Clasificación de Actividad de la Población con discapacidad en Argentina

AQUINO Irupé – VACCARELLO Fabrizio – VALENCIA Joel

Universidad Tecnológica Nacional – Facultad Regional de Buenos Aires

Ciencia de Datos - 2020

Abstract

El objetivo de este proyecto es identificar cuales son las variables, o combinación de ellas, que mejor explican la 'Condición de Actividad' de la Población con discapacidades de Argentina y crear un modelo de aprendizaje supervisado que posibilite la clasificación de los mismos según su condición.

1. INTRODUCCIÓN

Este trabajo fue realizado en el mes de octubre de 2020 sobre los datos de una serie de encuestas realizadas a través de la Dirección Nacional de Estadísticas Sociales y de Población y de la Dirección de Estadísticas Poblacionales en las localidades urbanas de 5.000 y más habitantes de todo el territorio nacional, las entrevistas fueron directas a través de dispositivos digitales. Dicha base permite obtener estimaciones sobre la población con algún tipo de discapacidades a nivel nacional y tiene como finalidad servir a todas aquellas personas, instituciones gubernamentales, empresas o centros de investigación que requieran información sobre la temática abordada.

2. OBJETIVO

Objetivo general: cuantificar a la población con dificultades para ver; oír; caminar o subir escaleras; agarrar y levantar objetos con los brazos o manos; atenderse por sí misma su cuidado personal, por ejemplo, para bañarse, vestirse o comer solo/a; hablar o comunicarse; entender lo que se le dice; aprender cosas; recordar o concentrarse; controlar su comportamiento; y, solo para la población de 5 a 12 años, jugar con otros niños de su edad.

Objetivo específico: en el siguiente trabajo utilizamos distintos modelos de clasificación con datos obtenidos del INDEC [1] con el fin de clasificar a las personas con falta de capacidad según su 'Condición de Actividad'.

[1] https://www.indec.gob.ar/indec/web/Institucional-Indec-BasesDeDatos-7

2. DESCRIPCIÓN DEL DATASET

Se utilizó para la realización de este trabajo el dataset otorgado por la página del INDEC[1].

Se encuentra conformado por 82.327 samples y 35 features, de los cuales las features estná conformados por: ID de cada encuesta, ponderador, cantidad de personas en el hogar, cantidad de hogares con alguna persona con discapacidad, cantidad de personas discapacitada por hogar, tipo de hogar, número de orden, relación de parentesco con la persona que presenta discapacidad, sexo, edad, edad agrupada, imputado b, imputado c, imputado d, personas de 6 años y más con dificultad, Personas con dificultad considerando toda la población, Cantidad de dificultades, Cantidad y tipo de dificultad, Tenencia de certificado de discapacidad, Comienzo de la dificultad, Edad de inicio de la dificultad, Causa de la dificultad, Cobertura de salud, Recibe jubilación o pensión, Tipo de beneficio que recibe, Sabe leer y escribir, Asiste a establecimiento educativo, Modalidad educativa de asistencia actual y pasada, Máximo nivel educativo alcanzado, Principal motivo de no cursar actualmente/nunca cursó, Convive en pareja, Estado civil legal, Motivo principal por el que no busco trabajo, Condición de actividad, Categoría ocupacional.

Este Dataset inicial contaba con muchas variables que consideramos no eran necesarias para explicar la distribución de 'Condición de Actividad' de la población con discapacidad, ya que no contaban con información relevante para realizar la clasificación (eran datos del hogar de cada persona o datos obtenidos una vez realizada la clasificación de actividad). Por lo cual se decidió realizar un filtrado del dataset y quedarse con las variables consideradas relevantes.

Siendo las componentes principales: Cantidad de personas en el hogar, Cantidad de personas con dificultad en el hogar, tipo de hogar, Grupos de edad, Cantidad de dificultades, Cantidad y tipo de dificultad, Edad de inicio de la dificultad, Cobertura de salud, Tipo de beneficio que recibe, Sabe leer y escribir, Causa de la dificultad, Máximo nivel educativo alcanzado, Motivo principal por el que no busco trabajo, Condición de actividad.

3. ANÁLISIS EXPLORATORIO DE DATOS

Para lograr analizar los datos a estudiar y convertirlos en información, el primer paso que se realizó fue la importación de los datos, es decir que cargar los datos del Dataset. El formato original se encontraba en CSV (Comma Separated Values, Valores Separados por Comas).

Una vez incorporados los datos, se importaron las librerías necesarias ([2]pandas,[3]numpy, matplotlib.pyplot,[4]seaborn, warnings, warnings.filterwarnings) para llevar adelante el Análisis Exploratorio de Datos (EDA: Exploratory Data Analysis).

[2] https://pandas.pydata.org/ ; [3] https://numpy.org/ ; [4] https://seaborn.pydata.org/

Cuando se tiene este de tipo de conjuntos de datos de alta dimensión y se utilizan todos para la creación de modelos de ML puede ocasionar:

- Ruido para el cual el modelo de ML puede tener un rendimiento extremadamente bajo, el modelo tarda más tiempo en entrenarse.
- Asignación de recursos innecesarios para estas características.

Por todo esto, se debe implementar la selección de características, con el objetivo de mejorar el rendimiento de predicción de los predictors.

Filtramos nuestro dataset por los criterios más importantes detallados en "DESCRIPCIÓN DEL DATASET" y luego:

- Filtramos el Dataset para las personas con dificultades (discapacidad) que son el grupo que vamos a estudiar.
- Filtramos el Dataset por Edad y dejamos los grupos que tienen personas en edad Laboral "14 a 39 años" y "40 a 64 años", quedando un dataset de 4389 samples y 13 features.
- Eliminamos los valores Nans en la columna 'Condición de actividad'.
- Rellenamos la columna de "máximo nivel educativo alcanzado" con "no específica nivel".
- Rellenamos la columna de "edad de inicio" con la moda.

