Random Forest para identificar los factores sociodemográficos asociados al uso de Internet en el Perú

Random Forest to identify sociodemographic factors associated with the use of Internet in Perú

Jorge Brian Alarcón Flores y María Estela Ponce Aruneri

Facultad de Ciencias Matemáticas, Universidad Nacional Mayor de San Marcos, Lima, Perú.

E-mail: jbrianaf@gmail.com, mepa@unmsm.edu.pe

Recibido el 10 de julio del 2016; revisado el 14 de noviembre del 2016 y aceptado el 15 de diciembre del 2016

DOI: https://doi.org/10.33017/RevECIPeru2016.0008/

Resumen

La sociedad de hoy en día se encuentra viviendo una etapa de constantes cambios, debido en gran medida a la introducción de nuevas tecnologías en el quehacer de la vida cotidiana; es por ello, que diversos líderes mundiales afirman que el uso de las Tecnologías de la Información y Comunicación (TIC) tienen un rol fundamental en el desarrollo de las naciones. Pero es Internet con más de 200 millones de usuarios a nivel global, que ha llegado a posicionarse como una de las TIC con mayor crecimiento tecnológico en los últimos años, llegando incluso a ser considerado como el medio de comunicación más popular en toda la historia de la humanidad. El Perú es un país emergente, que puede encontrar en estas herramientas tecnológicas el camino para convertirse en una sociedad de la información, ayudando a conseguir mejores oportunidades económicas y sociales para todos sus habitantes. Nuestro objetivo en la presente investigación es identificar los factores sociodemográficos asociados al uso de Internet en el Perú. Aplicamos el modelo de minería de datos de clasificación supervisada random forest, a la base de datos de la Encuesta Residencial de Servicios de Telecomunicaciones (ERESTEL) 2014, realizada por el Organismo Superior de Inversión Privada en Telecomunicaciones (OSIPTEL). La muestra fue de 14 626 hogares en los 24 departamentos del Perú, la cual fue aplicada a 42 046 personas de dichos hogares. El modelo propuesto nos permite identificar la edad, nivel educativo, departamento de procedencia y nivel socioeconómico como los factores sociodemográficos prioritarios para el uso de internet en nuestro país. El modelo clasificó correctamente al 83% de las personas. Esperamos que estos resultados contribuyan a la formulación de las políticas sociales y económicas ligadas a la accesibilidad y manejo de tecnologías en nuestro país, particularmente al uso del internet.

Descriptores: *random forest, factores, internet, tecnologías de la información y comunicación.*

Abstract

Society today is experiencing a period of constant change, due largely to the introduction of new technologies in the work of everyday life; It is for this reason that many world leaders say the use of Information Technology and Communication (ICT) play a fundamental role in the development of nations. But it is the Internet with more than 200 million users globally, which has come to position itself as one of ICT more technological growth in recent years, even to be considered as the most popular means of communication throughout history humanity. Peru is an emerging country, which can be found in these technological tools the way to

become an information society, helping to achieve better economic and social opportunities for all its inhabitants. Our goal in this research is to identify sociodemographic factors associated with the use of Internet in Peru. We apply the data mining model supervised classification random forest, to the database of the Residential Telecommunications Services Survey (Erestel) 2014, conducted by the Superior Agency for Private Investment in Telecommunications (Osiptel). The sample consisted of 14,626 households in the 24 departments of Peru, which was applied to 42,046 people in these households. The proposed model allows us to identify the age, level of education, department of origin and socio-economic level as the socio-demographic factors priority for the use of internet in our country. The model correctly classified 83% of people. We hope that these results will contribute to the formulation of economic and social policies related to accessibility and management of technologies in our country, particularly to the use of the internet.

Keywords: *random forest, factors, internet, information and communication technologies*

1. Introducción

En el año 2003, durante la Cumbre Mundial sobre la Sociedad de la Información realizada en Ginebra, los más importantes líderes mundiales declararon: "Somos plenamente conscientes de que las ventajas de la revolución de la tecnología de la información están en la actualidad desigualmente distribuidas entre los países desarrollados y en desarrollo, así como dentro de las sociedades" [1]. Hoy, 12 años después en el Perú, la situación no ha mejorado como se esperaba, si bien se incrementaron los indicadores de accesibilidad del uso de las TIC en nuestro país, aún sigue existiendo grandes brechas entre las personas que utilizan y no utilizan estas tecnologías, según el ranking de la XIV edición del Informe Global de Tecnologías de la Información 2015, Perú ocupa el puesto 90 de 143 economías mundiales [2]. Como el acceso a la información inmediata, principalmente ligado al uso de Internet, puede ofrecer claras ventajas competitivas, tanto en el ámbito económico, profesional, social, entre otros, además de contribuir directamente con el desarrollo de una nación; consideramos importante identificar los factores sociodemográficos que se encuentran asociados al uso de Internet en el Perú.

