The main objective was to identify women who in 2019 were at risk of losing their jobs in 2020, proposing that more than 18% would be in this situation based on historical trends. The results revealed that 34.68% of women employed in 2019 were at risk. The analysis reveals that the most influential variables in predicting women's job loss are related to their workplaces and household characteristics. Additional robustness tests of the model maintained consistent results, suggesting high vulnerability. The research underscores the need to deepen the study of female employment and explore effective public policies to mitigate impacts in future crises. It is recommended that advanced methodologies be employed for a more complete understanding of labor dynamics and their impact on public policies.

Keywords: Employment vulnerability, Women, Urban Peru, Female labor force participation, Peruvian labor market

Índice de contenido

Índice de tablas

Índice de figuras

Introducción

La incorporación de la mujer en el mercado laboral ha sido una revolución silenciosa. Esta, permitió la estabilidad de la participación laboral femenina en la última década; o, al menos, hasta la llegada de la pandemia del Covid-19 en el 2020. Pues, si bien del 2010 al 2019, la tasa de actividad femenina se mantuvo oscilando en 64.2%, esta disminuyó en 11.2% en el 2020 (Rivera et al., 2021). Esto nos lleva a pensar que si bien es cierto que ahora una considerable parte de la masa laboral está compuesta de mujeres (OIT, 2019), el trasfondo de esta incorporación es un tema alarmante.

La relevancia de esta situación en el mercado laboral adquiere particular importancia en la población peruana, pues el empleo y la calidad de este aún presentan desafíos en el Perú. El mercado laboral peruano es caracterizado por ser informal y precario (Herrera & Rosas, 2003), lo cual contribuye a que las remuneraciones sean más limitadas e inestables.

La situación del mercado laboral no solo influye en la esfera ocupacional de los peruanos y peruanas, sino que también tiene un impacto directo en sus niveles de bienestar. Esto se debe a que, en principio, los shocks en los ingresos de un hogar, que influyen en los niveles de bienestar de los hogares en el corto y posiblemente en el largo plazo (Lockshin & Ravallion, 2000), están determinados por las remuneraciones de los trabajadores que lo componen.

Una manera de evaluar esta posible pérdida de bienestar relacionada al empleo es investigar la vulnerabilidad de este, es decir, investigar el riesgo de la pérdida de bienestar a partir de posibles transiciones laborales. Para la presente investigación, estas transiciones abarcaran aquellas que se dan del empleo al desempleo o inactividad. De esta manera, definiremos a la vulnerabilidad del empleo como el riesgo de la pérdida del empleo, un fenómeno anteriormente estudiado en la coyuntura peruana (Herrera & Hidalgo, 2002).

Es relevante notar que los efectos de las transiciones laborales no son iguales para todos, sino que están diferenciados por las desigualdades preexistentes (Durán, 2022). No diferenciar lo suficiente entre los segmentos de individuos y hogares podría estar escondiendo evidencia importante para las poblaciones en mayor riesgo (Durán, 2022).

En particular, es relevante estudiar a las mujeres, un grupo vulnerable fuera y dentro de la esfera laboral, para así poder entender las relaciones laborales

específicas que las afectan de manera específica. Y es que, a raíz de las limitaciones que se les impone en el hogar debido de la división sexual del trabajo doméstico; y la marcada segregación laboral en el mercado laboral (Barba et al., 1997), las mujeres experimentan de manera dispar las dinámicas del mercado laboral que sus pares hombres.

En este sentido, estudiar la vulnerabilidad del empleo de las mujeres implica no solo investigar el riesgo de la pérdida de bienestar de estas mismas, un grupo especialmente expuesto a las condiciones precarias del mercado laboral y de la sociedad; sino también de los hogares a los que pertenecen, como señalan Lockshin & Ravallion (2000).

Para lograr estudiar este fenómeno, podemos considerar que entender la vulnerabilidad es un ejercicio que se puede realizar de reversa: conocer quiénes podrían perder su puesto de trabajo implica identificar a aquellos que ya lo hicieron y a aquellos que se encuentran en riesgo de hacerlo. Realizar este ejercicio demanda un tipo de metodología que nos permita establecer una relación que no sea de casualidad. En cambio, debe ser una relación de predicción de una condición no observable que puede materializarse en el futuro, como lo es la pérdida del empleo.

Muchos problemas de política, como la identificación de los jóvenes con mayor riesgo de violencia o la identificación de los profesores que tendrán el mayor valor agregado en los alumnos para su asignación en los colegios (Kleinberg et al., 2015), se han beneficiado de estos tipos de metodología pues sus resultados tienen grandes efectos en el bienestar de la sociedad, por ejemplo, mediante la relocalización de recursos y aplicación de políticas preventivas (Kleinberg et al., 2015).

De esta manera, el objetivo de esta investigación es identificar a las mujeres en el Perú urbano del 2019 que son vulnerables a la pérdida de su empleo. Con ello, identificamos el porcentaje de mujeres de la PEA urbana pertenecientes al 2019 que tienen una alta probabilidad de perder su empleo, así como también caracterizamos a estas mujeres y encontramos las características de mayor relevancia para la predicción. La hipótesis de la presente investigación será que la proporción de mujeres del 2019 que perderán su empleo es mayor del 18%, que es el porcentaje promedio de mujeres que tienen esta transición entre dos años del 2016 y 2018.

Con este objetivo, se tendrá como fin entrenar a un modelo estadístico disperso de Machine Learning, conocido como Penalized Logit, usando la muestra de remuestreo undersampling para predecir el riesgo de desempleo de las mujeres. Para

este proceso se usará los paneles bianuales apilados construidos a partir de la Encuesta Nacional de Hogares (ENAHO) Panel del 2016 al 2020. Esta base de datos estará compuesta por las mujeres pertenecientes a la Población Económicamente Activa (PEA) y sus características recopiladas por los diferentes módulos (variables del hogar, del empleo e ingresos, de salud, de educación, de sumarias del hogar y de miembros del hogar) con el fin de usar la mayor cantidad de información posible.

Con la finalidad de obtener una base de datos que nos permita mejorar el ajuste del modelo, en paralelo a lo realizado por INEI (2020), la base con una cantidad vasta de variables preseleccionadas según la literatura pasará por etapas secuenciales: descarte por porcentaje de missings que poseen las variables; generación de las categorías de las variables categóricas a dummies, que consiste en el proceso de one-hot-encoding; limitación de los valores outliers de las variables continuas, o winsorizing, y la pre-selección de variables mediante el análisis de sus correlaciones. Este último paso será un filtro que nos permite seleccionar variables previas a la estimación del modelo disperso, el cual también cumple con esta función. Estas variables pasarán por un proceso de estandarización previo al entrenamiento y evaluación del modelo.

Luego de esto, se realizará la división de datos en la muestra de entrenamiento (training set) y la muestra de prueba (test set) con el objetivo de evitar la contaminación de los datos al momento de entrenar el modelo. Si bien todos los paneles bianuales apilados pertenecen a las observaciones cuyos años iniciales son del 2016 al 2019, solo las observaciones con año inicial del 2016 al 2018 se usarán como parte de la muestra de entrenamiento. Para la muestra de validación se seleccionará sólo a las observaciones cuyo año inicial es el 2019, pues es sobre estas mujeres que se desea saber si pierden o no su empleo en el 2020.

Con el propósito de encontrar el precio al cual se intercambia la varianza y el sesgo en el término regularizador del modelo (λ) y tomando en consideración la complejidad distribucional de la muestra, aplicaremos al primer grupo de observaciones el método conocido como Stratified Group Cross Validation (CV). Asimismo, para tomar en cuenta el desbalance de la muestra, usaremos un método de remuestreo llamado undersampling.

Esto nos ayudará a reducir la sobre predicción de la clase mayoritaria de nuestra variable resultado (desempleo/inactividad o empleo). Se usará el área debajo de la curva ROC para evaluar el modelo, ya que balancea los verdaderos y falsos

positivos. Esto asegura una detección eficaz de casos relevantemente vulnerables y una asignación óptima de recursos. El resultado de este entrenamiento será hallar el hiper parámetro ideal.

Con este insumo, realizaremos la estimación de la curva ROC para poder hallar el umbral óptimo que optimice el rendimiento del modelo. Con estos dos parámetros, realizaremos la estimación del modelo disperso Logit Penalizado usando la muestra de validación. Esto nos permitirá la evaluación del rendimiento del modelo al comparar lo que se predijo en el 2020 con las mujeres a la PEA urbana en el 2019, con lo que realmente pasó. Limitar nuestra muestra de validación a las observaciones del panel con año inicial 2019 nos permite que esta validación sea efectivamente out-of-sample y out-of-time.

Asimismo, con el resultado de la estimación podremos identificar a las mujeres de la PEA urbana pertenecientes al 2019 que perderían su empleo en el 2020 según el modelo, y así conocer el porcentaje que representan. Esto nos permite comprobar si es que nuestra hipótesis nula era correcta, es decir, comprobar si es que más del 18% de estas mujeres hubiera perdido su empleo entre estos años. Asimismo, nos permite crear un perfil de estas mujeres, encontrando cuáles fueron las características que demostraron ser más relevantes en el proceso de predicción.

Para evaluar la robustez de esta estimación, usaremos otro método de remuestreo llamado Synthetic minority over sampling technique o SMOTE en el proceso de entrenamiento del modelo; mientras que, para evaluar la sensibilidad del modelo, cambiaremos el umbral que se usa para la clasificación del posible estado laboral de las mujeres. El objetivo de estas pruebas será comprobar que sus resultados aún sustentan la hipótesis nula.

Esta investigación busca aportar en la literatura relacionada al tema del empleo en el contexto peruano, en especial, en tiempo de crisis; busca aportar en el avance metodológico de la identificación de la vulnerabilidad del empleo, y busca aportar en la toma de decisiones de política pública para prevenir y atenuar la pérdida de empleo en las mujeres en el área urbana.

El presente estudio contribuye en la literatura pues permite profundizar en el limitado estudio de la pérdida del empleo en tiempos de incertidumbre (2020). Más aún, al restringirse a ser una investigación sobre la población femenina y en el contexto urbano, abarca poblaciones no estudiadas de manera desagregada y cuyas condiciones son usualmente precarias en el mercado laboral (Lavado & Campos,

2021).

Asimismo, esta investigación aporta a la literatura de metodologías para la estimación de la vulnerabilidad usando relevantes modelos de machine learning y métodos de remuestreo (Blagus & Lusa, 2013). Esto, con el fin de usar el modelo más adecuado para el ejercicio de clasificación, así como para lidiar con los problemas inherentes del contexto laboral peruano, como el desbalance de clases en el estado laboral de las mujeres. Particularmente, este estudio colabora en el debate de la importancia del uso de métricas adecuadas que nos permitan políticas públicas eficientes.

En términos de política, contribuye a la mejor ejecución y planificación de políticas que promuevan roles de género equitativos. Pues, al encontrar cuántas son las mujeres vulnerables, es posible planificar y asignar recursos para mitigar y prevenir su pérdida de bienestar. De la misma manera, identificar qué caracteriza a estas es esencial para la focalización de programas. Más aún, encontrar un subgrupo de variables clave que permiten caracterizar y clasificar a estas mujeres nos permite obtener resultados similares con menor información (Kleinberg et al., 2015). En situaciones de recuperación ante una crisis, todos estos logros permiten el uso eficiente del presupuesto público del Estado peruano.

Capitulo 1. Marco teórico

Bosch y Esteban-Pretel (2012) construyen un modelo de tiempo continuo de búsqueda y emparejamiento en el mercado laboral para tratar de explicar las transiciones laborales cuando existen empleos formales e informales. Este modelo, que utilizaremos como marco teórico de la presente investigación, busca estar alineado con 3 hechos estilizados que están presentes en el contexto peruano.

Uno de estos es la existencia de importantes traslados entre los trabajos formales regulados y los informales no regulados cuando se produce un ajuste en el ciclo económico (i). En este modelo, la tasa de desempleo es baja, mientras que el empleo informal conforma la mayor parte del empleo. Dado todo esto, la probabilidad de transición del sector informal al formal es altamente procíclica, pero menos volátil que la transición del desempleo.

Además, el modelo también busca estar alineado con que (ii) gran parte de las transiciones del desempleo pueden ser atribuidas a los cambios en la tasa de separación del trabajo. Más aún, la volatilidad del desempleo está explicada por los cambios en esta tasa asociados a los empleos informales. Por último, también buscar estar alineada con el hecho de que (iii) las tasas de transición del desempleo o empleo informal al empleo formal acaparan la mayor cantidad de transiciones que se dan relacionados a este y son procíclicas. No obstante, los cambios en el empleo formal suceden dentro de grupos similares de industrias, ocupaciones y sexo. En esta medida, la firma podrá tomar la decisión endógena de contratar a trabajadores de manera formal o informal, dependiendo de la calidad del emparejamiento y los *trade-offs* entre ambos.

Estas características pueden verse presente en la realidad peruana. Como hemos mencionado, un periodo de crecimiento aumenta las posibilidades de obtener un empleo formal (Morales et al., 2010; Rodriguez & Rodriguez, 2011). Dado que ha sido comprobado para Perú la presencia de un mercado laboral altamente informal que explique una baja tasa de desempleo (Morales et al., 2010), la probabilidad de transicionar de un empleo al desempleo o inactividad es particularmente alta en zonas urbanas (Herrera & Rosas, 2003; Morales et al., 2010; Rodriguez & Rodriguez, 2011).

A pesar de que se espera que, para la población vulnerable, como las mujeres, exista una nueva transición al empleo informal (Garavito, 2010; Herrera & Rosas, 2003); vale la pena recordar que esto es condicional a que las mujeres busquen empleo, a la disponibilidad de vacantes de puestos de trabajos y a que se cumplan los

requisitos para estos. Como hemos revisado, muchas mujeres priorizan las labores domésticas a las labores productivas (Barba et al., 1997) o no cuentan con las demandas del mercado laboral por estar envueltas en condiciones que disminuyen su crecimiento educativo y profesional (Garavito, 2010).

El modelo de Bosch y Esteban-Pretel (2012) se construye a partir de Mortensen y Pissarides (1994a), quienes plantean un modelo de creación y destrucción de empleo endógenos usando el enfoque de emparejamiento para encontrar el equilibrio del desempleo, y para la determinación del salario, modelando la búsqueda de empleo. Al igual que Mortensen y Pissarides (1994b), existen, por un lado, los desempleados que buscan empleo activamente y aquellos que ya se encuentran trabajando; mientras que por el otro lado se encuentran las empresas que tienen puestos de trabajo ocupados, o que están vacantes y en búsqueda de un trabajador que la cubra. El modelo también considera que los trabajadores informales puedan estar en la búsqueda de empleos formales con mejores condiciones.

Ambos, trabajadores y firmas, invertirán tiempo y recursos para conseguir un empleo o llenar la vacante del puesto de trabajo (ii). Sin embargo, existirán condiciones que ambos impongan (las condiciones del trabajo y las cualidades de productividad del trabajador) que determinarán que el mercado de desempleados y puestos de trabajo vacantes se llene o no (Mortensen y Pissarides, 1994a). Debido a la existencia de fallas en el mercado, como información asimétrica, este mercado de trabajadores formales e informales y vacantes no se vaciará completamente. El salario de reserva, la ganancia esperada de la búsqueda u otras variables ejemplifican estas fallas.

El modelo (iii) permitirá transiciones directas del sector formal al informal. Los trabajadores y empresas escogen su relación contractual óptima dadas las cualidades del emparejamiento, las características del posible empleo formal o informal y los incentivos provistos por los hacedores de política. El empleo formal es definido como altamente productivo, mientras el empleo informal no solo es bajamente productivo, sino que también es voluntario e integrado con el sector formal de la economía. También se sostiene que en la parte expansiva del ciclo se producen shocks de productividad, se da una mayor creación de vacantes y el incremento de emparejamientos entre estos y los trabajadores. Asimismo, se da menos destrucción de trabajos (Mortensen y Pissarides, 1994a).

En la parte expansiva del ciclo también aumenta el uso de contratos formales, lo cual incrementa el emparejamiento en trabajos formales para los desempleados y

trabajadores informales. Los trabajadores formales trabajan en empresas reguladas por el Estado, mientras que las informales no. En este sentido, se considera a estos últimos como autoempleados que no son profesionales y trabajadores asalariados que no tienen una tarjeta de trabajo; y que existen ya que para algunas empresas es óptimo no tener regulaciones laborales.

Los emparejamientos en el mercado laboral entre una empresa y un trabajador se dan de manera aleatoria y de acuerdo a la función $m = m(u, v)$, en el cual u representa el numero de desempleados y v el numero de vacantes disponibles. El emparejamiento entre una vacante y un desempleado se da según el proceso de Poisson con tasa de acontecimiento $q(\theta) = m(u, v) \ / \ v$, donde la presión del mercado es $\theta = v/u$; mientras que la tasa de acontecimiento de una vacante es de $q(\theta) = m(u, v) \ / \ u$ con el mismo proceso.

La productividad del trabajador es revelada sólo después que se produce el emparejamiento y esto genera heterogeneidad entre los trabajadores. Dependiendo de esta, la firma decide ofrecer un trabajo formal o informal. La productividad está compuesta por p y ε, un componente agregado común a todas las firmas de la misma economía y un componente idiosincrático que es particular a cada empresa y es extraído de la distribución $G: \left[\varepsilon_{min}^{G}, \varepsilon_{max}^{G}\right] \rightarrow [0, 1]$.

Las firmas que incurren en contratos formales tienen que cumplir con las regulaciones laborales, es decir, con c ,el costo de contratación de establecerse la relación; τ, la tasa de impuestos, y F, el costo de despido. Existe una tasa de acontecimiento mayor de un contrato informal para los trabajadores informales, aunque, debido a la existencia de regulaciones laborales, este tipo de contrato sea menos estable. Sin embargo, se tendrá mayor estabilidad en tanto se permita que los trabajadores informales puedan estar buscando nuevas oportunidades laborales. En esta medida, existe una tasa de acontecimiento de que una firma sea descubierta con empleo informales, $\emptyset$; y una multa σ que debe pagar luego de terminar la relación laboral informal.

Los nuevos emparejamientos, que son $i.i.d,$ entre firmas y en el tiempo, determinan nuevos niveles de productividad de acuerdo con un proceso de Poisson con tasa de acontecimiento λ_i tal que es mayor para los trabajadores formales que informales. Ambos, $j \in \{f, i\},$ salen de una distribución diferente para cada uno, $H_j: \left[\varepsilon_{min}^{H_j}, \varepsilon_{max}^{H_j}\right] \rightarrow [0, 1]$, lo cual permite su variabilidad de acontecimiento ante shocks.

Al igual que en Mortensen y Pissarides (1994), será la presencia de shocks idiosincráticos de productividad lo que determinará los flujos del empleo. En comparación con este modelo, Bosch y Esteban-Pretel (2012) no solo presenta flujos entre empleo y desempleo, sino también entre empleos formales e informales.

Primero se determinan los trabajos de manera endógena, luego los flujos entre empleo formales e informales en los cuales los trabajadores informales buscaran mejores oportunidades. En esta medida, a partir del problema de la firma y del trabajador, podremos determinar los salarios y excedentes óptimos tal que podemos determinar el estado estacionario.

1.1. El problema de la firma

El problema de la firma este compuesto por la determinación del valor de una vacancia, así como de las relaciones laborales formales e informales para la firma. V es el valor presente descontado de publicación de una vacante para una firma, $J_f^l(\varepsilon)$ el valor de un trabajo formal ocupado, tal que $l \in \{n, 0\}$ identifican un emparejamiento nuevo o ya establecido; y $J_i(\varepsilon)$ es el valor de un trabajo informal.

En este sentido, podemos plantear el valor presente de abrir una vacante a una tasa r como equivalente al costo de apertura k, la ponderación de la tasa de que la firma se encuentre con potenciales empleados $q(\theta)$ con una de los tres posibles resultados en el valor generados de este: la formación de una nueva relaciona laboral formal del contrato y su costo, $J_f^n(\varepsilon') - c$; contratación informal $J_i(\varepsilon')$, o retener la vacante, V. Asimismo, existe la probabilidad de que no se de este encuentro tal que se mantenga la vacante, $q(\theta)V$. Esta ecuación la podemos plantear de la siguiente manera:

$$rV = -k + q(\theta) \int_{\varepsilon_{min}^G}^{\varepsilon_{max}^G} \max \left[J_f^n(\varepsilon') - c, J_i(\varepsilon'), V\right] d\, G(\varepsilon') - \, q(\theta)V$$

Cabe resaltar que la existencia de un costo de una nueva relación laboral c nos permite diferenciar entre la el valor de una relación formal nueva $J_f^n(\varepsilon)$ y una existente $J_f^o(\varepsilon)$ que tendrán salarios diferentes puesto que no existirá un costo de despido si es que la nueva relación laboral formal no se establece. Por otro lado, el valor presente de la relación laboral formal a la tasa r para la firma se puede plantear como:

$$rJ_f^l(\varepsilon) = p + \varepsilon - (1+\tau)\, w_f^l\,(\varepsilon) + \lambda_f \int_{\varepsilon_{min}^{H_f}}^{\varepsilon_{max}^{H_f}} \max\left[J_f^0(\varepsilon'), J_i(\varepsilon') - F, V - F\right] d\, H_f(\varepsilon') - \lambda_f J_f^l(\varepsilon)$$

$$s.a.\, l \in \{n, o\}$$

Este valor es equivalente a la suma de la productividad general p, la productividad específica ε, el costo de los salarios y sus impuestos a la tasa τ, la ponderación de la probabilidad de aparición de un nuevo nivel de productividad idiosincrático formal λ_f que se distribuye de acuerdo a un proceso de Poisson con el valor de uno de los tres posibles escenarios: mantener la relación formal, $J_f^n(\varepsilon')$, el reemplazo de esta por una relación informal, $J_i(\varepsilon')$, y la disolución del emparejamiento y la creación de una vacancia luego de pagar el costo de despido.

Asimismo, también podemos estimar el valor presente de una relación laboral informal a la tasa r como el equivalente de la suma de p, ε, el gasto del salario $w_i\,(\varepsilon)$ y la ponderación de la probabilidad de aparición de un nuevo nivel de productividad idiosincrático informal λ_i que se distribuye de acuerdo a un proceso de Poisson con el valor de uno de los escenarios: pagar el costo de contratación y formalizar la relación, el mantener la relación informal, $J_i(\varepsilon')$, y la disolución del emparejamiento y la creación de una vacancia. Sin embargo, también existe la posibilidad de que los trabajadores informales cambien a trabajos de mayor calidad a la tasa η , o que, según proceso de Poisson, exista la posibilidad que un empleo informal sea descubierto a la tasa ϕ y paguen una multa σ. Esto se presenta en la siguiente ecuación:

$$rJ_i(\varepsilon) = p + \varepsilon - w_i\,(\varepsilon) + \lambda_i \int_{\varepsilon_{min}^{H_i}}^{\varepsilon_{max}^{H_i}} \max\left[J_f^n(\varepsilon') - c, J_i(\varepsilon'), V\right] d\, H_i(\varepsilon') - \lambda_f J_i(\varepsilon') + \eta(V - J_i(\varepsilon')) + \phi(V - J_i(\varepsilon')) - \phi\sigma$$

1.2. El problema del trabajador

Por otro lado, también tenemos las ecuaciones que representan el valor de un emparejamiento y el desempleo para los trabajadores. El valor presente del desempleo U para los trabajadores a la tasa r es equivalente al ingreso o valor que

obtiene el individuo en el desempleo, b; la ponderación de la tasa de encuentro de empleo $\theta q(\theta)$ con el valor de uno de los posibles escenarios laborales: el salario de trabajador formal $W_f^n(\varepsilon')$, de uno informal $W_i(\varepsilon')$ o mantenerse desempleado. Esto se puede ver representado en la siguiente ecuación:

$$rU = b + \theta q(\theta) \int_{\varepsilon_{min}^G}^{\varepsilon_{max}^G} \max\left[W_f^n(\varepsilon'), W_i(\varepsilon'), U\right] d\, G(\varepsilon') - \, q(\theta)U$$

Dependiendo del tipo de contrato y la productividad del emparejamiento que reciba gozara de un tipo de ganancias. Por un lado, de obtener un contrato formal, el valor presente de estas ganancias a la tasa r será el salario formal $w_f^l(\varepsilon)$ más la probabilidad de un cambio idiosincrático de la productividad λ_f con el valor de la ganancia de uno de los tres escenarios: mantener el tipo de contrato con la firma $W_f^0(\varepsilon')$**,** tratar de cambiarlo $W_i(\varepsilon')$, o mantenerse desempleado, U. Esto se puede denotar en la siguiente ecuación:

$$rW_f^l = w_f^l(\varepsilon) + \lambda_f \int_{\varepsilon_{min}^{H_f}}^{\varepsilon_{max}^{H_f}} \max\left[W_f^0(\varepsilon'), W_i(\varepsilon'), U\right] d\, H_f(\varepsilon') - \lambda_f W_f^l(\varepsilon), l \in \{n, 0\}$$

Por otro lado, de obtener un contrato informal, el valor de este trabajo será igual al salario $w_i(\varepsilon)$ más la probabilidad de un cambio idiosincrático de la productividad λ_i con el valor de la ganancia de uno de los tres escenarios: mantener el tipo de contrato con la firma $W_i(\varepsilon')$ **,** tratar de cambiarlo $W_f^n(\varepsilon')$, o mantenerse desempleado, U. Sin embargo, también se encuentra la posibilidad de que busquen empleos mientras están trabajando, teniendo una tasa de emparejamiento $\chi\theta q(\theta)$, con una eficiencia menor a la de los trabajadores formales tal que $\chi < 1$, o que exista una ruptura laboral a la tasa ϕ debido a la fiscalización.

Es posible contextualizar estas ecuaciones a las decisiones sobre el valor del desempleo y el empleo informal para las mujeres peruanas en el área urbana. El ingreso que recibe del desempleo (b) será escaso o nulo puesto que sus precarias condiciones laborales no se lo permitirán (Barba et al., 1997). Sin embargo, un ingreso implícito por sus labores domésticas no remuneradas en el hogar estará presente en la ecuación.

Las mujeres considerarán la posibilidad de obtener el salario de empleo formal $W_f^n(\varepsilon')$; sin embargo, dada la exclusión social a la que ha sido expuesta, su capital humano será limitado (Garavito, 2010). Esto hará que las probabilidades de obtener el salario de un empleo informal $W_i(\varepsilon')$ sean altas y que lo sean aún más en períodos recesivos. Asimismo, dadas las condiciones de flexibilidad que un empleo informal brinda, puede que las mujeres tengan que sacrificar condiciones laborales decentes por esto.

Más aún, las mujeres están condicionadas a priorizar las labores en el hogar (Barba et al., 1997). En este sentido, mantenerse en un empleo informal, o en el desempleo o inactividad puede resultar más valioso para ellas. Sin embargo, esto sólo podrá ser sostenido en el tiempo si es que el hogar aún tiene niveles de ingreso lo suficientemente altos como para la subsistencia (Jaramillo y Ñopo 2020 en Durán, 2022).

En este sentido, podemos argumentar que, si la necesidad de la realización de labores domésticas es alta y se cuentan con los recursos para ello en el hogar, entonces valoraran más estar desempleadas o en inactividad. Sin embargo, si es que alguna de estas condiciones no se cumple, se verán en la necesidad de buscar trabajo. Sin embargo, este emparejamiento se dará si es que ambas partes están satisfechas con las condiciones.

Con esto podemos definir los excedentes y los salarios que nos servirán para definir el estado estacionario. Los excedentes son las ganancias netas luego de que se realicen los emparejamientos y, debido a la existencia de costos de contratación y de despido, serán diferentes según el tipo de contrato que se tenga. Entonces, podemos definir al excedente del emparejamiento de los trabajadores formales ya establecido $S_f^n(\varepsilon)$, de los nuevo $S_f^0(\varepsilon)$, y de los trabajadores informales $S_i(\varepsilon)$ como:

$$S_f^n(\varepsilon) = (J_f^l(\varepsilon) - c) + W_f^n(\varepsilon) - V - U$$
$$S_f^0(\varepsilon) = J_f^0(\varepsilon) + W_f^0(\varepsilon) - (V - F) - U$$
$$S_i(\varepsilon) = J_i(\varepsilon) + W_i\,(\varepsilon) - V - U$$

Los salarios son elegidos a partir de un problema de negoción con solución de Nash cuando sucede un encuentro entre una firma y el trabajador o cuando se da un shock idiosincrático. Es decir, son las firmas y los empleadores los que definen el tipo de contrato que se forma, así como cuando se destruye de formarse uno. En esta

solución influyen el poder de negociación del trabajador formal β_f y el informal β_i y los impuestos en los salarios formales. $\beta_i < \beta_f$. Así, tenemos que los excedentes para los nuevos trabajadores formales, para los antiguos trabajadores formales y para los trabajadores informales serán distribuidos respectivamente como:

$$(J_f^l(\varepsilon) - c) - V = \frac{(1-\beta_f)(1+\tau)}{\beta_f}\ (W_f^n\ (\varepsilon) - U)$$

$$J_f^0(\varepsilon) - (V - F) = \frac{(1-\beta_f)(1+\tau)}{\beta_f}\ (W_f^0\ (\varepsilon) - U)$$

$$J_i(\varepsilon) = \frac{(1-\beta_i)}{\beta_i}\ (W_i\ (\varepsilon) - U)$$

Con estas podemos obtener los salarios para los nuevos trabajadores formales, para los antiguos trabajadores formales y para los trabajadores informales:

$$W_f^n\ (\varepsilon) = \ (1-\beta_f)b + \frac{\beta_f}{(1+\tau)}[p + \varepsilon + \theta k - (r + \lambda_f)c - \lambda_f F]$$

$$W_f^0\ (\varepsilon) = \ (1-\beta_f)b + \frac{\beta_f}{(1+\tau)}[p + \varepsilon + \theta k]$$

$$W_i\ (\varepsilon) = \ (1-\beta_i)b + \beta_i[p + \varepsilon - \phi\sigma + (1-\chi)\theta k]$$

1.3. Equilibrio en el estado estacionario

Este equilibrio tiene 5 variables endógenas y 5 condiciones que lo determinan: una ecuación de la creación del empleo, las condiciones de los flujos de entrada y salida de trabajadores formales e informales, y las ecuaciones de destrucción de empleos formales e informales.

Para la creación de la ecuación de empleo, se establecen 2 condiciones con respecto al proceso de contratación. Estas nos indican que existirá un nivel de productividad idiosincrática ε_R que hará a la firma indiferente entre contratar un trabajador formal, ya sea uno que estaba desempleado o en el sector informal, e informal. Asimismo, nos indica que existirá un nivel de productividad idiosincrática ε_T

que hará a la firma indiferente entre transformar un trabajo formal a uno informal luego de pagar por el costo de despido y mantener un trabajado informal. Así, se plantean como:

$$J_f^n(\varepsilon_R) - c = J_i(\varepsilon_R)$$

$$J_f^n(\varepsilon_T) + F = J_i(\varepsilon_T)$$

Estas nos indican que existirá un nivel de productividad idiosincrática ε_R que hará a la firma indiferente entre contratar un trabajador formal, ya sea uno que estaba desempleado o en el sector informal, e informal. Asimismo, nos indica que existirá un nivel de productividad idiosincrática ε_T que hará a la firma indiferente entre transformar un trabajo formal a uno informal luego de pagar por el costo de despido y mantener un trabajado informal. Con respecto al proceso de ruptura laboral, tenemos 2 condiciones. Estas son las siguientes:

$$J_f^0\left(\varepsilon_{d_f}\right) + F = 0$$
$$J_i(\varepsilon_{d_i}) = 0$$

Estas nos indican que existe un nivel de productividad ε_{d_f} en el cual una relación laboral formal se hace 0, mientras que el nivel ε_{d_i} hace lo mismo para una relación informal. Por último, la quinta condición de entrada libre, $V = 0$, nos indica que la creación de vacancias por parte de las firmas se dará hasta cuando crear una vacante extra tenga un valor descontado de 0 tal que se iguala el costo esperado de la vacancia a su retorno esperado. Usando esta última condición más las dos anteriores, el equilibrio del modelo esta compuesta por las siguientes 5 ecuaciones:

$$\frac{1+\tau(1-\beta_f)}{1+\tau}\left(\frac{\varepsilon_R - \varepsilon_{d_f}}{r+\lambda_f} - c - F\right) = \frac{\varepsilon_R - \varepsilon_{d_f}}{r+\lambda_f+\chi\theta q(\theta)+\phi}$$

$$\frac{1+\tau(1-\beta_f)}{(1+\tau)(r+\lambda_f)}\left(\varepsilon_T - \varepsilon_{d_f}\right) = \frac{\varepsilon_T - \varepsilon_{d_f}}{r+\lambda_i+\chi\theta q(\theta)+\phi}$$

$$\varepsilon_{d_f} = -p + (1+\tau)b - rF + \frac{\beta_f}{(1-\beta_f)}\theta k$$

$$-\lambda_f \frac{1+\tau}{1+\tau(1-\beta_f)} \left[\int_{\varepsilon_{d_f}}^{\varepsilon_T} \mathrm{S_i}(\varepsilon')\mathrm{dH_f}(\varepsilon') + \int_{\varepsilon_T}^{\varepsilon_{max}^{H_f}} S_f^0(\varepsilon')\mathrm{dH_f}(\varepsilon') \right]$$

$$\varepsilon_{d_i} = -p + b + \phi\sigma + (1-\chi)\theta k - \lambda_i \left[\int_{\varepsilon_{d_i}}^{\varepsilon_R} \mathrm{S_i}(p,\varepsilon')\mathrm{dH_i}(\varepsilon') + \int_{\varepsilon_R}^{\varepsilon_{max}^{H}} S_f^n(p,\varepsilon')\mathrm{dH}_i(\varepsilon') \right]$$

$$\frac{k}{q(\theta)} = (1-\beta_i) \left[\int_{\varepsilon_{min}^{G}}^{\varepsilon_R} \mathrm{S_i}(\varepsilon')\mathrm{dG}(\varepsilon') + \int_{\varepsilon_R}^{\varepsilon_{max}^{G}} S_f^n(\varepsilon')\mathrm{dG}(\varepsilon') \right]$$

Capitulo 2. Revisión de literatura

La revisión de literatura estuvo compuesta por dos procesos de búsqueda. Por un lado, se realizó una búsqueda de publicaciones e investigaciones en bases de datos y páginas web institucionales relevantes para el tema. Este estuvo compuesto en su mayoría por *papers*, documentos de trabajo, tesis publicadas, etc.

Por otro lado, se realizó una revisión sistemática mediante el uso del metabuscador Scopus para la extracción de investigaciones *peer-reviewed*, siguiendo la metodología de Cozzubo y Herrera (2021). Para más detalle sobre el flujograma y distribución de los *papers* considerados en el proceso, así como los *papers* que fueron incluidos finalmente, revisar el Anexo A. La justificación para esta última metodología se debe a la brevedad del estudio en el Perú y en el mundo sobre la vulnerabilidad del empleo y, en específico, sobre este fenómeno en las mujeres.

Es relevante resaltar que, si bien estamos realizando una revisión de literatura de la vulnerabilidad del empleo, en la presente investigación nos centraremos en uno de sus elementos: la transición laboral del empleo al desempleo. Ambos componentes de la vulnerabilidad del empleo y las variables que los afectan parecen ser lo suficientemente diferentes como para decidir centrarnos solo en uno. Una definición acotada acerca del fenómeno que predecimos permitirá un uso adecuado de los modelos y métodos.

Así, con ambas metodologías de revisión de literatura, encontramos que las condiciones del empleo en el Perú, a nivel agregado y descompuesto, han sido ampliamente estudiadas de manera estática debido a que es una de las preocupaciones más alarmantes en la realidad peruana. Sin embargo, se ha estudiado en menor medida el empleo de manera dinámica, las características de aquellos vulnerables a la pérdida del empleo, y el empleo para grupos particulares de la población. Particularmente, es en esta intersección de huecos en la literatura en la cual la presente investigación busca presentar evidencia novedosa.

2.1. Transiciones laborales

El estudio de la visión dinámica del empleo en el Perú ha tratado de explicar las transiciones laborales de la PEA en el mercado laboral. Así, se ha estudiado entre qué partes del mercado laboral se han dado y cuáles son las características de los trabajadores que las han experimentado.