3.1 Nivel Educativo Alcanzado.

Consideramos, para este estudio, que el nivel educación alcanzado, es una condición que afecta a corto o largo plazo la actividad de ocupación de una persona. Pudimos obtener, de los entrevistados, que la mayor cantidad de personas inactivas, estaban correlacionados con un Máximo Nivel de Educación de "Hasta la primaria completa" únicamente (hay que considerar que esta observación no es una condición de causalidad).

	Máximo nivel educativo alcanzado	Condición de actividad	Cantidades
0	Educación integral	Desocupado	3
1	Educación integral	Inactivo	65
2	Educación integral	Ocupado	22
3	Hasta primario completo	Desocupado	76
4	Hasta primario completo	Inactivo	814
5	Hasta primario completo	Ocupado	627
6	No especifica nivel/ignorado	Desocupado	9
7	No especifica nivel/ignorado	Inactivo	315
8	No especifica nivel/ignorado	Ocupado	84
9	Secundario completo	Desocupado	50
10	Secundario completo	Inactivo	289
11	Secundario completo	Ocupado	373
12	Secundario incompleto	Desocupado	59
13	Secundario incompleto	Inactivo	500
14	Secundario incompleto	Ocupado	444
15	Superior no universitario, universitario y pos...	Desocupado	42
16	Superior no universitario, universitario y pos...	Inactivo	229
17	Superior no universitario, universitario y pos...	Ocupado	388

Figura 1: Cantidad de personas según la correlación entre "Máximo nivel educativo alcanzado" y su "Condición de Actividad"

Para este primer análisis, hemos realizado una correlación entre las variantes, obteniendo lo siguiente:

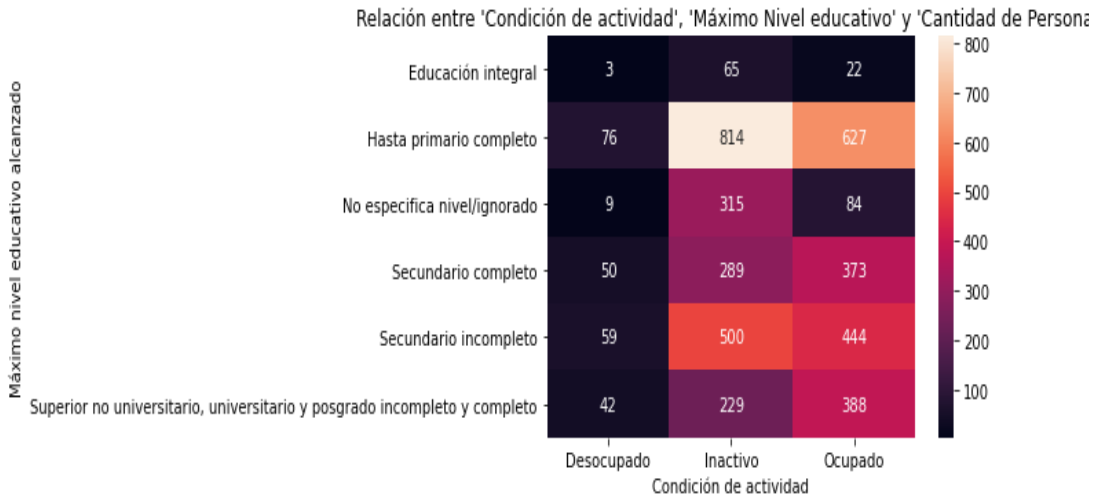

Figura 2: Heatmap de Relación entre "Condición de Actividad, Máximo nivel educativo y Cantidad de Personas"

Pudimos deducir de ambos cuadros realizados, que en general para todas las categorías hay una gran cantidad de 'Inactivos', excepto para la condición 'Superior no universitario, universitario y posgrado incompleto y completo' y 'Secundario Completo' donde es mayor la cantidad de gente 'Ocupada'.Por lo cual interpretamos que cuanto mayor sea el nivel alcanzado por la persona educativamente, hay razonablemente una mayor probabilidad de que esa persona ocupe la condición de "Ocupada".

Es notable la baja cantidad de muestras con la que cuenta la condición de "Desocupado", esto es debida a que el dataset ya tenía esta falencia. El motivo para esto podría haber sido que una errónea realización de la encuesta, o bien que la población en su mayoría decante en algunas de las otras dos clasificaciones.

Otro dato notado fue que la mayoría de la población con discapacidad del país tiene una formación educativa inconclusa.

Para poder sacar más conclusiones al respecto, hemos analizado cómo se encontraba segmentado nuestras muestras.

3.2 Sexo De La Persona.

El sexo de una persona, no debe ser un determinante a la hora de obtener o no un puesto laboral, y muchos menos con o sin falta de capacidad en la persona. Sin embargo, podemos observar a nivel macro, que la cantidad de mujeres 'Inactivas' supera a las 'Ocupadas', y que a la vez, las 'Inactivas' supera a los hombres.

En los hombres se puede notar justo lo contrario siendo mayor la concentración de hombres en la condición 'Ocupado', siendo esta mayor a las otras dos categorías juntas. Esto podría estar relacionado con la condición de muchos hombres del país de ser "Padres de Familia" siendo estos el único sustento/ingreso familiar.