Para muchos estudiosos de las nuevas tecnologías a nivel mundial, el Internet ha sido una de más grandes innovaciones del hombre, en la historia, como Manuel Castells [3] que menciona que *"Internet es la tecnología decisiva de la era de la información, del mismo modo que el motor eléctrico fue el vector de la transformación tecnológica durante la era industrial"*, concepto que cada vez se aproxima más a lo que vivimos en la actualidad, pues hoy en día Internet ya es mucho más que una herramienta de comunicación, interacción, o "la autopista de la información", como le gustaba decir a algunos expertos en el mundo digital durante la década del '90, sino que hoy en día Internet también se ha convertido en una fuente económica y de desarrollo social muy importante.

Pero, ¿Cómo se traducen estas teorías en términos reales? Muy simple: Internet genera trabajo, permite que las Pymes logren competitividad y profundicen su penetración en el mercado global. En los seis mercados emergentes más importantes del mundo, se estima que un 1.3% de los puestos laborales ya están relacionados con actividades online.

Para identificar los factores sociodemográficos asociados al uso de Internet en el Perú utilizaremos el modelo de Minería de Datos de clasificación supervisada random forest, los resultados proporcionan una fuente de información valiosa y confiable que deberán tomarse en consideración en las políticas sociales y económicas ligadas a la accesibilidad y manejo de tecnologías en nuestro país, como vía de progreso social y desarrollo nacional.

Como aplicación de los métodos de minería de datos, se tiene la investigación denominada "*Uso de internet en Chile: la otra brecha que nos divide*" [4] donde se utiliza el modelo Árboles de Clasificación-Regresión, que permite identificar las variables que explican por qué una persona se conecta a internet en Chile. El modelo de minería de datos random forest, se ha aplicado en diversas áreas de la ciencia y tecnología en diversos países, pero no al uso de las telecomunicaciones. Mostraremos la utilidad de aplicar el modelo, al trabajar simultáneamente con un conjunto de variables, y obtener un bajo porcentaje de error de clasificación.

2. Metodología

La fuente de datos de esta investigación es de tipo secundaria, estos fueron obtenidos a partir de la Encuesta Residencial de Servicios de Telecomunicaciones (ERESTEL) 2014 [5], realizada

por el Organismo Superior de Inversión Privada en Telecomunicaciones (OSIPTEL), la cual fue aplicada a 42 046 personas, ubicadas en una muestra de 14 626 hogares en los 24 departamentos del Perú, tanto en zonas urbanas como rurales. El tipo de muestreo aplicado para la selección de la muestra fue probabilística, multietápica, estratificada, por conglomerados estratificados implícitamente por nivel socio económico y de selección sistemática.

2.1. Random Forest

Propuesto inicialmente por Tin Kam Ho de Laboratorios Bell en 1995, posteriormente es desarrollado por Leo Breiman, quien en 2001 presenta el modelo totalmente desarrollado [6]. Random Forest surge como combinación de las técnicas de Classification And Regression Tree (CART) y Booststrap Aggregating (Bagging) para realizar la combinación de árboles predictores en la que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos. Este algoritmo mejora la precisión en la clasificación mediante la incorporación de aleatoriedad en la construcción de cada clasificador individual [7]. Esta aleatorización puede introducirse en la partición del espacio (construcción del árbol), así como en la muestra de entrenamiento.

2.1.1. Algoritmo [8]

El algoritmo Random Forest puede resumirse:

1. Para b= 1 a B:

   (a) Extraer una muestra bootstrap $\mathbf{Z}^*$ de tamaño N a partir de los datos de entrenamiento.