Por un lado, Morales et. al (2010) ha estudiado las transiciones entre sectores formales e informales. Así, se ha encontrado que no son muy frecuentes las

transiciones de trabajadores entre estos sectores en periodos de crecimiento sostenido: la tasa de rigidez informal se sostuvo en un promedio de 79,7% por 10 años. Sin embargo, en periodos de crecimiento económico, tienden a mejorar las oportunidades de empleo formal, pero los períodos recesivos conducen a una mayor vulnerabilidad laboral, con un aumento en la informalidad laboral (Morales et al., 2010). De la misma manera, pertenecer al sector de construcción o transporte, ser mujer, no ser jefe del hogar y tener bajos niveles educativos aumenta la probabilidad de pasar a la informalidad (Morales et al., 2010).

Por otro lado, autores como Herrera y Rosas (2003) y Rodriguez y Rodriguez (2011) han investigado las transiciones entre los estados del trabajador con respecto a estar fuera o dentro del mercado laboral. Así, han encontrado que la movilidad laboral es alta en Perú (Herrera & Rosas, 2003; Rodriguez & Rodriguez, 2011). Sin embargo, es diferente entre el área urbana y rural, siendo la movilidad laboral más alta en la primera (Herrera & Rosas, 2003; Mora, 2018). De la misma manera, también es diferente entre Lima Metropolitana y el resto del país (Morales et al., 2010).

Así, se ha encontrado que aquellos que se encuentren empleados se mantendrán así por un considerable periodo de tiempo; aquellos que caigan en el desempleo no se mantendrán así por mucho tiempo y pasaran a subemplearse, y que aquellos que sí se vuelvan desempleados crónicos pasarán a la condición de inactivos (Rodriguez & Rodriguez, 2011).

La transición más importante, especialmente en el área urbana, es del empleo a la inactividad o el desempleo (Herrera & Rosas, 2003; Morales et al., 2010; Rodriguez & Rodriguez, 2011). Esta se puede explicar en cierta medida porque la movilización que se da es la de la mano de obra secundaria (Herrera & Hidalgo, 2002) compuesta en su mayoría por mujeres, lo cual las hace más propensas a este tipo de transición (Garavito, 2010; Herrera & Rosas, 2003). Asimismo, se encuentra que a los trabajadores que pasan por un periodo de desempleo e inactividad se les reduce sus beneficios, como ingresos o derechos laborales, al encontrar un nuevo empleo (Yamada, 2007).

Sin embargo, ante la falta de alternativas en el empleo formal, la población que pertenece a sectores vulnerables (los más jóvenes, con niveles educativos más bajos, y a menudo trabajan en sectores inestables) suelen pasar al empleo informal para poder subsistir (Morales et al., 2010; Rodriguez & Rodriguez, 2011).

Ambos tipos de transiciones, entre estados del mercado laboral y entre sector

formal e informal, dependen de las características de los trabajadores (sociodemográficas, de su hogar y de capital humano) y de su posición en el mercado laboral (informal/formal, tipo de ocupación, empleado/desocupado/inactivo, etc.) al momento de realizarse las transiciones (Rodriguez & Rodriguez, 2011). Asimismo, la condición del ciclo económico en que se encuentran los trabajadores parece afectar ambos tipos de transiciones (Morales et al., 2010; Rodriguez & Rodriguez, 2011).

De esta forma, en cuanto a las características de los individuos, los que tienen menos años de educación y más jóvenes tienen más probabilidades de rotar entre estados del mercado laboral, especialmente hacia el desempleo (Mora, 2018; Rodriguez & Rodriguez, 2011). En el área urbana, ser del sexo femenino aumenta las probabilidades de entrar en la inactividad (Herrera & Rosas, 2003; Morales et al., 2010; Rodriguez & Rodriguez, 2011). Asimismo, pertenecer al sector comercio o de construcción aumenta esta probabilidad (Morales et al., 2010).

Un mayor número de trabajadores, así como de horas de trabajo afectan negativamente la probabilidad de desempleo. Asimismo, hay más probabilidades de pasar al desempleo con relación a mantenerse empleado si el trabajador se encuentra en un sector con mano de obra no calificada (Mora, 2018).

En cuanto a la temporalidad del mercado laboral, se ha encontrado que la recesión de 1997 estuvo relacionada con el aumento de transiciones del empleo a la inactividad, especialmente en el área urbana. Existió, en este periodo, una distinción más marcada por sexo en el área rural. Sin embargo, se encontró una mayor transición hacia el empleo por parte de las mujeres (Herrera & Rosas, 2003).

Esto no sucedió con la más reciente crisis del Covid-19, en la cual se vio una drástica transición del empleo al desempleo en el área urbana como rural, siendo mayor en la primera. En comparación con la crisis de 1997, existió una distinción más marcada en los niveles de empleo por sexo en el área urbana. En esta crisis, las trabajadoras mujeres no han podido retornar a sus niveles de empleo prepandémicos: la recuperación está pasando lentamente o no está pasando en lo absoluto (Durán, 2022).

Cabe mencionar que el problema del acceso a los datos de transiciones laborales ha sido solucionado parcialmente con la introducción de la ENAHO Panel en 1996 (Herrera & Rosas, 2003; Morales et al., 2010) y la Encuesta Permanente del Empleo (EPE) para Lima Metropolitana. Aun así, la primera es limitada en el periodo de seguimiento, mientras que la segunda en las características de los trabajadores y

sus hogares.

La presente investigación aporta nuevas perspectivas sobre las transiciones laborales del empleo al desempleo e inactividad en grupos demográficos específicos. Busca actualizar el conocimiento para la gran parte de la población que reside en el área urbana. Asimismo, busca estudiar este fenómeno en las mujeres, un grupo vulnerable dentro y fuera del mercado laboral. Ambos grupos demográficos no han sido estudiados particularmente en los últimos años de manera independiente o conjunta, y esta investigación busca llenar el vacío de literatura en esta intersección.

2.2. La vulnerabilidad del empleo

La vulnerabilidad es definida como el riesgo de un hogar a reducir su bienestar. Esta es medida como una probabilidad (existen hogares que tendrán poca o mucha probabilidad) que aumenta con el horizonte temporal (Pritchett et al., 2000). Esta definición ha sido usada al estudiar la pobreza y el mercado laboral. Por un lado, en Perú, se ha tratado de explicar el riesgo de caer en la pobreza monetaria a partir de la identificación de una línea de vulnerabilidad (Herrera & Cozzubo, 2016; INEI, 2020).

Sin embargo, también se ha usado esta definición para el mercado laboral en Perú y en otros países. Han existido 3 grupos estudios en los que se abordado la vulnerabilidad del empleo: estudiando a los trabajadores independientes y trabajadores familiares no pagados como empleados vulnerables (OIT, 2010), estudiando a aquellos con ingresos bajos y estudiando a aquellos a los que les falta cumplir con un conjunto de dimensiones relacionadas a la vulnerabilidad en el trabajo (Bazillier et al., 2015). Nosotros estudiaremos el elemento de pérdida del empleo que coincide con el último grupo de estudio mencionado.

Existen tres opciones para la identificación de grupos vulnerables en los mercados de trabajo para este grupo de estudio: estimar un indicador único, un set de indicadores, y el estado del trabajador en conjunto con indicadores adicionales como los sectoriales. Esta última aproximación es poderosa al entender cambios estructurales, a pesar de aún estar lejos de explicar la complejidad del mercado laboral (Sparreboom & de Gier, 2008).

Así, apelando a esta última manera de identificación y la multidimensionalidad de la vulnerabilidad del empleo, se ha interpretado a esta como “la probabilidad de que un trabajador pierda su empleo, o que encuentre una ocupación con condiciones de trabajo inferiores a las de su ocupación previa” (Garavito, 2010; Herrera & Hidalgo, 2002). En otros casos, se ha sido más explícito con esta definición al considerar la

última transición como una que va del sector formal al informal (Morales et al., 2010).

En comparación con el estudio de la vulnerabilidad a la pobreza monetaria, la vulnerabilidad del empleo permite dividir los indicadores o, en otras palabras, desagregar estas transiciones. Pues mientras que para la pobreza monetaria se estima la probabilidad de pasar de un ingreso monetario mínimo a otro (Herrera & Cozzubo, 2016; INEI, 2020), la vulnerabilidad del empleo estima la probabilidad de pasar de tener un empleo al desempleo, o de pasar de un empleo decente a uno no decente.

No obstante, esto no quiere decir que la vulnerabilidad a la pobreza monetaria y la vulnerabilidad del empleo no estén relacionadas. La inestabilidad laboral es una causa y expresión de la pobreza monetaria (Bocquier et al., 2010; Lockshin & Ravallion, 2000). En esta línea, según Herrera & Cozzubo (2016), la inestabilidad de los ingresos, producto de un mercado laboral mayormente informal y con actividades concentradas en el sector de producción primario en áreas urbanas, determina la naturaleza estructural de la vulnerabilidad en los hogares urbanos a la pobreza monetaria en Perú.

En cuanto a los determinantes de la vulnerabilidad del empleo en Perú, se ha encontrado que un menor nivel educativo y ser mujer están relacionado con un empleo vulnerable a nivel nacional (Garavito, 2010), así como también lo es trabajar en el sector comercio y construcción debido a la desaparición rápida de los negocios y la temporalidad de las obras en Lima Metropolitana, respectivamente (Herrera & Hidalgo, 2002). Sin embargo, también se ha encontrado que trabajadores con grados universitarios han sido vulnerables a la pérdida del empleo debido a la sobrecualificación (Herrera & Hidalgo, 2002).

Otro factor importante que tomar en cuenta es la relevancia de la ciclicidad de la economía. En esta revisión de documentos, se encontró que el ciclo económico o las condiciones agregadas de la economía son relevantes. De esta manera, se encontró que la vulnerabilidad se redujo entre 1998, año luego de la recesión de 1997, y 2008, año en cual aún estábamos en un periodo de crecimiento (Morales et al., 2010).

En otros países también se han estudiado estos fenómenos, aunque en mayor medida a partir de indicadores. En Latinoamérica algunos resultados coinciden con lo encontrado en Perú. Brasil sufre una vulnerabilidad considerable en su mercado de trabajo, especialmente en los sectores de agricultura y comercio (Sparreboom & de

Gier, 2008). También se relaciona menores niveles de educación con este fenómeno. Asimismo, Ecuador ha encontrado que un 67.7% de su población ocupada es vulnerable a perder la calidad de su empleo y solo un 1.25% no lo es por completo. El ser vulnerable a esta pérdida se relaciona al ser mujer, ser trabajador del sector rural y ser trabajador informal a este fenómeno (Villacís & Reis, 2015).

De la misma manera, para países en otros continentes, se ha encontrado similitudes con Perú. India, un país con también altas tasas de informalidad, se encuentra que más de la mitad de las trabajadoras informales sufren de vulnerabilidad multidimensional en el empleo (Kumar & Srivastava, 2021). En Pakistán, también se confirma la mayor vulnerabilidad de las mujeres (Sparreboom & de Gier, 2008).

En el continente africano, en Mauricio, un país de ingresos medios-altos se encuentra que el estar casado/a, tener mayor edad y un bajo nivel educativo aumenta la probabilidad de tener un empleo vulnerable. Este es considerado como aquel que se realiza de manera independiente y/o dentro del vínculo familiar (Gokhool et al., 2018). Adicionalmente, en Namibia, las mujeres poseen mayor vulnerabilidad en sus empleos (Sparreboom & de Gier, 2008). Asimismo, los empleos vulnerables son predominantes en los mercados laborales urbanos en el Oeste de África (Bocquier et al., 2010).

Adicionalmente, en Europa, se encuentra que los migrantes entre países que son altamente cualificados son más vulnerables que los nativos de los países a los que emigraron; mientras que los migrantes bajamente cualificados son menos vulnerables (Bazillier et al., 2015). Los sectores de turismo (Bazillier et al., 2015), construcción (Kumar & Srivastava, 2021), agricultura y comercio (Sparreboom & de Gier, 2008) son los que concentran mayores empleos vulnerables para este continente.

Se ha encontrado una disminución en la cantidad de trabajadores con empleos vulnerables en Brasil, Namibia y Pakistán, aunque con marcadas diferencias debido a las condiciones iniciales (Sparreboom & de Gier, 2008). Asimismo, se encontró que el crecimiento de la cantidad del empleo es solo importante para reducir la vulnerabilidad en los sectores de manufactura, construcción y comercio en estos países (Sparreboom & de Gier, 2008).

Debido a su nivel de inferencia y riqueza de datos, se ha usado la ENAHO Panel como base de datos principal para este tipo de estudios en Perú. Se han usado variables de los módulos de características de la persona, del hogar y de los ingresos

y gastos de hogar. En menor medida se ha usado la Encuesta Permanente del Empleo (EPE) y otras encuestas de este tipo para los diferentes países mencionados.

Como se mencionó, solo se está estudiando uno de los componentes de la vulnerabilidad del empleo en la presente investigación: la vulnerabilidad a la pérdida del empleo, pues en este se pierde bienestar debido a la reducción de ingresos laborales. Si bien en la presente investigación no vamos a ahondar en el otro elemento, es posible una ampliación de este pues también implica la pérdida de bienestar de los hogares peruanos. Para mayor detalle sobre la pérdida de la calidad del empleo, revisar el Anexo B.

Es importante no olvidar que la diversidad en estas bases de datos ha sido reciente en el caso de Perú y aún es mejorable su amplitud puesto que en otros países han podido examinar dimensiones del empleo más detalladas, como la capacidad de decidir sobre su manejo en el trabajo (Bazillier et al., 2015). Asimismo, estas bases de datos aún no han logrado ser usadas en toda su amplitud de variables, en parte, por la elección de la metodología.

Todos estos estudios de las transiciones laborales, la buena calidad del empleo y vulnerabilidad de este en el Perú, usan metodologías econométricas como logit o probit multinomial para determinar las características de quienes experimentan estos fenómenos. En otros países, adicionalmente se han usado indicadores multidimensionales para explicar este fenómeno a mayor profundidad.

Este documento pretende enriquecer la literatura sobre la vulnerabilidad usando modelos y métodos de Machine Learning. Al entender la vulnerabilidad en reversa y realizar un ejercicio de predicción en el mercado laboral peruano, este documento introduce en la literatura nuevas herramientas cuantitativas para medir y caracterizar a la vulnerabilidad que lidian con las particularidades del mercado laboral peruano, como la rigidez de la informalidad que crea un desbalance entre la tasa de empleo y desempleo.

2.3. Las mujeres en el mercado laboral

Las trabajadoras mujeres están inmersas en una serie de dinámicas que las diferencian de los trabajadores varones en el contexto peruano. Las condiciones laborales de las mujeres en el Perú tienen ciertas determinantes históricas y sociales consecuencia de la diferenciación por clase social, raza y género que han perdurado por años (Barba et al., 1997).

Han existido amplios cambios en la situación de la mujer en el mercado laboral,

pero estos siempre han estado condicionados a patrones. Estas pautas, a partir de supuestas diferencias sexuales, marcan una participación compleja de las mujeres en las funciones del hogar (ama de casa, madre, esposa) y en el mercado de trabajo, como una participación secundaria a la de los hombres (Barba et al., 1997).

Estos patrones se pueden ver en los recientes años y más aún en los periodos de crisis que ha afrontado el Perú. Por el lado de la restricción de las mujeres a unos sectores u ocupaciones del mercado laboral, Garavito señala que esta división sexual del mercado se puede deber, en parte, a la exclusión social (Garavito, 2010). En otras palabras, el menor acceso a la educación o capacitaciones por el que sufren las mujeres disminuye el capital humano que ellas pueden ofrecer al mercado, limitando su participación en el mercado laboral (Barba et al., 1997).

Por esta razón, las trabajadoras pobres se tienen que conformar con trabajos relacionados a características femeninas en los cuales serán explotadas debido a las pocas posibilidades de encontrar un empleo que le permita su autonomía económica y, más aún, su subsistencia (Barba et al., 1997).

Esto se puede ver en años recientes, pues existe una fuerte relación entre el trabajo familiar no remunerado y el nivel socioeconómico del hogar. Cuando el nivel socioeconómico del hogar es bajo, el trabajo familiar no remunerado femenino es alto. Esto implica que la vulnerabilidad de la mujer en el mercado laboral y la falta de independencia económica debido a la inestabilidad de un trabajo es mayor entre los hogares más pobres (Jaramillo y Ñopo 2020 en Durán, 2022).

Aun así, se ha encontrado que algunas de las mujeres que quedan desempleados tienen un mayor número de años de educación en comparación con los hombres y mujeres que mantienen su trabajo. Esto puede explicarse porque estas compiten con hombres calificados y/o los costos de oportunidad en el caso de las mujeres (Garavito, 2010).

Con respecto a otras restricciones que afectan la participación de la mujer en el mercado laboral, los roles que se le otorga dentro del hogar a la mujer la limitan a ser considerada como mano de obra secundaria en el mercado laboral. Es decir, su ingreso al mercado laboral es una acción complementaria a su rol de cuidadora en el hogar (Barba et al., 1997), lo cual implica que sólo entrará en el mercado cuando sea necesario para la supervivencia del hogar, y su entrada se verá limitada por la carga doméstica.

En este sentido, se ha visto que la mujer se inserta en el mercado laboral para apaciguar el efecto adverso de los mercados de trabajo en la mano de obra primaria (Herrera & Hidalgo, 2002). De esta manera, una mayor proporción de niños en los hogares implica para las mujeres una mayor probabilidad de abandonar el empleo por la inactividad. Asimismo, más niños implica una menor posibilidad al desempleo crónico para las familias más pobres en tanto no podrá compensar el costo de oportunidad de cuidar a los niños con sus bajos ingresos laborales (Herrera & Hidalgo, 2002).

Es por esto que también la participación de la mujer en el mercado laboral es menos constante. Así, se ha encontrado que la probabilidad de la inactividad aumenta cuando se es mujer, joven con pocos años de estudio. En el caso de las mujeres en Lima Metropolitana, el estar casada o conviviendo aumenta esta probabilidad (Garavito, 2010).

Más aún, en tiempos de crisis como lo fue la pandemia del Covid-19, esta parece incrementar. El confinamiento total o parcial implica mayor dedicación a la educación de los niños del hogar y mayor presencia de miembros del hogar por los cuales realizar labores de cuidado. Por estos motivos se puede explicar por qué las mujeres peruanas han dejado de buscar empleos o reducido su dedicación a trabajos pagados durante y luego de la pandemia (Durán, 2022).

En este sentido, la presente investigación propone avanzar en la comprensión de aquellas posibles características que componen el perfil de las mujeres vulnerables a la pérdida del empleo. Esta investigación, en comparación con la literatura previa, busca identificar y caracterizar a las mujeres vulnerables a la pérdida del empleo antes que pierdan su trabajo. Esto tiene implicancias de política importante pues nos permite focalizar programas que busquen mitigar y prevenir las pérdidas de bienestar del desempleo o la inactividad.

Capitulo 3. Marco conceptual

Para la definición de la vulnerabilidad del empleo, optamos por utilizar la definición usada en el mercado laboral peruano (Garavito, 2010; Herrera & Hidalgo, 2002) basados en la definición de vulnerabilidad por Pritchett et. al (2000). Es decir, definimos la vulnerabilidad del empleo como aquellas transiciones laborales que reduzcan el bienestar de la población o aquellas transiciones del empleo al desempleo o inactividad, y de un empleo de buena calidad a un empleo de mala calidad. Sin embargo, solo nos enfocaremos en el primero debido a la diferencia en tendencias y variables que mueven este y el segundo tipo de transición, como se resalta en el Anexo B.

Para la definición de la pérdida de empleo optamos por utilizar la definición de Garavito (2010). Es decir, mediremos la pérdida de empleo a partir de analizar las transiciones laborales que reduzcan el bienestar de la población: aquellas transiciones del empleo al desempleo o inactividad.

Incluimos las transiciones hacia la inactividad pues estas pueden suceder de manera voluntaria con mayor probabilidad para las mujeres. Considerando que salieron al mercado laboral con el fin de apoyar económicamente al hogar ante recortes en sus ingresos ante un periodo de crisis, es muy probable que las mujeres prefieran sacrificar sus trabajos por más tiempo en el hogar.

Esto implica que volverán a ser inactivas habiendo pasado este periodo, aun si esto reduce su bienestar como trabajadora (ingresos, horas trabajadas, estabilidad laboral). Aun así, puede que esto no se sostenga a largo plazo para los hogares más pobres, que necesitan mantener sus ingresos. Si bien no podemos conocer las preferencias de cada mujer, es necesario complejizar estas decisiones en esta población en específico y consideramos que podemos lograr esto al considerar tanto el desempleo como la inactividad como parte de la vulnerabilidad de la pérdida del empleo.

Capitulo 4. Hechos estilizados y bases de datos

4.1. Bases de datos

Para la siguiente sección, se hará uso de la base de datos de la Encuesta Nacional de Hogares (ENAHO) Panel del 2016 al 2020. Se usará esta por sus diversas cualidades, como su representatividad en diferentes niveles y su gran aporte en datos correspondientes a nivel de hogar y características del empleo. La ENAHO presenta representatividad a nivel nacional, por área (urbano/rural), departamental y a nivel de sexo, lo cual es beneficioso para nuestra investigación pues basamos centrarnos en la PEA urbana femenina.

En comparación con otras encuestas, como la Encuesta Permanente del Empleo (EPE), la ENAHO Panel recupera las características sociodemográficas, del hogar, de salud, del empleo, etc. en años consecutivos para un mismo grupo de personas. Esto nos permite no solo evaluar diferentes transiciones en el mercado laboral por el cual puede pasar un trabajador, sino que también nos permite profundizar en las características que este tiene de un considerable número de personas. De esta manera, usando estos módulos que recopilan información de hogares del 2016 al 2020, se crearon paneles bianuales de la PEA femenina urbana.

Se utilizarán las variables recopiladas de su módulo de características de la vivienda y del hogar, educación, salud, empleo e ingresos, sumarias del hogar y características de la persona. De estas, se seleccionó el amplio grupo de variables relevantes según la revisión de literatura con el fin de aún usar un número grande de variables que este tipo de modelos requiere, pero también de evitar lidiar con las variables que tienen missings condicionas a otras variables. Sin embargo, la base final que se usará para las estimaciones es resultado de un proceso de selección de variables que estuvo compuesto de 3 etapas. Estas serán explicadas más detalladamente en la sección de metodología.

4.2. Hechos estilizados

En la siguiente sección, se analizará la PEA urbana, PEA urbana ocupada y aquella solo compuesta de mujeres, la evolución en la tasa de empleo, desempleo e inactividad a nivel urbano por sexo, las transiciones entre el empleo y el desempleo o inactividad entre 2016 y 2020, así como las diferencias entre las mujeres que perdieron su empleo con las que no. Esto nos permitirá tener una visión más detenida de la vulnerabilidad a la pérdida del empleo de la población y de nuestra muestra.

Del apilamiento de los bases de datos panel bianuales, obtenemos a la PEA urbana, es decir, a la población con más de 14 años y que por lo tanto es considerada como en edad para trabajar. En las tablas del Anexo C, podemos ver que la pirámide poblacional es estable y regresiva, de acuerdo con las bajas tasas de natalidad de los últimos años. Asimismo, podemos ver que a lo largo de los años más del 50% de la PEA urbana se ha encontrado casada o conviviendo mientras que aproximadamente el 30% se ha encontrado soltero y el resto en divorciado, separado o viudo. En cuanto a la condición socioeconómica, en concordancia con la tendencia de los últimos años, la tasa de pobreza extrema de la PEA urbana se encuentra cercana al 1%, mientras que la pobreza no extrema se encuentra alrededor del 10%, siendo el restante de la PEA urbana población no pobre.

En cuanto al nivel educativo, casi el 45% de la PEA urbana ha cursado por lo menos el nivel secundario a lo largo de los años, mientras que alrededor del 15% ha cursado un nivel inferior, el nivel primario. De la misma manera, casi un 35% de la PEA urbana ha cruzado un grado educativo superior, universitario o no universitario; siendo la excepción aproximadamente un 2% de peruanos pertenecientes a la PEA urbana que han cruzado un nivel de postgrado.

La PEA urbana ocupada en estos años parece seguir estas mismas características en sus grupos etarios, estado civil, nivel de pobreza y nivel educativo. En cuanto sus características laborales, un 45% de la PEA urbana tiene uno niveles de ingreso entre 500 y 999 soles, rango que es menor al salario mínimo vital de los últimos años. Casi un 27% recibe igual o más que 1500 soles, mientras que el sobrante de aproximadamente un 18% recibe por su ingreso laboral entre 1000 y 1499 soles. Por último, aproximadamente un 66% de la PEA urbana ocupada tiene un empleo informal, mientras que el restante, un 33%, un empleo del tipo formal.

Ahora nos vamos a centrar en analizar la PEA urbana femenina pues es en este grupo de la población en que nos enfocaremos. Al igual que la PEA urbana, este grupo parece tener una pirámide poblacional estable y regresiva. También podemos ver que a lo largo de los años más del 47% de la PEA urbana femenina se ha encontrado casada o conviviendo, el 32% se ha encontrado soltero y el resto, 31%, divorciada, separada o viuda.

En cuanto a la condición socioeconómica, la tasa de pobreza extrema de la PEA urbana femenino se encuentra cercana al 1%, el de la pobreza no extrema en

aproximadamente un 10%, y el 89% restante de la PEA urbana femenina es no pobre. En cuanto al nivel educativo, casi el 43% de las mujeres que componen este grupo ha cursado el nivel secundario, alrededor del 18% ha cursado el nivel primario, y casi un 30% de la PEA urbana ha cruzado un grado educativo superior; siendo, nuevamente, un 2% de la PEA urbana femenina aquella que ha cruzado un nivel de posgrado.

La PEA urbana femenina ocupada en estos años es muy similar en sus grupos etarios, estado civil, nivel de pobreza y nivel educativo. En cuanto sus características laborales, casi un 55% de la PEA urbana femenina tiene niveles de ingreso laboral entre 0 y 999 soles, un 18% recibe igual o más que 1500 soles, mientras que el sobrante de aproximadamente un 30% recibe por su ingreso laboral entre 1000 y 1499 soles.

En comparación con la PEA urbana compuesta de hombres y mujeres, parece que sus ingresos se encuentran más concentrados en la parte inferior de la distribución. En concordancia con esto, casi un 71% de la PEA urbana femenina ocupada tiene un empleo informal, mientras que el restante, un 28%, un empleo del tipo formal, mostrando así una ligera concentración en el primer tipo de empleo para el caso de las mujeres.

Tabla 1.

Tabla resumen de las mujeres pertenecientes a la PEA urbana en su periodo inicial de los paneles bianuales del 2016 al 2019

	2016	2017	2018	2019	Total	2016	2017	2018	2019	Total
	%	%	%	%	%	N	N	N	N	N
Grupo etario										
14 años	2.00%	2.22%	2.21%	2.28%	2.18%	8,080	7,992	8,090	8,050	32,217
15 - 29 años	32.52%	32.74%	31.59%	29.77%	31.63%	8,080	7,992	8,090	8,050	32,217
30 - 44 años	29.27%	29.83%	30.12%	29.83%	29.77%	8,080	7,992	8,090	8,050	32,217
45 - 64 años	25.86%	25.56%	26.25%	27.43%	26.29%	8,080	7,992	8,090	8,050	32,217
>= 65 años	10.35%	9.66%	9.83%	10.69%	10.13%	8,080	7,992	8,090	8,050	32,217
Estado civil										
Conviviente	23.38%	23.87%	23.30%	22.01%	23.13%	8,080	7,992	8,090	8,050	32,217
Casado	25.82%	25.18%	24.67%	24.21%	24.95%	8,080	7,992	8,090	8,050	32,217
Divorciado	0.77%	0.53%	0.78%	0.93%	0.75%	8,080	7,992	8,090	8,050	32,217
Separado	11.50%	12.19%	13.40%	14.56%	12.95%	8,080	7,992	8,090	8,050	32,217
Soltero	32.66%	32.58%	32.25%	31.83%	32.32%	8,080	7,992	8,090	8,050	32,217
Pobreza										
Pobre extremo	0.76%	0.70%	0.80%	0.75%	0.75%	8,080	7,992	8,090	8,050	32,217
Pobre no extremo	11.11%	11.73%	12.67%	11.76%	11.83%	8,080	7,992	8,090	8,050	32,217
No pobre	88.13%	87.57%	86.53%	87.49%	87.42%	8,080	7,992	8,090	8,050	32,217
Nivel Educativo										
No nivel	4.75%	4.03%	3.73%	3.87%	4.09%	8,080	7,992	8,090	8,050	32,217
Primaria	19.90%	19.18%	18.33%	18.20%	18.88%	8,080	7,992	8,090	8,050	32,217
Secundaria	42.59%	43.93%	43.65%	43.00%	43.30%	8,080	7,992	8,090	8,050	32,217
Superior No Universitaria	14.96%	15.17%	15.44%	15.81%	15.35%	8,080	7,992	8,090	8,050	32,217
Superior Universitaria	16.33%	16.24%	17.53%	17.53%	16.93%	8,080	7,992	8,090	8,050	32,217
Postgraduado	1.47%	1.45%	1.32%	1.58%	1.46%	8,080	7,992	8,090	8,050	32,217

Fuente: ENAHO Panel 2016-2019, ENAHO Panel 2016-2020. INEI. Elaboración propia.

Tabla 2.

Tabla resumen de las mujeres ocupadas pertenecientes a la PEA urbana en su periodo inicial de los paneles bianuales del 2016 al 2019.

	2016	2017	2018	2019	Total	2016	2017	2018	2019	Total
	%	%	%	%	%	N	N	N	N	N
Grupo etario										
14 años	0.35%	0.75%	0.46%	0.53%	0.52%	4,750	4,676	4,722	4,824	18,973
15 - 29 años	26.22%	25.38%	25.02%	23.24%	24.92%	4,750	4,676	4,722	4,824	18,973
30 - 44 años	37.59%	37.07%	37.52%	36.94%	37.27%	4,750	4,676	4,722	4,824	18,973
45 - 64 años	30.91%	32.25%	31.90%	33.49%	32.18%	4,750	4,676	4,722	4,824	18,973
>= 65 años	4.93%	4.55%	5.11%	5.81%	5.11%	4,750	4,676	4,722	4,824	18,973
Estado civil										
Conviviente	25.26%	26.13%	26.52%	23.94%	25.45%	4,750	4,676	4,722	4,824	18,973
Casado	27.39%	27.25%	25.47%	25.35%	26.33%	4,750	4,676	4,722	4,824	18,973
Soltero	28.44%	26.29%	25.63%	26.05%	26.56%	4,750	4,676	4,722	4,824	18,973
Pobreza										
Pobre extremo	0.53%	0.72%	0.57%	0.42%	0.56%	4,750	4,676	4,722	4,824	18,973
Pobre no extremo	10.32%	9.78%	10.55%	10.67%	10.34%	4,750	4,676	4,722	4,824	18,973
No pobre	89.15%	89.49%	88.88%	88.91%	89.10%	4,750	4,676	4,722	4,824	18,973
Nivel educativo										
Primaria	19.37%	19.19%	18.24%	17.79%	18.62%	4,750	4,676	4,722	4,824	18,973
Secundaria	38.67%	40.07%	39.67%	38.98%	39.36%	4,750	4,676	4,722	4,824	18,973
Superior No Universitaria	18.12%	18.47%	18.39%	18.47%	18.37%	4,750	4,676	4,722	4,824	18,973
Superior Universitaria	17.61%	16.78%	18.75%	19.12%	18.09%	4,750	4,676	4,722	4,824	18,973
Postgraduado	2.28%	2.17%	2.07%	2.34%	2.21%	4,750	4,676	4,722	4,824	18,973
Ingreso laboral										
Sin ingresos	11.97%	12.13%	10.42%	11.23%	11.42%	4,750	4,676	4,722	4,824	18,973
<500 soles	30.02%	30.98%	30.33%	31.58%	30.75%	4,750	4,676	4,722	4,824	18,973
500 - 999 soles	25.70%	26.79%	26.73%	25.11%	26.08%	4,750	4,676	4,722	4,824	18,973
1000 - 1499 soles	12.93%	14.10%	14.09%	13.06%	13.55%	4,750	4,676	4,722	4,824	18,973
>=1500 soles	19.38%	15.99%	18.43%	19.03%	18.20%	4,750	4,676	4,722	4,824	18,973
Tipo de empleo										
Empleo informal	71.80%	72.87%	70.85%	70.99%	71.61%	4,750	4,676	4,722	4,824	18,973
Empleo formal	28.20%	27.13%	29.15%	29.01%	28.39%	4,750	4,676	4,722	4,824	18,973

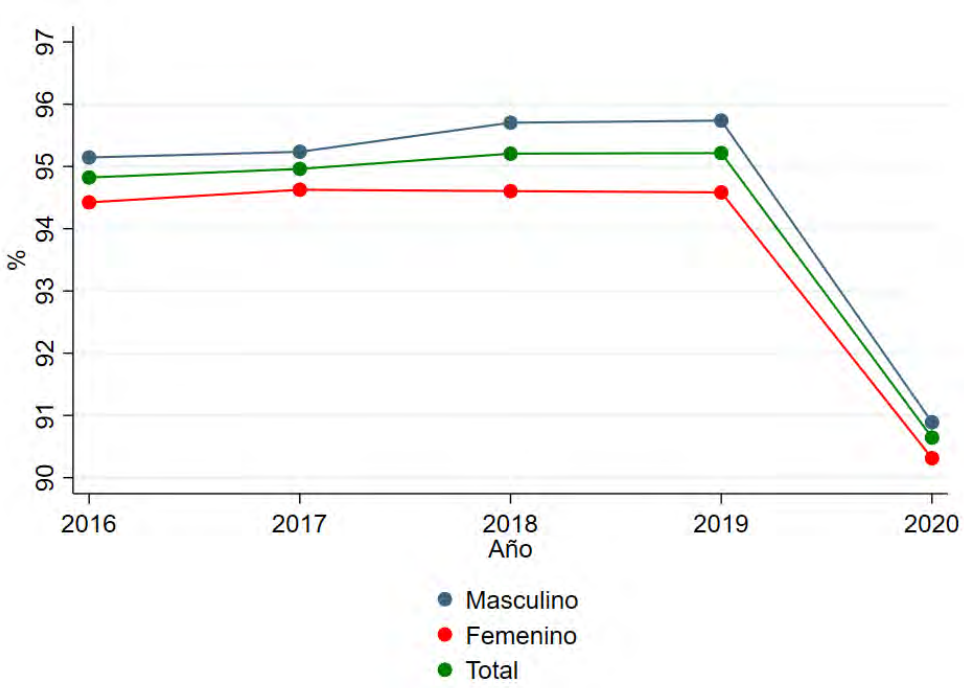
Fuente: ENAHO Panel 2016-2019, ENAHO Panel 2016-2020. INEI. Elaboración propia.

La evolución agregada del empleo en el área urbana puede resumirse en la tasa del empleo. En comparación con la tasa total y la tasa masculina, la tasa femenina se ha mantenido por debajo del 95% hasta antes del 2020. Antes de este año todas eran superiores a 94%; sin embargo, estas ocultan que la mayoría de estos empleos son informales. Aun así, en el 2020 se produjo una reducción abrupta del empleo para la población urbana peruana de casi el 5%.

En todos estos años, la tasa de empleo femenina siempre ha sido menor con respecto a la población total y de los hombres. La participación secundaria de las mujeres en el mercado laboral, así como estar inmersa en las funciones del hogar pueden ser las razones por las que su presencia en el mercado laboral, formal o informal es reducida.

Figura 1.

Tasa de empleo urbana (%) por sexo del 2016-2020.



Fuente: ENAHO Panel 2016-2020. INEI. Elaboración propia.