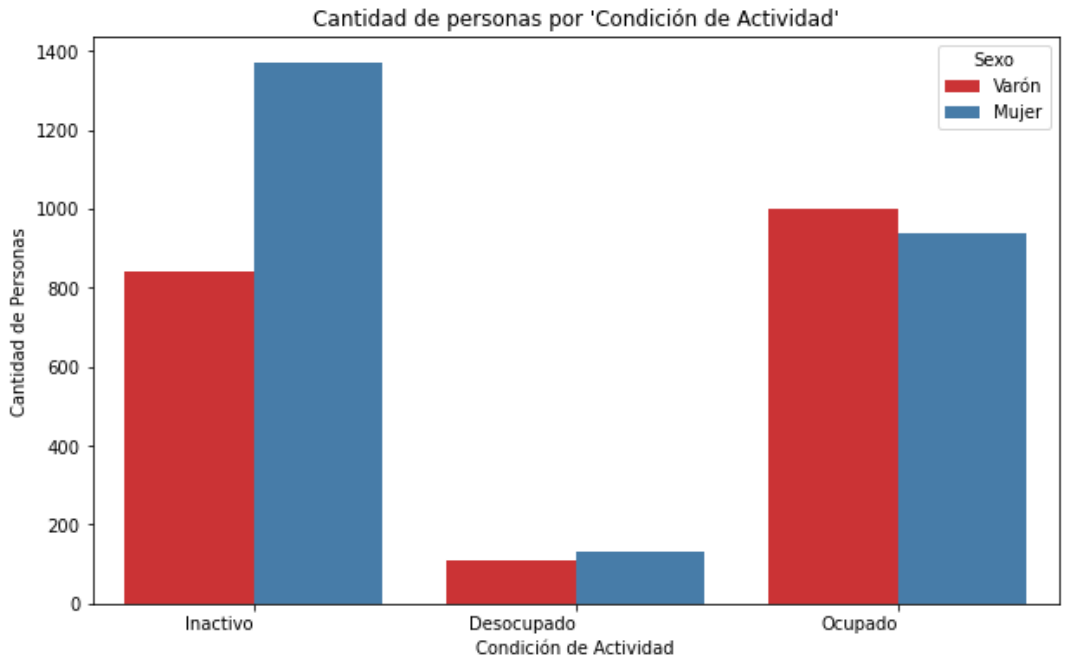

Figura 3

3.3 Cantidades De Discapacidades.

Para analizar más este comportamiento, entre nuestras muestras, analizamos la cantidad de dificultades que presentan las personas discapacitadas. Como se esperaba la gran mayoría de las población tiene únicamente una sola dificultad. Sin embargo, al aumentar la cantidad de dificultades luego de "2 dificultades" no decae la cantidad poblacional, como se puede ver hay un aumento entre las cantidades de 4 y 3 dificultades. Además se encuentra mayor cantidad de hombres en cada categoría que de mujeres.

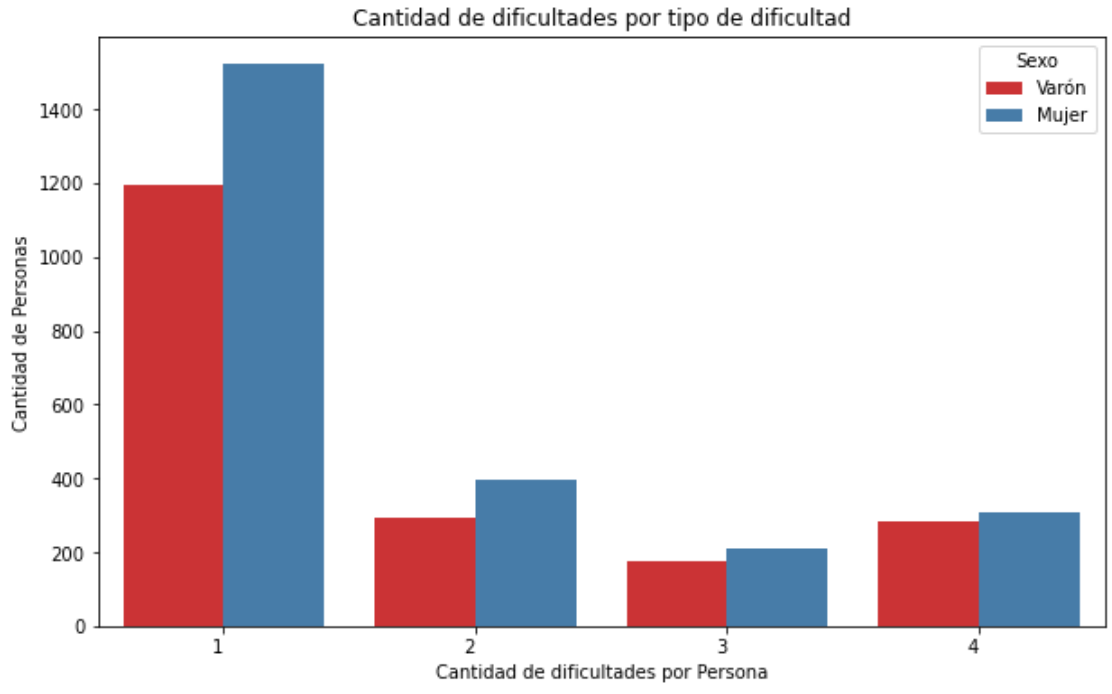

Figura 4

3.4 Distribución de la Cantidad de Dificultades

En el gráfico se puede observar que la distribución de cantidad de dificultades se encuentra en proporciones equivalentes en los distintos rangos de edades. Este análisis se realizó sobre la población que según los rangos de edades del dataset se encuentran dentro de la "Población Activa".

Si ahora se analiza únicamente la variabilidad de la falta de capacidad entre rango de edades, entre los 40-64 años, hay una relación equitativa entre los sexos. No obstante, el rango de "14 a 39 años" cuenta con una distribución mayor en el caso de los hombres con discapacidad.