   (b) Construir un árbol random forest T_b para los datos bootstrap, repitiendo de forma recursiva los siguientes pasos para cada nodo terminal del árbol, hasta que se alcanza el mínimo nodo de tamaño n_{min}.

   i. Seleccione m variables al azar de las p variables.

   ii. Elige la mejor variable / punto de división entre las m variables.

   iii. Dividir el nodo en dos nodos hijas.

2. Salida del conjunto de árboles $\{T_b\}_1^B$

Para hacer una predicción en un nuevo punto x:

$\hat{C}_b(x)$ es la predicción de la clase del b-ésimo árbol random forest.
Entonces $\hat{C}_{rf}^B(x) = voto\ mayoritario\ \{\hat{C}_b(x)\}_1^B$.

2.1.2. Metodología Random Forest

La figura 1 muestra el esquema resumido de la metodología random forest [9] [10].

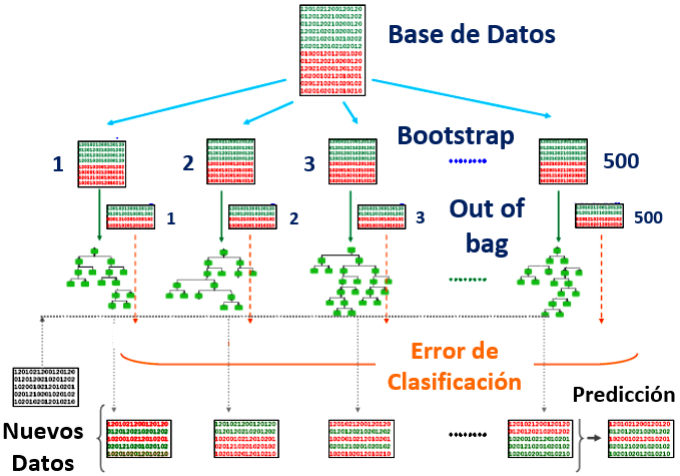

Figura 1: Esquema de la metodología Random Forest.

- No se genera un único árbol, sino un gran número de ellos, éstos se construyen a partir de muchos conjuntos de datos similares generados mediante bootstrap (remuestreo con reposición) de la muestra original, así se consigue corregir el error de predicción debido a la selección específica del conjunto de datos y disponer para cada árbol de una muestra independiente para la estimación del error de clasificación, puesto que aproximadamente un tercio de la muestra original queda excluida de cada muestra generada por bootstrap.

- La aleatoriedad en este modelo es introducida para cada división de un nodo, es decir no se selecciona la mejor variable de entre todas, sino que se selecciona al azar un subconjunto de variables del tamaño especificado y se restringe la selección de la variable a este subconjunto. De esta forma se incluye una mayor variabilidad de árboles y se reduce la dependencia del resultado con las divisiones precedentes.

- Random Forest establece rankings de importancia de las variables en la predicción de la variable respuesta.

Para identificar los factores sociodemográficos asociados al uso de Internet en el Perú, se tendrán en consideración dos medidas de importancia para las variables:

- Mean decrease accuracy (MDA), basada en el aporte de la variable al error de clasificación (porcentaje o número de incorrectamente clasificados). El error de clasificación de cada árbol se calcula a partir de la parte de la muestra que ha quedado excluida de la submuestra utilizada en la construcción del árbol, generada por remuestreo; la diferencia, el número de correctamente clasificados (R_{OOB}). Para calcular la importancia de cada una de las variables que aparecen en un árbol, se permutan aleatoriamente los valores de esa variable, dejando intactos el resto de variables, y se vuelven a clasificar los mismos individuos según el mismo árbol pero ahora con la variable permutada. La importancia en ese árbol se calcula como la diferencia entre el número de clasificaciones correctas antes de la permutación (R_{OOB}) y después de la permutación (R_{perm}) resultante. Finalmente se calcula la medida MDA, como la media de estas diferencias en todos los b-ésimos árboles, $b = 1, \dots, B$ en donde interviene la variable. [11]

$$MDA = \frac{1}{B}\sum_{b=1}^{B}(R_{OOB} - R_{perm})$$

- Mean decrease Gini (MDG), calculada a partir del índice de Gini. Éste es el criterio que se utiliza para seleccionar la variable en cada partición en la construcción de los árboles y que constituye una disminución de esta medida. La importancia de una variable en un árbol se mide como la suma de los decrementos atribuidos a esa variable y la importancia final, como la media en todos los árboles.

$$MDG = 1 - \sum_{i=1}^{C}(p_i)^2$$

Donde p_i es la frecuencia relativa de la clase C en el modelo.

