Predicción de la pobreza en Argentina

Bogliolo, Nicolás Alejandro; Chas, Santiago Martín & Mariño, Matías

Abstract
En el siguiente trabajo utilizamos distintos modelos de clasificación aplicados a la Encuesta Permanente de Hogares (EPH), con el fin de predecir la pobreza de un individuo a través de predictores no monetarios. El mejor resultado obtenido fue utilizando el modelo de Logistic Regression, logrando un accuracy del 82%.

Keywords
Logistic Regression – KNN – SVM – Pobreza

1 INTRODUCCIÓN

El presente trabajo fue realizado en el mes de octubre del 2019 sobre los datos del primer trimestre del mismo año. La Encuesta Permanente de Hogares es realizada por el INDEC, cuyo propósito es el relevamiento sistemático y permanente de los datos referidos a las características demográficas y socioeconómicas fundamentales de la población, vinculadas a la fuerza de trabajo. Su temática está orientada hacia la caracterización de la situación social integral de los individuos y los hogares, aunque los datos más difundidos son los relacionados con el mercado laboral.

En este sentido la Encuesta Permanente de Hogares Continua pretende conocer y caracterizar la situación de las personas y de los hogares -por ser éstos los núcleos básicos de convivencia en donde los individuos se asocian- según su lugar en la estructura social.

Actualmente en Argentina, la pobreza se mide únicamente por los Ingresos que se posee cuando en realidad es un factor multidimensional en el cual afectan distintas variables como la edad, ocupación, familia, acceso a determinados bienes, etc.

2 DESCRIPCIÓN DEL DATASET

Se decidió realizar el trabajo sobre la encuesta individual.

El dataset seleccionado está formado por 178 features y 59369 samples, de los cuales cada feature representa las preguntas realizadas a los individuos y cada sample representa un individuo.

Del dataset mencionado, se optó por conservar las siguientes features: Sexo, Edad, Estado Civil, Tipo de Cobertura, Alfabetismo, Estudiante, Nivel Educativo, Nacionalidad, Condición Laboral, Categoría Ocupacional, Categoría de Inactividad, Monto de Ingreso Individual, Monto de Ingreso Familiar, Monto de Ingreso Per Cápita, Habitantes por Hogar. Esto se debidio a que estas variables logran capotar diferentes dimensiones socioeconómicas de los individuos encuestados

Sobre estas features se realizó el reemplazo de los valores numéricos de las mediciones por su categoría correspondiente de acuerdo a la bibliografía correspondiente a la encuesta. Además de esto se realizó una recategorización sobre algunas de las features según criterio adoptado por los Autores.

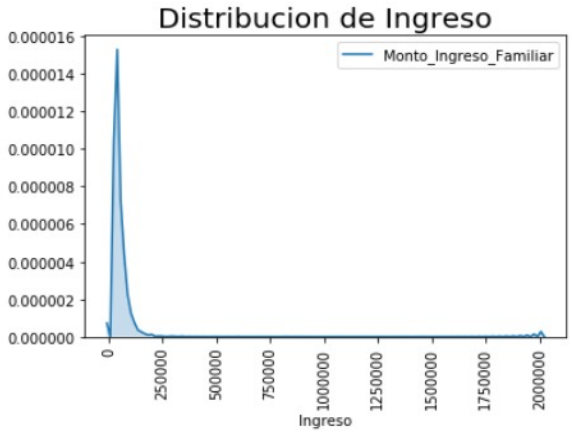
Luego de realizado lo anterior, se crearon nuevas features como: Banda de edad, Clase Social y Pobreza. Esta última hace referencia a si la persona es considerada pobre según los criterios adoptados por el INDEC que contemplan el Ingreso Familiar per Cápita y la Región de residencia de la persona.

Finalmente, se realizó una limpieza que consistió en eliminar los samples con valores nulos y con Monto Ingreso Familiar 0, obteniendo un Dataset de 46501 samples y 21 features.

3 ANÁLISIS EXPLORATORIO DE DATOS

3.1 Distribución del Ingreso Familiar

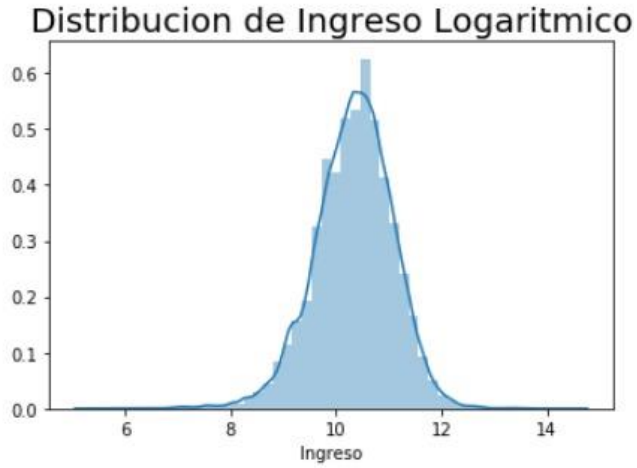
Se decidió realizar un gráfico con el fin de observar el comportamiento de la distribución de los ingresos individuales entre los encuestados.