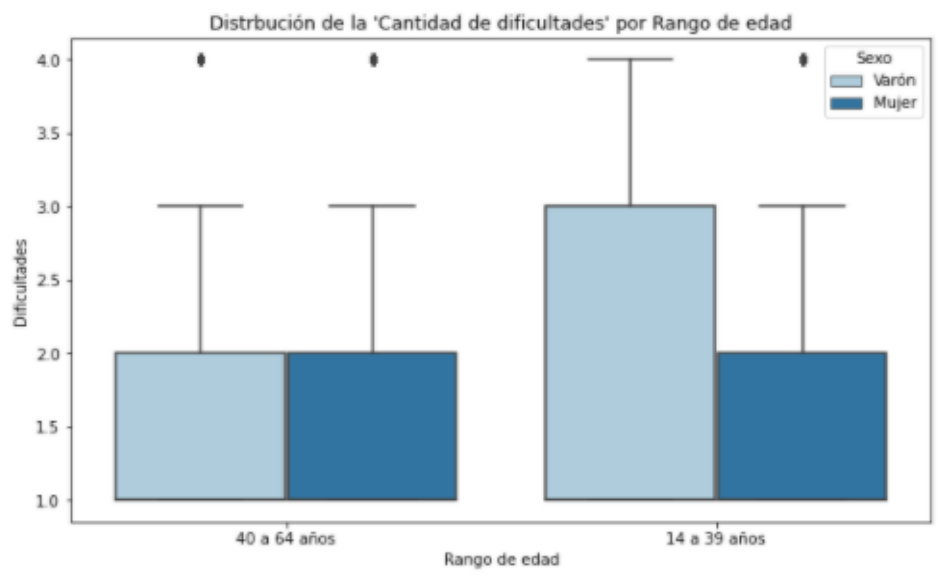

Figura 5

3.5 Tipo De Vivienda.

Finalmente, consideramos que el entorno donde se encuentra una persona, determina en cierto nivel su actividad laboral. Para este criterio, nos hemos basado en la vivienda del individuo, ya que es una necesidad básica a atender.

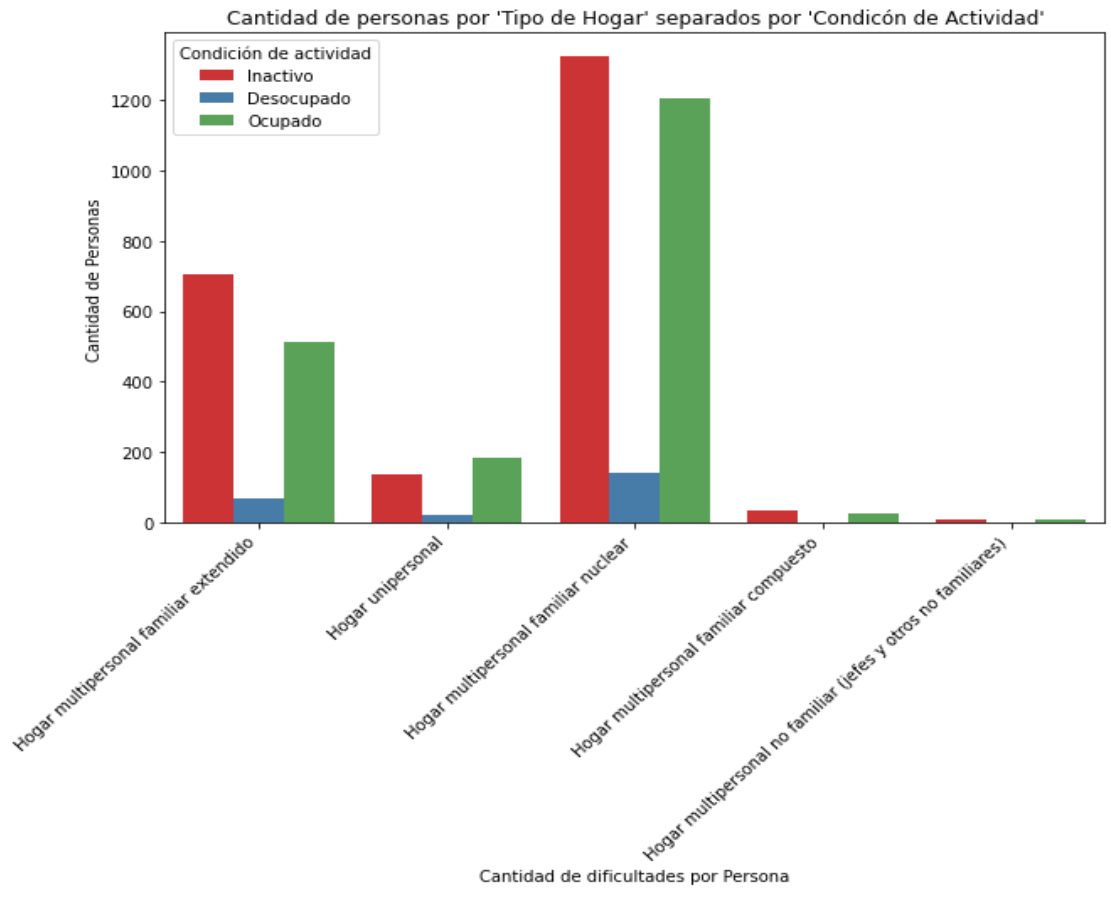

Figura 6

Del gráfico se puede visualizar que las personas con falta de discapacidad que se encuentran en un ambiente familiar, tienen incremento leve de inactividad al igual que las personas que se encuentran en un ambiente familiar externo.

Sin embargo, dentro las personas con discapacidad que viven unipersonalmente, hay un incremento en la actividad laboral. Esto seguramente sea debido a la necesidad de autosustento por el motivo de vivir solos.