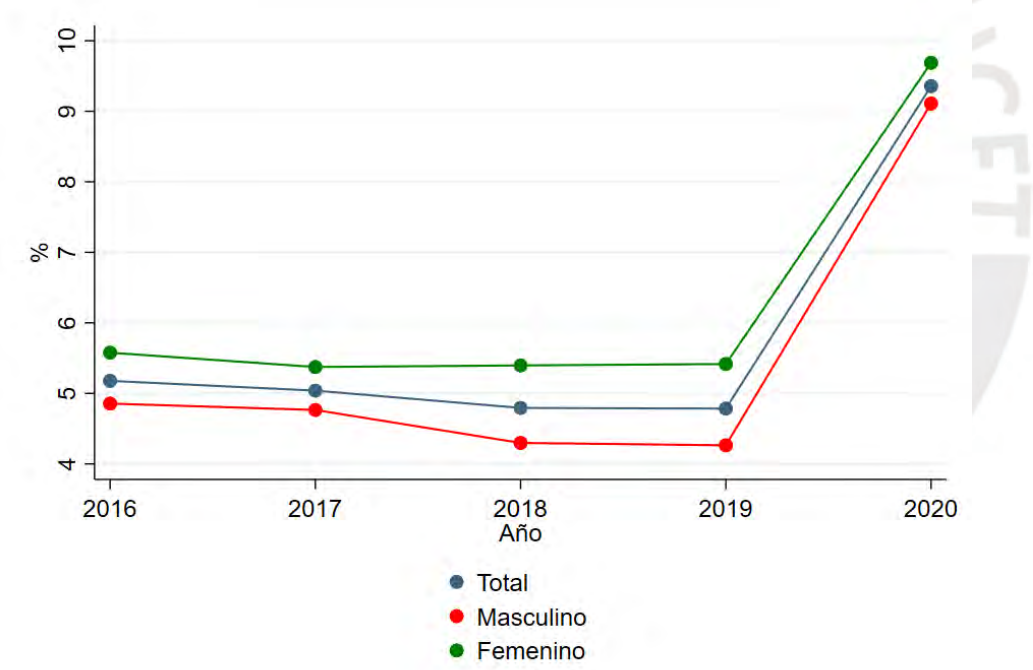
La tasa de desempleo urbano muestra cómo ha ido evolucionando el cambio de estado empleado a desempleado. La tasa de desempleo urbana de las mujeres se ha mantenido por encima de la de los hombres, aunque ha variado más que el de las mujeres. Las tres tasas, por lo general, se encuentran entre 4% y 6% lo cual refleja un

hecho conocido del mercado laboral peruano: no existe una alta tasa de desempleo puesto que la mayoría se subempleada. Esto cambió parcialmente con la pandemia en el 2020, año en el que todas las tasas fueron superiores al 9%, siendo la de las mujeres la mayor de todas.

Asimismo, esta tasa de desempleo femenino ha sido mayor con respecto a la tasa de desempleo urbana total. Esto puede deberse a su concentración en sectores poco productivos y su inestabilidad en el mercado laboral atribuida a su rol en el hogar. Esto se mantuvo incluso con la pandemia en el 2020, en la cual la tasa aumentó a casi 10%. Cabe tomar en cuenta que esta solo considera al desempleo oculto y abierto, mas no a la inactividad que consideramos en esta investigación como parte del "desempleo" en el caso de las mujeres.

Figura 2.

Tasa de desempleo urbana (%) por sexo del 2016-2020.



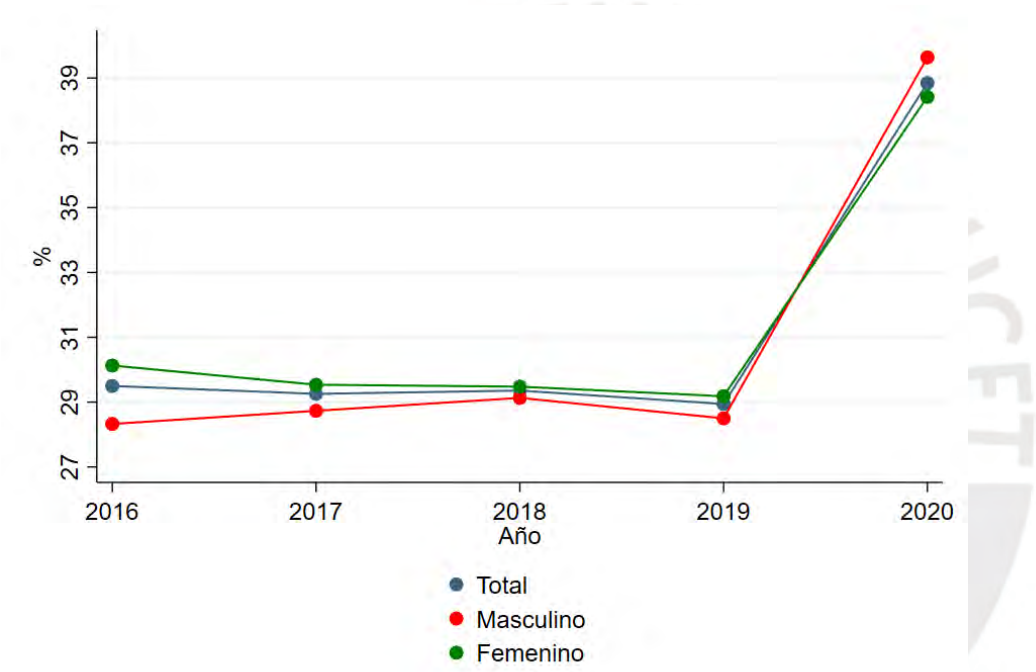
Fuente: ENAHO Panel 2016-2020. INEI. Elaboración propia.

La tasa de inactividad urbana tiene un rango distinto a la del desempleo urbano. Sin embargo, similar a la tasa de desempleo urbana, la tasa perteneciente a las mujeres es superior a la de hombres y a la total hasta antes del 2020. En este año, todas crecieron al menos 8% y llegaron a casi el 39%. Asimismo, parece que con el tiempo han ido recortándose las diferencias en la participación laboral por sexo puesto

que las tendencias se han ido acercando. Esto puede deberse a la progresiva pero importante incorporación de la mujer en el mercado laboral como mano de obra secundaria ante la necesidad de mayores ingresos en el hogar. De la misma manera, vale recalcar que, para nuestra definición de "vulnerabilidad" incluiremos a las mujeres que transaccionan del empleo a la inactividad, Esto ya que no es que las mujeres no estén buscando activamente empleo porque no quieren, sino porque no tienen el tiempo ni los recursos para insertarse en el mercado laboral.

Figura 3.

Tasa de inactividad urbana (%) por sexo del 2016-2020.



Fuente: ENAHO Panel 2016-2020. INEI. Elaboración propia.

Ahora, es necesario recordar que estas medidas son indicadores que recogen información de un momento en el tiempo. Es decir, almacenan el resultado de flujos de trabajadores que han sucedido en el tiempo. Por ejemplo, en la tasa de desempleo de determinado año se encuentran a los trabajadores que tienen una transición de empleo a desempleo en este periodo, así como a los trabajadores que se encontraron desempleados en ese determinado año y el anterior. Como nosotros vamos a analizar las transiciones de pérdida de empleo, es decir, del empleo al desempleo o la inactividad; es importante considerar el análisis de las transiciones de un año al otro que tengamos en nuestros paneles bianuales.

Para analizar cómo se dan las transiciones entre los diferentes estados, empleo y desempleo/inactividad, entre los dos tiempos; se construirá 4 matrices de transición con la PEA urbana femenina, una por cada panel bianual y una con toda la muestra en general. Esto nos permitirá establecer el porcentaje de mujeres que han sido vulnerables en el empleo en estos años.

Según el fenómeno que estamos analizando, existen 4 tipos de transiciones posibles. Para las mujeres que tenían empleo en el primer periodo, ocurrieron dos posibles eventos en el siguiente periodo: lo mantuvieron (De empleado a empleado) o lo perdieron (De empleado a desempleado o inactivo). Asimismo, para las mujeres que no tenían empleo, ocurrieron dos posibles eventos: mantuvieron su desempleo (De desempleado a inactivo a desempleado o inactivo) o encontraron uno (De desempleado o inactivo a empleado).

Tabla 3. Matrices de transición (%) de la PEA urbana femenina, paneles bianuales 2016-2019.

Panel	Pierden su empleo	Mantienen pérdida de empleo	Mantienen empleo	Encuentran un empleo	% de no empleados que encuentran empleo	% de empleados que pierden empleo
2016-2017	10.49%	30.22%	48.26%	11.03%	26.74%	17.85%
2017-2018	10.18%	29.98%	48.30%	11.53%	27.77%	17.41%
2018-2019	9.64%	29.65%	48.77%	11.93%	28.69%	16.50%
2016-2019	7.57%	22.45%	36.31%	8.62%	27.74%	17.50%

Fuente: ENAHO Panel 2016-2020. INEI. Elaboración propia.

A lo largo de los paneles bianuales construidos, podemos ver que aproximadamente el porcentaje de la PEA urbana femenina que pierde su empleo es aproximadamente un 10% en los años anteriores al 2020. La transición del 2019 al 2020, a raíz de la pandemia y sus efectos en el mercado laboral, parece haber tenido consecuencias severas en el mercado laboral urbano y, en especial, aquella parte que está conformada por mujeres.

Esto se puede notar igualmente al ver que el porcentaje de mujeres que pasaban del desempleo/inactividad al empleo era muy similar al de la transición reversa, mientras que entre 2019 y 2020, solo un 7.8% de desempleados/inactivos obtuvo un empleo.

Ahora, a nosotros nos interesa conocer el porcentaje de mujeres que tenían un empleo en el primer periodo del panel bianual al que pertenecen, y que lo perdieron en el segundo periodo de este. Como podemos notar, aproximadamente un 18% de las mujeres pertenecientes a la PEA urbana que tenían un empleo en el periodo inicial, lo perdieron en el siguiente. Este porcentaje aumentó en la transición del 2019 al 2020, siendo de aproximadamente un 33%, lo que es casi el doble de los anteriores porcentajes.

Tabla 4. Matriz de transición (%), panel bianual 2019-2020.

Panel	Pierden su empleo	Mantienen su pérdida de empleo	Mantiene empleo	Encuentra un empleo	% de D/I que encuentran empleo	% de E que pierden empleo
2019-2020	20.10%	32.24%	39.85%	7.80%	19.49%	33.53%

Fuente: ENAHO 2016-2020. INEI. Elaboración propia.

Al comparar estas cifras con lo que se encontró con la PEA urbana masculina notamos el contraste en las tasas. el porcentaje de hombres en el área urbana que perdían su empleo no pasaba del 7% antes del 2020. Una cantidad un poco mayor, de aproximadamente 8% realizaba la transición inversa.

Tabla 5. Matrices de transición (%) de la PEA urbana masculina, paneles bianuales 2016-2019.

Panel	Pierden su empleo	Mantienen pérdida de empleo	Mantienen empleo	Encuentran un empleo	% de no empleados que encuentran empleo	% de empleados que pierden empleo
2016-2017	6.35%	14.39%	71.06%	8.20%	36.29%	8.20%
2017-2018	6.23%	14.25%	71.86%	7.66%	34.94%	7.98%
2018-2019	6.50%	14.69%	71.02%	7.78%	34.61%	8.39%
2019-2020	15.60%	16.99%	61.42%	5.99%	26.08%	20.25%

Fuente: ENAHO Panel 2016-2020. INEI. Elaboración propia.

Antes del 2020, el porcentaje de hombres que tenían un empleo en el primer periodo del panel bianual al que pertenecen, y que lo perdieron en el segundo periodo de este estaba también alrededor del 8%, lo cual es casi la mitad de la tasa que tenían las mujeres en este año. Todos estos porcentajes empeoran en el 2020, pero aun así

no llegan a superar las tasas de pérdida de empleo o el porcentaje de aquellos empleados que pierden su empleo de las mujeres.

Por último, comparamos a las mujeres que tenían un empleo al inicio del periodo de su panel bianual y que lo perdieron (De E a D/I) con las que lo mantuvieron (De E a E). Parece que la distribución de la pirámide poblacional entre estos dos grupos varia pues los porcentajes en los diferentes grupos etarios son significativamente diferentes: Las mujeres que mantiene su empleo de un periodo a otro parecen tener una distribución concentrada en edades mayores. Además, parece que las mujeres que conservan su empleo tienden a estar casadas en una proporción marginalmente mayor que las que lo pierden.

En cuanto a los niveles de pobreza, si bien las diferencias son sutiles, parece que las mujeres que mantienen su empleo tienden a ser no pobres en una proporción marginalmente mayor que las que lo pierden. Por último, la distribución en los niveles educativos entre ambos grupos parece ser diferente también.

En concordancia con la diferencia de edad, se puede ver que las mujeres de la PEA urbana que no pierden su empleo de un periodo a otro suelen alcanzar mayores niveles educativos pues los porcentajes de aquellas que alcanza el nivel secundario, el nivel superior no universitario, el nivel superior universitario y el nivel de posgrado es mayor significativamente de las mujeres que pierden su empleo de un periodo a otro.

Esta investigación plantea que la proporción de mujeres del 2019 que perderán su empleo en el 2020 es mayor de 18%. Esto se probará después de realizarse la predicción *out-of-sample y out-of-time.* luego de clasificar a las mujeres en el 2019 como vulnerables o no, se realizará una prueba cuya hipótesis nula con respecto al porcentaje de mujeres vulnerables a la pérdida de su empleo p será Ho:$p>0.18$ y su hipótesis alternativa era Ha:$p\leq 0.18$. Por último, se espera que el modelo se valide de manera *out-of-time* usando la información disponible de las mujeres que pertenecen a la PEA en el 2020. Es decir, se espera que al menos el 50% de las mujeres que han sido clasificadas con empleos vulnerables a su pérdida en el 2019 hayan efectivamente perdido su empleo el año siguiente.

Tabla 6. Comparación entre mujeres perdieron su empleo y las que no entre el 2016 al 2019.

	De E a D/I		De E a E			
	Media	N	Media	N	Diff	P-value
Grupo etario						
14 años	0.01	4,030	0.00	14,943	0.01	0.00
15 - 29 años	0.31	4,030	0.19	14,943	0.12	0.00
30 - 44 años	0.27	4,030	0.34	14,943	- 0.07	0.00
45 - 64 años	0.29	4,030	0.40	14,943	- 0.11	0.00
>= 65 años	0.12	4,030	0.07	14,943	0.06	0.00
Estado civil						
Conviviente	0.25	4,030	0.24	14,943	0.01	0.11
Casado	0.26	4,030	0.29	14,943	- 0.03	0.00
Viudo	0.06	4,030	0.05	14,943	0.01	0.02
Divorciado	0.01	4,030	0.01	14,943	- 0.00	0.07
Separado	0.13	4,030	0.19	14,943	- 0.05	0.00
Soltero	0.28	4,030	0.22	14,943	0.06	0.00
Pobreza						
Pobre extremo	0.01	4,030	0.00	14,943	0.00	0.01
Pobre no extremo	0.10	4,030	0.08	14,943	0.02	0.00
No pobre	0.89	4,030	0.91	14,943	- 0.02	0.00
Nivel educativo						
No nivel	0.05	4,030	0.04	14,943	0.02	0.00
Primaria	0.21	4,030	0.20	14,943	0.01	0.30
Secundaria	0.41	4,030	0.35	14,943	0.06	0.00
Superior No Universitaria	0.16	4,030	0.19	14,943	- 0.03	0.00
Superior Universitaria	0.16	4,030	0.19	14,943	- 0.03	0.00
Postgraduado	0.01	4,030	0.03	14,943	- 0.03	0.00

Fuente: ENAHO Panel 2016-2020. INEI. Elaboración propia.

Capitulo 5. Metodología

5.1. Modelo

Para los problemas de predicción, sólo es requerido que obtengamos un $\hat{Y}$ que sea similar al Y generado por su proceso generador de datos, es decir, aquel Y que realmente se presenta en el contexto que estamos intentando predecir. Por esto, solo se requiere que $\hat{Y}$ tenga un bajo error, más no que los coeficientes sean insesgados. En este sentido, las técnicas empíricas estándares de Economía no son los más adecuados para los desafíos que presenta la predicción puesto que se enfocan en buscar los coeficientes no sesgados (Kleinberg et al., 2015).

De esta manera, para un ejercicio de predicción, si tenemos una cantidad N de observaciones (y_i, x_i), con P variables en x_i; debemos usar estas observaciones pertenecientes a una muestra para elegir una función $f \in F$ y así poder predecir el valor de $\hat{y}$ de un nuevo punto de los datos $(y_i, \widetilde{x_i})$, es decir, fuera de la muestra. Para ello, el objetivo es minimizar la función de pérdida que podemos plantear como $\left(y - \hat{f}(x)\right)^2$.

Esta función f puede ser lineal como no, pues existen métodos como los árboles de decisiones que nos permiten construir secuencias de funciones que serán integradas en diferentes etapas de la predicción. Sin embargo, en esta investigación, nos centraremos en el uso de un modelo lineal pues este resulta conveniente para poder lograr los objetivos planteados.

En este sentido, este es beneficioso pues permite la predicción de una variable respuesta "y" cuantitativa y estará bajo la suposición que la distribución de los errores es Gausiana (Kleinberg et al., 2015), lo cual nos permite predecir adecuadamente la probabilidad de ser vulnerable de una mujer.

Este tipo de modelo, además, permite una mayor interpretabilidad pues es muy similar a lo que normalmente se usa en los modelos clásicos de Economía (por ejemplo, el modelo de mínimos cuadrados ordinarios o MCO). Por último, este tipo de modelos ha sido usado anteriormente en otras investigaciones que han tenido objetivos similares, como el mapa de vulnerabilidad de la pobreza monetaria (INEI, 2020), lo cual resulta ventajoso ya que contamos con un marco de referencia para su aplicación.

Bajo estas condiciones, teniendo como nuestro objetivo es aproximar la variable respuesta, y, proponemos iniciar planteando una combinación lineal de predictores tal que se puede escribir como:

$$y(x_i) = \beta_0 + \sum_{j=1}^{P} x_{ij}\beta_j \quad (1)$$

Este es un modelo de regresión lineal, como el MCO, con parámetros desconocidos como β_o y $\beta = (\beta_1, \beta_2, \dots, \beta_p)$, con término de error e_i, que trata de estimar los primeros mediante la minimización de la función objetiva de mínimos cuadrados:

$$\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_{ij})^2 \quad (2)$$

En este caso, se tiene que $\hat{f}(x) = \widehat{\beta_0} - \sum_{j=1}^{P} x_{ij}\widehat{\beta_{ıJ}}$. El método MCO minimiza los errores dentro de la muestra escogiendo los mejores estimadores lineales insesgados. Sin embargo, para la predicción, nuestro objetivo no es este, sino el de desempeñarnos bien *fuera* de la muestra.

En este sentido, el modelo MCO no es el más apto para un problema de predicción en términos de eficiencia debido a que, al priorizar la insesgadez, tiende a presentar una alta varianza cuando se aplica a diferentes grupos de observaciones. Esto se debe, en parte, a que en este modelo no se eliminan coeficientes en su estimación.

Esto se puede ver de manera más detallada si descomponemos el error esperado de la función estimada $\hat{f}(x)$:

$$\begin{aligned} E_D[(y - \hat{f}(x))^2] &= \sigma^2 + E\left[\hat{f}(x) - f(x)\right]^2 + \{E[f(x)^2] - E[f(x)]^2\} \quad (3) \\ &= \sigma^2 + Sesgo\left[\hat{f}(x)\right]^2 + Var\left[\hat{f}(x)\right] \end{aligned}$$

Aquí, el f varía de muestra en muestra puesto que se abarca un diferente grupo de observaciones en cada una de ellas y esto produce que exista la varianza, el primer término. Cabe notar que, si existe un cambio en la varianza este debe ser compensado con un cambio en el sesgo para mantener el error esperado en su mismo valor (James et al., 2013). De esta manera, debido a que el MCO asegura la insesgadez dentro de

la muestra, es imposible realizar un intercambio con la varianza, creando problemas fuera de la muestra (Kleinberg et al., 2015).

Asimismo, si elegimos un modelo de mayor complejidad, obtenemos más datos y con ello un menor sesgo. No obstante, el aumento de la varianza se dará debido a que, para abarcar más datos, el modelo considerará más coeficientes y no reducirá su valor o los eliminará. Es decir, ninguno de los elementos del vector $\beta = (\beta_1, \beta_2, \dots, \beta_p)$ podrán tener la posibilidad de tener el valor de 0.

De la misma manera, este modelo hará la interpretación de los coeficientes más compleja para los lectores. Si es que existe un número grande de variables y no se descarta alguno, será retador interpretar el aporte de cada una de las variables y describirlas a los lectores y posibles tomadores de decisión en la política pública.

Particularmente, si es que el número de variables es mayor que el número de observaciones $(P > N)$, entonces puede que no sean únicos los valores que estos coeficientes pueden tomar. Esto implica que el modelo sea muy ajustado a los datos de la muestra, lo cual impedirá tener la varianza suficiente para poder predecir con observaciones fuera de la muestra.

Por todo esto, es más conveniente seleccionar al grupo de variables que presenta los mayores efectos (Hastie et al., 2015). En este sentido, existe la necesidad de restringir el proceso de estimación. Más aún, la precisión de la predicción se puede mejorar al reducir los valores de los coeficientes o incluso estableciéndose como cero, lo que se conoce como el encogimiento de coeficientes (o "*shinkage*").

Los métodos de *Machine Learning* fueron creados para maximizar la predicción al proveer una manera empírica de realizar el intercambio entre el sesgo y la varianza. Estos modelos lineales que no solo minimizan una función objetivo que solo se enfoca en el sesgo tienen la forma:

$$\widehat{f_{ML}} = \sum_{i=1}^{N} \left(y_i - \hat{f}(x_i)\right)^2 + \lambda R(f) \qquad (4)$$

En esta ecuación, $R(f)$ es un regularizador que penaliza la función que genera varianza. Este término $R(f)$ pertenece a un set de funciones, $F_c = \{f \mid R(f) < c\}$, el cual crea predicciones que tengan una mayor variabilidad cuando tenemos un límite c más grande. Para un modelo lineal, un c más grande permite coeficientes más grandes y, con ello, predicciones variadas en diferentes muestras, que implica mayor varianza.

La ecuación que minimizar incluye implícitamente el término de sesgo y el error en la muestra, y el término de varianza, $R(f)$. Al incorporar este último término, estamos sesgando a propósito la estimación para considerar la varianza. Este término adicional depende de β, lo cual implica que el β^* ya no será el mejor estimador lineal insesgado que se obtiene de resolver un MCO.

En este término, λ puede ser considerado como el precio al cual se intercambia la varianza y el sesgo. MCO es un caso especial pues, como mencionamos, no existe posibilidad al intercambio por lo que el valor de λ es cero. De esta manera, ponemos un valor infinito relativo al sesgo $\left(\frac{1}{\lambda} = \infty\right)$ (Kleinberg et al., 2015). Si no es cero, entonces es posible ponerle un precio al intercambio entre varianza y sesgo.

Un modelo estadístico disperso es uno en el cual solo un pequeño número relativo de parámetros o predictores son esenciales para predecir una variable (Hastie et al., 2015). Por ejemplo, el modelo Lasso, perteneciente a esta familia, combina la pérdida de los mínimos cuadrados (MCO) con una restricción l_1. Este modelo nos permite una manera automática de seleccionar variables y un problema de optimización convexo que puede ser eficientemente resuelto.

El modelo Lasso plantea la solución a un problema de minimización de N observaciones $\{(x_i, y_i)\}_{i=1}^{N}$ cuya solución está planteada como (β). En su forma no matricial, tenemos el modelo en forma Lagrangiana tal que puede ser expresado como:

$$\frac{1}{2N}\sum_{i=1}^{N}\left(y_i - \sum_{i=1}^{P} x_{ij}\beta_j\right)^2 + \lambda\sum_{i=1}^{P}|\beta_j| \qquad (5)$$

Aquí, $\left|\left|\beta\right|\right|_1 = \sum_{i=1}^{p}|\beta_j|$ es la forma en l_1 de los parámetros β y λ es el parámetro determinado por el usuario. Este segundo término incluido puede ser interpretado como la suma absoluta de coeficientes que participan en el problema de minimización y está limitado por el valor de λ. En este sentido, coeficientes más encogidos o con valores cercanos a cero corresponden a un modelo más restringido que se dará mientras más alto sea el valor de λ.

El modelo Lasso aumenta la capacidad predictiva del modelo al seleccionar los coeficientes más relevantes. Esta norma, l_1, permite que el modelo Lasso nos dé vectores solución dispersos, es decir, que algunas de sus coordenadas sean 0. Esto, a su vez, permite una interpretación más fácil, lo cual es útil pues mejora la comunicación de los resultados.

Asimismo, este modelo nos permite tener un problema convexo, lo cual nos simplifica los esfuerzos computacionales tanto como lo hace el supuesto de dispersión. Esto no ocurre para otros parámetros y más aún, para normas de mayores grados como l_2 que convierten a los problemas en no convexos, lo cual dificulta terriblemente el problema de minimización pues hace muy costosa su solución en términos computacionales (Hastie et al., 2015).

Adicionalmente, este modelo cumple con el principio de "apuesta por la dispersión", el cual nos plantea cómo el supuesto de encogimiento de parámetros es necesario para llegar a una solución. Este explica que, si tuviéramos más variables que observaciones y el verdadero modelo no es disperso, entonces no se podría tener una estimación adecuada de los parámetros puesto que el número de observaciones es muy pequeño y, ergo, no tendríamos solución a estimar. Sin embargo, si el modelo es disperso y sólo un número de variables son diferentes de cero es posible estimar los parámetros adecuados si es que este número es menor al número de observaciones (Hastie et al., 2015).

Las derivaciones de esta función para cada x_j componen un sistema de ecuaciones las cuales pueden ser expresadas en su forma no matricial e individual para cada x_j como:

$$\frac{1}{N}\sum_{i=1}^{N} x_{ij}\left(y_i - \sum_{i=1}^{P} x_{ij}\beta_j\right) + \lambda\sum_{i=1}^{p}|\beta_j| = 0 \qquad (6)$$

El valor λ de la restricción controla la complejidad del modelo pues mientras más pequeño sea, da espacio para más parámetros; mientras que, si es más grande, el modelo está más restringido en cuantos β_j puede usar y por lo tanto menos ajustado a los datos, simulando de esta manera el intercambio entre sesgo-varianza (Kleinberg et al., 2015). Para ver la transformación a la forma matricial en caso lo prefiera o una explicación más detallada de la optimización del modelo Lasso, revise el Anexo D.

Debido a que estamos usando una regresión Lasso, la predicción de y que se nos estime, $\hat{y}$, será una variable continua. Sin embargo, estos valores no necesariamente se encontrarán en el rango $[0,1]$ puesto que, al igual que un modelo MCO, el rango de la distribución de $\hat{y}$ no se limita entre estos valores. Estos resultados serán controversiales si es que el objetivo es que el modelo estime la probabilidad, cuyo rango es restringido.

Más aún, para evaluar el modelo fuera de la muestra, es necesario que $\hat{y}$ sea una variable dicotómica. En este sentido, si se quiere diferenciar a las mujeres pierden el empleo y las mujeres lo mantienen a partir de su probabilidad predicha, es esencial establecer un umbral. Cuando el valor de la probabilidad sea menor a este valor, la etiqueta será de "empleada", mientras que si toma un valor por encima será de "desempleada".

Para determinar cuál es el valor del umbral para la probabilidad, usaremos inicialmente el valor preestablecido de 0.5. Esto quiere decir que si para una mujer, su probabilidad predicha de caer en el desempleo es menor que 0.5, entonces se le pondrá la etiqueta de "empleada" (0) en la variable dicotómica $\hat{y}$. Pero si es que la probabilidad predicha es mayor, entonces se le pondrá la etiqueta de "desempleada" (1). Cabe mencionar que si bien para otros estudios de similar metodología se han usado valores empíricos para el umbral (INEI, 2020), en nuestra investigación se hará uso del 0.5.

Las ventajas de esta aproximación se basan en la muestra y en la naturaleza de esta investigación. Nuestra base de datos no incluye a todas las mujeres del Perú debido a la limpieza de los datos mencionada previamente. En este sentido, tiene cierta inconsistencia usar un umbral basado en porcentajes de una muestra que no es la nuestra. Asimismo, al ser una primera aproximación a este problema, usar valores preestablecidos nos asegura la posibilidad de comparabilidad con otros posibles estudios. Esta elección no es tan diferente a la que se realiza con el umbral de 0.5 en una estimación de un modelo *logit* normal.

Es necesario mencionar que aparte del establecimiento de un valor empírico o de un valor preestablecido, también existen otras herramientas estadísticas para la estimación del umbral ideal cuyo proceso no es muy diferente a los que se siguen para la elección de otros hiper parámetros. Más adelante, cuando se expliquen las maneras

de evaluar el desempeño de un modelo, se mencionará como se acoplan estos pues el umbral que utilizaremos vendrá finalmente de la información de los datos.

Aun así, la predicción de un modelo Lasso aún presenta ciertos problemas para nuestro problema de investigación. Si bien estamos estableciendo el umbral, esto no soluciona que la predicción esté fuera del rango $[0,1]$. Para considerar estos detalles metodológicos es necesario que exploremos más la función $\hat{f}(x)$. En paralelo a lo que conocemos en Economía, cuando queremos restringir el rango de la variable a estimar en un rango entre $[0,1]$, realizamos un cambio en la función de enlace. En nuestro caso, estableceremos la conexión a partir de la ratio de *log verosimilitud*, sin embargo, esta no es la única manera de hacerlo. Para más detalle sobre la justificación teórica usando una variable latente I^*, revise el Anexo E.

Entonces, si sabemos que p_i es la probabilidad de que suceda un evento ($y_i = 1$) , o en nuestro caso que la mujer i este desempleado de un año a otro; y $1 - p$ la probabilidad de que no suceda este evento ($y_i = 0$), podemos plantear la ocurrencia de estos eventos como:

$$Pr\,[y_i = 1 \mid x] = p_i \quad \text{y} \quad Pr\,[y_i = 0 \mid x] = 1 - p_i$$

Para hallar la función de probabilidad de y_i , debemos tomar en cuenta ambas posibilidades. Ante esto, podemos plantear la función de probabilidad de y_i como:

$$f(y_i) = p_i^{y_i}(1 - p_i)^{1-y_i} \tag{7}$$

Sin embargo, sabemos que la probabilidad de que y_i suceda o no depende una serie de variables, el respectivo grado de influencia que tienen sobre esta variable (lo que conocemos como x y β) y la relación que tienen con la variable respuesta y. De esta manera, podemos plantear:

$$p_i = Pr\,[y_i = 1|x]$$
$$p_i = Pr\,[y_i = 1|x] = F(x\beta) \tag{8}$$

La relación de x y β con y , o "función de enlace", puede ser planteada como $g(x\beta)$. En nuestro caso, para cumplir con las mismas propiedades que el MCO que usamos en el modelo Lasso, la función de enlace g debe ser tal que se cumpla $g: R \rightarrow R$, sea una función lineal y sea estrictamente monótona. Asimismo, tomando en cuenta las características de nuestra variable objetivo, esta debe tener una función de distribución que nos permita limitar los valores de y_i entre $[0,1]$.

El modelo logístico, según Hastie et. al (2015), modela la ratio de log verosimilitud de la variable como una combinación lineal tal que podemos plantearlo como:

$$log\ \frac{Pr\ Pr\ [y_i = 1 \mid x]}{Pr\ Pr\ [y_i = 0 \mid x]} = x\beta \tag{9}$$

Entonces, si queremos encontrar $p_i = Pr\ [y_i = 1|x]$, podemos realizar los siguientes cambios:

$$\frac{Pr\ [y_i = 1 \mid x]}{Pr\ [y_i = 0 \mid x]} = e^{x\beta} \tag{10}$$

$$Pr\ [y_i = 1 \mid x]\ = e^{x\beta} Pr\ [y_i = 0 \mid x] \tag{11}$$

Tal que, si desarrollamos la probabilidad de y un poco más para obtener el denominador, $Pr\ [y_i = 0 \mid x]$, que es la probabilidad de que no suceda el evento:

$$Pr\ [y \mid x] = Pr\ [y_i = 1 \mid x] + Pr\ [y_i = 0 \mid x] \tag{12}$$

$$\frac{Pr\ [y \mid x]}{Pr\ [y_i = 0 \mid x]} = \frac{Pr\ [y_i = 1 \mid x]}{Pr\ [y_i = 0 \mid x]} + \frac{Pr\ [y_i = 0 \mid x]}{Pr\ [y_i = 0 \mid x]} \tag{13}$$

Reemplazando los valores que ya conocemos anteriormente

$$\frac{Pr\ [y \mid x]}{Pr\ [y_i = 0 \mid x]} = e^{x\beta} + 1 \tag{14}$$

Así, si sabemos que efectivamente y sucedió, tal que $Pr\,[y \mid x] = 1$, entonces podemos encontrar que la probabilidad de que no suceda el evento es:

$$Pr\,[y_i = 0 \mid x] = \frac{1}{e^{x\beta} + 1} \tag{15}$$

Con esto, obtenemos que la probabilidad de que suceda el evento es:

$$Pr\,[y_i = 1 \mid x] = \frac{e^{x\beta}}{e^{x\beta} + 1} \tag{16}$$

El uso de la función logística dentro del campo de Economía y de *Machine Learning* es popular debido a que cumple con las condiciones mencionadas previamente. Así, si reemplazamos lo que encontramos en la equivalencia pasada tenemos que:

$$p_i = F(x\beta) = \frac{e^{x\beta}}{1 + e^{x\beta}} = \frac{1}{1 + e^{-x\beta}} \tag{17}$$

Tal que, ahora podemos plantear la función de probabilidad de y_i como:

$$f(y_i) = p_i^{y_i}(1 - p_i)^{1-y_i} \tag{18}$$

$$f(y_i) = \left(\frac{e^{x\beta}}{1 + e^{x\beta}}\right)^{y_i} \left(1 - \frac{e^{x\beta}}{1 + e^{x\beta}}\right)^{1-y_i} \tag{19}$$

Con esta función de probabilidad, podemos encontrar los β que maximizan la probabilidad de que y_i suceda y su ratio de log verosimilitud esté planteada por una combinación lineal compuesta $x\beta$. Sin embargo, esto solo es útil para la observación i, y no para todas las observaciones de y. En este sentido, vamos a usar el método de Máxima Verosimilitud, el cual nos permite construir una función conjunta de la probabilidad de todas las observaciones de y (Hastie et al., 2015).

Para construir la función de longverosimilitud, debemos tener en cuenta que para cada observación $i = 1, \dots, n$, tenemos que se cumple:

$$f(y_i) = p_i^{y_i}(1 - p_i)^{1-y_i} \tag{20}$$

Es esta manera, para la primera observación se cumple que su probabilidad es $Pr\ Pr\ [y_i = y_1 \mid x] = p_i^{y_1}(1-p_i)^{1-y_1}$. Entonces, para obtener la probabilidad conjunta de estas $i = 1, \dots, n,$ suponiendo que estos sus probabilidades son independientes y distribuidas idénticamente, podemos plantear a función de verosimilitud es la siguiente:

$$L = f(\beta; y_1, \dots, y_n \mid x_1, \dots, x_n) \tag{21}$$
$$= \left\{p_i^{y_i}(1-p_i)^{1-y_1}\right\} \dots \{p_i^{y_n}(1-p_i)^{1-y_n}\}$$

$$L = f(\beta; y_1, \dots, y_n \mid x_1, \dots, x_n) = \prod_{i=1}^{n}(1-p_i)^{1-y_i} \tag{22}$$

Aplicando el logaritmo a esto, obtenemos la función de longverosimilitud. Ejecutar esta transformación no implica algún problema pues la función sigue siendo monótonamente creciente. De esta manera, tenemos su forma como:

$$ln\ L(\beta) = log\ f(\beta; y_1, \dots, y_n \mid x_1, \dots, x_n) \tag{23}$$
$$= log\ \left\{p_i^{y_i}(1-p_i)^{1-y_1}\right\} + \cdots$$
$$+ log\{p_i^{y_n}(1-p_i)^{1-y_n}\}$$

$$ln\ L(\beta) = y_1\ ln\ ln\ (p_i)\ + (1-p_i)\ + \cdots + y_n\ ln\ (p_i) \tag{24}$$
$$+ (1-p_i)$$

Expresándolo como una sumatoria pues tenemos término en común entre todas las n observaciones, podemos plantear la función de longverosimilitud como:

$$ln\ L(\beta) = \sum_{i=1}^{n} y_i ln\ (p_i)\ + \sum_{i=1}^{n}(1-y_i)\ ln\ (1-p_i) \tag{25}$$

$$ln\,L(\beta) = \sum_{i=1}^{n} y_i\, ln\,(p_i) + \sum_{i=1}^{n} ln\,(1-p_i) - \sum_{i=1}^{n} y_i\, ln\,(1-p_i) \tag{26}$$

$$ln\,L(\beta) = \sum_{i=1}^{n} y_i\, ln\left(\frac{p_i}{1-p_i}\right) + \sum_{i=1}^{n} ln\,(1-p_i) \tag{27}$$

Finalmente, en esta ecuación, podemos ver que el objetivo de maximizar la función de longverosimilitud es maximizar la probabilidad de que los eventos sean definidos por las funciones que la conforman.

