PREDICCIÓN DEL SUBEMPLEO EN LIMA Y CALLAO

Luis Angel Pari Tesillo, Teófilo Chambilla Aquino

1. Introducción

En Lima Metropolitana existen 7 millones personas que tienen la edad para desempeñar una actividad económica (PET). De este total, el 69 % (4 millones) integran la Población Económicamente Activa (PEA) y el restante 30.9 % (2 millones) la Población Económicamente Inactiva (PEI), que agrupa a las personas que no participan en la actividad económica ni como ocupados ni desocupados[2].

Sin embargo, cuando el crecimiento del empleo es insuficiente, la falta de empleo no siempre se manifiesta a través del alto desempleo, sino en la caída de las tasa de participación laboral y en el mantenimiento de muchos empleos de baja calidad (Trabajos con más 8 horas e ingresos menores a la canasta básica de alimentos), lo que en adelante denominaremos como subempleado. Estos datos debieran encender una luz de precaución, aun cuando se confié en que el crecimiento económico durará, porque sugieren que las restricciones para salir buscar y conseguir un empleo están resurgiendo, en particular en Lima y Callao.

Esto lleva al estudio más certero acerca de la idiosincrasia de los individuos. Con cuyos datos muéstrelas se trabajará aplicándoles técnicas de minería de datos (Clustering, árboles de decisión, etc), para descubrir patrones de información ocultas en los datos de la Encuesta Permanente de Empleo.

2. Motivación

La motivación que nos llevó a elegir este tema, primero es que en Perú nunca se aplicó el proceso de minería de datos a la Encuesta Permanente de Empleo, solo se han realizados trabajos estadísticos. Por otro lado, el análisis inteligente de los datos, abre un mundo de posibilidades para muchas industrias, que muchas veces no es explotado. Por medio del presente trabajo queremos aportar con nuestro pequeño grano de arena, descubriendo patrones que pueda ayudar a tomar decisiones a los entes competentes, empresas y el público en general.

3. Definición del Problema

La falta de un modelo predictivo no permite identificar patrones que conllevan a la generación del subempleo y el aumento de trabajo de baja calidad, en la Encuesta Permanente del Empleo EPE.

4. Hipótesis

Se puede descubrir patrones de información oculta sobre subempleo en los datos de la Encuesta Permanente de Empleo.

5. Objetivo General

El objetivo principal del proyecto a desarrollar es caracterizar y descubrir la situación del empleo en Lima y Callao a través de la aplicación de técnicas Minería de Datos.

5.1. Objetivos Específicos

1. Describir la composición del empleo en Lima y Callo.
2. Etiquetar a los empleados, a partir de sus ingresos y horas trabajadas en subempleado o no subempleado.
3. Determinar cuál es la probabilidad de que una persona sea subempleado a partir de sus atributos.

6. Descripción del dataset de la Encuesta permanente de Empleado

La Encuesta Permanente de Empleo, es una investigación estadística continua que genera indicadores mensuales (por trimestre móvil), que permiten conocer la evolución del empleo e ingreso de los hogares de Lima Metropolitana. La finalidad de esta encuesta es obtener información base para estimar los principales indicadores del mercado laboral.

Características generales de dataset:

País:	Perú
Cobertura:	Provincia De Lima y La Provincia Constitucional Del Callao
Colección:	Encuesta de hogares.
Tipo De Estudio:	Encuesta de empleo.
Años:	2001 a 2015
Productor:	Instituto Nacional de Estadística e Informática del Perú.
Nro. de Archivos:	178 en formato .sav hasta Marzo del 2015
Nro. de Registros:	3 204 000 hasta Marzo del 2015.
Total de Variables:	92
Web Site:	http://iinei.inei.gob.pe/microdatos