En relación a las personas que viven en hogares “multipersonal familiar compuesto”y “multipersonal no familiar” no fuimos capaces de realizar ninguna conclusión debido a que son categorías que cuentan con poca cantidad de muestras como para aseverar alguna afirmación.

3.6 Rango De Edades.

La edad de inicio de la dificultad también es una variable que consideramos que podría permitir predecir la condición que determina la actividad de ocupación de una persona. En base a los resultados observados se puede decir que para inicios de la dificultad a una edad más avanzada hay mayor cantidad de personas que cumplen la condición de “Ocupado”.

Contrariamente las personas que tuvieron su inicio de las dificultades a temprana edad están más relacionadas con la condición de “Inactivos”, pudiendo deberse esto a una mala inserción en el mercado laboral. Esto podría deberse a falta de incentivos o ayudas a edad temprana.

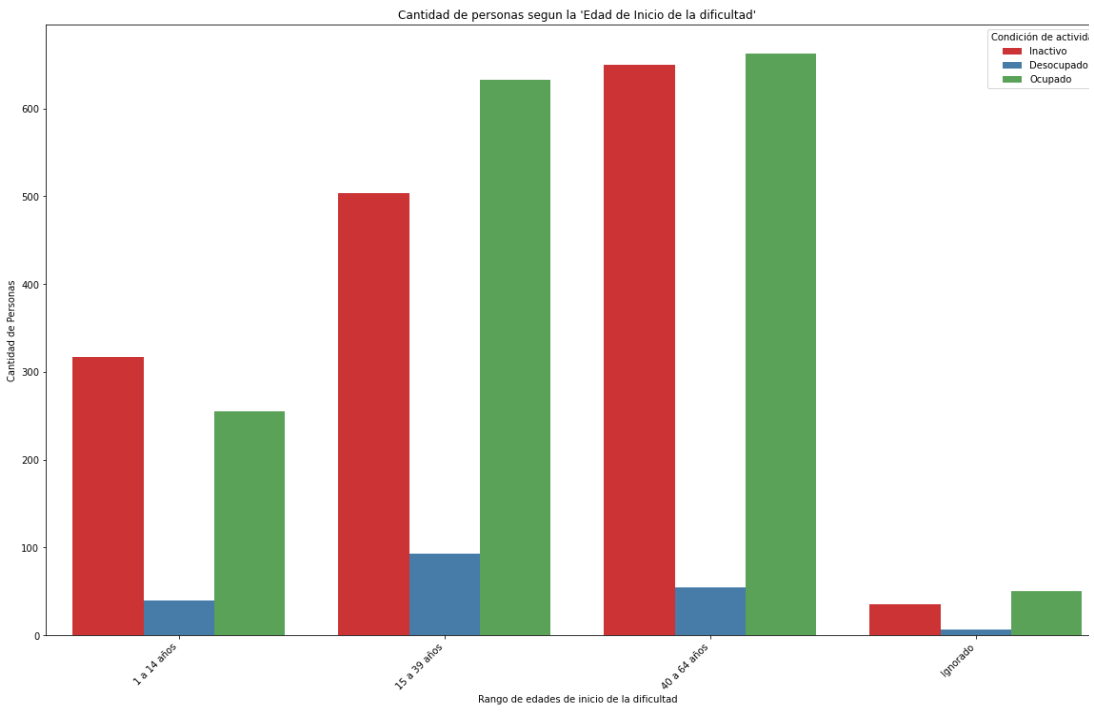

Figura 7

3.7 Tipo de Dificultad

Estudiamos nuestros datos en base al tipo de dificultad. Encontramos que la mayor parte de la población se reparte en dos tipos: “Dificultad visual y motora”.