Para la aplicación y obtención de resultados del modelo Random Forest se utilizará el paquete Rattle del software libre R Project for Statistical Computing. [12]

2.2. Evaluación del modelo

2.2.1 Matriz de error

Muestra los resultados de la muestra contra los resultados clasificados por el modelo. Se espera encontrar porcentaje de errores de clasificación similares, en la fase de entrenamiento, convalidación y de prueba para considerar que modelo es eficiente.

2.2.2. Curva ROC

Evalúa la capacidad del modelo para clasificar correctamente. Un valor de 1 significa que el método es perfecto; un valor de 0.5 indica que el método no es útil, y valores intermedios miden la capacidad del método para discriminar. [13]

2.3. Variables

Variable dependiente:

- Uso de internet: Las categorías de esta variable son: Si (usa internet), No (no usa internet).

Variables independientes:

- Edad: Se categorizó en 5 grupos; Generación Z (15-20 años), Millennials (21-34 años), Generación X (35-49 años), Boomers (50-64 años) y la Generación Silenciosa (65 a más años). [14]

- Nivel de estudios: las categorías consideradas en esta variable son, sin estudios, nivel primario, nivel secundario, superior técnico y superior universitario.

- Departamento: las categorías son los 24 departamentos del Perú.

- Nivel socioeconómico: esta variable se categorizó según lo indicado por el Instituto Nacional de Estadística e Informática (INEI) [16], nivel socioeconómico A, B, C, D y E.

- Nivel de pobreza: se consideró la clasificación del Instituto Nacional de

Estadística e Informática (INEI) pobreza extrema, pobreza no extrema y no pobre.

- Estado ocupacional: Si (se encuentra trabajando actualmente), No (no se encuentra trabajando actualmente).
- Lenguaje/Idioma: las categorías consideradas son castellano, quechua, aymara u otras.
- Área sociodemográfica: se consideró rural y urbano.
- Lee y escribe: Si (lee y escribe) y No (no lee y no escribe).
- Género: consideramos masculino y femenino.

3. Resultados

En relación a los entrevistados en esta investigación, podemos indicar que el 76,2% provienen de áreas urbanas dispersas a lo largo del territorio nacional, además en relación a los niveles de pobreza, se encuentra que el 35% de ellos son considerados económicamente pobres, siendo el nivel socioeconómico D, el predominante entre entrevistados con un 37,9%.

El 18,2% de los entrevistados se encuentran ubicados en el departamento de Lima, con un 14,3% proveniente de los diferentes distritos de Lima Metropolitana y el 3,9% del resto de provincias de Lima; seguido por el departamento de Tumbes con un 7,1% y Puno con 5,6%.

El 51,4% son mujeres, el 90,6% hablan castellano y sólo un 25.6% de los encuestados tienen estudios superiores, ya sean universitarios o técnicos; finalmente aún existen peruanos que no saben leer y escribir, ellos son alrededor del 3,2% ubicados en las zonas rurales de nuestro país.

La generación de Millennials (21 a 34 años) y generación X (35 a 49 años) son los grupos de edad con más entrevistados, 26,9% y 25,6% respectivamente.

El 45,1% de los entrevistados hace uso de Internet a nivel nacional.

3.1 Random Forest

3.1.1 Construcción del Modelo

Para la etapa de entrenamiento se trabajó con el 70% de la muestra (29 432). El error obtenido en la etapa de entrenamiento es de 17,39%.

Durante esta etapa de construcción del modelo, se crearon 500 árboles. Al producirse la división en cada nudo del árbol se usaron 10 predictores (variables independientes).

Tabla 1: Matriz de confusión o error en la etapa de entrenamiento

Uso de internet		Modelo		Error
		No	Sí	
Muestra	No	13683	2579	0,1585906
	Sí	2539	10631	0,1927866

Del total de personas que indicaron que no usan internet, 13683 fueron clasificadas de manera correcta por el modelo. En el caso de las personas que indicaron que si usan internet, 10631 fueron clasificadas correctamente.