Como $p_i = Pr\ Pr\,[x]$ es la función logística, entonces podemos reemplazar en ella tal que:

$$ln\,L(\beta) = \sum_{i=1}^{n} y_i\, ln\left(\frac{\frac{e^{x\beta}}{1+e^{x\,\beta}}}{1-\frac{e^{x\beta}}{1+e^{x\,\beta}}}\right) + \sum_{i=1}^{n} y_i\, ln\left(1-\frac{e^{x\beta}}{1+e^{x\,\beta}}\right) \tag{28}$$

$$ln\,L(\beta) = \sum_{i=1}^{n} y_i\, ln\left(\frac{\frac{e^{x\beta}}{1+e^{x\,\beta}}}{\frac{1+e^{x\,\beta}-e^{x\beta}}{1+e^{x\,\beta}}}\right) + \sum_{i=1}^{n} y_i\, ln\left(\frac{1+e^{x\beta}-e^{x\beta}}{1+e^{x\,\beta}}\right) \tag{29}$$

Despejando los términos $e^{x\beta}$:

$$ln\,L(\beta) = \sum_{i=1}^{n} y_i\, ln\left(e^{x\beta}\right) + \sum_{i=1}^{n} y_i ln\left(\frac{1}{1+e^{x\,\beta}}\right) \tag{30}$$

$$ln\,L(\beta) = \sum_{i=1}^{n} y_i(x\beta) + \sum_{i=1}^{n} [ln\,(1) - ln\left(1+e^{x\beta}\right)] \tag{31}$$

$$ln\, L(\beta) = \sum_{i=1}^{n} y_i(x\beta) - \sum_{i=1}^{n} ln\,(1 + e^{x\beta}) \tag{32}$$

Entonces, si p es el número de variables y N de observaciones, si tenemos un problema en el cual existen muchas variables o incluso existen más variables que observaciones $(p > N)$; es necesario penalizar la función por un término que permita materializar el intercambio entre sesgo-varianza (Hastie et al., 2015). Así, es posible añadir el regularizador que conocemos, $\left|\left|\beta\right|\right|_1 = \sum_{i=1}^{p} |\beta_j|$ en la forma en l_1 de los parámetros β; y λ, que es el parámetro que terminar el precio entre sesgo-varianza.

De manera paralela al modelo Lasso, el modelo de Logit Penalizado nos plantea la solución a un problema de maximización de N observaciones $\{(x_i, y_i)\}_{i=1}^{N}$ cuya solución está planteada como (β) tal que es:

$$\max_{\beta} \{ \sum_{i=1}^{n} y_i(x\beta) - \sum_{i=1}^{n} ln\,(1 + e^{x\beta})) + \lambda \left|\left|\beta\right|\right|_1 \} \tag{33}$$

Añadir este término de penalización que ya conocemos, $\lambda \left|\left|\beta\right|\right|_1$,nos permite asegurar que el modelo también pertenezca a la familiar de modelos dispersos puesto que permitirá seleccionar variables y tener un problema de optimización convexo.

Si derivamos esta ecuación para encontrar sus primeras condiciones de orden, podemos ver forman un sistema de ecuaciones no lineales en β. Así, para el modelo logístico presentado, tenemos este sistema como:

$$\frac{L(\beta)}{\partial \beta} = \sum_{i=1}^{n} [y_i - \frac{e^{x\beta}}{1 + e^{x\beta}}] x_i + \lambda \frac{\partial \left|\left|\beta\right|\right|_1}{\partial \beta} \tag{34}$$

En comparación con el modelo Lasso, usamos el algoritmo de Newton Raphson para poder optimizar las condiciones iniciales puesto que no nos encontramos en un problema lineal. Para su mayor desarrollo, revisar el Anexo F junto con las demás notas hechas en su desarrollo.

Para la siguiente investigación, el modelo *logit* penalizado será el que usaremos para comprobar nuestra hipótesis, así como el que usaremos para las pruebas de robustez y sensibilidad. Este modelo es beneficioso puesto a parte de solucionar el problema del rango de y, también tiene vectores solución dispersos y se queda con las variables relevantes. Esto mejora su capacidad predictiva, así como la interpretación de los resultados y nos permite mantener las ventajas del principio de la apuesta por la dispersión mencionada anteriormente. Sin embargo, su mayor desventaja es que, al ser no lineal, el costo computacional de estimarlo será mayor al que de un Lasso.

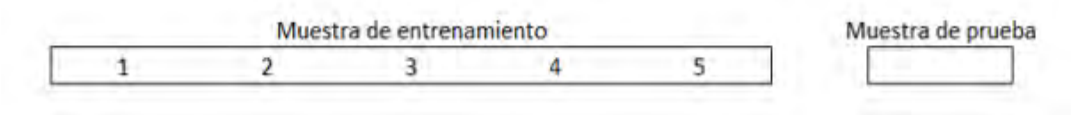
5.2. Validación cruzada

Los métodos de ML proveen una manera ordenada de predecir $\hat{y}$ en la cual se puede usar la base de datos misma para decidir cómo se realizará el intercambio sesgo-varianza. También proveen una manera de predecir $\hat{y}$ usando un conjunto extenso de variables y formas funcionales. Sin embargo, se debe tomar en cuenta que como estos coeficientes están ajustados para predecir $\hat{y}$, estos no pueden aferrarse a las normas de interpretación comunes como sí lo son los insesgados (Kleinberg et al., 2015).

Ahora, como queremos evaluar el ajuste del modelo fuera de la muestra, es necesario dividir un grupo de observaciones que cumpla el papel de muestra, y otro con el rol de "fuera" de esta. A este proceso de particionar los datos en dos muestras se le conoce como *data splitting*. El *training set* o muestra de entrenamiento es aquel conjunto de observaciones que se usará para entrenar al modelo y obtener el parámetro que reduzca el error de predicción de la mejor manera. Por otro lado, el *test set* o muestra de prueba es el conjunto de observaciones que se usará para evaluar cómo este parámetro elegido rinde fuera de la muestra (Picard & Berk, 1990).

Figura 4.

Data Splitting.



Fuente: Elaboración propia.

Lo ideal es que estos conjuntos sean disjuntos para evitar un problema conocido como *data leakage*. Esto sucede cuando observaciones que pertenecen al *training set* también pertenecen al *test set* y, de esta manera, son consideradas al momento de evaluar el rendimiento del modelo. No es ideal que esto suceda puesto que el *test set* tiene como objetivo simular los los datos de "la vida real", que son aquellos datos que no tenemos disponibles y no podemos considerar al momento de entrenar el modelo.

El *data leakage* puede producir serios asuntos al momento de medir el rendimiento del modelo fuera de la muestra; en particular, tendríamos indicadores que nos indicarían de un rendimiento muy alto pues estamos evaluando al modelo con datos con los cuales ya ha sido entrenado y, por lo tanto, con los cuales acertara. Las consecuencias de este rendimiento sobreestimado se verán al enfrentarse con los datos de la vida real, teniendo un bajo nivel de rendimiento.

La precisión en la predicción, medida por el error de predicción en el *test set*, es el error de generalización y depende de las observaciones que componen al *test set* y el valor del parámetro λ. El error en el *test set* será inflado si el valor de λ es muy alto, pues evitará que el modelo captura las señales que brindan los datos; mientras que también lo será si el valor de λ es muy pequeño pues genera sobreajuste.

Para encontrar el valor ideal de λ y a la vez obtener un modelo en el cual algunos coeficientes sean cero, debemos usar *Cross-Validation (CV)* (Hastie et al., 2015). En este proceso, usando solo los datos del *training set*, tenemos que generar muestras de entrenamiento y de prueba artificiales (estos últimos conocidos como *validation set*), dividiendo aleatoriamente los datos de este *set* en pliegues y estimando su desempeño entre las muestras aleatorias. Así, minimiza el estimado del error de predicción esperado en la muestra de entrenamiento (Hastie et al., 2015). Este procedimiento simula el cambio en la varianza que se presenta cuando alteramos las observaciones que componen la muestra en la que entrenamos el modelo.

Hallar el valor ideal de λ se da antes de la evaluación de la precisión del modelo pues primero debemos conocer sus parámetros, λ y los β_j; los cuales se han usado en el *CV*. Luego de haber obtenido λ, podemos recién comparar nuestro modelo con otros modelos evaluando el error de predicción del modelo entrenado, con los parámetros óptimos, λ y β_j, en el *test set*. Es decir, al momento de realizar el *CV* nos

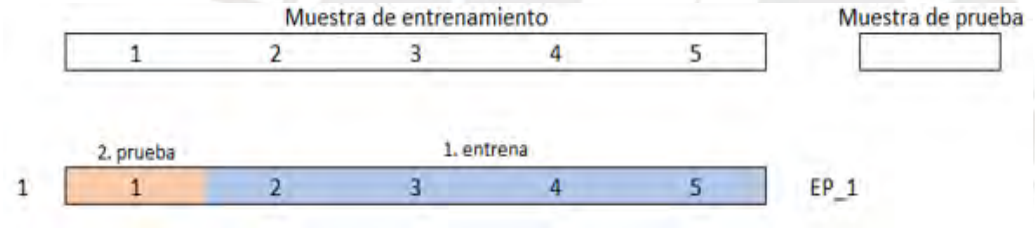
restringimos a usar el *training set*. Asimismo, el resultado que obtenemos del *CV* es el λ óptimo; no los β_j.

El *Cross-Validation* es un método no paramétrico que estima el error fuera de la muestra de manera implícita tal que podemos expresarlo como $Err = E[L(Y, \hat{f}(X)]$. Descrito de otra forma, busca estimar la diferencia entre el valor real fuera de la muestra (*validation set*) de una variable Y con el que fue predicho a partir de una función entrenada $\hat{f}(X)$ con las observaciones X que compone al *training set*.

Para esto, se generan dinámicamente dos muestras: una de entrenamiento (*training set, de color azul en la Figura 5*) que solo se emplea para construir el modelo de predicción, y otra de validación (*validation set, de color naranja en el Figura 5)* que solo se emplea para estimar el nivel de ajuste del modelo usando cierto λ y los β_j estimados en el *training set*.

Figura 5.

5 Cross Validation.



Fuente: Elaboración propia.

El *validation set* se consigue separando entre un 20% y 30% de los datos de manera aleatoria normalmente, y el resto de las observaciones componen al *training set* de la primera división. Para no perder eficiencia por recortar la cantidad de datos, el proceso *CV* usa un pliegue (*fold*) de la muestra para estimar el modelo y lo que sobra para la validación, siendo estas fracciones cortes secuenciales de todos los datos. Si se tiene K pliegues, entonces se le dominara al *CV* como *K-fold CV* (Hastie et al., 2015).

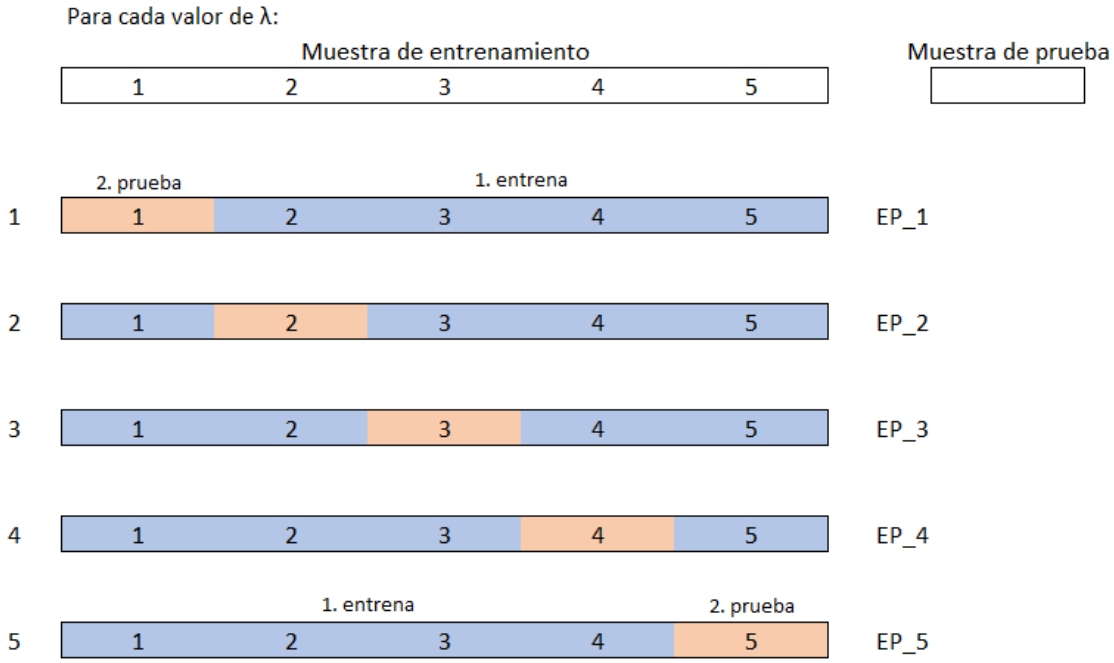
El procedimiento para realizar este método paramétrico es el siguiente. Primero, debemos dividir los datos que componen la muestra de entrenamiento en un número $K > 1$ de pliegues o conjuntos, de los cuales uno de ellos toma el rol de muestra de prueba (*validation set*) y los demás se agrupan y toman el rol de muestra

de entrenamiento (*training set*). Interpolamos cual de estos grupos toma el rol de *validation set* tal que cada pliegue pueda tener la oportunidad de jugar este rol.

Usando un modelo de Machine Learning, estimamos los errores medios de predicción al cuadrado para cada uno de los valores de λ (Hastie et al., 2015). Es decir, para un valor de, estimamos el error de predicción en los diferentes K validation sets que se tienen dentro de la muestra de entrenamiento y estimamos un promedio de esos errores de predicción. Gráficamente, podemos verlo en la Figura 5.

Figura 6.

5 Fold Cross Validation.



Fuente: Elaboración propia.

Como se puede notar, este proceso es repetido K veces para cada valor de λ; tal que se le da la posibilidad a cada K pliegue de tener el rol de muestra de prueba, mientras que los $K-1$ pliegues, en conjunto, juegan el rol de muestra de entrenamiento en esta submuestra de los datos que tenemos. Así, para un rango de valores de λ, tenemos K estimados del error de predicción por cada valor de λ que pertenece a este (Hastie et al., 2015). Para las $k = 1,..,K$ particiones, siendo la función estimada sobre la base de datos sin el $K-ésimo$ pliegue, $\hat{f}^{-k}$; el promedio estimado *CV* del error de predicción hallado es:

$$CV = \sum_{i=1}^{K} L(y_i, \hat{f}^{-K}(x_i)) \tag{35}$$

Dado un set de modelo estimados con diferentes parámetros β, tenemos a $\hat{f}^{-k}(x, \beta)$ como la $\beta - ésima$ estimación del modelo con el $k - ésimo$ pliegue de los datos removido. Así, se puede definir al promedio del error de predicción para cierto λ como:

$$CV = \sum_{i=1}^{K} L(y_i, \hat{f}^{-K}(x_i, \beta)) \tag{36}$$

Al promediar estos errores de predicción que pertenecen a un mismo λ pero con diferentes β estimados dependiendo del pliegue que se use como muestra de prueba, podemos producir una curva del error de *CV* en función de λ. La función $CV(\lambda)$ es un estimado del error de validación y si la graficamos para todos los valores de λ, se puede encontrar que el parámetro $\hat{\lambda}$ que minimiza este error. Esta se verá presentada en las figuras del valor promedio del AUC ROC en las muestras de entrenamiento y validación según lambda que presentemos en la sección de resultados.

Así, el modelo que elegiremos será $\hat{f}^{-k}(x, \hat{\lambda})$ y este se estimará con todas las observaciones de la muestra de entrenamiento que se asignaron en el *data splitting*. Siguiendo a la literatura, se usará un $K = 5$ tal que λ será elegido por un proceso de $5\ fold\ CV$ para un conjunto de valores de λ. Para poder evaluar qué tan bien acertó el modelo fuera de la muestra de entrenamiento, usaremos la comparación entre el valor predicho, $\hat{y}$, y el valor real, y_{real} (Davis & Goadrich, 2006; James et al., 2017). Los modelos de clasificación dan como resultado una variable dicotómica con etiquetas que son positivas (1) o negativas (0). En nuestro caso, el valor de 1 se interpreta como la etiqueta de “desempleo”, mientras que el valor de 0 como “empleo”.

Las decisiones que toma el modelo pueden ser resumidas en una matriz de confusión o tabla de contingencia. Esta tiene 4 categorías que comparan la condición predicha por el modelo y la condición real de la observación. Los verdaderos positivos (VP) son las observaciones positivas que fueron clasificados de manera correcta, mientras que los falsos positivos (FP) son aquellas observaciones negativas que fueron erróneamente clasificados como positivas (James et al., 2017). De la misma

manera, los verdaderos negativos (VP) son las observaciones negativas que fueron correctamente clasificadas como tal; mientras que los falsos positivos (FP) son las observaciones positivas que fueron equivocadamente clasificadas como negativas (James et al., 2017).

Estas decisiones son muy similares a los resultados al realizar una prueba de hipótesis con una hipótesis nula (Ho) y una hipótesis alternativa (Ha). Los verdaderos positivos nos ayudan a entender cuántas veces acertó el modelo en el evento "desempleo"; mientras que los falsos positivos representan las "falsas alarmas", lo cual es similar al error de tipo I, que sucede cuando aceptamos la hipótesis alternativa siendo falsa.

Asimismo, los verdaderos negativos nos ayudan entender cuántas veces el modelo rechazo correctamente el evento "desempleo", es decir, cuanto acertó en detectar el evento de "empleo"; mientras que los falsos negativos representan las faltas en las observaciones positivas, lo cual es similar al error de tipo II, que sucede cuando rechazamos la hipótesis alternativa siendo verdadera (James et al., 2017). Esto se puede resumir en la siguiente tabla:

Tabla 7.

Matriz de confusión.

		Condición predicha	
		Negativo (NP)	Positivo (PP)
Condición real	Negativo (N)	Verdadero negativo (VN)	Falso positivo (FP)
	Positivo (P)	Falso negativo (FN)	Verdadero positivo (VP)

Fuente: Elaboración propia. Adaptado de Davis & Goadrich (2006).

Si sumamos los verdaderos positivos (VP) y los falsos positivos (FP), obtenemos todos los verdaderos reales en nuestra muestra, y si sumamos los verdaderos negativos (VN) y los falsos negativos (FN), obtenemos los todos los falsos reales de nuestra muestra. Estas matrices serán presentadas para los resultados del *logit* penalizado, así como sus variantes estimadas para las pruebas de robustez y sensibilidad.

Existen diversos indicadores, en función de los aspectos específicos que se

desean analizar, para la evaluación del modelo[1] (James et al., 2017). El valor predictivo positivo, o *precision*, es el porcentaje de todos los positivos predichos que realmente eran positivos (VP/PP), mientras que la tasa de las falsas omisiones es el porcentaje de los negativos predichos que eran en verdad negativos (FN/PN) (James et al., 2017).

La evaluación del modelo, en tanto depende de la clasificación que dé, está sujeta al umbral. Uno más alto implica que menos observaciones serán clasificadas como positivas, lo que reducirá los FP; pero también implica que los FN incrementaron debido a que será más difícil clasificar los positivos reales como tales. Pasará lo inverso si es que decidimos que el umbral sea menor: los FN se reducirán, mientras que los FP aumentarán. Notaremos el efecto de esto en los resultados de la prueba de robustez y sensibilidad.

Por un lado, podemos establecer un umbral predeterminado (0.5); sin embargo, también es posible incorporar el valor del umbral al evaluar un modelo. La curva de la característica operativa del receptor o *ROC* por sus siglas en inglés muestra que la relación entre la tasa de positivos verdadero (VP/P) y la tasa de los falsos positivos (FV/N) a lo largo de diferentes valores del umbral (Davis & Goadrich, 2006).

El escenario ideal en la curva ROC es que esta se acerque al lado izquierdo superior pues en este la tasa de verdaderos positivos es del 100% mientras que la tasa de falsos positivos es 0%, lo cual implicaría que todos los positivos predichos son efectivamente positivos (Davis & Goadrich, 2006).

En nuestro caso, usaremos dos indicadores que cumpla con 2 condiciones: aquellos en la cual el valor que nos dé en la muestra de entrenamiento se parezca más entre a la muestra de prueba, y aquellos que nos permita evaluar qué tan bien acierta el modelo en predecir los positivos, que este caso es el evento de “desempleo” (Davis & Goadrich, 2006).

La primera condición nos permitirá omitir el problema de *overfitting* anteriormente mencionado, pues no estaremos sobre ajustando el modelo a los datos

[1] La tasa de positivos verdadero, o recall, es la probabilidad de detección de positivos y es el porcentaje de todos los verdaderos que fueron detectados correctamente (VP/P). La tasa de los falsos positivos es la probabilidad de “falsa alarma” y es el porcentaje de todos los negativos que fueron erróneamente clasificados como positivos (FV/N). La tasa de verdaderos negativos, o specify, es el porcentaje de todos los negativos que fueron detectados correctamente (VN/N) (James et al., 2017). La tasa de los negativos positivos es la probabilidad de “falsa alarma” y es el porcentaje de todos los negativos que fueron erróneamente clasificados como positivos (FV/N).

de entrenamiento, lo cual permitirá una mayor variabilidad fuera de la muestra; mientras que la segunda condición nos permitirá evaluar qué tan bien cumple el modelo su objetivo principal, que es predecir el desempleo en la PEA femenina urbana con objetivos de política.

Para nuestra investigación, la métrica que usaremos para evaluar el modelo será el área debajo de la curva ROC (o AUC ROC por sus siglas en inglés). Esto debido a que representa un equilibrio entre la tasa de verdaderos positivos (VP) y la de falsos positivos (FP) las cuales cumplen con las condiciones necesarias para la evaluación dentro y fuera de la muestra.

Por un lado, la tasa de VP nos ayuda a asegurarnos que se está optimizando la detección de los positivos, es decir, que las mujeres que son propensas a perder el empleo están siendo detectadas. En términos de política, esto nos asegura que la focalización de los recursos está dándose a las mujeres que los necesitan. Por el otro, la tasa de FP nos ayuda a asegurarnos que se está optimizando la asignación de estos recursos pues el grupo seleccionado está compuesto de positivos verdaderos en su mayoría.

Por otro lado, para poder hallar aquel λ ideal que nos ayuda a minimizar el error de predicción, es necesario tener un conjunto de posibles valores de los cuales evaluaremos cuál es el mejor usando *CV*. Existen diferentes maneras para esto, pero siempre fijando un rango establecido de posibles valores de λ. Estos valores pueden determinarse mediante una secuencia con un valor de saltos determinado, *grid search*, o de manera aleatoria, lo que se conoce como *random search* (James et al., 2017).

En la presente investigación, usaremos la primera aproximación, *grid search,* puesto que no solo nos permite una evaluación más ordenada de los posibles valores, sino que también nos permite ver la evaluación del valor de la métrica para evaluar el error de predicción que nos ayude a escoger el λ que minimice el promedio de estos (James et al., 2017). Para establecer el rango en el que evaluaremos, empezaremos desde que el valor de λ es 0, es decir, simulando cómo si no fuera un modelo disperso. Así, probaremos diferentes rangos de posibles valores de λ hasta encontrar aquel que nos permite visualizar un máximo valor y que cumpla con las 2 condiciones mencionadas.

Adicionalmente, si bien podemos clasificar las observaciones en pliegues de manera aleatoria en el *CV*, también podemos determinar, a partir de las características

de las observaciones, cuál es la distribución conjunta en cada uno de los pliegues. En otras palabras, también podemos estratificar las observaciones que poseemos para que, de esta manera, cada pliegue contenga una muestra muy similar a la otra en estas características (Diamantidis et al., 2000).

Esto presenta 2 ventajas en comparación con una aproximación aleatoria. Por un lado, los pliegues serían lo más cercano a ser representativos de la muestra en las características que consideremos. Si bien no estamos usando los pesos muestrales al distribuir las observaciones en los pliegues (Diamantidis et al., 2000), estas variables pueden ser niveles de inferencia válidos para los paneles de nuestra muestra; presentando una varianza lo suficientemente alta. Por ejemplo, la variable que describe las regiones a la que pertenece una mujer tienen más del 15% de varianza en nuestra muestra. En general, no es común el uso de pesos muestrales al momento de revisar la literatura del *CV* en la literatura. Por ello, dejamos al uso de pesos en *Stratified-CV* como agenda de investigación.

Asimismo, otro beneficio de no optar por una aproximación aleatoria es que la distribución conjunta será similar entre la muestra de tratamiento y de validación. Así, el rendimiento será mejor pues, si asumimos que cuando la muestra de entrenamiento en la que se entrena el modelo es similar a la muestra de validación, la probabilidad de una predicción correcta es mayor (Diamantidis et al., 2000). Esto no constituye un aspecto negativo ni introduce sesgo puesto que la muestra de prueba es verdaderamente una simulación de las observaciones de la población total y de la que no se puede hacer uso puesto que nos servirá para evaluar al modelo (Diamantidis et al., 2000).

Asimismo, considerar la estratificación de las observaciones a partir de las "clases" de la variable respuesta, y, presenta esta misma ventaja. Para entrenar a nuestro modelo es importante considerar el desbalance anual que existe entre ambos posibles resultados, que es que la mujer esté desempleada o empleada. Al igual que para las variables X, esto no representa un problema pues, para evaluar al modelo, estaríamos simulando la distribución de la variable y para la población "real".

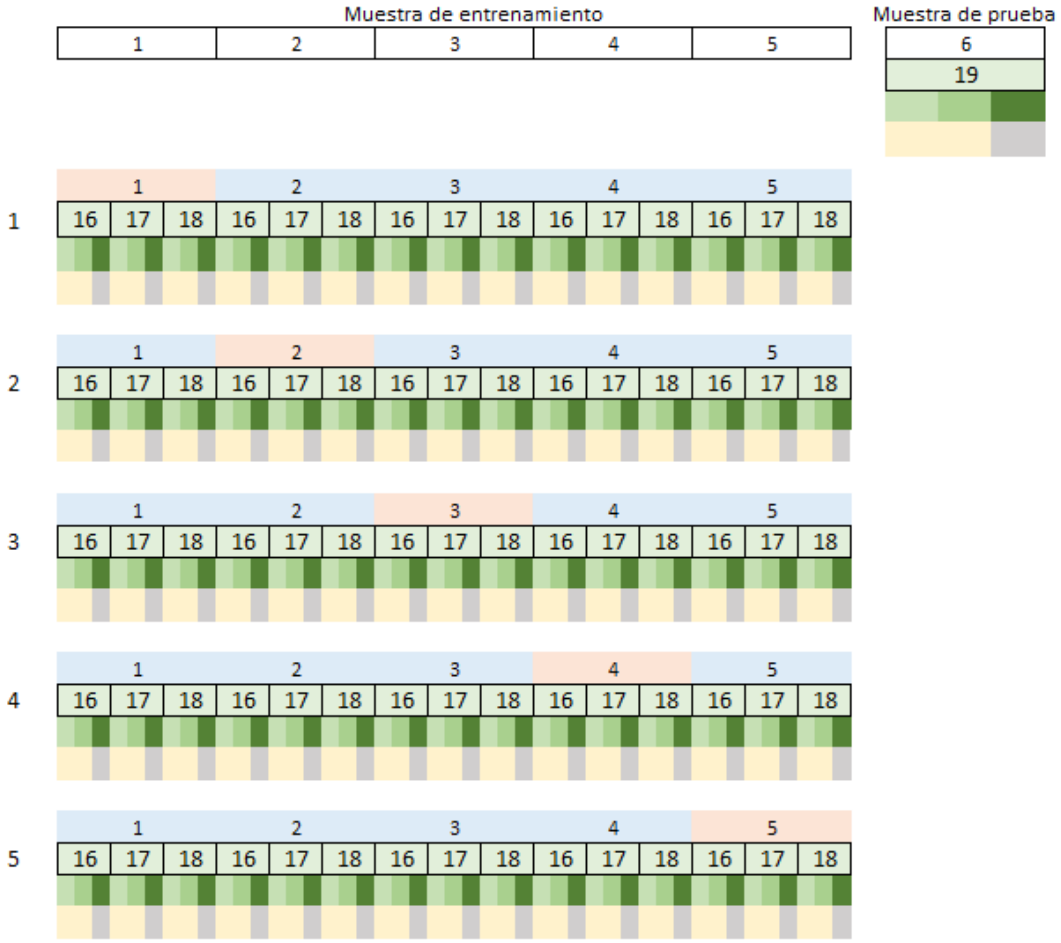
Para la presente investigación, el año inicial al que pertenece la mujer y la región a la que pertenece serán las variables X que usaremos para estratificarlas. Primero, es importante tomar en cuenta la temporalidad de las observaciones pues el ciclo económico afecta el mercado laboral. Asimismo, el mercado laboral varía entre

diferentes regiones debido a las actividades económicas a las que se dedican. Asimismo, usaremos la variable Y para también estratificar los pliegues en base a sus clases (empleada o desempleada) y así tomar en cuenta el desbalance.

Tenemos en cuenta que existen otras variables que no necesariamente permiten niveles de inferencia pero que se pueden usar para agrupar a la muestra, como la probabilidad de selección, departamentos, etc; sin embargo, consideramos que, al ser este una investigación exploratoria haciendo uso de nueva metodología, estas variables son suficientes para hacer frente a la estacionalidad del mercado laboral por años y regiones. Por ello, consideramos que sería ideal si es que en futuras investigaciones consideran usar estas variables.

Figura 7.

Stratified Group Cross Validation.



Fuente: Elaboración propia.

De esta manera, hemos considerado que la muestra puede dividirse, considerando estas características que pertenecen a la matriz X de nuestra ecuación, podemos dividir a nuestra muestra en un total de 12 “grupos”: las 3 regiones del Perú presentes en los 4 años (2016, 2017, 2018 y 2019) a los que pertenece nuestra muestra.

Considerando esto, haremos uso del *Stratified Group Cross Validation*, que es una variante del *CV-K fold* en la medida que los pliegues están estratificados con grupos no solapados y balanceados por la clase a la que pertenece. Esta validación cruzada busca construir pliegues conservando el porcentaje de cada clase en cada grupo que se forma. Gráficamente, podríamos presentar esta validación cruzada como la estratificación por año-region (grupos representados por las *áreas verdes en la Figura 6*) y los valores de la variable a predecir (*clases representadas por las areas amarillo y gris en la Figura 6*) en los pliegues que conforman la muestra de entrenamiento y el pliegue de la muestra de prueba.

Con este tipo de *CV* estamos realizando lo que se conoce como *validation out-of-time* o validación fuera de tiempo. Esto pues estamos usando un año en específico, 2019, para la muestra de prueba o *test set* . Esto es beneficioso para la investigación pues nos permite cumplir con el objetivo de predecir el porcentaje de mujeres que quieren su empleo en este año usando la información de las mujeres en años pasados.

5.3. Procesamiento de datos

Como se mencionó en la sección de base de datos, las características recopiladas y pre-seleccionadas de las mujeres que pueden ser representadas como la matriz X pasaron un proceso compuesto de 3 etapas. La primera etapa fue un descarte por cantidad de observaciones faltantes; mientras que la segunda etapa fue ingeniera de datos, con un proceso de *winsorizing,* o transformación de los datos a las variables continuas y el proceso de *one-hot-encoding* a las variables categóricas. Así, la tercera etapa fue la de selección de variables por su correlación bivariada.

Primero, se descartó a las variables que contaban con más del 90% de observaciones faltantes o *missings.* Esto ya que este tipo de variables normalmente eran preguntas complementarias que estaban condicionadas a respuestas anteriores y que necesitaban de un procesamiento más detenido. Asimismo, para el objetivo de esta investigación, el análisis de los *missings* en las respuestas no genera una ganancia adicional en términos de información.

Segundo, se limitó sus valores extremos o *outlier* al percentil 95 a las variables continuas de las que pasaron el primer filtro para evitar que afecten el entrenamiento del modelo. A este proceso se le conoce como *winsorizing*. Paralelamente, se les realizó el proceso de *one-hot-encoding* a las variables categóricas. En otras palabras, se les creó una variable binaria por cada categoría de tal manera que cada variable representa una característica particular de una dimensión de la mujer. Como se verá en la sección de resultado, esto nos permite conocer cuál es el perfil de una mujer vulnerable.

Finalmente, para descartar a las variables que no aportan a la predicción del modelo por ser muy similares, se realizó un análisis de la correlación de variables. Así, se seleccionó a los pares de variables que tenían una correlación alta, estableciendo un umbral de 0.8 para esta definición. Se descartó aquella variable que estaba menos correlacionada con la variable a predecir.

Antes de proceder a la estimación de los parámetros, se estandarizaron las variables ya que esto nos asegura que podamos interpretar el peso como la importancia de las variables en las estimaciones. Cabe mencionar que las variables que señalan los departamentos y años de la mujer en su periodo inicial, que eran 29, fueron excluidos de este proceso puesto que eran necesarias para el *Stratified Group Cross Validation* que se usará luego para lidiar con la distribución de la muestra de entrenamiento.

5.4. Desbalance de clases

Asimismo, es importante mencionar que la base de datos que hemos construido es no balanceada en las clases que componen y. Es decir, la distribución de las observaciones entre el tipo de etiqueta que tiene la variable y , mantiene el empleo o lo pierde, está bastante desequilibrada. Con mayor detalle, solo el 11.74% de nuestra muestra que componen a las mujeres en su periodo inicial del 2016 al 2019 tiene el valor de 1 en su variable respuesta, es decir, está clasificada como que pierde el empleo; mientras que solo el 10.48% para aquellas que pertenecen del 2016 al 2018.

Esto genera un problema pues el modelo tenderá a predecir mejor la clase mayoritaria debido a que existe mayor probabilidad de que atine en esta, ignorando a la clase minoritaria. En la medida que existan más observaciones de mujeres que no perdieron su empleo, será más factible que el modelo se entrene mejor en predecir este evento. El resultado será que el error de estimación que evalúe estas dos

clasificaciones tendrá un valor alto pues considerará la buena predicción de la clase mayoritaria y no tendremos un buen ajuste para la clase minoritaria, la cual es nuestro objetivo en esta investigación (Blagus & Lusa, 2013).

Asimismo, estamos usando una base de datos con alta dimensionalidad, lo cual eleva el sesgo de clasificación a la clase mayoritaria, incluso cuando no existe mucha diferencia entre estas. Esto pues, para la clase minoritaria existen mayores diferencias entre la muestra de entrenamiento y la población "real" debido a la mayor variabilidad entre muestras que se usan para entrenar el modelo. Por ello, el modelo no podrá ser entrenado en datos que representen adecuadamente a la verdadera proporción de la clase minoritaria en la población real (Blagus & Lusa, 2013).

En este sentido, haremos uso de un método de remuestreo que nos permita solucionar parcialmente este problema de desbalance en el entrenamiento del modelo. Estos métodos tienen el objetivo de balancear las clases en la muestra de entrenamiento para poder evitar los problemas mencionados (Blagus & Lusa, 2013; He & García, 2009).

Algunas tienen el objetivo de aumentar las observaciones de la clase minoritaria (*oversampling*) o reducir las observaciones de la clase mayoritaria (*undersampling*) tal que podamos tener una muestra más balanceada (Blagus & Lusa, 2013). Usar alguna de estas técnicas causará cambios en la distribución de clases que tenemos al entrenar el modelo; sin embargo, no causará cambios en el porcentaje que presentemos como el resultado para comprobar la hipótesis en esta investigación puesto estará basado en el porcentaje de las observaciones que tengamos.

Por un lado, el *undersampling* remueve un subconjunto de observaciones de la muestra de manera aleatoria de la clase mayoritaria con el objetivo de obtener el balance entre clases. Es decir, reduce las observaciones que componen la clase mayoritaria, sin cambiar la cantidad de observaciones de la clase minoritaria (Blagus & Lusa, 2013). Por otro lado, el *oversampling* genera observaciones sintéticas de la clase minoritaria usando información disponible, sin cambiar la cantidad de observaciones de la clase mayoritaria. En otras palabras, aumenta las observaciones que pertenecen a la clase minoritaria, sin cambiar la cantidad de observaciones de la clase mayoritaria (Blagus & Lusa, 2013).

