




Predicción de clases sociales en encuesta de vivienda 2020

Jesús Iván Ruíz Martínez



Encuesta nacional de vivienda 2020



Encuesta nacional de vivienda

La ENVI 2020 se llevó a cabo del 26 de octubre al 18 de diciembre del 2020, con el objetivo de producir información estadística sobre las características de la vivienda en México que permita generar un panorama amplio sobre la situación de la vivienda en el país, necesidades y demanda de la población al respecto.

- Demandas y necesidades de vivienda
 - Características del hogar
- Características de los residentes del hogar
 - Segunda vivienda
 - Gasto en viviendas secundarias
 - Características de la vivienda

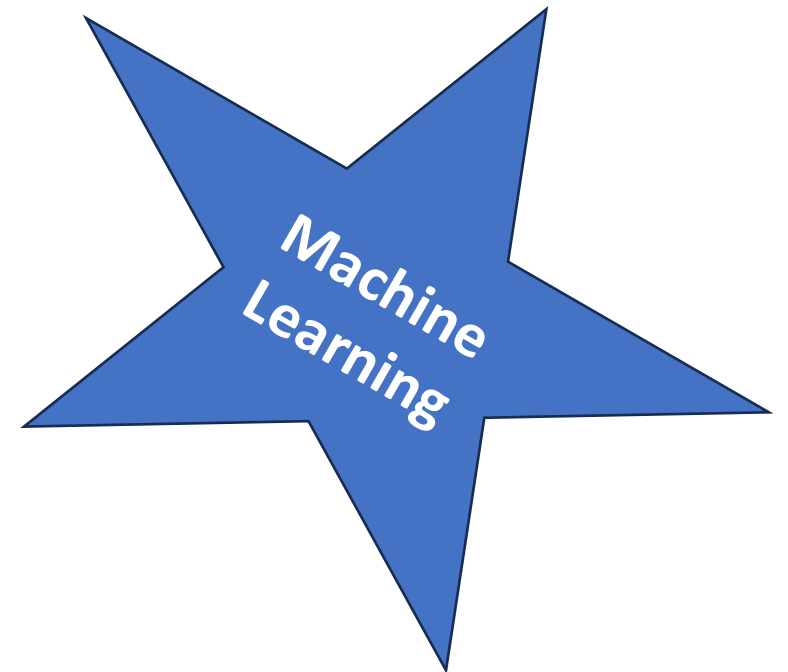
Características de los residentes del hogar

Objetivo general

Predecir la influencia de diferentes características en orden de saber cuales tienen mayor influencia para ser de una clase social más alta.

Objetivos específicos

- Adquisición de base de datos de recurso abierto ENVI 2020.
- Base de datos montada en mariadb.
- Estandarización de datos
- Implementación de distintos clasificadores
- Estimación de precisiones



Que datos utilizaremos?

Clave	Preguntas	Tipo de respuesta
ENT	Localidad perteneciente	1,2,...,32
SEXO	Sexo de la persona dueña o encargada de la vivienda?	1, 2
EDAD	Edad del dueño o encargado de la vivienda?	1, 2 , .. , 99
P2_5	Habla alguna lengua indígena?	1, 2, 9
P2_8	Actualmente vive en (estado civil)	1, 2, .., 6
P3_1	La persona encargada, trabaja, estudia, es pensionado, etc...	1, 2, ..., 8
P3_3	El trabajo de la semana pasada fue trabajador, empleado, empleador, jornalero...	1, 2, ..., 5
P3_4	Cuanto gana por su actividad?	00000, ..., 99999

Montura de base de datos

- Crear el contenedor de docker para mariadb
- Subir los archivos CSV con mysql workbench
- Sustituir celdas vacias con NULL
- Eliminar filas donde el ingreso reportado fue 0 o "no contesto"
- Homogenizar columna de "ingresos"
- Quitar columnas innecesarias