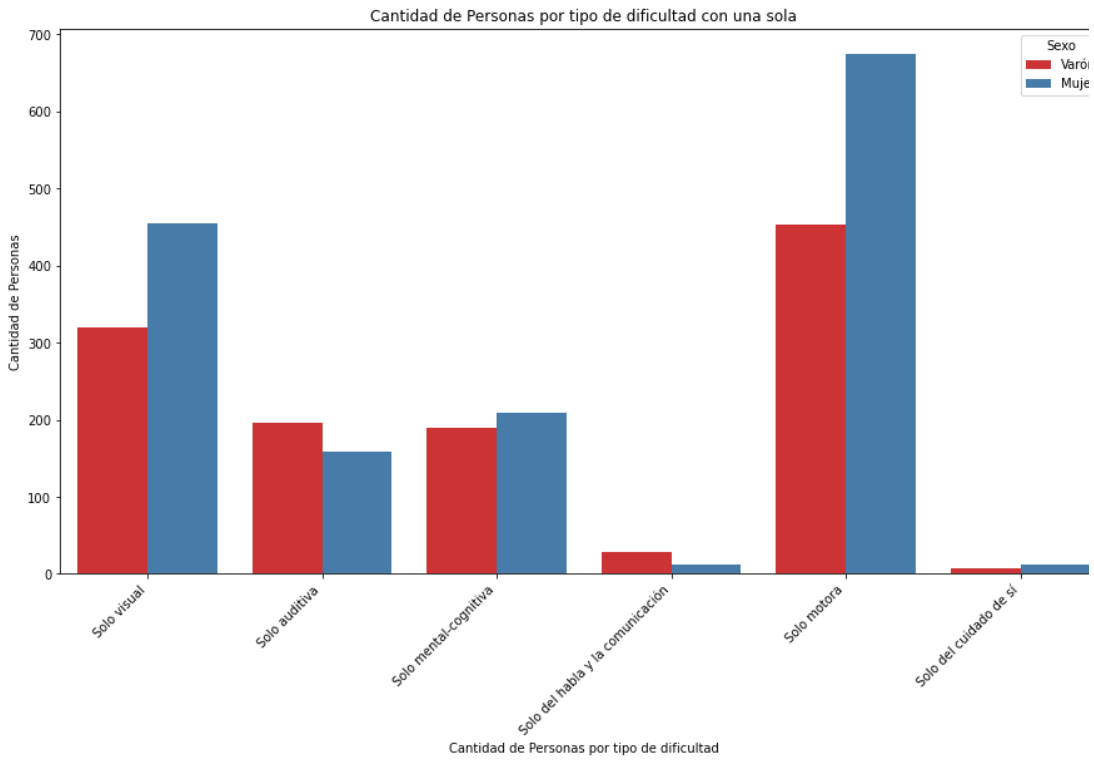

Figura 8

3.8 Cobertura de Salud

Estudiamos la relación entre el tipo de cobertura de salud que tienen las personas en relación con la cantidad de dificultades que tiene y su ‘Condición de Actividad’.

Notamos que las personas con mayor dificultad promedio se reúnen en el grupo que recibe un “Programa o plan estatal de salud”. Esto puede deberse a que al tener mayores dificultades el Estado le brinda una mejor cobertura de salud.

También relacionamos que para la “Condición de Actividad”, los grupos de personas que tienen en promedio mayor cantidad de dificultades se encuentran en ‘Inactivos’, en cambio, las personas con menor promedio de dificultades son las que se encuentran ‘Ocupadas’. Relacionamos esto, con que las personas al tener menor cantidad de dificultades, tienen menos barreras para realizar actividades/ tareas requeridas para insertarse en el mercado laboral.

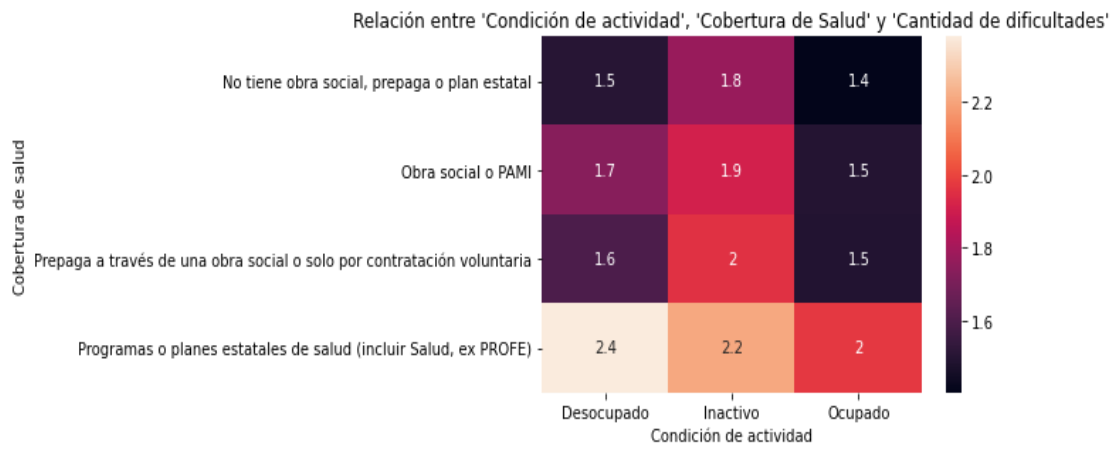

Figura 9: Heatmap de Relación entre “Condición de Actividad, Cobertura de salud y Cantidad de Dificultades promedio”

4. MATERIALES Y MÉTODOS

Con el fin de clasificar a la persona en “Inactiva, Desocupada u Ocupada” según las input la condición de analfabetismo (si sabe leer y/o escribir), el ambiente/entorno familiar donde vive, el máximo nivel de educación alcanzado, cantidad de discapacidad que presenta y el sexo de la persona. , se utilizaron los siguientes modelos para la realización de machine learning:

Logistic Regression; KNN y SVM.

Para implementar estos modelos, primero escalamos los datos con MinMaxScaler [5], ya que otorga mejores resultados en las features con gran cantidad de ceros en sus columnas.

Esta función transforma las variables escalando cada variable a un rango determinado. Este estimador escala y traduce cada variable individualmente de manera que se encuentre en el rango dado en el conjunto de entrenamiento, p.ej. entre cero y uno.

Funcion Min Max Scaler:

$$Xscaled = \frac{X - Xmin}{Xmas - Xmin}$$

Luego dividimos nuestros datos en train y test, tomando un 20% de nuestros datos como testeo, para que el clasificador aprendiera la regla de decisión utilizando el train set (samples + labels) para que luego pudiese clasificar las muestras de test (sin tener en cuenta las labels de test). Una vez obtenidas las labels del modelo de clasificación medimos la exactitud (Accuracy) de clasificación en testeo y realizamos una matriz de confusión para cada modelo realizado.