Características de las variables

Las variables se pueden agrupar en Empleo e Ingreso. El grupo Empleo se puede agrupar en los siguientes subgrupos: *Ocupado* (Agrupa a las variables que cuantifican la información referente a las ocupaciones de los encuestados en el periodo de estudio), *Desocupado* (Agrupa a las variables que contienen información sobre desocupación, si buscó o no trabajo y qué está haciendo para buscar trabajo, etc), *Dependiente* (Agrupa a las variables relacionadas a las personas que tienen una ocupación dependiente. Presenta información de ingreso, tipo de actividad, etc) e *independiente* (Agrupa a las variables relacionadas a las personas que tienen una ocupación independiente. Presenta información de ingreso, tipo de actividad, etc). El grupo Ingreso se puede agrupar en los siguientes subgrupos: *Monetario* (Agrupa a las variables con las que obtenemos información de los ingresos monetarios que obtienen las personas) y *En especie* (Agrupa a las variables que presenta información de los pagos en especie que reciben las personas).

Descripción del Desarrollo realizado y por realizar:

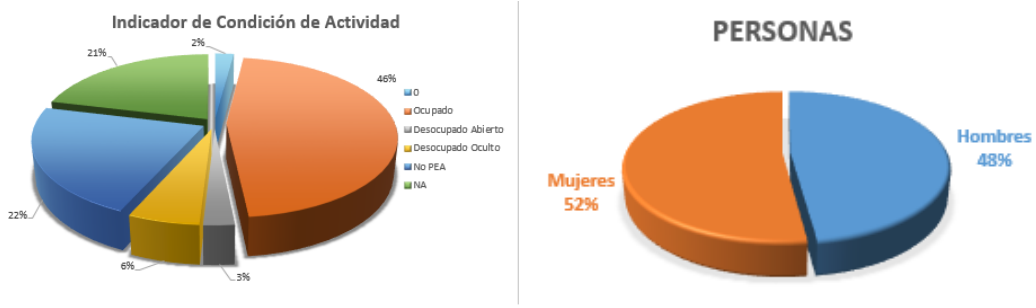
1. Desde la web site de la INEI, se ha descargado los 178 archivo en formato .sav, acumulando 512 MB.

2. Se ha descargado IBM SPSS Statistics 20 trial para poder visualizar los archivo .sav, donde se ha identificado que existe gran cantidad valores nulos (NA), el campo esta rellenado con un guion(-) o punto(.).

   Como los datos son semestrales, entonces los archivos vienen empaquetados por tres meses, pero sucede que en la mayoría de los paquetes se repites los datos. Por ejemplo, los archivos Trim-Ene-Feb-Mar14.sav y Trim-Feb-Mar-Abr14.sav, en el segundo archivo se repiten datos de los meses de Enero y febrero, para estos se ha revisado los datos de forma detallada reduciendo el número de archivos a 57 de los 178 que se disponía inicialmente.

3. Dentro de los 92 atributos que dispone la dataset. Se agrupó por categorías tal como:
   - Nivel Académico: Nivel educativo, años de estudio, grado de estudio.
   - Ingreso: Número total de ingreso por trabajos independientes y dependientes.
   - Socio Demográficos: horas trabajas durante la semana (por días), tipo de ocupación (Empleador, trabajador independiente, empleado).

4. Sea realizado una exploración inicial de los diferentes archivos. Los gráficos mostrados corresponden al semestre Octubre, Noviembre y Diciembre del 2014. La Figura 1 (a), nos indica que la clase más representadas son los ocupados, clase que sera parte de nuestro estudio a realizar y el 23 % son los NA por lo que claramente tenemos trabajo que realizar con respecto a la limpieza de datos. La Figura 1(b), nos indica que la clase más representada son las mujeres.



(a) Indicador de Condición de Actividad (b) Muestra representativa de instancias

Figura 1: Muestras

5. Se realizará la limpieza datos dando especial atención a los valores NULL, guion y puntos.

6. Procederemos a juntar los 57 archivos en una sola base de datos y su respectiva limpieza de datos, para posteriormente realizar experimentos con RStudio y Weka, aplicando las técnicas de minería de datos aprendidas en el curso.