Un caso específico de este último método de remuestreo es *Synthetic minority over-sampling tecnhique* o SMOTE. Esta es una técnica de sobremuestreo que no

crea simples duplicados de las observaciones de la clase minoritaria, sino que usa las características de diferentes muestras de esta para generar nuevas observaciones muy similares a las que ya se tienen de esta clase (Blagus & Lusa, 2013).

Para explicar mejor estos métodos, es necesario establecer algunos términos. Si x_{ij} es el valor de la variable $j-ésima$ para la $i-ésima$ muestra compuesta de observaciones tales que estas pertenecen a una clase c donde $c = 1{,}0$ en la variable y_i; $x_i \in X$ es el conjunto de características de las observaciones que componen una muestra. Asimismo, la distribución gaussiana de la media μ y la desviación estándar σ está definida como $N(\mu, \sigma)$ y mientras que la distribución uniforme está definida entre $[0{,}1]$ con tal que tenemos $U(0{,}1)$.

Tomando en cuenta esto, es posible definir el método de *undersampling* y *oversampling* aleatorio. Considerando $m = |S|$ ejemplos de muestras de entrenamiento, donde $S = \{(x_i, y_i)\}$ tal que $i = 1, \dots, m$, definimos los $S_{min} \subset S$ y $S_{may} \subset S$. De esta manera, S_{min} es el conjunto de observaciones de la clase minoritaria, en S y S_{may} es el conjunto de la clase mayoritaria mientras se cumpla que $S_{min} \cap S_{may} = \{\Phi\}$ y $S_{min} \cup S_{may} = \{S\}$.

Por un lado, el *oversampling* aleatorio implica añadir un subgrupo E de observaciones del conjunto de la clase minoritaria. Este subgrupo no es más que el resultado de realizar un muestreo aleatorio de observaciones de S_{min} y replicándolos en S, aumentando así su tamaño. Por otro lado, el *undersampling* aleatorio implica seleccionar aleatoria un subgrupo de observaciones de la clase mayoritaria para luego removerlos de S tal que ahora $|S| = |S_{min}| + |S_{may}| - |E|$.

Si bien por un lado nos ayudan a reducir el problema de una muestra no balanceada, también presenta sus limitaciones ya que puede sesgar el proceso de aprendizaje del modelo al momento de entrenarlo. Por ejemplo, usar el *undersampling* puede generar que, al remover información, el modelo puede no considerar ciertas señales características de la clase mayoritaria. Asimismo, usar el *oversampling* puede generar overfitting debido a la repetición de las observaciones de la clase minoritaria. Sin embargo, esto se puede resolver si la selección del subgrupo ya no es aleatoria.

En el caso del SMOTE, para el conjunto de variables que componen las observaciones de una muestra (X), cinco observaciones con las menores distancias euclidianas entre ellas son seleccionadas usando el método de vecinos cercanos o *nearest neighbours*. Una de estas cinco observaciones (X^R) es seleccionada

aleatoriamente tal que se usará para poder generar las observaciones que componen a la muestra SMOTE de observaciones nuevas. Si la muestra de la clase minoritaria (n_{min}) está compuesta por menos de cinco observaciones, entonces se selecciona $n_{min} - 1$ observaciones con las menores distancias euclidianas (Blagus & Lusa, 2013).

De esta manera, siendo u elegido aleatoriamente de $U(0,1)$ tal que es igual para todas las variables que componen una muestra pero que varía entre muestra *SMOTE* que se genera, podemos plantear al nuevo conjunto de observaciones que componen a la muestra *SMOTE* como:

$$S = X + u\,(X^R - X) \tag{37}$$

Que el u cumpla estas propiedades con referencia a la muestra *SMOTE* creada este en línea con las observaciones de las muestras originales usadas para generarlas, X^R y X (Blagus & Lusa, 2013). Debido al tipo de generación de observaciones, las propiedades teóricas de la muestra *SMOTE* que han sido construidas, que son incluidas en la muestra minoritaria, cumplen ciertas características que importan dependiendo del modelo a evaluar.

Por un lado, el valor esperado de la clase minoritaria considerando la nueva muestra SMOTE no cambia, pero sí reduce su variabilidad. Esto tiene consideraciones prácticas mínimas en los modelos que basan sus predicciones usando el valor de las medias o sus variaciones, como es el modelo Lasso. Asimismo, debido a que genera una reducción en la variabilidad puesto que las observaciones son construidas unas a partir de otras, tiene impacto en la selección de variables. Debido al cambio en la variabilidad, los *p-values* de las variables son más pequeños para la muestra de la clase minoritaria que considera a la muestra *SMOTE* en comparación con la que no la considera (Blagus & Lusa, 2013).

Adicionalmente, el *SMOTE* produce correlación entre algunas de las observaciones que componen la muestra debido a que justamente son creadas sintéticamente por un conjunto de ellas; sin embargo, no genera una correlación entre las variables de estas observaciones. En términos prácticos, esto implica que *SMOTE* es problemático para los clasificadores que asumen independencia entre las distribuciones de las muestras, como lo es el *logit penalizado* (Blagus & Lusa, 2013).

Por último, debido a que generamos observaciones muy similares, *SMOTE* modifica la distancia euclidiana entre las muestras de entrenamiento y la muestra de la clase minoritaria que toman en cuenta la muestra *SMOTE* generada. Debido a que las observaciones muestras de entrenamiento son más cercanas a las observaciones muestras SMOTE generadas que a las observaciones de la muestra original, se genera un sesgo de clasificación (Blagus & Lusa, 2013).

Este involucra los modelos que usan la distancia euclidiana para medir la cercanía entre muestras, tenderán a clasificar más las observaciones como pertenecientes a la clase minoritaria pues la composición de las muestras de entrenamiento será muy similar a las observaciones de la clase minoritaria creadas (Blagus & Lusa, 2013). Ni el modelo Lasso ni el logit penalizado son parte de la familia de este tipo de modelos.

Para la presente investigación estimaremos el modelo *logit* penalizado usando el método de remuestreo de *undersampling* para predecir qué mujeres serán vulnerables en el 2019. Primero realizaremos el proceso de *CV* para poder hallar el λ óptimo usando un umbral preestablecido (0.5) para clasificar a las mujeres como vulnerables. Con este valor, buscaremos hallar el umbral óptimo estimando el área debajo de la curva ROC usando también la muestra de entrenamiento. Así, con estos 2 hiperparametros, estimaremos el modelo *logit* penalizado en la muestra de prueba.

Sin embargo, para poner a prueba la robustez de estos resultados realizaremos cambios en ciertos supuestos. El primer cambio será probar el *SMOTE* como método de remuestreo en el modelo *logit* penalizado (1). Asimismo, con el λ óptimo, estimaremos el modelo nuevamente con el valor predeterminado de 0.5 y comprobaremos si es que los resultados se mantienen para ambos métodos de remuestreo con el mismo modelo (2-3).

Capitulo 6. Resultados

En la siguiente sección redactamos los resultados encontrados en el orden en que se realizaron los pasos. Para comenzar, presentaremos los resultados del *Stratified Group Cross-Validation* y la elección de la λ que tenga la mayor precisión usando la PEA femenina urbana. Esto se realizará usando la muestra de entrenamiento de los datos del 2016 al 2018 de nuestra base de datos previamente mencionada.

Seguiremos con la presentación y estimación del modelo *Logit Penalizado* usando *Undersampling* con el λ elegido. Esto pues aún necesitamos estimar su nivel de ajuste en la muestra de prueba, es decir, su nivel de rendimiento *out-of-sample*. Una vez analizado el modelo, pasaremos a estimar la predicción de la PEA femenina urbana vulnerable que pertenece al 2019. Usando el hiper parámetro elegido λ así como los parámetros β, se usarán los datos panel del 2019 para estimar quiénes son aquellas mujeres que perderán su empleo el año siguiente según nuestro modelo.

Con esta predicción, exploramos el perfil de las mujeres vulnerables a la pérdida de su empleo en el 2019. Describiremos sus principales características educativas, de salud, de empleo, del hogar y equipamiento, de vivienda, de servicios básicos y de TIC para poder crear un perfil. Las variables que se usarán serán aquellas que son de mayor influencia para la predicción. De igual manera, realizaremos una prueba de hipótesis del porcentaje de mujeres de la PEA femenina urbana en el 2019 que se predijo que en el 2020 perdería su empleo. Como hemos planteado anteriormente, para la presente investigación planteamos que esta será superior al promedio de años pasados debido a la implicancia de la crisis en el mercado laboral (18%).

Haciendo esto, compararemos el estado de la PEA femenina urbana del 2019 en el 2020, verificando si es que sucedió lo que predijo el modelo. Es decir, compararemos el estado predicho de las mujeres que, según este, perderían su empleo o pasarían a la inactividad en el 2020 y lo que realmente pasó con ellas en este año usando los datos panel. Esto nos permitirá no solo comprobar el nivel de ajuste del modelo *out-of-sample*, sino también comprobar el nivel de ajuste *out-of-time*, pues estaríamos usando datos de un año que no se usaron para entrenar el modelo.

Cabe mencionar que para esto se creó una variable dicotómica, y, que tendría el valor de 1 si es que, en un panel bianual, una mujer que tuviera empleo en el periodo

inicial lo perdería en el segundo periodo (ya sea por desempleo o inactividad); mientras que tendría el valor de 0 si es que mantendría este empleo. Luego de esto, nos quedamos con la observación del periodo inicial al que pertenece la mujer en su panel bianual del 2016 al 2020.

Como ya se mencionó, la base de datos ha sido el resultado del apilamiento de datos panel del 2016 al 2020 de los diferentes módulos de la ENAHO Panel. Para un resumen de las variables que se incluirán en el modelo, revise el Anexo G. Luego de seleccionar las variables más relevantes según la literatura, esta base de datos ha pasado por tres etapas secuenciales: el descarte por porcentaje de *missings*, la ingeniería de datos, que consiste en el proceso de *one-hot-encoding* a las variables categóricas y de *winsorizing* de las variables continuas, y la selección de variables mediante el análisis de sus correlaciones.

La base, después del proceso de construcción de variables y selección, consiste en un total de 23,234 observaciones panel (solo nos quedamos con la observación inicial) y 107 variables, que incluyen el identificador único. La primera etapa, que consistía en el tratamiento de *missings,* nos llevó a descartar un total de 18 variables, pues estas poseían más de un 90% de *missings*.

Luego de descartar estas variables, el *winsorizing* consistió en reemplazar el valor de los *outliers* de las variables continuas con el valor promedio, lo cual nos permite no descartar los valores que podrían generar distorsiones. El *one-hot-encoding* aumentó la cantidad de variables a 239 debido a transformar las opciones de las variables categóricas en variables individuales, incluyendo los *missings*.

De la misma manera, la selección de variables mediante el análisis de sus correlaciones redujo la cantidad de variables a 216. Se evaluaron los pares con alta correlación (mayor o igual a 80%) y seleccionamos a aquella del par con mayor correlación con la variable Ypues contribuiría con la estimación. La limpieza de estas variables nos asegura tener una muestra con las variables más relevantes para la predicción.

Luego de esto, se descartaron las observaciones que tenían *missings* restantes con el fin de poder correr los modelos. Este proceso dejó un total de 18,293 observaciones y 224 variables que se usaron para los siguientes pasos. Cabe mencionar que las variables de identificación de año, de la región natural y su interacción, así como la variable de identificación de observación no fueron consideradas dentro de este proceso pues son necesarias para el proceso de *CV*.

6.1. Cross Validation

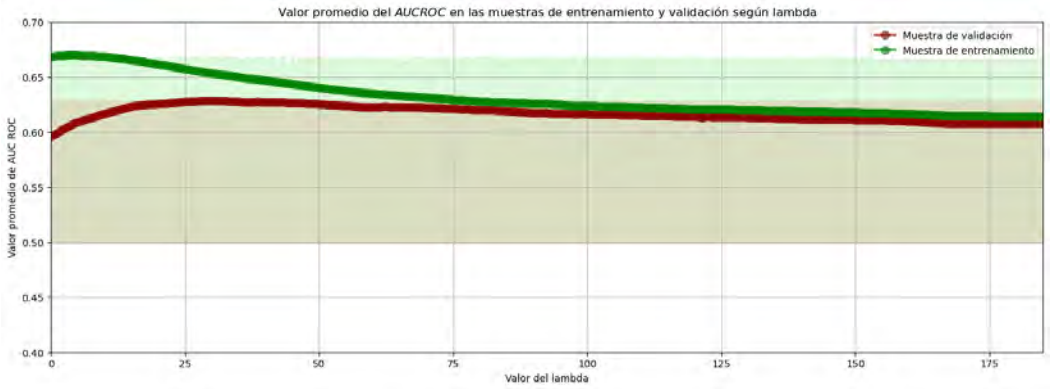
Empezamos con el proceso de *Stratified Group Cross Validation*, en el cual evaluamos cómo evoluciona la métrica de *la curva ROC* a lo largo de un rango de valor de λ, así como la evolución de los valores que toman los β. Para escoger el rango de λ, se empezó fijando el valor mínimo en un valor muy cercano a 0, y alternando diferentes valores hasta que se podía ver una clara caída constante en el área bajo la curva (*AUROC*) En este caso, el rango de este valor irá de 0.001 a 500 con unos 1000 valores intermedios.

A continuación, presentamos la evolución del valor promedio del *área de la curva ROC* en cada muestra de validación evaluada en el proceso de *Stratified Group Cross Validation*. Al examinar la gráfica, podemos notar que cuando λ es 30.03, el promedio de *AUROC* toma el valor más alto posible (62.85%) en la muestra de validación.

El valor del promedio de esta medida suele oscilar entre valores mayores al 50% según sus bandas de confianza. Aun así, es posible notar que existen 2 disminuciones en los extremos de la figura. Como es de esperarse, el promedio de esta medida toma valores más altos de manera sistemática en la muestra de entrenamiento (56.04%) que en la de prueba (55.07%), aunque sus intervalos de confianza se traslapan. Como se mencionó anteriormente, esto es una prueba de que no existe un sobreajuste del modelo al usar la muestra de entrenamiento.

Figura 8.

Valor promedio del AUC ROC en las muestras de entrenamiento y validación según λ usando Undersampling.



Fuente: Elaboración propia.

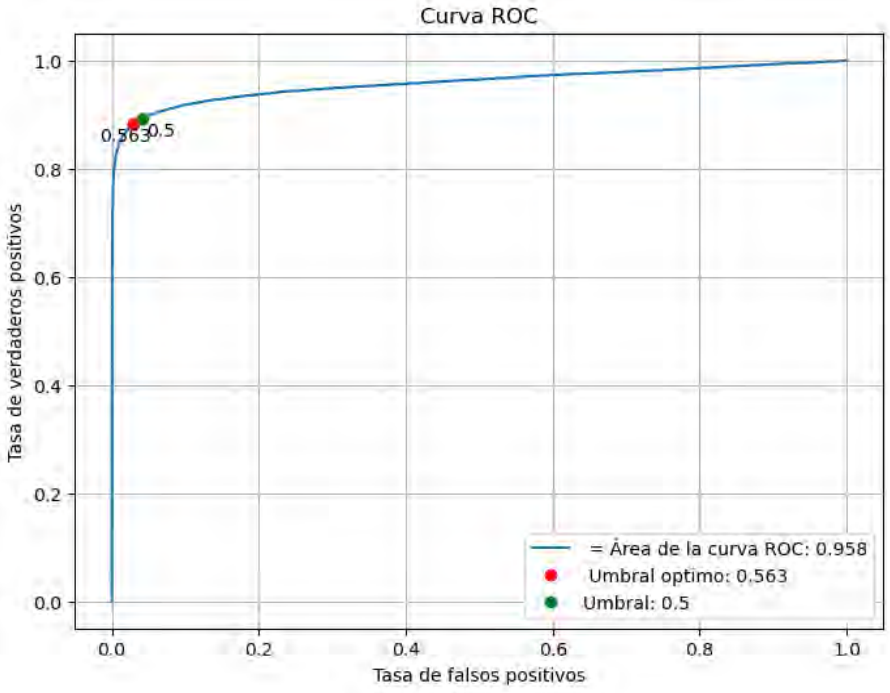
6.2. Elección del umbral

Hasta ahora, hemos realizado la elección del λ teniendo como fijado el umbral de 0.5 para clasificar a las mujeres entre vulnerables y no vulnerables. Sin embargo, es posible hallar el umbral que maximice el *AUROC* realizando una serie de iteraciones con los posibles valores que puede tomar el umbral en una serie de submuestras de nuestra muestra.

Usando el λ óptimo encontrado en el modelo logit penalizado con el método de remuestreo *undersampling*, estimamos la curva ROC para encontrar cual es el valor del umbral que optimiza su área para unas 1,000,000 submuestras que mantengan la misma distribución de las clases. Como podemos notar, sin importar que umbral usemos, el λ óptimo nos garantiza un área de un 0.96, lo cual significa que en un 96% de mujeres fuera de la muestra son clasificadas de manera adecuada usando este valor.

Figura 9.

Curva ROC con el umbra óptimo y el umbral hallado usando Undersampling.



Fuente: Elaboración propia.

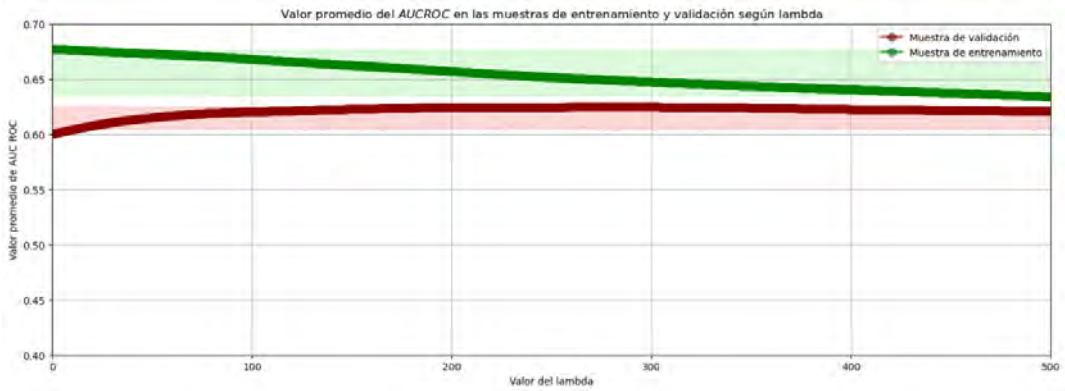
Ahora, podemos notar que si bien el valor de 0.5 para el umbral nos da una ratio aceptable entre la tasa de VP y la tasa de FP, el valor óptimo del umbral que maximiza

el área de la curva ROC es de 0.56 (0.5634). Como hemos mencionado en la sección de metodología, la necesidad de un umbral más alto refleja que el modelo debe clasificar menos observaciones como positivas, lo cual reducirá los FP. Asimismo, esto también significa un aumento de los FN ya que será más costoso clasificar los positivos reales como tales.

Usando el método de remuestreo SMOTE con el mismo modelo para encontrar a la curva ROC nos da un resultado similar: la necesidad de tener un umbral mayor para clasificar a las mujeres como vulnerables. En el proceso de Stratified Group Cross Validation, vemos que cuando es 286.28, el promedio de la curva ROC toma el valor más alto posible (62.6%). El valor del promedio de esta métrica es normalmente mayor al 60% según sus bandas de confianza. Aun así, es posible notar que existen 2 disminuciones en los extremos del gráfico. El promedio de esta medida toma valores más altos de manera sistemática en la muestra de entrenamiento (65.28%) que en la de prueba (62.12%), aunque esta vez sus intervalos de confianza no se traslapan, lo cual nos podría manifestar un problema de sobreajuste usando SMOTE.

Figura 10.

Valor promedio del AUC ROC en las muestras de entrenamiento y validación según λ usando SMOTE.



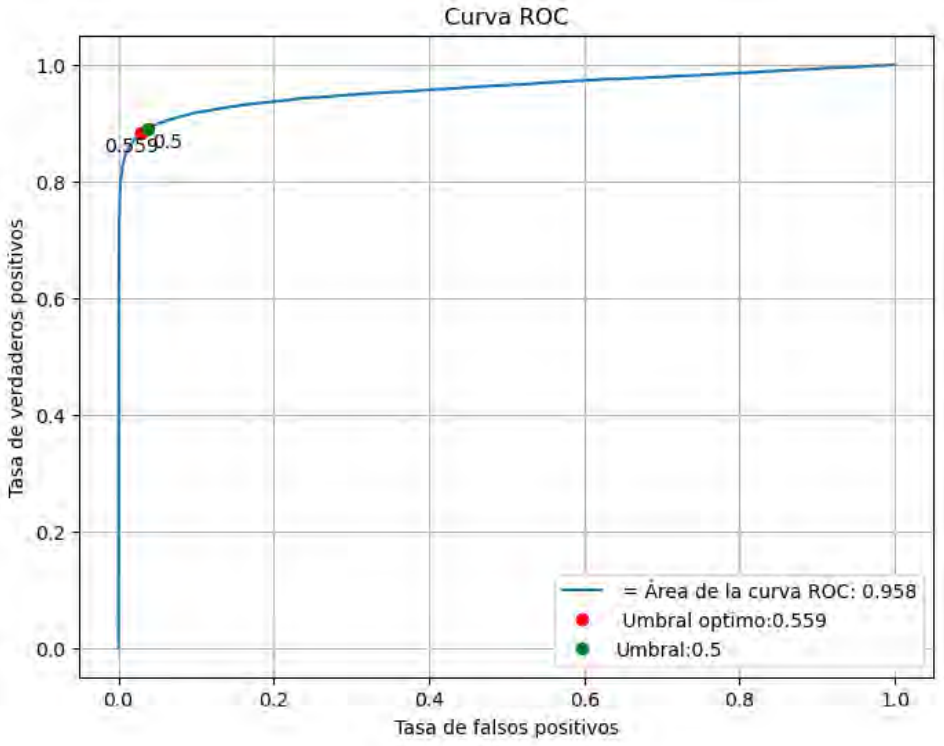
Fuente: Elaboración propia.

Usando este valor de λ, estimamos, para unas 1,000,000 submuestras que mantengan con el mismo porcentaje de clases, la curva ROC para y encontramos que 0.56 (0.5586) también es optimizar el *AUROC*. En este sentido, encontramos que el valor óptimo para el umbral que utilizará el modelo para clasificar a las mujeres entre vulnerables o no es 0.56.

Esto nos demuestra que, si bien ambos métodos de remuestreo buscan lidiar con el desequilibrio de las clases, aún existe la posibilidad de clasificar de mejor manera la vulnerabilidad del empleo en las mujeres. Ahora, con ambos parámetros encontrados, estimaremos el modelo logit penalizado usando el *undersampling* como método de remuestreo.

Figura 11.

Curva ROC con el umbra óptimo y el umbral hallado usando SMOTE.



Fuente: Elaboración propia.

6.3. Estimación del modelo logit penalizado con un umbral de 0.56 usando el método de remuestreo undersampling

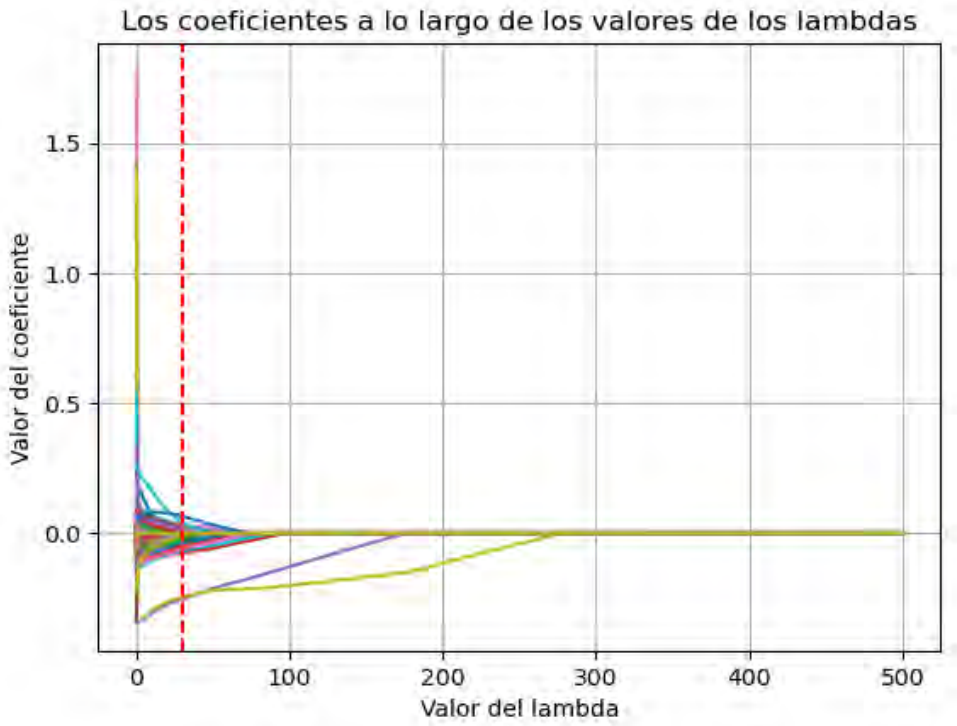
Con los resultados previos también podemos ver cómo evolucionan los valores de los coeficientes β según aumenta el término de regularización λ. Mientras mayor sea este, se espera que los valores de β sean menores debido a que será más costoso mantener un coeficiente diferente a 0. En la figura 11, la mayoría de las líneas se acerca al valor de 0 mientras avanzamos en este.

Las variables que persisten a lo largo de este aumento (es decir, que son diferente de 0) de la penalización del modelo hasta llegar al valor óptimo de la λ son 210. De estas, las 15 con valores absolutos mayores son: el ingreso laboral total, el

ingreso promedio mensual proveniente de trabajo, la edad en años cumplidos, si es jefa del hogar, agricultura, pesca y minería como los sectores de su ocupación principal, si reside en el departamento de Cusco, el ingreso secundario total y si es que tiene limitaciones de forma permanente para entender o aprender.

Figura 12.

Los coeficientes a lo largo de los valores de las lambdas para el modelo principal.



Fuente: Elaboración propia.

Asimismo, también lo son si el tamaño de la empresa en la que trabajaba es de 1 a 10 trabajadores, si es una trabajadora independiente en la actividad principal y secundaria, el nivel de educación secundaria, ingreso per cápita mensual a precios de Lima de la transferencia monetaria Bono Gas, los gastos en muebles y enseres reales y el último gasto mensual por consumo de agua.

Su resiliencia al aumento del parámetro de penalización las consagra como las variables más importantes para predecir el estado de desempleo o inactividad de las mujeres. Particularmente, los coeficientes de las variables con un valor absoluto mayor al 10%, son el ingreso promedio mensual proveniente de trabajo y el ingreso laboral total del año inicial del panel en el que se encontraba la mujer.

Los valores que alcanzan en el λ óptima y la correlación positiva o negativa con

la variable resultado se pueden ver en la siguiente tabla. Solo 4 de ellas tienen una relación positiva con la vulnerabilidad del empleo: trabajar en el sector agricultura, pesca y minería en su ocupación principal, si trabaja en una microempresa, si tiene el nivel de educación secundaria y un mayor gasto mensual por consumo de agua. Todas las demás variables indican que, de aumentar su valor o suceder, reducirán la probabilidad de ser vulnerable a la pérdida del empleo de la mujer.

Tabla 8.

Valor del coeficiente de las variables más relevantes para predecir la vulnerabilidad en el modelo principal.

N	Nombre de variable	Valor absoluto	Relación
1	El ingreso laboral total	0.253	-
2	El ingreso promedio mensual proveniente de trabajo	0.246	-
3	La edad en años cumplidos	0.075	-
4	Si es jefa del hogar	0.067	-
5	Agricultura, pesca y minería como los sectores de su ocupación principal	0.066	+
6	Si reside en Cusco	0.061	-
7	El ingreso secundario total	0.060	-
8	Si reside en Ayacucho	0.056	-
9	Si es que tiene limitaciones de forma permanente para entender o aprender (concentrarse)	0.049	-
10	Si el tamaño de la empresa en la que trabajaba es de 1 a 10 trabajadores	0.041	+
11	Si es una trabajadora independiente en la actividad principal y secundaria	0.035	-
12	El nivel de educación secundaria	0.035	+
13	Ingreso per cápita mensual a precios de Lima de la transferencia monetaria Bono Gas	0.034	-
14	Los gastos en muebles y enseres reales	0.034	-
15	El último gasto mensual por consumo de agua	0.032	+

Fuente: Elaboración propia.

Existen variables que tienen una relación más directa y simple con la vulnerabilidad. Todas las variables con correlación positiva, ya sea por el sector o el

tamaño que tiene la empresa en la que trabajan, el nivel educativo que tiene o mayores gastos de improviso, nos demuestran que situaciones más precarias dentro y fuera del mercado laboral aumenta la probabilidad de vulnerabilidad del empleo. Esto se alinea con lo encontrado en la revisión de literatura. Asimismo, existen variables con correlación negativa que son más claras de explicar: mayores ingresos laborales o provenientes del Estado, que también implicar mayores gastos premeditados; o una mayor edad con posibilidad de experiencia laboral, reducirán la probabilidad de ser vulnerable a la pérdida del empleo.

Sin embargo, existen variables que nos recuerdan que si bien la pérdida del empleo puede ser voluntaria, esta solo se mantendrá si es que es posible subsistir sin los ingresos de un empleo o subempleo. En este sentido, podemos ver que poblaciones muy indefensas, mujeres que son jefas del hogar, aquellas que tienen limitaciones de forma permanente para entender o aprender (concentrarse), que tienen dos empleos independientes o residen en los departamentos más pobres del país, no podrán mantenerse en esta situación de desempleo o inactividad y por ello su probabilidad de mantenerse sin empleo por un periodo de tiempo largo será menor.

Es relevante mencionar que en el proceso de búsqueda del λ, el umbral inicial que se estableció para clasificar a las observaciones fue 0.5. Sin embargo, es importante recordar que escogimos a la curva ROC como métrica del rendimiento del modelo. Esto implica que, se escogió al λ que, con relación a todos los posibles umbrales posibles, en promedio tenía un valor mayor para el área debajo de la curva ROC.

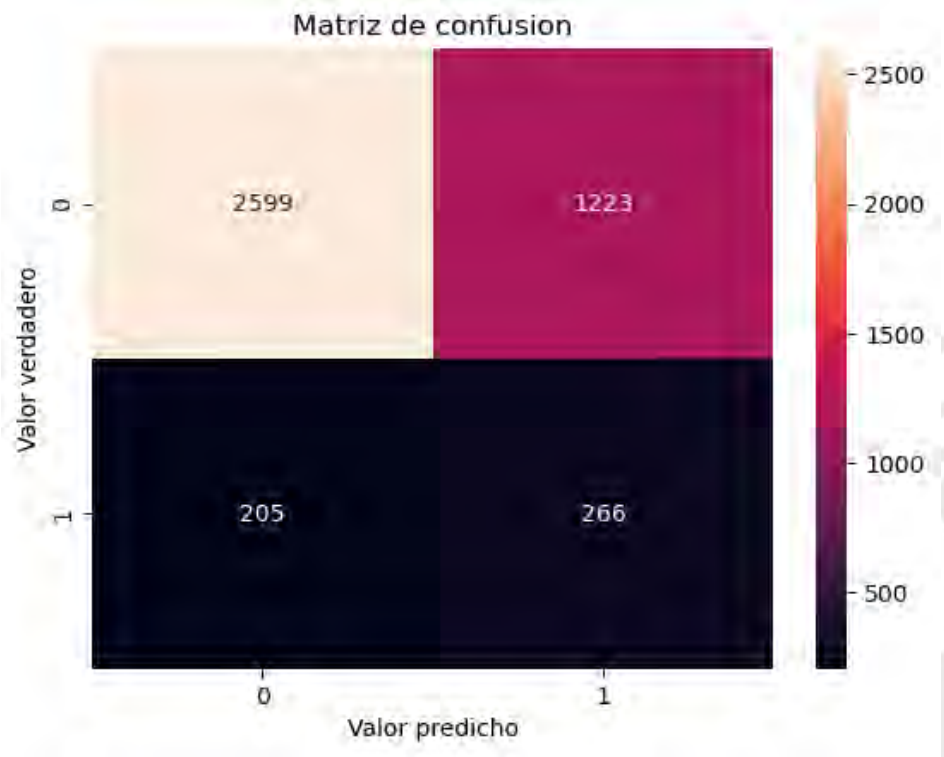
Ahora, procedemos a estimar el valor de los β y por ende nuestra variable dependiente usando la muestra de prueba que nos permitirá clasificar a las mujeres vulnerables. Este resultado responderá a nuestra hipótesis y calcular el ajuste del modelo *out-of-sample* usando el estado laboral en el 2020 de las observaciones panel cuyo año inicial es del 2019. En esta estimación, seguiremos con el supuesto de que el umbral óptimo para clasificar a las observaciones es el 0.56. Posteriormente estimaremos el modelo con otro valor en las pruebas de robustez.

Así, construimos la matriz de confusión en nuestra muestra de prueba, el 2019, para evaluar el desempeño del modelo de manera más detenida. Con estos valores, tenemos que el valor del área de la curva ROC es 62.23%. Esto quiere decir que, usando estos parámetros, casi un 62% de mujeres fuera de la muestra son clasificadas

de manera adecuada. Este valor corresponde a un valor de *precision* de 92.69% para la mantención del empleo, mientras que de un 17.86% para la pérdida de empleo; mientras que corresponde a un valor de *recall* de 68.00% para la mantención del empleo, mientras que de un 56.48% para la pérdida de empleo.

Figura 13.

Matriz de confusión del modelo principal.



Fuente: Elaboración propia.

El valor de esta medida y las medidas que lo componen nos revela que el modelo está sobreestimado el porcentaje de mujeres que perderán su empleo, puesto que existe una considerable cantidad de positivos predichos que en verdad son negativos. Sin embargo, también significa que la mayoría de las mujeres que en verdad sufrirán de una pérdida de bienestar ligada al empleo están siendo identificadas con el algoritmo.

En términos de política, cumplimos con el objetivo de encontrar a más de la mitad de las mujeres que posiblemente pierdan su empleo (casi un 60%) y poder asistirle antes que pierdan su nivel de bienestar; sin embargo, como clasificamos a mujeres que no pertenecen a esta categoría, podemos gastar recursos en población que no necesariamente perderá su nivel de bienestar.

No obstante, esto no significa necesariamente que esta ayuda sea desperdiciada pues puede que esta población viva en condiciones precarias y los programas de apoyo que se asigne mejoren su bienestar. Este es un punto para explorar a mayor profundidad en estudios posteriores, pues si bien la clasificación final no es precisa, no significa que estas mujeres no compartan características de precariedad similar a las que sí lo hacen. Vale recordar que muchas de estas mujeres se encuentran bajo el perfil de nivel de educación primario y con un pozo ciego o negro como baño o servicio higiénico.

De esta manera, el modelo predice que un 34.68% de las mujeres de la PEA urbana del 2019 perderán su empleo el siguiente año, estando en un intervalo de confianza de entre 33.26% y 36.11%. Esto nos comprueba que el modelo predice que más del 18% de las mujeres en la PEA urbana en el 2019 podrían haber perdido el empleo en el 2020 dadas sus características. Comparándolo con el porcentaje que encontramos para las mujeres que pierden el empleo en nuestra muestra, 33.53%, encontramos que el porcentaje estimado está dentro de los posibles valores que este estadístico puede tener.

Si bien nuestro modelo logró aproximarse al porcentaje real, no logró identificar a todas las mujeres vulnerables, aunque sí a más de la mayoría. Una de las posibles razones de la sobreestimación de las mujeres vulnerables puede ser la necesidad de más información. Por ejemplo, en este estudio no incluimos datos con respecto al estado de salud de las madres y niños menores, variables determinantes en el estado laboral de la mujer en el Perú y que la Encuesta Demográfica y de Salud Familiar (ENDES) puede brindar.

Asimismo, otro de los posibles cambios en nuestra estimación sería la elección del modelo. Debido a la elección del *logit* penalizado, para este ejercicio exploratorio; se dio un recorte en las observaciones con *missings* incluidas en las estimaciones. Otros modelos más complejos, como los árboles de decisión o *decision tree*, si admiten este tipo de valores, permitiendo incluir la varianza que estas observaciones traen de las otras variables sin *missing*.