Este gráfico muestra que la distribución se encuentra inclinada hacia la izquierda, la mayor densidad de individuos posee un ingreso inferior a los $25.000 mensuales. También se observa que superado el monto de ingresos mencionado, hay una alta dispersión de ingresos hasta llegar a un máximo de $2.000.000 con una pequeña concentración cerca de este valor.

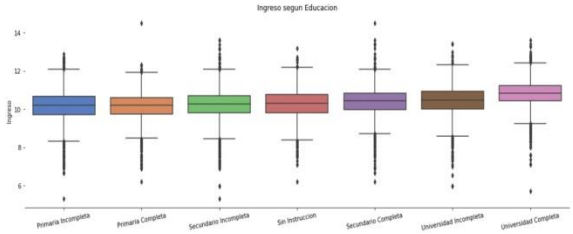
Debido a esto, se optó por aplicar una función logarítmica a los montos de ingreso individuales obteniendo una distribución que se asemeja a una normal.

Representado por el siguiente gráfico:



3.2 Nivel de ingresos según nivel educativo

Otra realización que consideramos pertinente analizar fue la existente entre el nivel de ingresos y el nivel educativo del individuo (utilizando la función logarítmica aplicada a los ingresos).

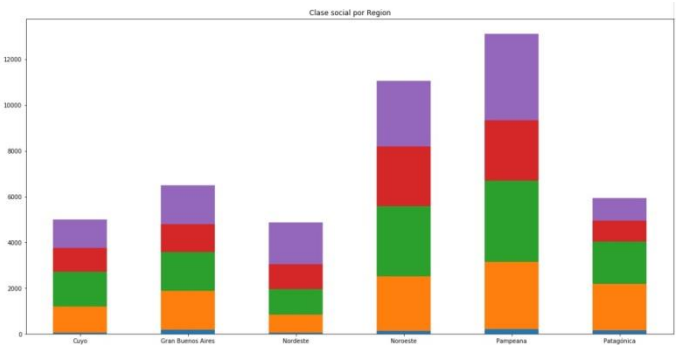


El gráfico está ordenado de acuerdo a la mediana de los ingresos de las distintas categorías.

Es importante destacar que a diferencia de lo esperado aquellas personas que no poseen instrucción alguna están justo en la mitad. Además se puede ver un gran salto en el ingreso de las personas con Universidad Completa.

3.3 Clase social de acuerdo a la región

Luego lo que se buscó fue mostrar la cantidad de personas que hay en cada clase social, discriminadas por región.



La mayor concentración de gente ABC1 se encuentra en el Gran Buenos Aires, mientras que la mayor cantidad de gente D2/E se ubica en la región pampeana. Este gráfico también observar la distribución de personas encuestadas según la región.

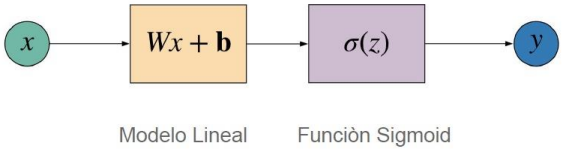
4 MATERIALES Y MÉTODOS

Con el fin de clasificar a la persona en "Pobre/No Pobre" según los predictores no monetarios ya mencionados, se utilizaron los siguientes modelos:

Logistic Regression; KNN y SVM.

4.1 Logistic Regression

Es un clasificador lineal compuesta por una regression lineal, precedida de una función activación sigmoide, por lo cual el output es binario y no continuo. A cada muestra clasificada le asigna una probabilidad de pertenecer a cada clase existente en el problema. Si esta es mayor a cierto treshold (0,5) entonces pertenece a esta clase, en caso contrario viceversa.

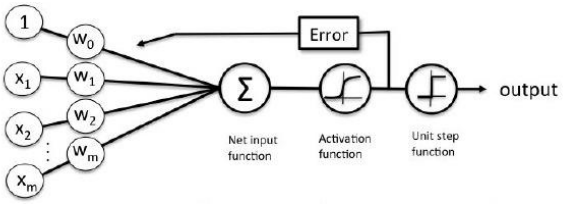


El regresor logístico aprende de un parámetro interno por cada dimensión del vector de entrada (vector W). Calcula el gradiente del error de clasificación y trata de minimizarlo. La probabilidad de la clase Yi viene dada por la siguiente función

$$P(y_i|x) = \sigma(w^t x)$$

La función sigmoide

$$\sigma = \frac{1}{1 + \exp(-x)}$$



Lo interesante de este modelo es que resulta útil para capturar relaciones lineales en los datos, si es que existen.