Variable de
predicción



Implementación algoritmos de ML

**Regresión
lineal**

**Regresión
logística**

Accuracy

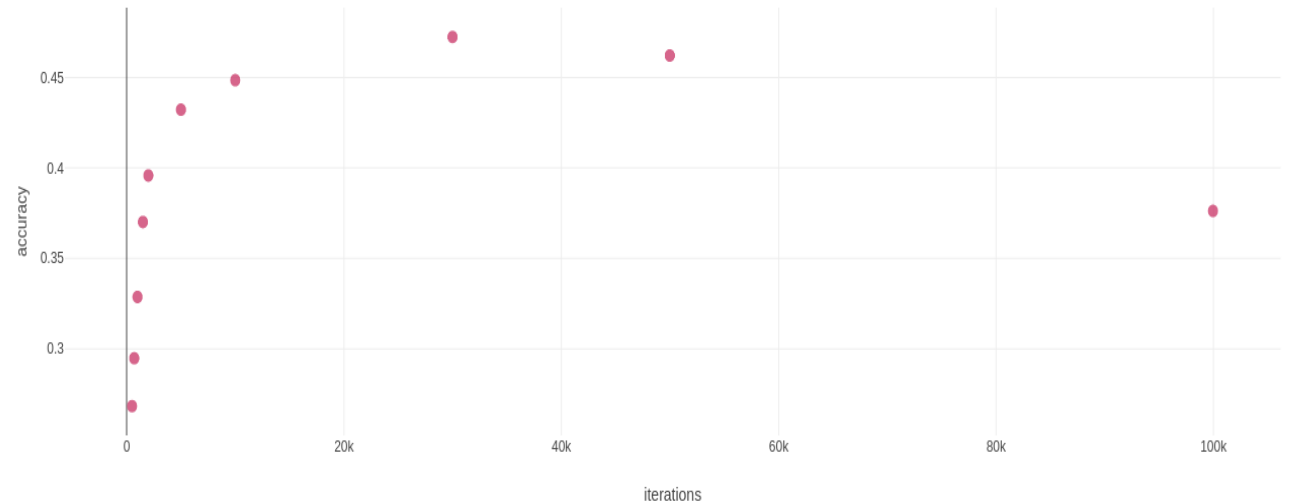
**Gaussian
Mixture**

**Multi layer
perceptron**

Decision tree

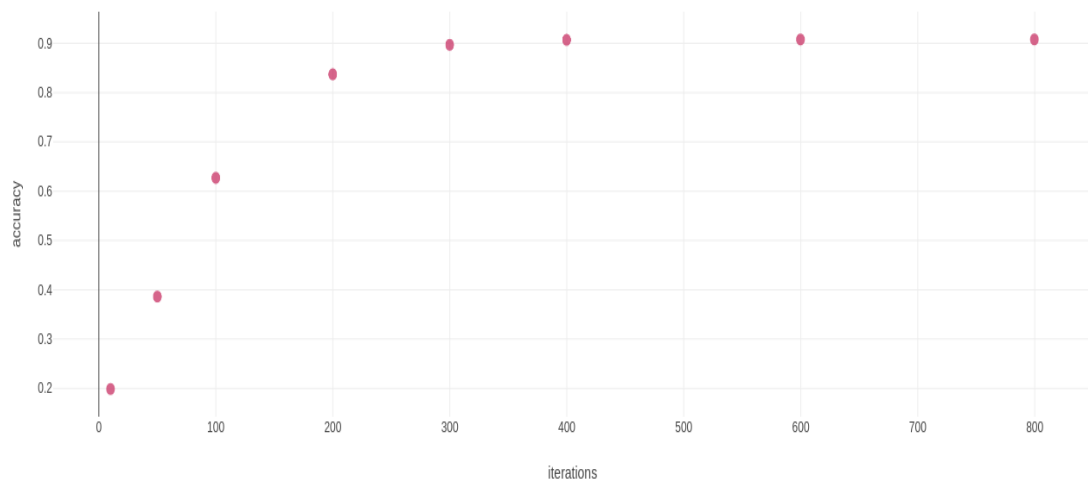
Regresión Lineal

```
PROBLEMAS 7 SALIDA CONSOLA DE DEPURACIÓN TERMINAL
ivan@ivan-A320AM4-M3D:~$ /bin/python3 /home/ivan/Documentos/Pr
No GPU/TPU found, falling back to CPU. (Set TF_CPP_MIN_LOG_LEV
Características ordenadas por importancia:
ENT: -0.5174124240875244
P3_3: -0.505962073802948
P2_5: -2.6231679916381836
SEX0: -0.12121573090553284
EDAD: 0.22568193078041077
P3_1: 0.6485194563865662
P2_8: -0.1943260282278061
Classification accuracy: 47.24%
ivan@ivan-A320AM4-M3D:~$
```