[5] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

4.1 Logistic Regression [6]

La regresión logística es un enfoque de modelado matemático que puede ser usado para describir la relación de varias variables X a una variable dependiente dicotómica. También son posibles otros enfoques de modelización, pero la regresión logística es, con mucho, el modelo más popular procedimiento utilizado para analizar los datos epidemiológicos cuando la medida de la enfermedad es dicotómica.[3]

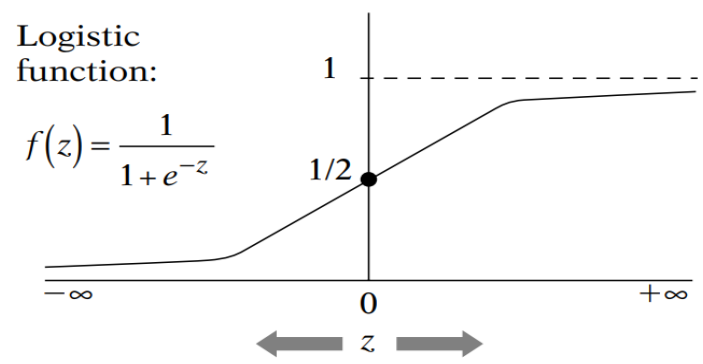

Figura 10

[6] http://www.fao.org/tempref/AG/Reserved/PPLPF/ftpOUT/Gianluca/stats/Logistic%20Regression,%20A%20Self-Learning%20Text,%202Ed%20(Statistics%20For%20Biology%20And%20Health)%20(David%20G%20Kleinbaum,%20Mitchell%20Klein)%200387953973.pdf

Es un clasificador lineal compuesta por una regresión 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 threshold (0,5) entonces pertenece a esta clase, en caso contrario viceversa.

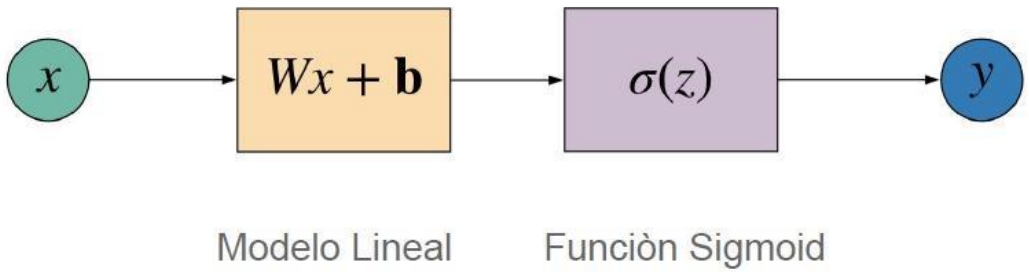

Figura 11: Imagen extraída de apuntes de la Cluster AI

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:

$$P(Yi \mid X) = \sigma(W^t, X)$$

La función sigmoide:

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

[5] Apuntes de Cluster AI Clase 05/clusterai2020_clase05_regresion_prez.pdf

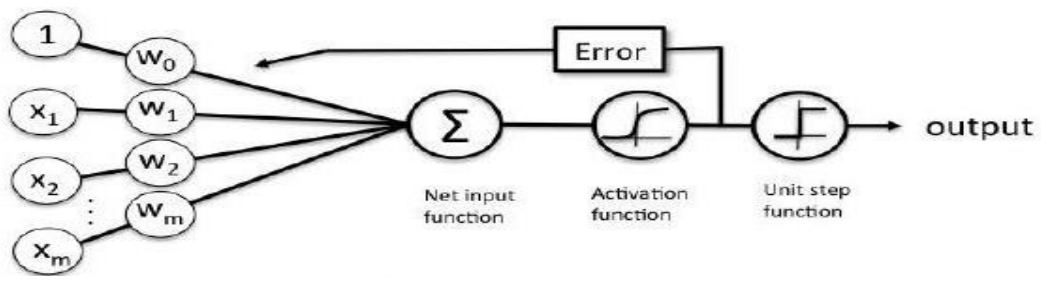

Figura 12: Imagen extraída de apuntes de la Cluster AI

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

4.2 K-Nearest Neighbors [7]

Es un modelo de clasificación en el cual un nuevo dato es agrupado según K vecinos más cercanos. Para esto se calcula la distancia del elemento nuevo a los existentes y se ordenan para seleccionar a qué grupo pertenecen. Uno de los hiper parámetros del modelo es determinar la cantidad de K vecinos. En el método KNN se clasifica cada nuevo set de pesajes según corresponda los K vecinos más cercanos de una comuna u otra. El cálculo de la distancia que utilizamos es el default, que es la Euclidiana que se corresponde a la fórmula. La ventaja del modelo KNN 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.

$$d\,(P1, P2) = \sqrt{(X2 - X1)^2 + (Y2 - Y1)^2}$$

[7] Hastie, Trevor.; Friedman, J. H. (Jerome H.) (2001). The elements of statistical learning : data mining, inference, and prediction : with 200 full-color illustrations

4.3 SVM: Super Vector Machine

Se trata de un clasificador lineal, el cual busca un hiperplano que maximiza 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. Se busca maximizar el margen de los datos generados con el hiperplano , haciendo que la mayor parte de las muestras caigan cerca del plano. El margen separador queda definido por “s” muestras, llamadas support vectors.

En nuestro modelo utilizamos un Kernel (función de similaridad entre muestras) Gaussiano, para determinar una frontera no lineal de clasificación.

$$K\ gaussiano\ (Xi, Xj) = exp\left(-\frac{|Xi - Xj|^2}{2\sigma^2}\right)$$

5 RESULTADOS

La matrix de confusión es un modelo de clasificación de ML con el cual vamos a predecir nuestra variable categorica, para nuestro caso queremos a averiguar si la persona se encuentra desocupada , ocupada o inactiva , es una metrica intuitiva y sencilla que se utliza para encontrar la precision de nuestro modelo que tiene 3 tipos de salidas

El accuracy es una métrica para evaluar nuestro modelo de clasificación. Informalmente, la **exactitud** es la fracción de predicciones que el modelo realizó correctamente. Formalmente : accuracy = (número de predicciones correctas)/(numero total de predicciones)
La exactitud resulta ser de 0.64 o 64% (64 predicciones correctas de 100 ejemplos totales). Eso significa que nuestro clasificador está haciendo un buen trabajo en la identificación personas que se encuentran ocupadas, desocupadas o inactivas
Utilizando las herramientas y modelos descritos, obtuvimos una clasificación de variables con una exactitud (Accuracy) 63% promedio entre los 3 modelos estudiados. Siendo nuestro mejor modelo el KNN-Classifier.