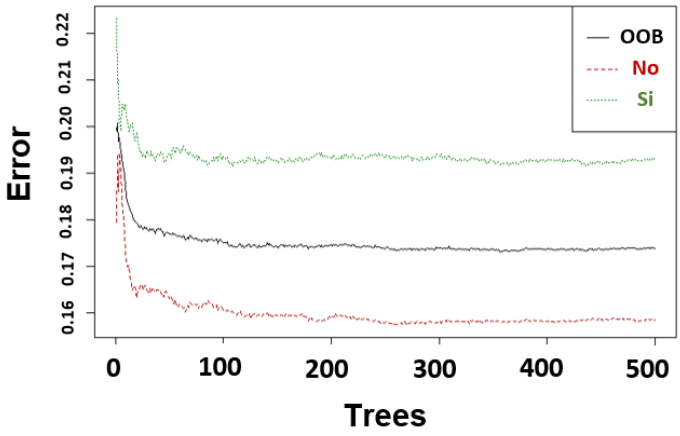

Figura 2 Tasas de Error del Modelo Random Forest en la etapa de entrenamiento

En la Figura 2 se observa que las tasas de error de clasificación del modelo fueron menores para la clasificación de personas que indicaron que no usan internet.

3.1.2 Evaluación del Modelo

Random Forest nos brinda un conjunto de opciones para evaluar el comportamiento del modelo, para esta investigación se desarrollará la validación del modelo mediante la Matriz de Error y la Curva de ROC. [15]

- Matriz de Error:

Tabla 3: Matriz de Error (muestra convalidación)

Uso de internet		Modelo		Error
		No	Sí	
Muestra	No	0,47	0,08	0,15
	Sí	0,09	0,37	0,19

Para esta fase se trabajó con el 15% de los datos (6307).El error de clasificación promedio es del 17%.

Tabla 4: Matriz de Error (muestra prueba)

Uso de internet		Modelo		Error
		No	Sí	
Muestra	No	0,45	0,09	0,16
	Sí	0,09	0,38	0,18

Se utilizó 15% final de los datos (6307). El error de clasificación promedio es del 17%.

- Curva ROC

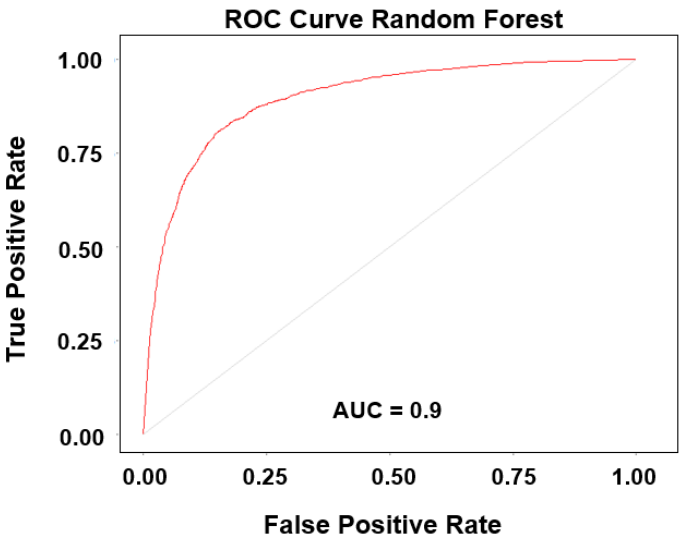

Figura 3: Curva ROC (muestra de prueba)

La Figura 3 la curva ROC proporciona un valor de 0,9, para el modelo random forest utilizado para identificar los factores sociodemográficos asociados al uso de Internet en el Perú.

3.1.3 Importancia de las variables

Nos permite determinar qué variables juegan un papel fundamental en la predicción.

En la Figura 4 se presentan valoraciones de las variables por "importancia", es decir el nivel de influencia de una independiente sobre la variable objetivo. Cuanto mayor sea la magnitud de la valoración, "más importante" será la variable correspondiente.

4. Discusión e interpretaciones

La Tabla 4 muestra que el modelo clasifica correctamente al 83% de los entrevistados.

La muestra de convalidación y prueba (Tabla 3 y 4) muestran 17% de error de clasificación, lo que indica que el modelo es eficiente.

La curva ROC (Figura 3) muestra que el modelo random forest tiene una alta capacidad para clasificar correctamente a la población bajo estudio.