7. Se describirá la composición del empleo en lima y callao mediante los siguientes atributos: ¿Trabajando en algún negocio propio o de un familiar? (P2041), ¿Ofreciendo algún servicio? (P2042), ¿Haciendo algo en casa para vender?(P2043), ¿Vendiendo productos de belleza, ropa, joyas,etc.? (P2044), ¿Haciendo prácticas pagadas en un centro de trabajo?(P2045), ¿Trabajando para un hogar particular? (P2046), ¿Fabricando algún producto? (P2047), ¿Realizando labores en la chacra o cuidado de animales? (P2048), ¿Ayudando a un familiar sin remuneración?

(P2049) ¿Otra? (P20410), ¿Ud. se desempeñó en su ocupación principal o negocio cómo? (P206). Para este punto se utilizara Rstudio, para realizar un análisis de las variables que nos permita visualizar la composición del empleo en Lima y Callao.

8. Se etiquetará a los empleados en función a sus ingresos (Menores a la canasta básica de alimentos) y horas trabajadas (Mayores a 8 horas diarias) en SubEmpleado o no subempleado. Por consecuencia se creará un nuevo atributo que contendrá estas etiquetas, este atributo será nuestra clase para nuestro estudio.

7. Trabajos Relacionados

Denis Lizazo Torres [3] presenta un trabajo en base a la información generada por la Encuesta Permanente de Hogares del año 2009, del Observatorio Social de la Universidad del Litoral en Santa Fe, Argentina. Este trabajo nos muestra que utilizando los algoritmos de clasificación se puede identificar la distribución de número de personas por familia, utilizando como atributo clase tipo de familia (Familias tipo, familia Unipersonales, etc.).

Por otro lado, Luis Alfonso Cutro [1] presenta un trabajo sobre extracción de patrones socio demográficos, educativos y de ingresos de la provincia de Corrientes, Argentina. Basándose en la información de la Encuesta Permanente de Hogares (EPH) del primer trimestre 2003 al primer trimestre 2007. El estudio se realizó con las herramientas: Manejo de base de datos, MultiPlataforma IBM DB2 UDB; Elaboración del Almacén de Datos(Data Warehouse), suite IMB DB2 UDB WorkGroup Server 8.1.0; Análisis de datos, IBM DB2 Intelligent Miner for Data. El enfoque de este trabajo fue predecir el ingreso de los jefes y jefas de hogar que cuenten con seguro social público.

Creemos que nuestro trabajo estará basado en el trabajo de Luis Alfonso Cutro [1] pero con un enfoque diferente, debido a que realizaremos un análisis más profundo de la estructura y características del empleo, ademas crearemos un modelo predictivo para determinar el subempleo en Lima y Callao.

Realizando una comparación entre los dataset EPH de Argentina y PEP de Perú ambos están orientados a obtener indicadores sobre empleo e ingreso para el seguimiento y análisis del mercado laboral. Pero, son completamente diferentes, tanto en la estructuración de las preguntas, numero de atributos, numero de instancias.

Referencias

[1] Cutro, A. Minería de datos aplicada a la encuesta permanente de hogares. *Trabajo Final de Aplicación de la Licenciatura en Sistemas de Información dirigido por el Prof. David Luis la Red Martínez. Corrientes. Argentina* (2008).

[2] del Concejo de Ministros, P. Situacion del mercado laboral en lima metropolitana, 2013. `http://www.pcm.gob.pe/wp-content/uploads/2014/01/02-Empleo-Oct-Nov-Dic-2013.pdf`.

[3] Torres, D. L., Meyer, R. D., and Cárdenas, V. T. Minería de datos en la encuesta permanente de hogares 2009, universidad nacional del litoral, argentina. *Revista Ingeniería Industrial*, 1 (2011), 19–28.