Matriz de confusión observada para el modelo KNN-CLassifier:

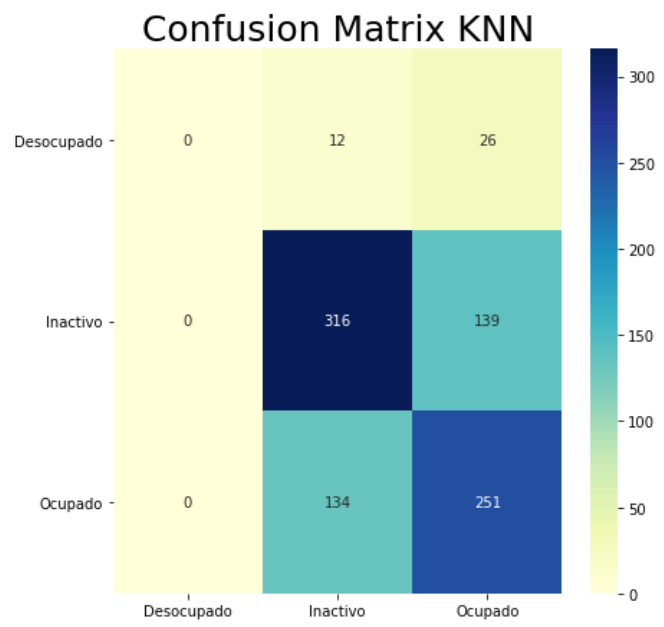

Figura 13

Esta matriz tiene en su diagonal principal las muestras resultantes verdaderas positivas. En los demás cuadrantes muestra la cantidad de variables dependientes mal asignadas por cada categoría. Como se puede ver la Clase ‘Desocupado’ resultó no tener ninguna variable bien asignada, esto se debió a la poca cantidad de muestras en el Dataset para que el modelo sea capaz de entrenarse con ellas.

Los resultados de precisión obtenidos fueron:

MODELOS	ACCURACY
KNN-Classifier	**64,58 %**
Support Vector Machine	**63,55 %**
Logistic Regression	**62,76 %**

6. CONCLUSIONES

Luego de recurrir a diferentes estrategias de balance de datos y explorar diferentes modelos, pudimos llegar a un modelo basado en KNN que obtuvo los mejores resultados. Nuestro modelo arroja un Accuracy del 64, 58 % para clasificar entre las 3 muestras diferentes, por lo tanto podemos inferir que los datos tienen distribuciones de sus atributos y features distintos. Sin embargo sigue sin ser una medida de exactitud muy alta, esto puede ser debido a que para las clases 'Ocupado' y 'Inactivo' muchos de los atributos asociados son los mismos para cada una, sin tener una notoria diferencia. Además se pudo notar que la clase ‘Desocupado’ no pudo ser entrenada por ninguno de los modelos desarrollados, siendo sus resultados para la matriz de confusión siempre ‘Falsos’.

Pudimos concluir que de contar con más muestras de la clase ‘Desocupado’ nuestro modelo se podría haber entrenado mejor. La creación del pipeline de este modelo hace que en los próximos años, se pueda usar como base para nuevos estudios. Sin embargo, no creemos que la exactitud de ningún modelo de clasificación/predicción en base el Dataset analizado aumente mucho por encima de las alcanzadas, ya que los atributos asociados a cada clase son muy parecidos.

En lo personal, este trabajo nos permitió adquirir y profundizar el uso de herramientas como Python. A la vez nos permitió implementar conceptos de Machine Learning a casos reales y cercanos de la sociedad. Pudimos obtener resultados concretos sobre la ocupación de las personas discapacitadas y el comportamiento de cada persona tomando en cuenta distintos factores que forman de cada una de ellas. El estudio nos permitió tomar mayor conciencia de las problemáticas que sufren las personas discapacitadas en la sociedad.

7. REFERENCIAS

- Indec (2020) . Consultado el 15/10/2020. Disponible en: https://www.indec.gob.ar/ftp/cuadros/menusuperior/enpd/estudio_discapacidad_manual_base_datos_usuario.pd
- Indec (2020) . Consultado el 15/10/2020. Disponible en: https://www.indec.gob.ar/indec/web/Institucional-Indec-BasesDeDatos-7
- Indec (2020) . Consultado el 15/10/2020. Disponible en: https://www.indec.gob.ar/ftp/cuadros/menusuperior/enpd/estudio_discapacidad_nota_tecnica.pdf
- Apuntes Cluster IA. 2020 - Cátedra de Ciencia de Datos UTN - FRBA. Consultado el 16/11/2020
- Página Oficial Numpy. Consultado el 16/11/2020. Disponible en: https://pandas.pydata.org/
- Página Oficial Pandas. Consultado el 16/11/2020. Disponible en: https://pandas.pydata.org/
- Página Oficial Seaborn. Consultado el 16/11/2020. Disponible en: https://seaborn.pydata.org/
- Página Oficial Sklearn Preprocessing MinMaxScaler. Consultado el 16/11/2020 https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html
- Hastie, Trevor.; Friedman, J. H. (Jerome H.) (2001). The elements of statistical learning : data mining, inference, and prediction : with 200 full-color illustrations. Consultado el 16/11/2020
- Logistic Regression. Consultado el 16/11/2020. Disponible en: http://www.fao.org/tempref/AG/Reserved/PPLPF/ftpOUT/Gianluca/stats/Logistic%20Regression,%20A%20Self-Learning%20Text,%202Ed%20(Statistics%20For%20Biology%20And%20Health)%20(David%20G%20Kleinbaum,%20Mitchell%20Klein)%200387953973.pdf