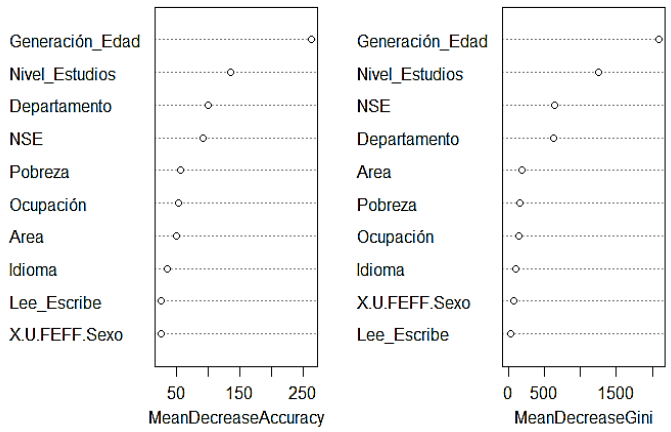

Figura 4: Medidas de Importancia de las variables

Las medidas mean decrease accuracy y de gini, identifican a la edad y nivel de estudios, como los factores más importantes asociados al uso de Internet en el Perú; luego le siguen departamento de procedencia y nivel socioeconómico.

Con el fin de conocer los patrones de usos de Internet en el Perú, se realizó el análisis estadístico

de las variables más importantes obtenidas con el modelo random forest:

- Edad: las generaciones de personas más jóvenes, como la generación Z (De 15 a 20 años) con un 79,3% y 59,3% entre los Millennials (De 21 a 24 años), son las que más utilizan internet.

  En la generación silenciosa (De 65 a más años) el 6,4% de los entrevistados usan internet en Perú.
  Conforme se incrementan las edades de los entrevistados disminuye el porcentaje que utiliza internet en el Perú.

- Nivel de estudios se encuentra claramente asociado al uso de internet en el Perú. Entre las personas que no cuentan con ningún tipo de estudios sólo el 2,1% utiliza internet, mientras que las personas que tienen estudios universitarios el 82,1%.

- Departamento de procedencia: los departamentos con mayor porcentaje de usuarios de internet se encuentran ubicados en la Región Costa del Perú, entre ellos Tacna con 59,8% y Lima (Metropolitana con 57,1% y en provincias 52%), mientras que Loreto y Amazonas son los departamentos con menor cantidad de usuarios de internet, con 26% y 27,3% respectivamente.

- Nivel Socioeconómico: el uso de internet en el Perú es mayor en niveles socioeconómicos altos, el nivel “A” cuenta con un 85,3% de usuarios, el nivel “E” cuenta con el menor porcentaje 28,4 de usuarios de internet en el Perú. Podemos afirmar que a mayor nivel socioeconómico es más probable que un habitante peruano haga uso de internet.

Los resultados obtenidos en esta investigación se encuentran bastante alineados con otras investigaciones, como la de “*Uso de internet en Chile: la otra brecha que nos divide*” donde se encuentran como variables más importantes en el uso de internet, nivel de escolaridad y edad de las personas, similar a lo obtenido en esta investigación, el uso de internet en el Perú se encuentra principalmente ligado a variables relacionadas a la edad y el nivel de educación de las personas.

5. Conclusiones

Los factores sociodemográficos asociados al uso de internet en el Perú, de mayor importancia, son la generación de edad a la que pertenece la persona, junto al máximo nivel educativo alcanzado, además del departamento de procedencia y el nivel sociodemográfico.

Son los más jóvenes, la generación Z (de 15 a 20 años) y los Millennials (de 21 a 24 años) los grupos de edad que más usan internet en el Perú, no es coincidencia, este es el resultado de generaciones que crecieron con la aparición de internet en la humanidad, y que conocen que es una herramienta que usada de manera correcta, puede ser un portal para encontrar mejores oportunidades de desarrollo.

Las autoridades gubernamentales pueden lograr que el internet en el Perú, sea mejor aprovechado por los jóvenes; por ejemplo con capacitaciones virtuales y gratuitas en diversas temáticas a nivel profesional y técnico, que las certificaciones otorgadas signifiquen un valor agregado a sus hojas de vida, además de desarrollar sus capacidades y habilidades, lo que les permitirá tener mayores y mejores oportunidades durante el proceso de inserción en el mercado laboral.