6.4. Pruebas de robustez

Como se mencionó previamente, las pruebas de robustez y sensibilidad buscarán poner a prueba nuestros resultados hallados en el modelo previamente establecido. En estos análisis buscaremos cambiar al método de remuestreo, como

prueba de robustez; y el umbral que se usa para clasificar a mujeres como vulnerables, como prueba de sensibilidad.

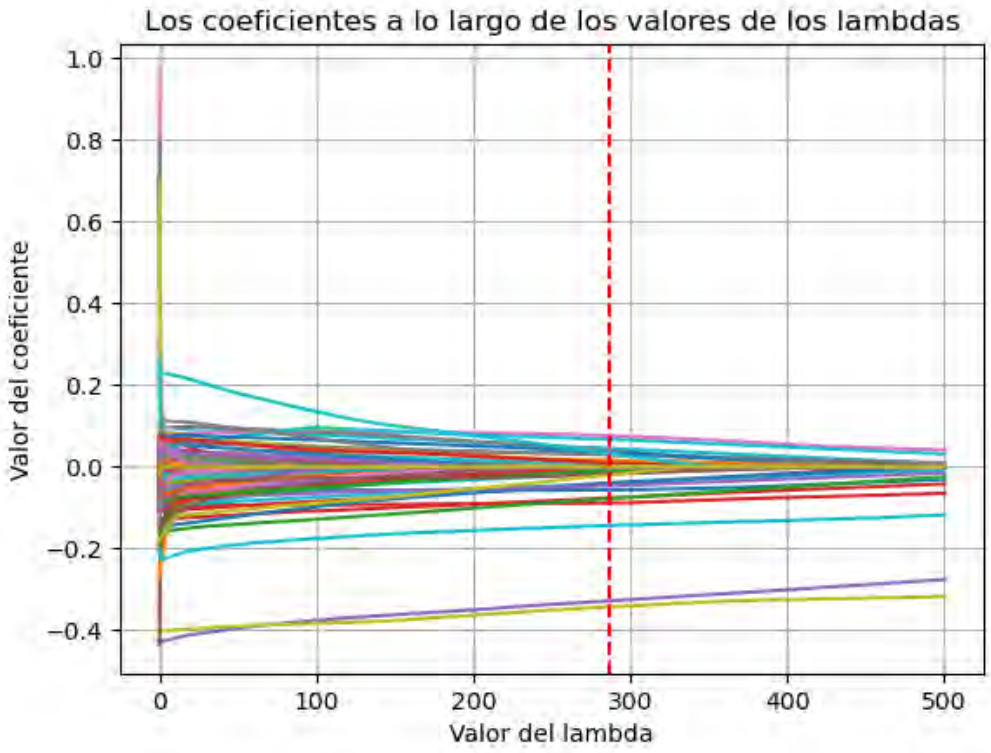
6.4.1. Estimación del modelo logit penalizado usando el método de remuestreo SMOTE con el umbral de 0.56

De la misma manera en que vimos cómo evoluciona el valor de los β según el valor de λ, podremos ver cómo evoluciona el valor de los valores β. Como se puede ver en el gráfico, la mayoría de las líneas se acerca al valor de 0 representado en el eje X mientras avanzamos en el eje X.

Las variables que son resilientes a lo largo del aumento del término de regularización λ hasta su valor óptimo son 286. Las 15 variables con valores absolutos mayores son: el ingreso laboral total, el ingreso promedio mensual proveniente de trabajo, si el tamaño de la empresa en la que trabajaba es de 1 a 10 trabajadores, si es jefa del hogar, ingreso per cápita mensual a precios de Lima de la transferencia monetaria de Otros programas, si es una trabajadora independiente en la actividad principal y secundaria y la edad en años cumplidos.

Figura 14.

Los coeficientes a lo largo de los valores de las lambdas para el modelo de la prueba de robustez 1.



Fuente: Elaboración propia.

También lo son comercio como el sector de su ocupación principal, si vive en un estrato geográfico de 100,000 a 499,999 habitantes, si es que tiene limitaciones de forma permanente para entender o aprender, el nivel de educación superior no universitaria, el ingreso secundario total, el último gasto mensual por consumo de agua, el ingreso per cápita mensual a precios de Lima monetario por trabajo secundario y si es que es trabajadora independiente en la actividad secundaria.

Tabla 9.

Valor del coeficiente de las variables más relevantes para predecir la vulnerabilidad en pruebas de robustez.

N	Nombre de variable	Valor absoluto	Relación
1	El ingreso laboral total	0.406	-
2	El ingreso promedio mensual proveniente de trabajo	0.394	-
3	Si el tamaño de la empresa en la que trabajaba es de 1 a 10 trabajadores	0.202	+
4	Si es jefa del hogar	0.200	-
5	Ingreso per cápita mensual a precios de Lima de la transferencia monetaria de Otros programas	0.146	-
6	Si es una trabajadora independiente en la actividad principal y secundaria	0.132	-
7	La edad en años cumplidos	0.122	-
8	Comercio como el sector de su ocupación principal	0.113	-
9	Si vive en un estrato geográfico de 100,000 a 499,999 habitantes	0.103	+
10	Si es que tiene limitaciones de forma permanente para entender o aprender (concentrarse)	0.102	-
11	El nivel de educación superior no universitaria	0.097	+
12	El ingreso secundario total	0.095	-
13	El último gasto mensual por consumo de agua	0.089	+
14	El ingreso per cápita mensual a precios de Lima monetario por trabajo secundario	0.089	-
15	Si es que es trabajadora independiente en la actividad secundaria.	0.085	-

Fuente: Elaboración propia.

Las 10 variables con más altos valores tienen valores absolutos mayores al 10%, lo cual no sucedía al usar el *undersampling*. Particularmente, las variables con

valores mayores al 20% son el ingreso promedio mensual proveniente de trabajo, el Ingreso laboral total del año inicial del panel en el que se encontraba la mujer, el tamaño de la empresa de “1 a 10 trabajadores” y la edad en años cumplidos. Todas estas variables también pertenecen al grupo de variables con mayor peso en la predicción anterior.

Los valores que alcanzan en el λ óptimo y la correlación positiva o negativa con la variable resultado se pueden ver en la siguiente tabla. Solo 4 de ellas tienen una relación positiva con la vulnerabilidad del empleo: trabajar en el sector agricultura, pesca y minería en su ocupación principal, si trabaja en una microempresa, si tiene el nivel de educación secundaria y un mayor gasto mensual por consumo de agua. Todas las demás variables indican que, de aumentar su valor o suceder, reducirán la probabilidad de ser vulnerable a la pérdida del empleo de la mujer.

Al igual que en el anterior modelo, casi todas las variables con correlación positiva, ya sea por el tamaño que tiene la empresa en la que trabajan, el nivel educativo que tiene o mayores gastos de improviso, nos demuestran la relación entre precariedad y vulnerabilidad. Asimismo, existen variables con correlación negativa con este mismo supuesto detrás: mayores ingresos laborales o provenientes del Estado, o una mayor edad con posibilidad de experiencia laboral, reducirán la probabilidad de ser vulnerable a la pérdida del empleo.

Similar a la anterior estimación, una mujer solo se podrá mantener sin empleo si es que es posible subsistir sin estos ingresos. Mujeres que son jefas del hogar, aquellas que tienen limitaciones de forma permanente para entender o aprender, que tienen uno o dos empleos independientes, cuyos empleos se dan en sectores inestables o que residen en grandes ciudades en el cual este tipo de sectores abunda, tendrán una probabilidad muy baja de mantenerse sin empleo.

De esta manera, encontramos que el valor del área de la curva ROC es 63.87%. Este corresponde a un valor de *precision* de 93.47% para la mantención del empleo, mientras que de un 17.92% para la pérdida de empleo (similar al que el anterior modelo); mientras que corresponde a un valor de *recall* de 64.05% para la mantención del empleo, mientras que de un 63.69% para la pérdida de empleo. Este modelo también sobreestima el porcentaje de mujeres que perderán su empleo, pero también logra identificar a la mayoría de las mujeres que en verdad sufrirán de una pérdida de bienestar.

Figura 15.

Matriz de confusión para el modelo de la prueba de robustez 1.

Fuente: Elaboración propia.

Nuevamente, cumplimos con el objetivo de encontrar a más de la mitad de las mujeres que posiblemente pierdan su empleo (63.87%); aun así, también clasificamos a una gran cantidad de mujeres que no son vulnerables como si lo fueran. De esta manera, el modelo predice que un 38.99% de las mujeres de la PEA urbana del 2019 perderán su empleo el siguiente año, con un intervalo de confianza entre 37.53% y 40.45%. En comparación con el verdadero valor, este estadístico se encuentra por encima de él por aproximadamente 4%. Es relevante recordar que, aparte de que el método de remuestreo no era el adecuado para usar con *logit* penalizado, este modelo tenía un valor del λ que estaba sobre ajustado, haciéndolo menos exacto en su predicción fuera de la muestra.

Esto nos comprueba que aún con un nuevo método de remuestreo, se predice que más del 18% de las mujeres en la PEA urbana en el 2019 podrían haber perdido el empleo en el 2020. Más aún, nos demuestra que el modelo con ambos parámetros escogidos, el umbral y λ, no se ve drásticamente afectado por el método de remuestreo que se use con el fin de solucionar el problema de desbalance de clases.

Ahora, evaluaremos el modelo logit penalizado con ambos métodos de

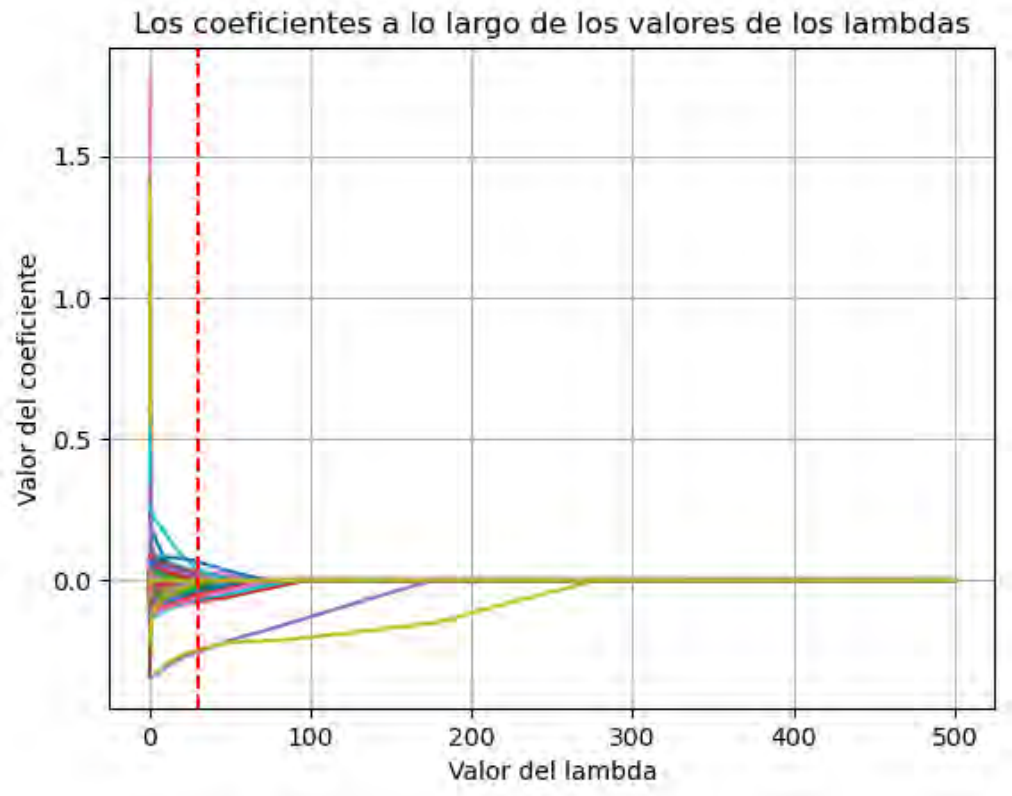
remuestreo (*undersampling* y SMOTE) para verificar que nuestros resultados originales aún se mantienen si es que cambiamos el umbral con el que clasificamos. En este caso, usaremos el umbral "estándar" que anteriormente mencionamos (0.5). Esto nos permitirá poner a prueba como varía la clasificación del modelo permitiendo una mayor cantidad de positivos, que podrían ser verdaderos o falsos.

6.4.2. Estimación del modelo logit penalizado usando el método de remuestreo undersampling con el umbral de 0.5.

Usando nuevamente el modelo logit penalizado con el método de *undersampling* pero con el umbral de 0.5, encontramos que las variables que son resilientes a lo largo del aumento del término de regularización λ hasta su valor óptimo son también 210. Las 15 variables con valores absolutos mayores, así como las variables con un valor absoluto mayor al 10% son las mismas que en la anterior estimación. Los resultados hasta esta parte del proceso son idénticos a la del primer modelo debido a que en el proceso de *CV* el umbral no determina qué variables tienen coeficientes absolutos más grandes, el λ si y este no cambia.

Figura 16.

Los coeficientes a lo largos de los valores de las lambdas para el modelo de la prueba de robustez 2.

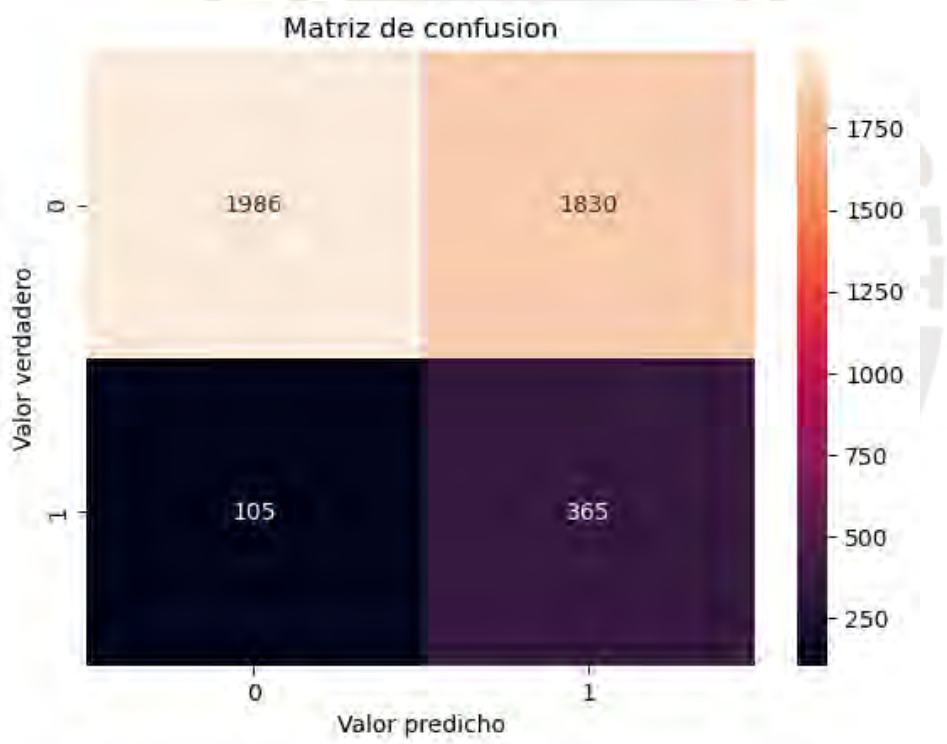


Fuente: Elaboración propia.

De esta manera, encontramos que, usando la prueba de robustez y sensibilidad, el valor del área de la curva ROC para el modelo logit penalizado usando *undersampling* es 64.84%. Este corresponde a un valor de *precision* de 95.09% para la mantención del empleo, mientras que de un 16.53% para la pérdida de empleo; mientras que corresponde a un valor de *recall* de 51.13% para la mantención del empleo, mientras que de un 78.56% para la pérdida de empleo. Al igual que todos nuestros modelos, este sobreestima el porcentaje de mujeres vulnerables a la pérdida de su empleo, aunque logrando identificar a la mayoría de las mujeres que efectivamente perderán su empleo. Aun así, se pueden ver cambios sutiles en los porcentajes que componen a la métrica y en las que la componen.

Figura 17.

Matriz de confusión para el modelo de la prueba de robustez 2.



Fuente: Elaboración propia.

Aun cambiando el umbral, el objetivo de encontrar a más de la mitad de las mujeres que posiblemente pierdan su empleo es cumplido. Asimismo, el modelo predice que un 52.13% de las mujeres de la PEA urbana del 2019 perderán su empleo el siguiente año. Este porcentaje tiene un intervalo de confianza de entre 50.64% y 53.63%. Esta estimación es mayor a la que encontramos con el umbral de 0.56 pero también cumple no rechazar nuestra hipótesis nula: que este porcentaje será mayor a

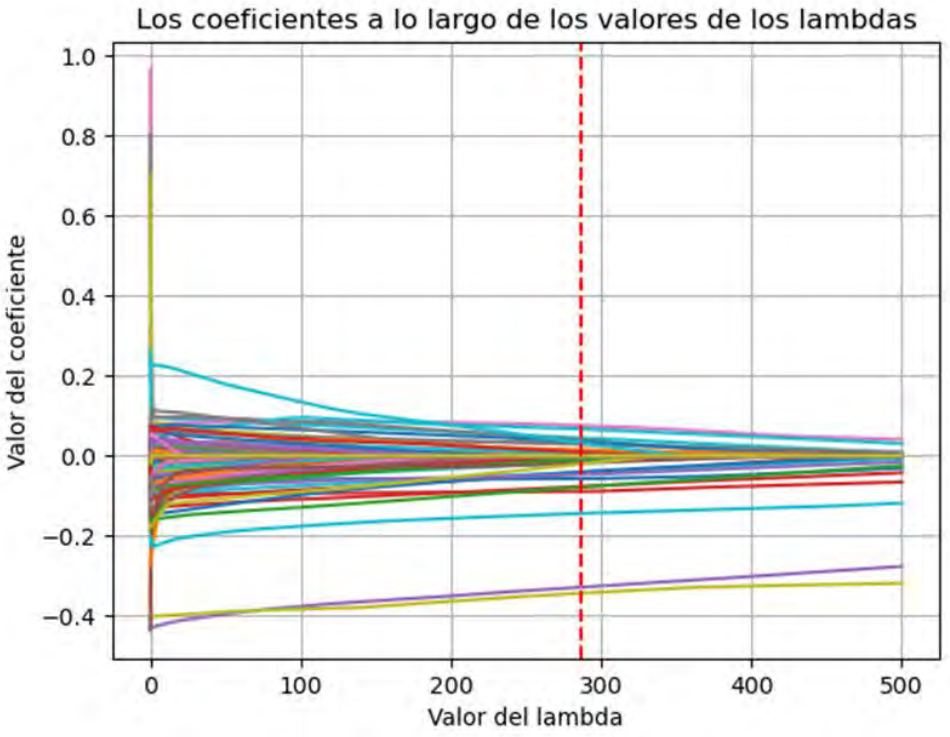
18%. Más aún, nos resalta la importancia que tiene el umbral al momento de clasificar las probabilidades estimadas de las mujeres.

6.4.3. Estimación del modelo logit penalizado usando el método de remuestreo *SMOTE* con el umbral de 0.5

En paralelo, también podemos evaluar las variables que mantienen un coeficiente diferente a 0 a lo largo del aumento de λ, así como de las variables con valores absolutos mayores y con aquellas con valor absoluto mayor al 10%. Como sucedió anteriormente con el método de remuestreo *undersampling*, los resultados son idénticos a la de la anterior modelo debido a que diferentes valores del umbral fueron probados al usar el área de la curva ROC en el proceso de entrenamiento.

Figura 18.

Los coeficientes a lo largo de los valores de las lambdas para el modelo de la prueba de robustez 3.



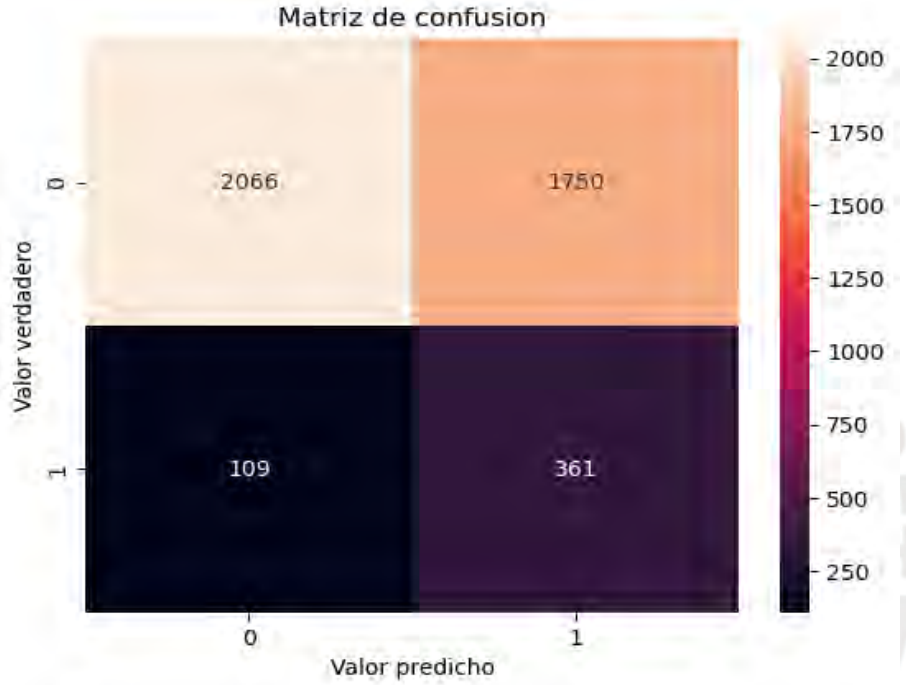
Fuente: Elaboración propia.

Al evaluar el modelo con el método de SMOTE en la muestra de entrenamiento hallamos que el valor del área de la curva ROC es 65.45%. Para la mantención del empleo tenemos un valor de *precision* de 95.04%, mientras que para la pérdida de empleo este es de 17.03%. Para la mantención del empleo tenemos un *recall* de 53.61%; mientras que para la pérdida de empleo este es de un 77.28%.

Este resultado también sobreestima la cantidad de mujeres vulnerables a la pérdida de su empleo en el 2019. Sin embargo, al igual que en anteriores modelos, este también logra predecir a la mayoría de las mujeres de la PEA urbana en el 2019 que perderán su empleo en el 2020. En comparación con el modelo que usaba el umbral de 0.56, existen diferencias en las métricas de evaluación del modelo.

Figura 19.

Matriz de confusión para el modelo de la prueba de robustez 3.



Fuente: Elaboración propia.

Nuevamente, se logra encontrar a más de la mitad de las mujeres que posiblemente pierdan su empleo cumplido. Finalmente, el modelo predice un 49.98% de las mujeres de la PEA urbana del 2019 perderán su empleo el 2020, estando este porcentaje entre los intervalos de confianza de 48.28% y 51.27%.

Las pruebas de robustez y sensibilidad a partir de la alteración de otros parámetros también se realizaron. En retrospectiva, el cambio del umbral para seleccionar a las variables que más se correlacionaron, la imputación en las variables antes de descartar observaciones, el uso de un modelo Lasso y otras métricas para la evaluación del modelo fueron variaciones que se intentaron con el fin de probar la consistencia de los resultados. Estos no fueron incluidos debido a que no eran

métodos, indicadores o estrategias de procesamiento de datos que se alineaban con el objetivo de la investigación.

Reducir o aumentar el umbral en 10% para la selección de variables en la tercera etapa del procesamiento no cambia radicalmente la lista de variables que entraron en el proceso de modelación. Usar otro set de métricas para evaluar al modelo, como precisión o recall, hubiera dejado de lado el objetivo de encontrar una estadística que sea útil para la política pública. Como se explicó en la sección metodológica, un modelo *Lasso* hubiera resultado en estimaciones fuera del rango que nosotros necesitábamos para clasificar la vulnerabilidad de las mujeres.

Asimismo, a pesar de tener aproximadamente un 20% de observaciones que aún tenían *missings* antes del modelamiento, la imputación no era tan relevante ya que no buscamos la insesgadez o precisión en nuestros coeficientes, sino que logren ajustarse adecuadamente a la predicción.

Más aún, debido a los bajos coeficientes que obtuvieron los predictores, un proceso de imputación simple hubiera quebrantado con esta débil correlación al crear datos en la media. Lo que sí se puede considerar es la posible pérdida en la varianza que se pudo haber tenido debido a la omisión de estas variables. En este sentido, un espacio de mejora en esta investigación será realizar imputaciones múltiples o más ajustadas a los datos.

Tabla 10.

Resumen de las métricas.

		Mantiene su empleo		Pierden su empleo		
Tipo	Modelo	P (%)	R (%)	P (%)	R (%)	AUC ROC (%)
Original	LP + Undersampling + 0.56 de umbral	92.69	68.00	17.86	56.48	62.23
Pruebas de robustez y sensibilidad	LP + SMOTE + 0.56 de umbral	93.47	64.05	17.92	63.69	63.87
	LP + Undersampling + 0.5 de umbral	95.09	51.13	16.53	78.56	64.84
	LP + SMOTE + 0.5 de umbral	95.04	53.61	17.03	77.28	65.45

Fuente: Elaboración propia.

Un resumen de todas las métricas de las pruebas de robustez y sensibilidad en comparación con el modelo principal planteado se puede ver en las siguientes tablas.

Podemos notar que al usar el umbral ideal con cualquiera de los dos métodos de remuestreo el rendimiento no cambia radicalmente, ni tampoco lo hace el porcentaje de mujeres que se predice que perderán su empleo.

Adicionalmente podemos ver que existen diferencias cuando cambiamos el umbral. No solo cambia el porcentaje de mujeres que se consideran como vulnerables, sino también las métricas que se usan para medir su eficiencia. Si bien por un lado obtenemos una mayor área debajo de la curva ROC, encontramos que la métrica de *precision* para el evento de la pérdida del empleo va empeorando. Esta es la que nos permite evaluar que tan bien está prediciendo a las mujeres vulnerables.

Lo que no se puede discutir es que en todos los casos el porcentaje de mujeres en la PEA urbana del 2019 que son vulnerables a la pérdida del empleo son más del 18% y están cerca o por encima del 33% del valor real de nuestra muestra. Si bien cuando cambiamos el umbral las estadísticas suben entre 10% y 20%, parecen no cambiar drásticamente cuando solo modificamos el método de remuestreo. Asimismo, todas estimaciones llegan a predecir a más del 50% de las mujeres que realmente pierden su empleo.

Tabla 11.

Resumen de los resultados de las estimaciones.

Tipo	Modelo	Porcentaje (%)	Intervalo inferior (%)	Intervalo superior (%)
Original	LP + Undersampling + 0.56 de umbral	34.68	33.26	36.11
Pruebas de robustez y sensibilidad	LP + SMOTE + 0.56 de umbral	38.99	37.53	40.45
	LP + Undersampling + 0.5 de umbral	52.13	50.64	53.63
	LP + SMOTE + 0.5 de umbral	49.98	48.28	51.27

Elaboración propia.

Sin embargo, al comparar estos resultados con la realidad, encontramos que si bien el porcentaje del modelo original es muy cercano aún sabemos que nuestro modelo no detectó a un porcentaje de estas mujeres. En todos los modelos, logramos encontrar a casi más del 60% de mujeres que efectivamente pierden su empleo o pasan a la inactividad. Por ello, consideramos que hay espacio para mejorar estas estimaciones, ya sea incluyendo datos acerca del contexto del hogar y salud de las

mujeres y niños; o aplicando modelos de *Machine Learning* que impliquen un menor grado de procesamiento de los datos.

De la misma manera, vale la pena recordar que existe también un patrón en las variables más importantes para predecir el modelo. El ingreso promedio mensual proveniente de trabajo, el ingreso laboral total, si es jefa del hogar, el tamaño de la empresa en la que trabajaba de “1 a 10 trabajadores”, la edad en años cumplidos, si es una trabajadora independiente en la actividad principal y secundaria y si es que tiene limitaciones de forma permanente para entender o aprender son las variables que se han repetido en casi todos los modelos.

En concordancia con la revisión de literatura, estas representan a las características de las mujeres que están ligadas a la dimensión laboral de su vida (cuál es el tamaño de la empresa en la que trabajan, los ingresos de su trabajo y del hogar, su tipo de trabajo en la actividad principal y secundaria) pero también a dimensión de su hogar y personal (su rol en el hogar, su edad, sus limitaciones).

Adicionalmente, algunas tienen relación más directa con la vulnerabilidad del empleo: mayor precariedad dentro y fuera del mercado laboral implica más probabilidad de vulnerabilidad del empleo en el largo plazo. Sin embargo, también nos muestra que la población más indefensa dentro de ese grupo no podrá costearse no tener los ingresos del empleo o subempleo. Asimismo, el tamaño de los coeficientes en el modelo original cambia cuando usamos un diferente λ y método de remuestreo, pero no diferente umbral.

Existen más coeficientes relevantes con mayor valor absoluto en el segundo grupo (10) de modelos que en el primero (2), estos pueden llegar a 0.4 en comparación de 0.25. Para las variables que se repiten entre ambos grupos de modelos, la dirección de correlación no cambia, solo el grado. Aun así, la mayoría de las variables fuera y dentro de este grupo para todos los modelos tienen coeficientes muy bajos. Esto nos demuestra que si bien existe un grupo de variables relevantes que individualmente aportan a la predicción de la vulnerabilidad, el aporte del conjunto de variables parece ser mayor para esta tarea, haciéndolo un fenómeno que puede ser estudiado en su multidimensionalidad.

Tabla 12.

Resumen de variables más relevantes para la predicción de la vulnerabilidad según los modelos

	Undersampling + 0.56 y 0.5 de umbral		SMOTE + 0.56 de umbral y 0.5 de umbral	
N	Nombre de variable	Valor	Nombre de variable	Valor
1	El ingreso laboral total	- 0.253	El ingreso laboral total	- 0.406
2	El ingreso promedio mensual proveniente de trabajo	- 0.246	El ingreso promedio mensual proveniente de trabajo	- 0.394
3	La edad en años cumplidos	- 0.075	Si el tamaño de la empresa en la que trabajaba es de 1 a 10 trabajadores	+ 0.202
4	Si es jefa del hogar	- 0.067	Si es jefa del hogar	- 0.200
5	Agricultura, pesca y minería como los sectores de su ocupación principal	+ 0.066	Ingreso per cápita mensual a precios de Lima de la transferencia monetaria de Otros Programas	- 0.146
6	Si reside en Cusco	- 0.061	Si es una trabajadora independiente en la actividad principal y secundaria	- 0.132
7	El ingreso secundario total	- 0.060	La edad en años cumplidos	- 0.122
8	Si reside en Ayacucho	- 0.056	Comercio como el sector de su ocupación principal	- 0.113
9	Si es que tiene limitaciones de forma permanente para entender o aprender	- 0.049	Si vive en un estrato geográfico de 100,000 a 499,999 habitantes	+ 0.103
10	Si el tamaño de la empresa en la que trabajaba es de 1 a 10 trabajadores	+ 0.041	Si es que tiene limitaciones de forma permanente para entender o aprender	- 0.102
11	Si es una trabajadora independiente en la actividad principal y secundaria	- 0.035	El nivel de educación superior no universitaria	+ 0.097
12	El nivel de educación secundaria	+ 0.035	El ingreso secundario total	- 0.095
13	Ingreso per cápita mensual a precios de Lima de la transferencia monetaria Bono Gas	- 0.034	El último gasto mensual por consumo de agua	+ 0.089
14	Los gastos en muebles y enseres reales	- 0.034	El ingreso per cápita mensual a precios de Lima monetario por trabajo secundario	- 0.089
15	El último gasto mensual por consumo de agua	+ 0.032	Si es que la trabajadora independiente en la actividad secundaria.	- 0.085

Elaboración propia.

Conclusiones

El objetivo de esta investigación era predecir el porcentaje de mujeres en el Perú urbano que en el 2019 que pierden su empleo en el 2020. Con este fin, se entrenó un modelo *logit* penalizado usando el *undersampling* como método de remuestreo, para predecir la pérdida de empleo de las mujeres usando datos del 2016 al 2018. Se planteó que la proporción de mujeres con empleo vulnerable a la pérdida en el Perú en el 2019 sería mayor al 18%, que es igual al promedio del porcentaje de mujeres empleadas del 2016 al 2018 que perdieron su empleo de un año a otro.

Siendo nuestra hipótesis nula fue $Ho: > 0.18$ y nuestra hipótesis alternativa $Ho: p \leq 0.18$, no se pudo rechazar nuestra hipótesis alternativa al encontrar un valor de 37.1% con intervalos de confianza que no incluyen el valor de 18%. En otras palabras, podemos afirmar que, según el modelo, el porcentaje de mujeres en el 2020 que puede perder el empleo es mayor de 18%.

De la misma manera, se validó esta hipótesis mediante diferentes pruebas de robustez y sensibilidad alterando el método de remuestreo y el umbral para clasificar a las mujeres vulnerables. En todos estos casos se halló que el porcentaje seguía siendo mayor al 18%, aunque con diferentes porcentajes. Si es que alteramos el método de remuestreo con el umbral ideal, mantenemos casi igual los hallazgos con el método de remuestreo de *undersampling*. Asimismo, encontramos que, al alterar el umbral, los resultados son mayores al porcentaje que se encontró en el modelo original, pues tenemos valores entre el 40% y 50%, pero que aún siguen siendo mayor al 18%.

Asimismo, para evaluar al modelo podemos ver que la métrica elegida, el área debajo de la curva ROC, tiene un valor aceptable de 62.23% con nuestro modelo original. Esto nos demuestra que casi el 62% de las mujeres fuera de la muestra serán clasificadas en la clase ideal usando este modelo. Este valor se mantiene o incrementa ligeramente en las pruebas de robustez y sensibilidad, lo cual nos demuestra que la eficiencia del modelo para clasificar se mantiene en este porcentaje a pesar de alterar ciertos supuestos.

En términos de política, esto es beneficioso porque logramos encontrar a la mayoría de las mujeres que son vulnerables. Con estos resultados, logramos realizar la focalización de aquellos programas que busquen prevenir y mitigar la pérdida de bienestar que sucede con el desempleo o la inactividad en las mujeres.

Asimismo, a pesar de que ciertas mujeres no vulnerables son clasificadas como si lo fueran, los recursos que posiblemente se inviertan en ellas parecen no estar totalmente malgastados ya que estas mujeres parecen presentar similares carencias en el hogar y en el ámbito laboral que las que sí fueron clasificadas como vulnerables. Más aún, parece que algunas de las mujeres ya reciben apoyo económico de algún programa público al ser una de estas variables más relevantes para predecir el desempleo.

Esto nos lleva a reflexionar, además, sobre un hecho relevante: las mujeres son un grupo en condiciones precarias tanto dentro como fuera del mercado laboral. Pues, si bien puede que estas mujeres mal clasificadas no sean vulnerables en el empleo, son caracterizadas por las precariedades en su bienestar económico. Esto nos recuerda que la inestabilidad laboral es una causa y expresión de la pobreza monetaria.

Adicionalmente, podemos ver que existe un patrón en las variables que aportan más a la predicción de la probabilidad de la mujer de perder su empleo: variables que describen en qué tipo de lugar trabajan, pero también en qué tipo de hogares se encuentran. Así, podemos definir el perfil de las mujeres vulnerables al empleo: son aquellas que trabajan en empresas pequeñas, dependen de sus ingresos laborales, tienen un nivel bajo de educación, pocos años de experiencia laboral, y viven en áreas geográficas donde la pobreza es común.

Este ejercicio, además, nos lleva a también poder definir parcialmente a las mujeres con menor probabilidad de ser vulnerables en el empleo: son jefas de hogar, con alguna limitación para entender o estudiar y son trabajadoras independientes en su ocupación principal y secundaria. Este ejercicio de contraste nos muestra el otro lado del desempleo y la inactividad: sólo es posible mantener este estado si es que el hogar puede subsistir por un largo tiempo sin los ingresos de un empleo o subempleo. Las poblaciones más vulnerables, en el largo plazo, no podrán costearse esta situación y aceptarán empleos con probablemente condiciones laborales más deplorables.

Es relevante resaltar que estas mujeres vulnerables comparten características y la tarea de descripción de estas es valiosa para la exploración de problemáticas en las que no hemos profundizado aún lo suficiente, como lo es el empleo femenino. Un ejemplo de esto es que podemos ver características muy específicas y poco comunes como relevantes para predecir la vulnerabilidad, como lo son las limitaciones de forma

permanente para entender o aprender (concentrarse) o los gastos mensuales y no mensuales que realizan.