LR = 0.001, Interacciones variables

Regresión Logística

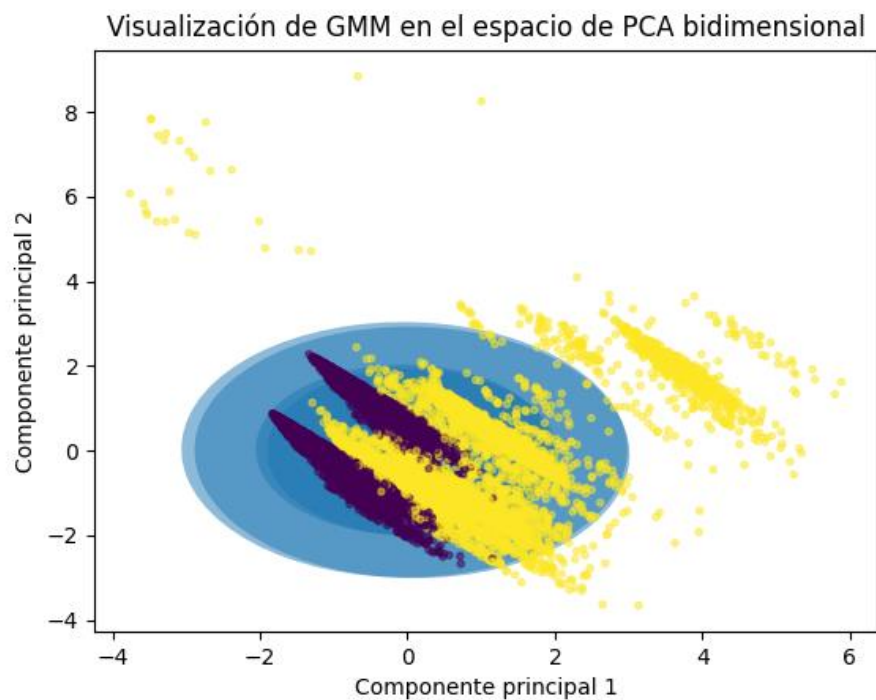


PROBLEMAS SALIDA CONSOLA DE DEPURACIÓN TERMINAL

```
Iteración 999: MSE de entrenamiento = 0.37886008620262146, MSE de validación = 0.38626885414123535
Iteración 1000: MSE de entrenamiento = 0.3788224160671234, MSE de validación = 0.38623106479644775
Características ordenadas por importancia:
P2_5: -2.4687838554382324
P3_3: -1.1046310663223267
SEX0: 1.0431225299835205
ENT: -0.7770714163780212
EDAD: 0.8411794304847717
P2_8: -0.2931194007396698
P3_1: 1.033476710319519
Precisión del modelo: 90.79%
ivan@ivan-A320AM4-M3D:~$
```

LR = 0.01, Interacciones variables

Gaussian Mixture

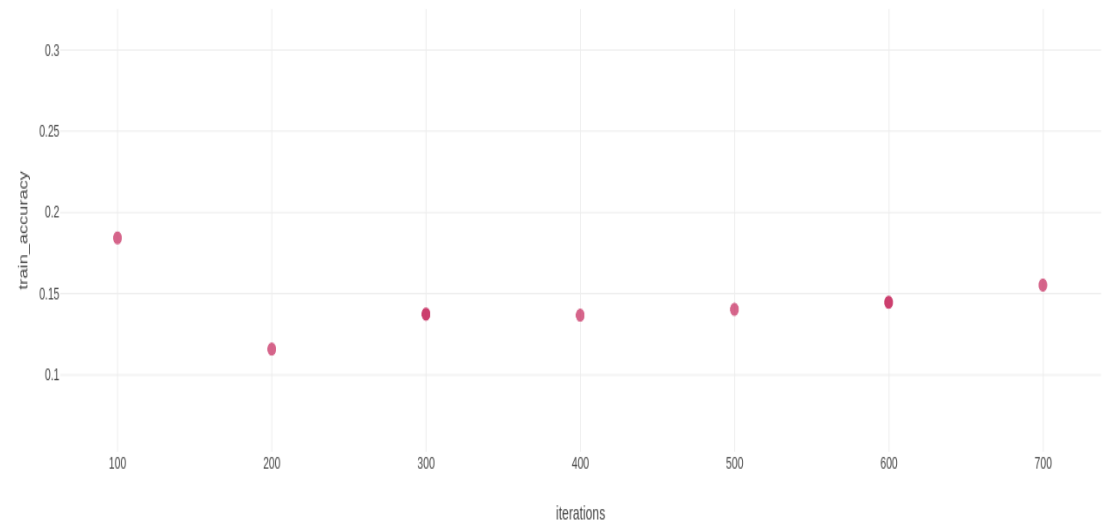


```
ivan@ivan-A320AM4-M3D:~$ /bin/python3 /home/ivan/Documentos/ProyectFE/Mixturemodel2.py
Característica ENT: 20.94%
Característica SEX0: 16.22%
Característica EDAD: 14.76%
Característica P2_5: 14.04%
Característica P2_8: 13.22%
Característica P3_1: 11.84%
Característica P3_3: 8.97%
/home/ivan/.local/lib/python3.10/site-packages/_distutils_hack/__init__.py:33: UserWarning:
  warnings.warn("Setuptools is replacing distutils.")
Precisión del modelo: 62.24%
ivan@ivan-A320AM4-M3D:~$
```