La influencia que tiene el nivel educativo en el uso de internet en el Perú, nos muestra las grandes brechas sociales que existen hoy en día en nuestro país, donde estudiar una carrera de educación superior depende de la disponibilidad económica de las personas. Como muestran nuestros resultados, personas con niveles educativos bajos, prácticamente no usan internet, perdiendo oportunidades e información que podrían tener a su alcance, y que podría significar mejoras sustanciales, tanto a nivel de conocimiento y por qué no hasta económico.

Finalmente, podemos decir que hoy en día internet se ha convertido en un medio que contribuye al desarrollo y progreso de toda sociedad, generando ventajas competitivas en las personas que realizan el uso óptimo de todas las herramientas disponibles en la red y siendo Perú un país en vías de progreso y desarrollo, tiene el gran reto de incrementar el acceso a internet en la población como medio de convertir a la sociedad peruana en una “Sociedad de la Información”, reduciendo las distancias sociales y geográficas, que les permita conocer las ventajas del uso de internet y aprovechen las oportunidades que

ofrece como herramienta de desarrollo personal, social y económico.

Referencias

[1] Cumbre Mundial sobre la Sociedad de la Información. Declaración de Principios. Documentos Finales. (2005) pp 9-12. https://www.itu.int/net/wsis/outcome/booklet-es.pdf

[2] Jhonson Cornell University, World Economic Forum. The Global Information Technology Report 2015. (2015) pp 30. https://www.itu.int/net/wsis/outcome/booklet-es.pdf

[3] M. Castells El impacto de Internet en la Sociedad: Una perspectiva global. (2014) pp 10-13. https://www.bbvaopenmind.com/wpcontent/uploads/2014/03/BBVA-Comunicaci%C3%B3n-Cultura Manuel-Castells-El-impacto-de-internet-en-la sociedad-una-perspectiva-global.pdf

[4] M. Stager, J. Núñez Uso de internet en Chile: la otra brecha que nos divide. (2015) pp 39-41. http://paisdigital.org/wpcontent/uploads/2015/07/Brecha-Digital-Internet-Estudio-Pa%C3%ADs-Digital CASEN.pdf

[5] OSIPTEL. Encuesta Residencial de Servicios de Telecomunicaciones (ERESTEL 2014). (2014) pp 4-14. https://www.osiptel.gob.pe/documentos/erestel-2014

[6] A. Montillo, University of Pennsylvania. Random Forest, Guest Lecture. (2009) pp 20-22. http://www.dabi.temple.edu/~hbling/8590.002/Montillo_RandomForests_4-2-2009.pdf

[7] L. Breiman, Random forests, Machine Learning. (2001) pp 5-32.

[8] Swiss Federal Institute of Technologhy Zurich. Applied Multivariate Statistics-Spring. (2012) pp 3-8. https://stat.ethz.ch/education/semesters/ss2012/ams/slides/v10.2.pdf

[9] K. Hultstrom, Image based Wheel detection using Random Forest Classification. (2013) pp 22-28. http://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=3457767&fileOId=3459875

[10] A. Criminisi, J. Shotton and E. Konukoglu Decision Forests for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning. (2011) pp 25-32. https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/decisionForests_MSR_TR_2011_114.pdf

[11] Z. Jones, F. Linder, Exploratory Data Analysis using Random Forests. (2012) pp 4-11. http://zmjones.com/static/papers/rfss_manuscript.pdf

[12] G. Williams, Data Mining with Rattle and R. The Art of Excavating Data for Knowledge Discover. Springer. USA. (2011) pp 266-289.

[13] G. Biau, Analysis of a Random Forests Model. (2012) pp 2-11. http://www.jmlr.org/papers/volume13/biau12a/biau12a.pdf

[14] Nielsen. Estilos de vidas generacionales. (2015) pp 2-3. https://www.nielsen.com/content/dam/nielsenglobal/latam/docs/reports/2016/EstilosdeVidaGeneracionales.pdf

[15] R. Aler, Evaluación y aprendizaje dependiente de la distribución y el coste. (2014) pp 29-34. http://ocw.uc3m.es/ingenieria-informatica/analisis-de-datos/transparencias/2%20ROC.pdf

[16] INEI. Características socioeconómicas de los hogares. (2007) pp 13-19. https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib0744/Libro.pdf