Debemos reconocer que esta investigación también presenta limitaciones y espacios de mayor investigación. Por un lado, las limitaciones se presentan en el costo computacional del entrenamiento del modelo debido a la cantidad de variables y observaciones para estimar el modelo y las diferentes pruebas de robustez y sensibilidad. Aquí solo se explicitan las más importantes que involucran al método de remuestreo para lidiar con el problema de desbalance de clases, así como una reducción en el umbral que nos permitirá verificar la eficacia del modelo al clasificar a las mujeres de la PEA urbana.

Más aún, reconocemos que, si bien logramos identificar a casi más del 60% de las mujeres que son vulnerables, nuestra estimación aún tiene espacios de mejora. En ese sentido, recomendamos incluir más información acerca de variables del hogar y salud de las mujeres y niños. Consideramos que otros modelos más complejos, como el *decision tree*, pueden evitar restringir la muestra a aquella que no tenga missings. Asimismo, en caso se quiera lidiar con estas usando imputación, consideramos que procesos más complejos pueden ser adecuados para el presente modelo y método de remuestreo.

Por otro lado, este estudio propone espacios de mayor indagación que nos permiten profundizar más en el estudio de otros tipos de transiciones en el mercado laboral y otros grupos poblacionales, así como también hacer uso de las ventajas que nos presenta la base de datos usada, la ENAHO Panel. Es posible un mayor aprovechamiento de la representatividad que nos brinda la ENAHO Panel incluyendo los pesos muestrales en el entrenamiento del modelo. Así también se puede incluir no solo años anteriores al 2016 de la ENAHO Panel; sino también una mayor cantidad de variables, de tal manera que podamos conseguir mayor cantidad de observaciones para predecir la pérdida del empleo.

Asimismo, aún existe mayor cabida para la predicción de distintas transiciones en el mercado laboral peruano. La ENAHO panel nos permite construir paneles de los trabajadores con transiciones con más de dos tiempos. Un análisis en el cual existan múltiples variables respuestas es posible usando ML. El uso de la misma metodología también nos permite estudiar otro tipo de transiciones, como aquellas que se dan de un empleo formal a un empleo informal o de un empleo en el sector formal a uno en

el sector informal. El estudio de la pérdida del empleo nos abre paso al de fenómenos más complejos pero que pueden ser abordados con una metodología similar a la planteada.

En esta investigación se usa un conjunto de mujeres que comparten características individuales y del hogar previas a la pandemia. El cambio que generó la pandemia en todas las dimensiones de esta población alienta en pensar una nueva estimación de este grupo ante este gran cambio. De la misma manera, aún hay espacio para el estudio de otros grupos poblacionales importantes en el mercado laboral. Sería interesante conocer más acerca de las transiciones en el área rural, en el cual no se dan muchas; asimismo, en la población joven de los últimos años que presenta flujos diferentes a los de generaciones pasadas.

Este estudio nos invita a reflexionar sobre cómo las decisiones de los parámetros que usamos en las estimaciones pueden tener implicancias políticas. Tanto la alteración del umbral como la elección del método de remuestreo nos generan cambios en la clasificación de mujeres entre vulnerables y no vulnerables. Esto implica también decidir que mujeres potencialmente podrían recibir posible apoyo o no, hecho que pudiera mejorar su bienestar y el de su hogar.

Más aún, la elección de las métricas con las que optimizan los modelos no solo también juega un rol en la clasificación, sino también en las consecuencias de ello a nivel de política pública (la adecuada distribución de los recursos). Durante las elecciones de esta investigación se ha buscado tomar elecciones que generen un equilibrio entre la formación de nuevo conocimiento y la eficiente ejecución de políticas con estos resultados.

De la misma manera, este estudio nos invita a expandir las definiciones para ciertos fenómenos que afectan particularmente a las mujeres. Considerar a la inactividad como vulnerabilidad del empleo permite que desde el Estado sea posible promover políticas que promuevan roles de género equitativos en hogares cuyas mujeres no trabajan de manera voluntaria por el doble turno que les tocaría. Al localizar a esos hogares, se podrán ejecutar estrategias que promuevan una repartición equitativa de las labores domésticas y de cuidado. Esto permitirá que las mujeres no tengan que sufrir las consecuencias inmediatas y a largo plazo de perder o dejar su empleo por centrarse en las labores domésticas.

En el caso en el que el empleo haya sido involuntario, permitirá promover

políticas que ayuden a la reinserción en el mercado laboral de estas mujeres que probablemente sigan buscando empleo. En ambos casos, permitirá mitigar el impacto negativo que viene con la reducción del ingreso en los hogares. Esto será de especial utilidad en los hogares más pobres, pues es en estos grupos en los que las desigualdades se acentúan.

Referencias bibliográficas

Barba, M., Estrada, M., & Godoy, R. (1997). Género y trabajo femenino en el Peru. *Revista Latino-Americana de Enfermagem*, *5*(2), 23–31. https://doi.org/10.1590/S0104-11691997000200004

Bazillier, R., Boboc, C., & Calavrezo, O. (2015). *Employment vulnerability in Europe: Is there a migration effect?* https://halshs.archives-ouvertes.fr/halshs-01203755

Blagus, R., & Lusa, L. (2013). *SMOTE for high-dimensional class-imbalanced data* (Vol. 14). http://www.biomedcentral.com/1471-2105/14/106

Bocquier, P., Nordman, C. J., & Vescovo, A. (2010). Employment vulnerability and earnings in Urban West Africa. *World Development*, *38*(9), 1297–1314. https://doi.org/10.1016/j.worlddev.2010.02.011

Bosch, M., & Esteban-Pretel, J. (2012). Job creation and job destruction in the presence of informal markets. *Journal of Development Economics*, *98*(2), 270–286. https://doi.org/10.1016/j.jdeveco.2011.08.004

Cozzubo, Á., & Herrera, J. (2021). Pobreza y desarrollo. In *Balance de Investigación en Políticas Públicas 2011-2016 y Agenda de Investigación 2017-2021* (pp. 494–587). CIES.

Davis, J., & Goadrich, M. (2006). The Relationship Between Precision-Recall and ROC Curves. In Department of Computer Sciences and Department of Biostatistic and Medical Informatics (Ed.), *The 23rd International Conference on Machine Learning*.

Diamantidis, N. A., Karlis, D., & Giakoumakis, E. A. (2000). Unsupervised stratification of cross-validation for accuracy estimation. In *Artificial Intelligence* (Vol. 116).

Durán, R. L. (2022). COVID-19 and heterogeneous vulnerabilities in the Peruvian labor market: implications for social inequalities and for gender gaps. *Economia Politica*, *39*(1), 129–156. https://doi.org/10.1007/s40888-021-00245-5

Gamero, J. (2018). De la Noción de Empleo Precario al Concepto de Trabajo Decente. *Derecho & Sociedad*, *0*(37), 117–125.

Garavito, C. (2010). Vulnerabilidad en el empleo, género y etnicidad en el Perú. *Economia*, *33*(66), 89–127.

Gokhool, S., Kasseeah, H., & Tandrayen-Ragoobur, V. (2018). Vulnerable employment in Mauritius: experience of an upper-middle-income country. *International Journal of Development Issues*, *17*(2), 187–204. https://doi.org/10.1108/IJDI-11-2017-0180

González, P., Sehnbruch, K., Apablaza, M., Méndez Pineda, R., & Arriagada, V. (2021). A Multidimensional Approach to Measuring Quality of Employment (QoE) Deprivation in Six Central American Countries. In *Social Indicators Research* (Issue 0123456789). Springer Netherlands. https://doi.org/10.1007/s11205-021-02648-0

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Herrera, J., & Cozzubo, Á. (2016). La vulnerabilidad de los Hogares a la Pobreza en el Perú, 2004-2014. *Departamento de Economía - PUCP*, *429*, 93.

Herrera, J., & Hidalgo, N. (2002). Vulnerabilidad del empleo en Lima. Un enfoque a partir de encuestas a hogares. In *Bulletin de l'Institut français d'études andines* (Issue 31 (3)). https://doi.org/10.4000/bifea.6705

Herrera, J., & Rosas, G. (2003). *Labor Market Transitions in Peru* (14).

INEI. (2012). *Perú: Calidad del Empleo y Mecanismos Colectivos de Integración Social, 2010*. 1–136. https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1017/libro.pdf

INEI. (2020). *Perú : Estimación de la Vulnerabilidad Económica a la Pobreza Monetaria. Metodología de cáculo y perfil sociodemográfico.*

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to Statistical Learning*. http://www.springer.com/series/417

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, *105*(5), 491–495. https://doi.org/10.1257/aer.p20151023

Kumar, N., & Srivastava, A. (2021). Measuring the Employment Vulnerability Among Female Workers in Uttar Pradesh. *Indian Journal of Human Development*, *15*(2), 307–322. https://doi.org/10.1177/09737030211040842

Lavado, P., & Campos, D. (2021). Empleo e Informalidad. In *Balance de Investigación en Políticas Públicas 2011-2016 y Agenda de Investigación 2017-2021* (pp. 438–492). CIES.

Lockshin, M., & Ravallion, M. (2000). *Short-Lived Shocks with Long-Lived Impacts?* (2459).

Mora, B. (2018). Transiciones del mercado laboral en Perú: 2013-2017. *Revista Yachay*, *1*(7), 316–321.

Morales, R., Rodriguez, J., Higa, M., & Montes, R. (2010). *Transiciones laborales, reformas estructurales y vulneranilidad laboral en el Perú (1998-2008)* (281).

Mortensen, D. T., & Pissarides, C. A. (1994a). Job creation and job destruction in the theory of unemployment. *Review of Economic Studies*, *61*(3), 397–415. https://doi.org/10.2307/2297896

Mortensen, D. T., & Pissarides, C. A. (1994b). Job creation and job destruction in the theory of unemployment. *Review of Economic Studies*, *61*(3), 397–415. https://doi.org/10.2307/2297896

OIT. (2002). *Panorama Laboral 2002*.

OIT. (2010). *Global Employment Trends* (Issue Enero).

OIT. (2019). *Mujeres en el mundo del trabajo*. 204. www.ilo.org/publns

Pritchett, L., Suryahadi, A., & Sumarto, S. (2000). Quantifying Vulnerability to Poverty. A proposed Measure, Applied to Indonesia. In *The World Bank* (Vol. 2437, Issue September).

Rivarola, R. (2019). *Dinámica del mercado laboral en el Perú urbano. Un análisis desde la calidad del empleo en los trabajadores dependientes.* [Tesis para optar por el título de licenciado en Economía]. PUCP.

Rivera, G., Arcondo, Z., & Tenorio, D. (2021). *Informe anual de la mujer en el mercado laboral 2020*.

Rodriguez, J., & Rodriguez, G. (2011). *Movilidad en los mercados laborales del Perú: 2007-2011*. 2007–2011.

Sparreboom, T., & de Gier, M. (2008). Assessing vulnerable employment: The role of the status and sector indicator in Pakistan, Namibia and Brazil. In *Employment Sector* (13; Issue 13).

Villacís, A., & Reis, M. (2015). Analysis of labor vulnerability and determinants of decent work. The case of Ecuador 2008-2011. *Revista de Economia Del Rosario*, *18*(2), 157–185. https://doi.org/10.12804/rev.econ.rosario.18.02.2015.01

Yamada, G. (2007). *Reinserción Laboral Adecuada: Dificultades e implicancias de política*

(82). http://repositorio.up.edu.pe/bitstream/handle/11354/229/DT78.pdf?sequence=1

Zavaleta, D., Moreno, C., & Santos, M. E. (2018). La medición de la pobreza multidimensional en América Latina. In S. Deneulin, J. Clausen, & A. Valencia (Eds.), *Aportes para el desarrollo humano en América Latina*. Instituto de Desarrollo Humano de América Latina.

Anexos

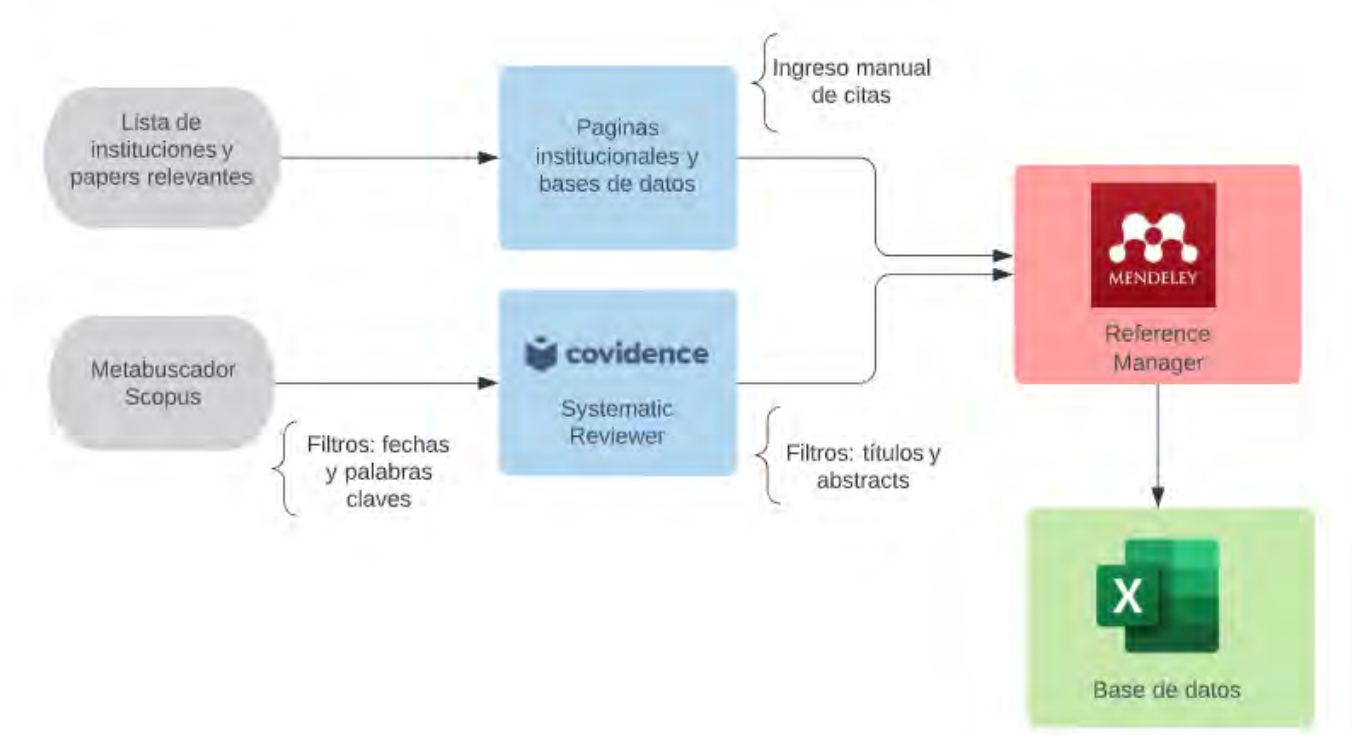
Anexo A. Revisión sistemática

a) Para recolección de investigaciones *peer-reviewed*, se realizó una la revisión sistemática basada en la metodología de Cozzubo y Herrera (2021), usando el metabuscador *Scopus* que localiza las investigaciones de otras bases de datos bibliográficas.
b) Se realizó la búsqueda en el metabuscador con palabras claves, frases y tokens relacionados al tema de la investigación: *("employment vulnerability" OR "vulnerable workers" OR "heterogeneous vulnerabilities") AND (woman OR female OR gender).* Asimismo, se limitó la búsqueda por la fecha de publicación del 2000 en adelante.
c) Las palabras claves, frases y tokens abarcar el tema de vulnerabilidad del empleo, pero, como ya fue mencionado antes, en la presente investigación nos centraremos en el componente de pérdida del empleo.
d) Las 95 fuentes encontradas fueron descargadas en archivos de referencia (.ris) e importados a Covidence, un software de revisión sistemática de literatura.
e) En Covidence, mediante la revisión de títulos y abstracts se seleccionaron 5 fuentes relevantes. Estas luego fueron exportadas al gestor bibliográfico Mendeley.
f) Para estas 5 fuentes se exportó el título, autor, revista/institución, año de la publicación, DOI, etc. a un documento xlsx. Estas fuentes fueron incluidas en la parte de revisión de literatura.

El siguiente diagrama de flujo presenta el proceso de búsqueda bibliográfica:

Figura 20.

Flujograma de revisión sistemática



Elaboración propia

Asimismo, la siguiente tabla mostrara el título, autor, revista/institución, año de la publicación y el DOI de los documentos que fueron seleccionadas e incluidas en la revisión de literatura:

Tabla 13.

Características de los documentos seleccionados por la revisión de literatura sistemática.

Titulo	Autores	Abstract
Measuring the Employment Vulnerability Among Female Workers in Uttar Pradesh	Kumar, N. P., & Srivastava, A.	This article attempts to measure employment vulnerability among women workers in Uttar Pradesh by constructing a multidimensional vulnerability index (MVI). The index is based on 23 dichotomous (binary) variables corresponding to various dimensions of vulnerability related to employment. A composite index of vulnerability is developed for each occupational category, sector of employment and gender. Here, MVI is the average of five indices which are computed for the respective dimensions of employment vulnerability. The findings suggest high levels of vulnerability among informal workers with the MVI values ranging from 0.087 (low) to 0.783 (high).The overall MVI (measured by principal component loading [PCA]) was 0.768 for the construction and domestic workers, followed by tailors (0.629) and garment workers (0.635). Appropriate policies are needed to help lift women from the cumulative neglect that they experience in unorganised labour market.
A Multidimensional Approach to Measuring Quality of Employment (QoE) Deprivation in Six Central American Countries	González, P., Sehnbruch, K., Apablaza, M., Méndez Pineda, R., & Arriagada, V	This paper proposes a methodology for measuring Quality of Employment (QoE) deprivation from a multidimensional perspective in six Central American countries (Costa Rica, El Salvador, Guatemala, Honduras, Nicaragua, and Panama) using a dataset specifically designed to measure employment conditions. Building on previous work on multidimensional poverty and employment indicators, the paper uses the Alkire/Foster (AF) method to construct a synthetic indicator of the QoE *at an individual level.* It selects four dimensions that must be considered as essential to QoE deprivation: income, job stability, job security and employment conditions. These dimensions then subdivide into several indicators, a threshold for each indicator and dimension is established before defining an overall cut-off line that allows for the calculation of composite levels of deprivation. The results generated by this indicator show that Central American countries can be divided into three distinct and robust performance groups in terms of their QoE deprivation. Overall, approximately 60% of the deprivation levels are attributable to non-income variables, such as occupational status and job tenure. The methodology used can allow policymakers to identify and focus on the most vulnerable workers in a labour market and highlights the fact that having a formal written contract is no guarantee of good job quality, particularly in the case of women.

COVID-19 and heterogeneous vulnerabilities in the Peruvian labor market: implications for social inequalities and for gender gaps	Durán, R. L.	Using quarterly data from the 2020 Peruvian National Household Survey (ENAHO), this paper estimates the differentiated impacts of the COVID-19 pandemic on a set of labor market indicators, such as labor participation, occupational categories, informality, and number of hours worked. The impacts are calculated from an individual perspective (effects on the activities of the heads of household and their spouses, distinguishing them according to sex) and also from a joint strategy perspective among the partners. The results indicate that the intersectionalities of vulnerability considered (rural/urban area, and those contained in the type of households and in the situation of single-parenting or two-parenting of household heads and their spouses) determine that women, who live in rural areas, have children and do not have a partner were the most affected by the global health crisis.
Vulnerable employment in Mauritius: experience of an upper-middle-income country	Gokhool, S., Kasseeah, H., & Tandrayen-Ragoobur, V	Purpose – The purpose of this paper is to shed light on the socio-economic characteristics of workers engaged in vulnerable jobs in Mauritius. The study has a particular focus on the gender and youth dimensions of vulnerable employment. The study also provides a pre-crisis and post-crisis analysis of vulnerable employment. Design/methodology/approach – The paper uses several waves of the continuous multi-purpose household survey, which is a high-quality individual-level data set, to study vulnerable employment. Several definitions of vulnerable employment are used to identify the workers employed in vulnerable jobs. These include “own-account” workers and “contributing family workers”. Findings – The results obtained suggest that women and young workers have a lower probability of being in vulnerable employment. Marital status, age and education are also important variables influencing the probability of being in vulnerable employment. Research limitations/implications – The paper has important policy implications regarding welfare and education policies. Appropriate mechanisms need to be put in place for the social protection and training of workers so that they do not end up in vulnerable jobs. Originality/value – This paper studies Mauritius as it is a small island economy vulnerable to external shocks. Vulnerable unemployment has often been understudied as the focus of many studies has been solely on employment, and the quality of employment has often not been considered
Analysis of labor vulnerability and determinants of decent work. The case of Ecuador 2008-2011	Reis, M., & Villacís, A.	El artículo hace un análisis de la vulnerabilidad del mercado laboral medida a través de la iniciativa de trabajo decente de la oit. Se aplica —para Ecuador en el período 2008-2011— un modelo derivado de la metodología de pobreza multidimensional que toma en cuenta tanto el bienestar económico como el bienestar social. Como resultados, se encuentra que, para el promedio del período, apenas el 1 % del total de la población ocupada tiene un trabajo decente, mientras que el 67,6 % tiene un trabajo considerado como no decente, en el cual su ingreso laboral no alcanza para cubrir sus necesidades básicas y posee más de cuatro carencias en sus derechos laborales. Los trabajadores vulnerables por mínimos estándares laborales representan el 31 % y los trabajadores vulnerables por salario, el 0,4 %. Mujeres, trabajadores del sector rural y trabajadores informales son los grupos más propensos a no tener un trabajo decente, mientras el trabajador con mayores años de escolaridad, o que trabaja en el sector público, tiene una mayor probabilidad de conseguir un trabajo decente. Para concluir, se discuten recomendaciones de política pública para el aumento del trabajo decente en Ecuador.

Fuente: Elaboración propia

Anexo B. Definición de la calidad del empleo

Algunos autores han tratado de describir las cualidades del empleo evaluando su calidad, debatiendo sobre cómo medirla y la evolución de esta en los últimos años en diferentes sectores demográficos. Así, se ha puesto en debate lo que significa un "empleo de calidad" y que dimensiones se deben incluir dentro del indicador que califique este, con el fin último de evaluar monitorear políticas públicas relacionadas a este (Gamero, 2018).

Parece haberse llegado a un consenso en que un empleo de calidad implica un "trabajo decente" otorgado por la OIT (2002) o una definición que parte de esta; implica ser lo opuesto a un "empleo precario". Así, un empleo decente se entiende como aquel que es productivo, en el cual son respetados los derechos laborales de los trabajadores y estos acceden a la protección y diálogo social (OIT, 2002).

Más aún, años después de esta definición el mismo OIT ha sido más explícito al mencionar que, en contraste con un empleo decente, un empleo precario es aquel que, con ingresos inadecuados, baja productividad, condiciones laborales precarias y no respeta los derechos fundamentales (OIT, 2010). Esta definición ha sido usada posteriormente por otros autores (Gamero, 2018; INEI, 2012; Rivarola, 2019) para demarcar, por oposición, lo que será un empleo decente.

Sin embargo, algunos autores han puesto en cuestión esta definición. Para Gamero (2018), un trabajo decente implica ir más allá del contexto laboral pues debe considerarse el balance entre el trabajo y el hogar, la participación en la administración de la empresa por parte del trabajador, entre otros elementos. Asimismo, también se ha debatido que la definición de la OIT es insuficiente y debe ser ampliada (Garavito, 2010).

Con ello se han construido a lo largo de los años indicadores que evalúen la calidad del empleo para los trabajadores (Gamero, 2018; Garavito, 2010; Herrera & Hidalgo, 2002; INEI, 2012; Rivarola, 2019) en diferentes grupos de la población peruana (a nivel de Lima Metropolitana o a nivel nacional). Estas tienen en común, debido en parte a las limitaciones de información disponible de las características del empleo, tres dimensiones relacionadas a la definición previamente mencionada.

La primera dimensión está relacionada a la estabilidad laboral y evalúa si es que el trabajador posee un contrato si es dependiente y/o si es que su unidad productiva está registrada en el Registro Único de contribuyentes en el caso de un trabajador independiente. La segunda dimensión (subempleo por ingresos) está

vinculada a la productividad del trabajo y evalúa si es que el trabajador produce lo suficiente para recibir la remuneración mínima vital (RMV). Por último, asociado una jornada laboral decente, se evalúa si es que el horario laboral del trabajador está limitado por un rango de horas a la semana (Gamero, 2018; Herrera & Hidalgo, 2002; INEI, 2012; Rivarola, 2019).

Asimismo, todos consideran que existe unas diferencias sustanciales entre la calidad del empleo en el área rural y urbano debido a la naturaleza de los sectores y las condiciones a las que se enfrentan los trabajadores (Rivarola, 2019). De la misma manera, se han establecido diferencias en las dimensiones a cumplir por parte de los trabajadores dependientes e independientes (Gamero, 2018; Garavito, 2010; INEI, 2012; Rivarola, 2019).

Si bien algunos indicadores han incluido dimensiones relacionadas a los derechos laborales, como el acceso a un servicio de salud o pensiones (Gamero, 2018; Garavito, 2010); algunos también las han omitido debido a que estas dimensiones pueden ser cubiertas no necesariamente por tener un empleo, sino a raíz de la ampliación de aseguramiento universal como política del Estado o porque se comportan de manera muy similar a la dimensión de estabilidad laboral (Rivarola, 2019).

Así, se ha encontrado que tener bajos niveles educativos y ser joven (Garavito, 2010; Rivarola, 2019) aumenta la probabilidad de tener un empleo de mala calidad; como también lo es el ser mujer (Herrera & Hidalgo, 2002). Adicionalmente, tener vinculación laboral con el sector comercio o construcción (Rivarola, 2019) o un menor ingreso no laboral reduce las probabilidades de un empleo de mala calidad (Garavito, 2010). La coyuntura macroeconómica afecta las transiciones del empleo, así como la calidad de este (Herrera & Hidalgo, 2002; Rivarola, 2019).

Como se puede notar, hay una sobreposición entre este subgrupo de la PEA con baja calidad del empleo con aquellos que son los más vulnerables en la sociedad tal que ambas, la condición precaria del empleo y del trabajador como ser humano, se retroalimentan. En particular, un grupo mayoritario dentro de esta intersección son las mujeres y jóvenes.

Vale la pena destacar que las dimensiones consideradas en Perú han sido también las más consideradas en otras partes de Latinoamérica (González et al., 2021). Esto se debe, al igual o en mayor medida que las transiciones, a la falta de información disponible sobre los empleos. A pesar de ser considerada como una

dimensión que influye en la medición de la pobreza, y, por ende, bienestar de las personas (Zavaleta et al., 2018), no se encuentran encuestas específicas para este tipo de medidas.

La definición de vulnerabilidad del empleo que elegimos implica también tener que determinar también qué es un empleo de buena calidad. De la misma manera, se puede usar la definición brindada por la OIT (2010) para entender a un empleo decente como aquel en que cuente con ingresos adecuados, alta productividad, condiciones laborales dignas y que se atenga a los derechos fundamentales de los trabajadores. Explícitamente, en concordancia en parte con Rivarola (2019), podemos plantear un índice que define que la calidad del empleo está compuesta por las 3 dimensiones más frecuentes en la literatura:

1. Subempleo por ingresos: un empleo de buena calidad otorgará por lo menos la RMV al trabajador el año en el que se encuentre laborando.
2. Subempleo por horas: un empleo de buena calidad implica no menos de 30 horas semanales, pero no más de 40.
3. Estabilidad laboral: un empleo de buena calidad implica la presencia de un contrato en el caso de las trabajadoras dependientes, mientras que para las trabajadoras independientes implica la tenencia del RUC de su unidad productiva.

Así, se podría considerar como un empleo de buena calidad a aquellos que cumplan con las 3 dimensiones; mientras que a uno de mala calidad si es que no cumple con al menos una de estas. La determinación de estas tres dimensiones presentadas está en línea con lo planteado no solo por entidad internacionales que defienden el empleo, como la OIT; sino porque también es la definición que se ha usado en distintas investigaciones laborales en el Perú.

Sin embargo, esta definición no está fija y ha cambiado a lo largo de los documentos revisados. Un empleo de calidad no se puede ni se debería limitar a cubrir una sola o varias, pero no todas las aristas del derecho de un empleo, no al menos si a lo que aspiramos es un estándar que le permita a los trabajadores tener una vida decente. Es decir, podemos tener en cuenta que más dimensiones pueden ser consideradas dependiendo del contexto en el que nos encontremos.

En otros estudios, estas dimensiones han cambiado a partir del tipo de población del mercado laboral a estudiar, del sector productivo que se ha tomado en cuenta, etc. Las dimensiones que se han tomado en cuenta pueden variar desde aquellas consideradas como objetivas, como tener un contrato o no; así como ser

subjetivas, es decir, basadas en la percepción del trabajador sobre su empleo. Más aún, no necesariamente se ha usado solo un indicador para evaluar la calidad del empleo, sino que también varios indicadores en conjunto (González et al., 2021; Kumar & Srivastava, 2021).

Las dimensiones que componen a la calidad del empleo, como hemos podido explicar, es un tema muy amplio y que dejamos como agenda de investigación en el presente documento. Alentamos a profundizar en el debate sobre el término de “calidad del empleo” pues al hacerlo estamos debatiendo también que implica cumplir con los estándares mínimos que un trabajador debe tener. En una sociedad con altos niveles de precariedad laboral como lo es la peruana, el discutir más acerca de qué es lo que consideramos necesario para vivir y trabajar es el primer paso para mejorar esta situación que nos afecta a todos.

Anexo C. Tablas resúmenes de la PEA urbana ocupada

Tabla 14.

Tabla resumen de la PEA urbana ocupada en su año inicial de las bases bianuales del 2016 al 2019.

	2016		2017		2018		2019		Total	
	N	%	N	%	N	%	N	%	N	%
Grupo etario										
14 años	10,063	0.33%	9,877	0.60%	9,937	0.46%	9,975	0.47%	39,852	0.47%
15 - 29 años	10,063	26.96%	9,877	26.10%	9,937	25.89%	9,975	26.00%	39,852	26.00%
30 - 44 años	10,063	36.45%	9,877	36.37%	9,937	36.40%	9,975	35.96%	39,852	35.96%
45 - 64 años	10,063	30.79%	9,877	31.90%	9,937	31.61%	9,975	31.95%	39,852	31.95%
>= 65 años	10,063	5.47%	9,877	5.03%	9,937	5.64%	9,975	5.62%	39,852	5.62%
Estado civil										
Conviviente	10,063	28.47%	9,877	29.17%	9,937	28.95%	9,975	28.46%	39,852	28.46%
Casado	10,063	29.11%	9,877	29.04%	9,937	27.42%	9,975	28.27%	39,852	28.27%
Divorciado	10,063	0.54%	9,877	0.34%	9,937	0.59%	9,975	0.56%	39,852	0.56%
Separado	10,063	9.11%	9,877	10.03%	9,937	11.53%	9,975	10.79%	39,852	10.79%
Soltero	10,063	30.19%	9,877	28.72%	9,937	28.78%	9,975	29.14%	39,852	29.14%
Pobreza										
Pobre extremo	10,063	0.67%	9,877	0.64%	9,937	0.62%	9,975	0.59%	39,852	0.59%
Pobre no extremo	10,063	10.74%	9,877	10.37%	9,937	11.06%	9,975	10.74%	39,852	10.74%
No pobre	10,063	88.59%	9,877	88.99%	9,937	88.31%	9,975	88.67%	39,852	88.67%
Nivel educativo										
No nivel	10,063	2.29%	9,877	1.86%	9,937	1.73%	9,975	1.95%	39,852	1.95%
Primaria	10,063	16.55%	9,877	16.19%	9,937	15.98%	9,975	16.01%	39,852	16.01%
Secundaria	10,063	44.88%	9,877	44.17%	9,937	43.61%	9,975	44.00%	39,852	44.00%
Superior No Universitaria	10,063	16.96%	9,877	17.59%	9,937	17.57%	9,975	17.53%	39,852	17.53%
Superior Universitaria	10,063	17.36%	9,877	18.02%	9,937	19.00%	9,975	18.36%	39,852	18.36%
Postgraduado	10,063	1.96%	9,877	2.17%	9,937	2.10%	9,975	2.16%	39,852	2.16%

Fuente: ENAHO Panel 2016-2019, ENAHO Panel 2016-2020. INEI. Elaboración propia.

Tabla 15.

Tabla resumen de la PEA urbana del apilamiento de las bases bianuales del 2016 al 2019.

	2016		2017		2018		2019		Total	
	N	%	N	%	N	%	N	%	N	%
Grupo etario										
14 años	15,293	1.99%	14,995	2.26%	15,124	2.29%	15,067	2.52%	60,479	2.27%
15 - 29 años	15,293	33.69%	14,995	33.52%	15,124	32.90%	15,067	32.13%	60,479	33.04%
30 - 44 años	15,293	29.04%	14,995	29.37%	15,124	29.44%	15,067	28.20%	60,479	29.01%
45 - 64 años	15,293	25.62%	14,995	25.74%	15,124	26.03%	15,067	27.15%	60,479	26.15%
>= 65 años	15,293	9.67%	14,995	9.12%	15,124	9.35%	15,067	10.00%	60,479	9.54%
Estado civil										
Conviviente	15,293	24.00%	14,995	24.63%	15,124	24.02%	15,067	23.01%	60,479	23.91%
Casado	15,293	26.49%	14,995	25.98%	15,124	25.13%	15,067	24.88%	60,479	25.60%
Viudo	15,293	3.92%	14,995	3.91%	15,124	3.91%	15,067	4.35%	60,479	4.03%
Divorciado	15,293	0.50%	14,995	0.37%	15,124	0.59%	15,067	0.69%	60,479	0.54%
Separado	15,293	8.08%	14,995	8.62%	15,124	9.63%	15,067	10.24%	60,479	9.16%
Soltero	15,293	37.00%	14,995	36.49%	15,124	36.72%	15,067	36.84%	60,479	36.76%
Pobreza										
Pobre extremo	15,293	0.79%	14,995	0.68%	15,124	0.77%	15,067	0.64%	60,479	0.72%
Pobre no extremo	15,293	11.20%	14,995	11.57%	15,124	12.31%	15,067	11.36%	60,479	11.61%
No pobre	15,293	88.01%	14,995	87.75%	15,124	86.92%	15,067	88.00%	60,479	87.67%
Nivel Educativo										
No nivel	15,293	3.00%	14,995	2.50%	15,124	2.43%	15,067	2.44%	60,479	2.59%
Primaria	15,293	17.20%	14,995	16.44%	15,124	16.10%	15,067	15.61%	60,479	16.32%
Secundaria	15,293	46.93%	14,995	46.54%	15,124	46.54%	15,067	46.12%	60,479	46.52%
Superior No Universitaria	15,293	14.42%	14,995	15.10%	15,124	15.00%	15,067	15.44%	60,479	15.00%
Superior Universitaria	15,293	17.01%	14,995	17.85%	15,124	18.40%	15,067	18.65%	60,479	17.99%
Postgraduado	15,293	1.45%	14,995	1.57%	15,124	1.53%	15,067	1.73%	60,479	1.57%

Fuente: ENAHO Panel 2016-2019, ENAHO Panel 2016-2020. INEI. Elaboración propia.