4.2 KNN

Es un modelo de clasificación en el cual un nuevo dato es agrupado según K vecinos más cerca de uno que del otro. Para esto se calcula la distancia de un elemento nuevo a los existentes y se ordenan para seleccionar a que grupo pertenecen. Uno de los hiperparametros del modelo es determinar la cantidad de K vecinos.

$$d_E(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

La ventaja del modelo k vecinos más cercanos es que al ser un método no paramétrico, se nutre de la existencia de no-linealidades en los datos, a diferencia del modelo de regresión logística.

4.3 SVM: Super Vector Machine

Se trata de un clasificador lineal, el cual busca un hiperplano que máxima el margen entre las clases. En el caso de que las clases no sean linealmente separables se acude al Soft Margin, un penalizador de muestras mal clasificadas, las cuales se penalizan con un Costo seleccionado por el usuario.

Este modelo calcula el mejor hiperplano dentro de las opciones posibles, lidiando con clases superpuestas mediante el Soft Margin ya mencionado

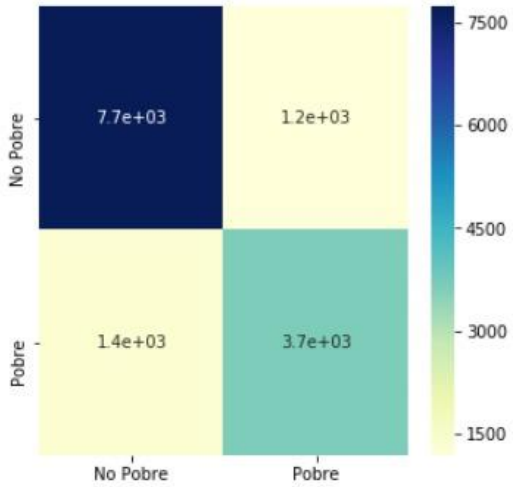
5 RESULTADOS

Los resultados obtenidos utilizando Logistic Regression fueron los siguientes, siendo este el modelo con mayor perfomance sobre los datos:

- Accuracy: 81,69%

$$Accuracy = \frac{True\ Positive + False\ Negative}{Total}$$

Confusion Matrix:



Esta matriz en el primer cuadrante (arriba a la dercha) tiene las muestras que resultaron verdaderos negativos; en el segundo los falsos positivos; en el tercero los falsos negativos y por último los verdaderos positivos.

Utilizando el modelo KNN, habiendo fijado la cantidad vecinos K = 5, se obtuvo un accuracy del 79,44%.

Finalmente se corrio un SVM sobre el Dataset. Luego de haber porbado con diferentes hiperparametros, el mayor Accuracy se obtuvo con un Kernel Lineal; Gamma automatico y C=1. Se obtuvo un accuracy del 81,64%.

Para la seleccione de los Hiperparametros del modelo, se realizaron distntas pruebas cambiando los valores correspondientes llegando a los mencionados como los optimos. No se pudo realizar un GridSearch debido a limitaciones computacionales

6 DISCUCIÓN YCONCLUSIONES

Se llego a la conclusión de que el modelo que mejor performa en el dataset trabajado fue Logistic Regression. Como se menciono se predijo con un 82% de accuracy si la persona en cuestión es pobre o no mediante las siguientes variables: Región, Sexo, Estado Civil, Alfabetismo, Nacionalidad, Educación, Estado Ocupacional, Banda de Edad y si actualmente es un estudiante.

En este sentido, los resultados sugieren que las caracterísitas demográficas, socioeconómicas y su estado ocupacional de los individuos son muy relevantes para explicar la situación de pobreza.

Para futuros análisis se podria evaliar la incorporacion de más variables existentes en la Encuesta Permanente de Hogares como las relacionadas a la vivienda, el tipo de empleo y los aglomerados de viviendas, entre otros.

Ademas, es altamente recomendable realizar un SVM con Gridsearch y CrossValidation que debido a limitaciones computacionales no se pudo realizar.

7 REFERENCIAS

1. Predicción de la pobreza en la Argentina usando Random Forest (https://aaep.org.ar/anales/works/works2016/cardinale.pdf)

2. Bibliografía de la Encuesta Permanente de Hogares (https://www.indec.gob.ar/ftp/cuadros/menusuperior/eph/EPH_nota_metodologica_1_trim_2019.pdf)

3. Informe de prensa del Indec: Incidencia de la pobreza y la indigencia (https://www.indec.gob.ar/uploads/informesdeprensa/eph_pobreza_01_19422F5FC20A.pdf)

4. Apuntes de Cluster AI.