No. De componentes = 2

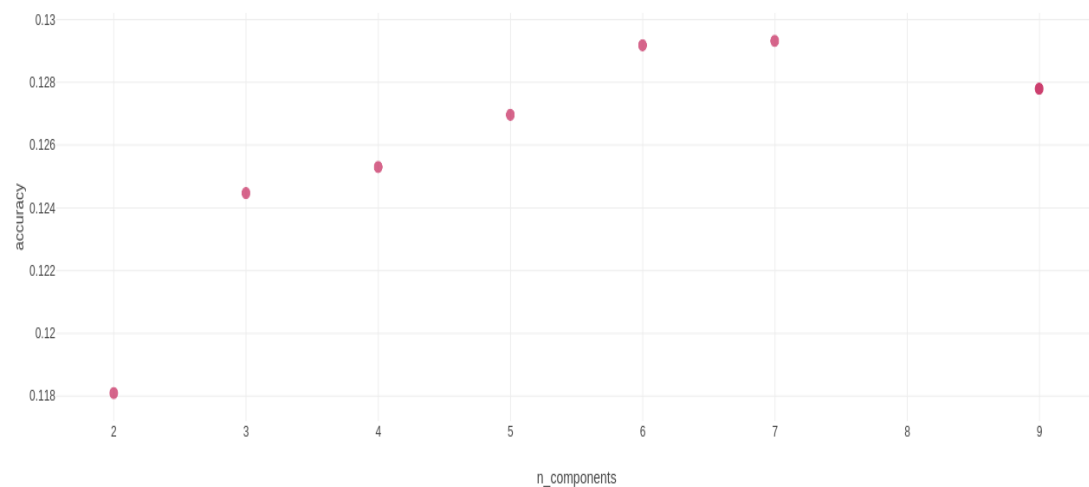
Multi layer perceptron

```
PROBLEMAS 8 SALIDA CONSOLA DE DEPURACIÓN TERMINAL
Iteration 570: Validation Accuracy = 0.1450
Iteration 580: Validation Accuracy = 0.1469
Iteration 590: Validation Accuracy = 0.1476
Importancia de las características:
ENT: 0.0021
SEX0: 0.0002
EDAD: -0.0031
P2_5: 0.0020
P2_8: 0.0095
P3_1: -0.0030
P3_3: -0.0012
ivan@ivan-A320AM4-M3D:~$
```



LR = 0.001, Interacciones variables

Decision tree



```
PROBLEMAS 9 SALIDA CONSOLA DE DEPURACIÓN TERMINAL
ivan@ivan-A320AM4-M3D:~$ /bin/python3 /home/ivan/Documentos/ProyectFE/Decisiontree1.py
No GPU/TPU found, falling back to CPU. (Set TF_CPP_MIN_LOG_LEVEL=0 and rerun for more info.)
Classification accuracy: 12.92%
Características ordenadas por importancia (árbol de decisión):
P3_3: 0.2956024884715493
P3_1: 0.26983758506014005
EDAD: 0.15245178031946452
ENT: 0.1030779478808462
SEX0: 0.10059877869121167
P2_5: 0.04554755495458299
P2_8: 0.03288386462220517
ivan@ivan-A320AM4-M3D:~$
```

Produndidad máxima 7

Conclusiones

- Algunos modelos tuvieron mejor performance al momento de realizar las predicciones
- Pese a que para algunos modelos su precision fue baja apuntaban a unos resultados similares
- Influye que estes trabajando para ser de una clase social alta
- Influye también la edad
- Influye la localidad donde se encuentre
- Influye negativamente hablar una lengua indígena