Anexo D. Estimaciones metodológicas del modelo Lasso

Si partimos de:

$$\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_{ij})^2 + \lambda\left|\left|\beta\right|\right|_1$$

Normalmente, se estandarizan los predictores x_i tal que se cumple que la media de cada variable es $\frac{1}{N}\sum_{i=1}^{N} x_{ij} = 0$ y para la varianza se cumple que $\frac{1}{N}\sum_{i=1}^{N} x_{ij}^2 = 1$. Esto nos ayuda puesto que ahora los coeficientes no dependen de las unidades en que se encuentran. Asimismo, por conveniencia, también estandarizaremos la variable respuesta, es decir, $\frac{1}{N}\sum_{i=1}^{N} y_{ij} = 0$. Esto nos permite omitir el intercepto β_0 en la optimización del modelo, y a la vez nos permite recuperar el intercepto no estandarizado. Suponiendo que existe un $\hat{\beta}$ en los datos estandarizados, tenemos que para los datos no centrados β es el mismo y el intercepto es:

$$\beta_0 = \underline{y} - \sum_{j=1}^{P} \underline{x_j}\widehat{\beta_j} \qquad (1)$$

Donde y $\left\{\underline{x_j}\right\}_1^p$ son las medias con los datos originales no estandarizados. De esta manera, podemos plantear el problema de manera matricial y en su forma lagrangiana. La restricción puede ser escrita en su forma reducida de norma como $\left|\left|\beta\right|\right|_1 \leq \lambda$, con un $\lambda > 0$; mientras que el vector respuesta $1\ x\ N$ puede ser descrito como $y = (y_1, \ldots, y_N)$ y la matriz $N\ x\ p$ con los $x_i \in R^p$ predictores en su i^{th} columna puede ser expresado como X. Así, el problema de minimización puede ser planteado como:

$$\{\frac{1}{2N}\left|\left|y - X\,\beta\right|\right|_2^2 + \lambda\left|\left|\beta\right|\right|_1\} \qquad (2)$$

Por la dualidad lagrangiana, existe una regla de correspondencia de 1 a 1 entre la restricción del problema y la forma lagrangiana. Por cada valor de λ en el rango en el cual los $\left|\left|\beta\right|\right|_1 \leq \lambda$ están activos, existe un valor de λ que da la misma solución de la forma lagrangiana que explicaremos más adelante pues nos sirve para profundizar en el intercambio varianza-sesgo.

En muchos casos, el término $\frac{1}{2N}$ es simplificado por 1 pues nos permite una estandarización de λ que hace que sus valores sean comparables en muestras de diferentes tamaños, a la vez que no nos genera diferencias en el problema de optimización y nos es útil para procedimiento posteriores.

De esta manera, las condiciones necesarias y suficientes para la solución de este problema convexo toman la forma de un sistema de ecuaciones que puede ser planteado como:

$$-\frac{1}{N}\langle x_j, y - X\,\beta\rangle + \lambda s_j = 0, j = 1, \dots, p \tag{3}$$

Aquí, cada s_j es un valor arbitrario igual a $signo(\beta_j)$ si es que $\beta_j \neq 0$ y su valor esta entre $[-1,1]$, el cual es la subgradiente para el valor absoluto de la función de valor. Este sistema tiene la forma de las condiciones de Karush-Kuhn-Tucker (KKT) y nos sirve, al ser expresado en la forma de subgradiente, para hallar su solución mediante algoritmos.

Aquí nos explayaremos en la derivación del modelo Lasso descrito anteriormente. A partir de la función objetivo, podemos ver que el modelo Lasso es un programa cuadrático con una restricción convexa. En este sentido, lo natural sería tomar la gradiente, o primera derivada, con respecto a β e igualar está a 0. Este procedimiento se repetiría para las $k - ésimas$ ecuaciones presentes, teniendo así un sistema de ecuaciones. Sin embargo, es necesario analizar las partes de este sistema de ecuaciones de manera detallada para una variable. Así, podemos dividir la ecuación en dos términos:

$$= -\frac{1}{N}\sum_{i=1}^{N} x_{ik}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \partial_{\beta_j}\lambda\sum_{i=1}^{P}|\beta_j| \tag{4}$$
$$= \partial_{\beta_k} RSS^{OLS}(\beta) + \partial_{\beta_k}\lambda\left|\left|\beta\right|\right|_1$$

Primero, podemos ver el primer término, $\partial_{\beta_k} RSS^{OLS}(\beta)$, para la variable k es:

$$= -\frac{1}{N}\sum_{i=1}^{N} x_{ik}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 \quad (5)$$

El cual, partiéndolo en partes, podemos plantear también como:

$$= -\frac{1}{N}\sum_{i=1}^{N} x_{ik}\left(y_i - \sum_{j\neq k}^{p} x_{ij}\beta_j - x_{ik}\beta_k\right)^2 \quad (6)$$

$$= \frac{1}{N}\sum_{i=1}^{N} x_{ik}\left(y_i - \sum_{j\neq k}^{p} x_{ij}\beta_j\right)^2 + \beta_k \sum_{i=1}^{N}(x_{ik})^2$$

Como podemos ver, $\sum_{i=1}^{P} \quad (x_{ij})^2$ es la varianza de x_{ik}. Como mencionamos anteriormente, esta es una variable esta estandarizada por lo cual este término sería igual a 1. De esta manera, podemos replantear esta parte de la primera derivada como:

$$= -\frac{1}{N}\sum_{i=1}^{N} x_{ik}\left(y_i - \sum_{j\neq k}^{p} x_{ij}\beta_j\right)^2 + \beta_k \sum_{i=1}^{N}(x_{ik})^2 \quad (7)$$

$$= -\rho_j + \beta_k\, z_k$$

Donde podemos obtener ρ_j con los datos que tenemos, mientras que z_k es una constante normalizada que toma el valor de 1 pero por motivos metodológicos dejaremos expresada de esta manera. Ahora, si analizamos el segundo término de la ecuación, $\partial_{\beta_k}\ \lambda|\beta_k|$, podemos notar que existe un problema pues la derivada de $|\beta_k|$ no está definida en cero. Por esta razón, vamos a usar el método del descenso de coordenadas para optimizar esta parte. Este método consiste en la optimización inteligente de coordenadas y el uso de subderivadas y subdiferenciales, que pasaremos a explicar a continuación partiendo del desarrollo que se tiene en Hastie et al. (2015).

Normalmente, cuando tenemos una función convexa y derivable con un límite inferior $f\colon R^p \to R$ que está delimitada entre $[a, b]$, podemos encontrar un x_0 que es el mínimo global si cumple con ciertas condiciones. Entonces, si f es una función

convexa, se cumple que para dos vectores β y β' en el dominio de f y con un escalar $s \in [0,1]$, tenemos que:

$$f\big(\beta(s)\big) = f(s\beta + (1-s)\beta') \leq sf(\beta) + (1-s)f(\beta')$$

Esto nos permite garantizar que todo mínimo local es, a la vez, mínimo global puesto que el resultado que salga de la función usando a la combinación lineal del valor ponderado de dos vectores como insumo, será siempre menor que la combinación lineal de los resultados ponderados de los vectores.

La relevancia de esto se justifica en que varios de los problemas de optimización tanto fuera como dentro de los modelos econométricos comunes hacen uso de funciones convexas con restricciones. En este sentido, Teniendo una función objetivo convexa de este tipo y diferenciable con un conjunto de restricción $\mathsf{C} \subset R^p$ tal que:

$$min_{\beta_o,\beta} f(\beta) \quad tal\ que\ \ \beta \in \mathsf{C} \tag{8}$$

Podemos hallar su mínimo global s$\beta^* \in \mathsf{C}$ si es que este cumple con su condición necesaria y suficiente:

$$\langle \nabla f(\beta^*), \beta - \beta^* \rangle \geq 0$$

Aquí se presenta la suficiencia considerando a $f(\beta) \geq f(\beta^*) + \langle \nabla f(\beta^*),\ \beta - \beta^* \rangle \geq f(\beta *)$, que se deriva de la convexidad de f y la condición de optimalidad. Así, podemos plantear un conjunto de restricciones C que pueden ser funciones convexas para cualquiera de estas funciones $g: R^p \to R$ que la componen y que cumple que $\beta \in R^p \mid g(\beta) \leq 0$. De esta manera, podemos plantear la función de optimización como:

$$min_{\beta_o,\beta} f(\beta) \quad tal\ que\ \ g_i(\beta) \leq 0 \ \ para\ cada\ j = 1, \dots, m \tag{9}$$

Otra manera de plantearla es en su forma lagrangiana, tal que si tenemos que esta función $L: R^P x\ R_+^m \to R$ la podemos plantear como:

$$L(\beta; \lambda) = f(\beta) + \sum_{j=1}^{m} \lambda_j g_j(\beta) \tag{10}$$

Donde los pesos λ son conocidos como multiplicadores e imponen una penalidad por la restricción. Si este conjunto de restricciones fuera diferenciable, lo que normalmente tendríamos que realizar para hallar f^*, el valor óptimo del problema de optimización, es, tras hallar el o los λ ideal, hallar aquel conjunto de β que minimice la función lagrangiana de cumplirse las funciones de restricción tal que $g_i(\beta) \leq 0$ para cada $j = 1, \dots, m$. De esta manera, en esta situación óptima, se cumpliría no solo las restricciones $g_j(\beta^*) \leq 0$, sino también que sería un resultado de igualar al gradiente de esta a 0 tal que:

$$0 = \nabla_\beta L(\beta^*; \lambda^*) = \nabla f(\beta^*) + \sum_{j=1}^{m} \lambda_j^* \nabla g_j(\beta^*) \tag{11}$$

De esta, se derivan dos condiciones conocidas como las condiciones de Karush-Kuhn Tucker. Así, $g_j(\beta^*) \leq 0$ es la viabilidad primaria, y $\lambda_j^* g_j(\beta^*)$ la de holgura complementaria garantizan que la solución sea un óptimo global siempre y cuando el problema de optimización cumpla con las condiciones de dualidad fuerte.

Sin embargo, como los autores de Hastie et al. (2015) mencionan, en el modelo Lasso tenemos una parte que no es diferenciable cuando al menos una coordenada de β_j es igual a 0, pero que si es una función convexa: $\sum_{i=1}^{P} |\beta_j|$. Con un procedimiento similar al anterior, podríamos intuir que un subgradiente es cualquier hiperplano normalizado que limita por debajo de la función Lagrangiana y es tangente con el punto mínimo β^*. Así, para f, podríamos plantear que $v \in \partial f(x)$ es un subgradiente de esta función si es que se cumple:

$$f(\beta') \geq f(\beta) + \langle z \ \beta' - \beta \rangle$$

En los puntos no derivables, el subdiferencial es un conjunto convexo que contiene todos los posibles subgradientes. Gráficamente, para la función del valor absoluto $f(\beta) = |\beta|$, podemos intuir que los planos que cumplen con estas condiciones son aquellas cuyas pendientes se encuentran entre $[-1,1]$. De tal manera, se tendría que:

$$\partial f(\beta) \{+1, \qquad si\ \beta > 0\ [-1, +1], \qquad si\ \beta = 0\ -1, \qquad si\ \beta < 0 \tag{12}$$

Considerando este cambio, podemos intuir que la condición de la gradiente de la función Lagrangiana equivalente a cero no es adecuada. Sin embargo, como en el paso anterior, podemos aplicar una generalización de la teoría de los Karush-Kuhn-Tucker (KKT), usándola condición modificada:

$$0 = \partial f(\beta^*) + \sum_{j=1}^{m} \lambda_j^* \partial g_j(\beta^*) \tag{13}$$

Así, como el subdiferencial es un conjunto, la anterior ecuación nos indica que un vector de ceros pertenece a la suma de los diferenciales. Con esta contextualización, podemos dividir a este segundo término de manera muy similar como hicimos en la primera, tal que obtenemos una función univariada que si podemos optimizar como:

$$\lambda \sum_{i=1}^{P} |\beta_j| = \lambda|\beta_k| + \lambda \sum_{j \neq k}^{P} |\beta_j| \tag{14}$$

Usando la definición de subdiferencial como un intervalo no vacío y cerrado $[a, b]$ donde a y b son límites de la derivada de β_k, podemos plantearla como:

$$\partial\beta_k \, \lambda \sum_{i=1}^{P} |\beta_j| = \partial\beta_k \, \lambda|\beta_k| = \begin{cases} -\lambda, & si\ \beta_k < 0 \\ [-\lambda, \lambda], & si\ \beta_k = 0 \\ \lambda, & si\ \beta_k > 0 \end{cases} \tag{15}$$

Así, podemos notar que el signo de λ depende de la posición que el valor absoluto de β_k tome con respecto a 0. De esta manera, uniendo ambas partes podemos expresar la siguiente ecuación como:

$$\begin{aligned} &= \partial_{\beta_k} RSS^{OLS}(\beta) + \partial_{\beta_k} \lambda \left|\left|\beta\right|\right|_1 \\ &= -\rho_j + \beta_k \, z_k + \ \partial_{\beta_j} \lambda \left|\left|\beta_j\right|\right| \end{aligned} \tag{16}$$

Tal que, planteando por los casos presentes en el segundo término de la ecuación, tenemos:

$$= \{-\rho_j + \beta_k\, z_k - \lambda, \quad si\ \beta_k < 0 \left[-\rho_j - \lambda, -\rho_j + \lambda\right], \quad si\ \beta_k = 0 \ -\rho_j + \beta_k\, z_k + \lambda, \quad si\ \beta_k > 0 \tag{17}$$

Ahora, tenemos que recalcar las 3 propiedades de la teoría subdiferencial planteados anteriormente:

Es posible diferenciar una función convexa en x_0 sí y solo si el conjunto de subdiferencial está constituido por solo un punto, el cual es la derivada de x_0.

El teorema de Moreau-Rockafellar sostiene que si f y g son funciones convexas con sus subdiferenciales ∂f y ∂g, entonces la subdiferencial $f + g$ es $\partial(f + g) = \partial f + \partial g$. La condición estacionaria sostiene que un punto x_0 es el mínimo global de una función convexa f si y solo si el cero pertenece al subdiferencial.

Para cumplir con la última propiedad, es necesario que, en el segundo caso, cuando $\beta_k = 0$, el intervalo contenga a 0 y así se pueda encontrar un mínimo global. De esta manera, se plantean las siguientes inecuaciones:

$$-\rho_j - \lambda \leq 0$$
$$-\rho_j + \lambda \leq 0$$
$$-\lambda \leq \rho_j \leq \lambda$$

Así, despejando β_k para el primer, $\beta_k < 0$, y el tercer caso, $\beta_k > 0$, y combinando con los resultados anteriores, obtenemos:

$$\beta_k = \{\frac{\rho_j + \lambda}{z_j}, \quad si\ \rho_j < -\lambda\ 0, \quad si\ -\lambda \leq \rho_j \leq \lambda\ \frac{\rho_j - \lambda}{z_j}, \quad si\ \rho_j > \lambda \tag{18}$$

De esta manera, reconocemos a esto como la función de umbral suave (*Soft Thresholding Function*): $\frac{1}{z_j}\, S(\rho_j, \lambda)$. Aquí, $\frac{1}{z_j}$ es una constante normalizada y que toma el valor de uno puesto que, como vimos, z_j es igual a 1 pues nuestros datos están normalizados. Así, podemos obtener:

$$\beta_k = \{-\rho_j + \lambda, \quad si\ \rho_j < -\lambda\ 0, \quad si\ -\lambda \leq \rho_j \leq \lambda\ \rho_j + \lambda, \quad si\ \rho_j > \lambda \tag{19}$$

Ahora, como planteamos en un inicio, esto solo sucede para un coeficiente β_k. Entonces, para hallar para todos los demás coeficientes, debemos aplicar el descenso de coordenadas cíclico. Este consiste en minimizar nuestra función objetivo con respecto a cada coordenada para cada una de las variables.

Primero, partiendo de que existe una función diferenciable $f: R^p \to R$, y si partimos de un punto x tal que $f(x)$ se minimiza a lo largo de cada eje de coordenadas, podemos encontrar un mínimo global pues se cumple que, para $e_i = (0, \dots, 1, \dots, 0) \in R^p$, siendo el $i - esimo$ vector estándar básico, podemos probar que se cumple $f(x + d * e_i) \geq f(x)\ para\ todo\ d, i \Rightarrow f(x) = f(z)$, es decir, que x es el vector que nos lleva al punto mínimo de la función en comparación con desviaciones de este pues en este punto se cumple que:

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x), \dots, \frac{\partial f}{\partial x_n}(x)\right) = 0 \tag{20}$$

Al minimizar nuestra función objetivo con respecto a cada coordenada para cada una de las variables, estamos usando el descenso de coordenadas. Este método también se puede usar para el tipo de funciones objetivos al que pertenece el modelo Lasso. Así, si tenemos $f(x) = g(x) + \sum_{i=1}^{n} \quad h_i(x_i)$ siendo g convexa y diferenciable (en nuestro caso, $\frac{1}{2N}\sum_{i=1}^{N} \quad \left(y_i - \sum_{j=1}^{p} \quad x_{ij}\beta_j\right)^2$); mientras cada h_i , llamada ahora la parte separable, es solo convexo (en nuestro caso, $\lambda \sum_{i=1}^{P} \quad |\beta_j|$), podemos encontrar que existe un mínimo global x pues se cumple que:

$$f(y) - f(x) \geq \nabla g(x)^T (y - x) + \sum_{i=1}^{n} [h_i(y_i) - h_i(x_i)] \tag{21}$$
$$= \sum_{i=1}^{n} \quad [\nabla_i g(x)(y_i - x_i) + h_i(y_i) - h_i(x_i)] \geq 0$$

De manera similar, podemos intuir que es posible minimizar nuestra función objetivo con respecto a cada coordenada para cada una de las variables. Empezando con un $x^{(0)} = (x_1^0, \dots, x_n^0)$ arbitrario, para cada $k = 1,2,3, \dots$, tenemos

$$x_1^k \in argmin_{x_1} f(x_1, x_2^{(k-1)}, x_3^{(k-1)}, \dots, x_n^{(k-1)})$$
$$x_2^k \in argmin_{x_2} f(x_1^{(k)}, x_2, x_3^{(k-1)}, \dots, x_n^{(k-1)})$$

$$x_3^k \in argmin_{x_3} f(x_1^{(k)}, x_2^{(2)}, x_3, \dots, x_n^{(k-1)})$$
$$\dots$$
$$x_n^k \in argmin_{x_n} f(x_1^{(k)}, x_2^{(2)}, x_3^{(k)}, \dots, x_n)$$

Así, para cada ronda $k+1$ se define $x^{(k+1)}$ de $x^{(k)}$ al resolver de manera interactiva un problema de optimización univariado que se puede generalizar como:

$$x_i^{(k+1)} \in argmin_w f(x_1^{(k+1)}, \dots, x_{i-1}^{(k+1)}, w, \dots, x_n^{(k-1)}) \quad (22)$$

Donde $x_i \coloneqq x_i - \alpha \frac{\partial f}{\partial x_i}(x)$. De esta manera se repite para cada variable x_i en x para $i = 1, \dots, n$. Note que el ρ_j puede ser rescrito como:

$$\rho_j = \sum_{i=1}^{m} x_j^{(i)} (y^{(i)} - \sum_{k \neq j}^{n} \beta_k x_k^{(i)}) = \sum_{i=1}^{m} x_j^{(i)} (y^{(i)} - y_{pred}^{(i)} + \beta_j x_j^{(i)}) \quad (23)$$

Así, al hallar el f, que es una función continua en un conjunto compacto de valores $\{x: f(x) \leq f(x^{(0)})\}$ y contienen a su mínimo; cualquier punto límite de $x^{(k)}$ para $k = 1,2,3, \dots$ es un mínimo de f puesto que si $x^{(k)}$ es una subserie que converge a x^* (Teorema de Bolzano-Weirstrass), entonces $f(x^{(k)})$ converge a f^* por la convergencia monótona.

Cabe mencionar que lo que intento probar aquí, además de la optimización del modelo Lasso, las metodologías, que son *sets* de herramientas para los investigadores, que se ha desarrollado en otras áreas familiares a la Economía – como Estadística- no son muy diferentes a las que aplicamos normalmente en nuestra área, por lo que es posible establecer una conversación entre ellas.

Anexo E. Justificación teórica del modelo logístico a partir de una variable latente

Un modelo teórico que nos permite incluir la función logística es aquella que proclama la existencia de una variable latente I^* que está detrás de cómo se comporta y_i (Gujarati & Porter, 2010). Así, tenemos una variable no observable I^* que determina el valor de y_i a partir de su posición ante un umbral. Si I^* supera el valor del umbral, que puede ser 0 por ejemplo; entonces $y_i = 1$, mientras que, si no lo hace, $y_i = 0$. Suponiendo $I^* = x\beta + u$, podemos plantear estas condiciones en un sistema de ecuaciones podemos:

$$y_i = \{1, \quad si\ I^* > 0\ \ 0, \quad si\ I^* < 0$$

$$y_i = \{1, \quad si\ \ x\beta + u > 0\ \ 0, \quad si\ \ \ x\beta + u < 0$$

Entonces, teniendo en cuenta esta variable no observable $I^* = x\beta + u$, para determinar la probabilidad p_i de que suceda $y_i = 1$, establecemos las siguientes equivalencias:

$$p_i = Pr\,[y_i = 1|x]$$
$$p_i = Pr\,[y_i = 1|x] = Pr\,[I^* > 0]$$
$$p_i = Pr\,[y_i = 1|x] = Pr\,[I^* > 0] = Pr\,[x\beta + u > 0]$$
$$p_i = Pr\,[y_i = 1|x] = Pr\,[I^* > 0] = Pr\,[x\beta + u > 0] = F(x\beta)$$

Donde sabemos que $F(x\beta)$ es la función logística, es decir, $F(x\beta) = \frac{1}{1+e^{-x\beta}}$.

Anexo F. Optimización del logit penalizado

El objetivo del algoritmo de Newton Raphson es maximizar $ln\ ln\ L(\beta)$ usando esquema de interacciones del escalar de β. Y lo que busca es una aproximación lineal de segundo orden de Taylor de la función de longverosimilitud alrededor de un punto, β_1 por ejemplo, tal que tenemos:

$$ln\ L(\beta)\ = ln\ L(\beta_1) + \frac{\partial ln\ (\beta_1)(\beta - \beta_1)}{\partial \beta} + \frac{1}{2}\frac{\partial^2 ln\ L(\beta_1)(\beta - \beta_1)^2}{\partial \beta^2}\ + \cdots$$

Como se quieren hallar los $\hat{\beta}$ óptimos, maximizamos esta aproximación tal que nos queda un problema como:

$$\{\ ln\ L(\beta_1) + \frac{\partial ln\ (\beta_1)(\beta - \beta_1)}{\partial \beta} + \frac{1}{2}\frac{\partial^2 ln\ L(\beta_1)(\beta - \beta_1)^2}{\partial \beta^2}\}$$

Con esto, en paralelo al modelo Lasso, podemos hallar las condiciones de primer orden a partir de derivar en función de β:

$$\frac{L(\beta_1)}{\partial \beta} = \frac{L(\beta_1)}{\partial \beta} + \frac{\partial^2 ln\ L(\beta_1)}{\partial \beta^2}\left(\widehat{\beta_2} - \beta_1\right) = 0$$

Tal que podemos ver que el supuesto $\widehat{\beta_2}$ ideal esta descrito como:

$$\widehat{\beta_2} = \beta_1 - \left[\frac{\partial^2 ln\ L(\beta_1)}{\partial \beta^2}\right]^{-1}\frac{\partial ln\ L(\beta_1)}{\partial \beta}$$

Si tomamos en cuenta que:

$$H(\beta_1) = \frac{\partial^2 \beta_o}{\partial \beta^2}$$

$$S(\beta_1) = \frac{\partial ln\ L(\beta_1)}{\partial \beta}$$

Donde $H(\beta_1)$ es la matriz de segundas derivadas de la función de longverosimilitud, la cual si invertimos no es más que la Hessiana inversa; mientras que $S(\beta_1)$ es la matriz gradiente de la función. Aquí, debemos recordar que para que la matriz Hessiana sea invertible, la función $ln\ ln\ L(\beta)$ debe ser doblemente diferenciable, lo cual se logra puesto que es una transformación monótona de una distribución logística.

De esta manera, podemos transcribir $\widehat{\beta_2}$ como:

$$\widehat{\beta_2} = \beta_1 - [H(\beta_1)]^{-1} S(\beta_1)$$

Esto se puede generalizar para los n interacciones como:

$$\widehat{\beta_{n+1}} = \widehat{\beta_n} - \left[H\left(\widehat{\beta_n}\right)\right]^{-1} S\left(\widehat{\beta_n}\right)$$

Esta acaba cuando $S\left(\widehat{\beta_n}\right) \approx 0$ pues $\widehat{\beta_{n+1}} \approx \widehat{\beta_n}$.

Los pasos para la interacción son los siguientes. Primero, se calcula la gradiente del primero punto que planteamos, β_o, tal que obtenemos $S(\beta_o)$. Si es que $S(\beta_o) \neq 0$, por lo cual aún no nos encontramos en el óptimo $\hat{\beta}$, entonces se prueba realizar esta operación con vectores vecinos a β_o, como sería β_1.

Cabe recordar que β_1 es la intersección entre el eje x de la tangente de la función en $(\beta_o, ln\ ln\ L(\beta_0))$. Sucesivamente, se prueba con las gradientes de los demás posibles soluciones, como $\beta_1, \beta_2, \ldots$; obteniéndose así $S(\beta_1), S(\beta_2), \ldots$. Cuando la gradiente de alguno de estos intentos sea igual a cero, entonces se detiene el algoritmo pues nos encontraremos en el óptimo y se cumple que $\widehat{\beta_{n+1}} \approx \widehat{\beta_n}$ y se suprime el término que posee al $S\left(\widehat{\beta_n}\right)$.

Entonces, para poder encontrar el óptimo $\hat{\beta}$, necesitamos las funciones que definan $H\left(\widehat{\beta_n}\right)$ y $S\left(\widehat{\beta_n}\right)$, lo cual no es más que la segunda y primera derivada de la función de longverosimilitud, respectivamente. Entonces, recordando que la primera derivada es la siguiente:

$$S(\beta) = \frac{L(\beta)}{\partial \beta} = \sum_{i=1}^{n} [y_i - \frac{e^{x\beta}}{1 + e^{x\beta}}] x_i + \lambda \frac{\partial \left|\left|\beta\right|\right|_1}{\partial \beta}$$

En esta, el algoritmo de Newton-Raphson puede ser aplicado con derivadas analíticas, lo cual haría que la conversión sea más rápida, aunque con las derivadas que conocemos también se lograría al mismo objetivo. Asimismo, para la segunda derivada podemos probar que esta es:

$$H(\beta) = \frac{L(\beta)}{\partial \beta} = -\sum_{i=1}^{n} \frac{e^{x\beta}}{1+e^{x\beta}} \left(1 - \frac{e^{x\beta}}{1+e^{x\beta}}\right) x_i^2 + \lambda \frac{\partial^2 \left|\left|\beta\right|\right|_1}{\partial \beta^2}$$

Como se puede ver, esta ecuación es independiente de y_i y siempre será definida negativa. En este sentido, Podemos afirmar que $ln\, L(\beta)$ es una cóncava global cuando se usa la función logística como función de enlace.

Anexo G. Variables de la base de datos

Tabla 16.

Variables incluidas en la base de datos

Nombre	Etiqueta	Modulo	Tipo
p300a	Lenguaje materno	Varios	Categórica
y_pri_dep	Ingreso primario dependiente	Creado	Continua
y_pri_indep	Ingreso primario independiente	Creado	Continua
y_pri	Ingreso primario total	Creado	Continua
y_sec_dep	Ingreso secundario dependiente	Creado	Continua
y_sec_ind	Ingreso secundario independiente	Creado	Continua
y_sec	Ingreso secundario total	Creado	Continua
y_mkt	Ingreso laboral total	Creado	Continua
sector	Sector de ocupación principal	Varios	Categórica
educ	Nivel de educación	Varios	Categórica
dpto	Departamento	Varios	Categórica
regnat	Regiones naturales	Varios	Categórica
dominio	Dominio geográfico	Creado	Categórica
pobre2	Pobreza monetaria total	Creado	Continua
gpcm	Gasto mensual promedio	Creado	Continua
ingtrabw	Ingreso promedio mensual proveniente de trabajo	Creado	Continua
ipcr_0	Ingreso per cápita mensual a precios de Lima monetario	Creado	Continua
ipcr_1	Ingreso per cápita mensual a precios de Lima monetario por trabajo	Creado	Continua
ipcr_2	Ingreso per cápita mensual a precios de Lima monetario por trabajo principal	Creado	Continua
ipcr_3	Ingreso per cápita mensual a precios de Lima monetario por trabajo secundario	Creado	Continua
ipcr_4	Ingreso per cápita mensual a precios de Lima pago en especie y autocon	Creado	Continua
ipcr_5	Ingreso per cápita mensual a precios de Lima pago extraordinario por trabajo	Creado	Continua
ipcr_6	Ingreso per cápita mensual a precios de Lima transferencia corriente	Creado	Continua
ipcr_7	Ingreso per cápita mensual a precios de Lima transferencia monetaria del país	Creado	Continua
ipcr_8	Ingreso per cápita mensual a precios de Lima transferencia monetaria extranjero	Creado	Continua
ipcr_9	Ingreso per cápita mensual a precios de Lima transferencia monetaria privada	Creado	Continua
ipcr_10	Ingreso per cápita mensual a precios de Lima transferencia monetaria Publica total	Creado	Continua
ipcr_11	Ingreso per cápita mensual a precios de Lima transferencia monetaria Publica Juntos	Creado	Continua
ipcr_12	Ingreso per cápita mensual a precios de Lima transferencia monetaria Publica Pensión65	Creado	Continua
ipcr_13	Ingreso per cápita mensual a precios de Lima transferencia monetaria Bono Gas	Creado	Continua

Nombre	Etiqueta	Modulo	Tipo
ipcr_14	Ingreso per cápita mensual a precios de Lima transferencia monetaria Beca 18	Creado	Continua
ipcr_15	Ingreso per cápita mensual a precios de Lima transferencia monetaria Otros Publica	Creado	Continua
ipcr_16	Ingreso per cápita mensual a precios de Lima renta	Creado	Continua
ipcr_17	Ingreso per cápita mensual a precios de Lima extraordinario	Creado	Continua
ipcr_18	Ingreso per cápita mensual a precios de Lima alquiler imputado	Creado	Continua
ipcr_19	Ingreso per cápita mensual a precios de Lima donación publica	Creado	Continua
ipcr_20	Ingreso per cápita mensual a precios de Lima donación privada	Creado	Continua
tamahno	Tamaño de la empresa	500	Categórica
ciiu_1d	Ramas de actividad según CIIU en empleo formal	500	Categórica
ciiu_6c	Ramas de actividad según CIIU en empleo informal	Creado	Categórica
p_relativa	Pobreza relativa	Creado	Categórica
p558e6	Inclusión financiera	300	Categórica
celular	Tenencia de celular	300	Categórica
gpgru1	Gastos en alimentos dentro del hogar real	Creado	Continua
gpgru2	Gastos en alimentos fuera del hogar real	Creado	Continua
gpgru3	Gastos en total en alimentos real	Creado	Continua
gpgru4	Gastos en vestido y calzado real	Creado	Continua
gpgru5	Gastos en alquiler de vivienda y combustible real	Creado	Continua
gpgru6	Gastos en muebles y enseres real	Creado	Continua
gpgru7	Gastos en cuidados de la salud real	Creado	Continua
gpgru8	Gastos en transportes y comunicaciones real	Creado	Continua
gpgru9	Gastos en esparcimiento diversión y cultura real	Creado	Continua
gpgru10	Gastos en otros gastos en bienes y servicios real	Creado	Continua
analfa	Analfabetismo	Creado	Categórica
p22	Además de esta vivienda, ¿existe otra vivienda en la que usted o algún miembro	100	Categórica
p101	Tipo de vivienda	100	Categórica
p102	El material predominante en las paredes exteriores es	100	Categórica
p103	El material predominante en los pisos es	100	Categórica
p103a	El material predominante en los techos es	100	Categórica
p104a	Cuantas habitaciones se usan para exclusivamente para dormir	100	Continua
p104b1	La vivienda cuenta con licencia de construcción	100	Categórica
p105a	La vivienda que ocupa su hogar es	100	Categórica
p106a	¿Esta vivienda tiene un título de propiedad?	100	Categórica
p110	el abastecimiento de agua en su hogar procede de:	100	rica
p111a	el baño o servicio higiénico que tiene su hogar está conectado a:	100	Categórica
p1141	su hogar tiene: teléfono (fijo)	100	Categórica
p1142	su hogar tiene: celular	100	Categórica
p1143	su hogar tiene: tv, cable	100	Categórica
p1144	su hogar tiene: internet	100	Categórica

Nombre	Etiqueta	Modulo	Tipo
p1145	su hogar tiene: telefono fijo, celular, tv, cable, internet	100	Categórica
i1172_01	(imputado, deflactado, anualizado) el último gasto mensual por consumo de: agua,	100	Continua
i1172_02	(imputado, deflactado, anualizado) el último gasto mensual por consumo de: elect	100	Continua
i1173_01	(imputado, deflactado, anualizado) el último gasto mensual por consumo de: agua,	100	Continua
i1173_02	(imputado, deflactado, anualizado) el último gasto mensual por consumo de: elect	100	Continua
nbi1_pobre	sí tiene más de una de las necesidades básicas del hogar	100	Categórica
p203	¿cuál es la relación de parentesco con el jefe(a) del hogar?	200	Categórica
p204	¿es miembro del hogar?	200	Categórica
p208a	¿qué edad tiene en años cumplidos? (en años)	200	Continua
p209	¿cuál es su estado civil o conyugal?	200	Categórica
p302	¿sabe leer y escribir? - respuesta espontánea	300	Categórica
p306	este año, ¿está matriculado en algún centro o programa de educación básica o superior?	300	Categórica
p314a	Acceso a internet en el anterior mes	300	Categórica
p401c	Tiene DNI	400	Categórica
p401f	hace 5 años,... ¿vivía en este distrito?	400	Categórica
p401h1	¿tiene ud. limitaciones de forma permanente, para: moverse o caminar, para usar	400	Categórica
p401h2	¿tiene ud. limitaciones de forma permanente, para: ver, ¿aún usando anteojos?	400	Categórica
p401h3	¿tiene ud. limitaciones de forma permanente, para: hablar o comunicarse, aún usa	400	Categórica
p401h4	¿tiene ud. limitaciones de forma permanente, para: oír, ¿aun usando audífonos?	400	Categórica
p401h5	¿tiene ud. limitaciones de forma permanente, para: entender o aprender (concentrarte)?	400	Categórica
p401h6	¿tiene ud. limitaciones de forma permanente, para: relacionarse con los demás?	400	Categórica
p4191	el sistema de prestación de seguro de salud al cual ud. está afiliado actualmente	400	Categórica
p4192	el sistema de prestación de seguro de salud al cual ud. está afiliado actualmente	400	Categórica
p4193	el sistema de prestación de seguro de salud al cual ud. este afiliado actualmente	400	Categórica
p4194	el sistema de prestación de seguro de salud al cual ud. este afiliado actualmente	400	Categórica
p4195	el sistema de prestación de seguro de salud al cual ud. este afiliado actualmente	400	Categórica
p4196	el sistema de prestación de seguro de salud al cual ud. este afiliado actualmente	400	Categórica
p4197	el sistema de prestación de seguro de salud al cual ud. este afiliado actualmente	400	Categórica
p4198	el sistema de prestación de seguro de salud al cual ud. este afiliado actualmente	400	Categórica
p501	la semana pasada, ¿tuvo ud. algún trabajo? (sin contar los quehaceres del hogar)?	500	Categórica

Nombre	Etiqueta	Modulo	Tipo
p506r4	¿a qué se dedica el negocio, organismo o empresa en la que trabajo en su ocupación?	500	Categórica
p510a1	el negocio o empresa donde trabaja, ¿se encuentra registrado en la sunat?	500	Categórica
p511a	bajo qué tipo de contrato	500	Categórica
p512b	en su trabajo, negocio o empresa incluyéndose ud., ¿laboraron…?	500	Categórica
p512a	número de personas	500	Continua
p513t	¿cuántas horas trabajó la semana pasada, en su ocupación principal, al día: total?	500	Continua
p523	en su ocupación principal, ¿a ud. le pagan	500	Categórica
p599	¿es un trabajador con ingreso independiente?	500	Categórica
ocu500	Indicador de la pea	500	Categórica
ocupinf	Situación de informalidad (ocu. principal)	500	Categórica
emplpsec	Empleo informal dentro y fuera del sector informal (ocup. Principal)	500	Categórica
p517	¿ud. se desempeñó en su ocupación secundaria o negocio como:	Creado	Categórica
n_edad_prim	Número de niños en el hogar con edad normativa en primaria	Creado	Continua
n_edad_sec	Número de niños en el hogar con edad normativa en secundaria	Creado	Continua
n_edad_esc	Número de niños en el hogar con edad normativa en el colegio	Creado	Continua
n_matr_prim	Número de niños con edad normativa matriculados primaria	Creado	Continua
n_matr_sec	Número de niños con edad normativa matriculados secundaria	Creado	Continua
n_matr_esc	Número de niños con edad matriculados en el colegio	Creado	Continua

Fuente: ENAHO Panel 2016-2019, ENAHO Panel 2016-2020. INEI. Elaboración propia