Análisis de la condición laboral del país según la Encuesta de Hogares 2019

Ivan Fernando Mujica Mamani
Universidad Católica Boliviana “San Pablo“, BOLIVIA
ifmm87@gmail.com
Profesor: Phd. Eduardo Morales

Abstract

La Encuesta de Hogares es un instrumento del Instituto Nacional de Estadística (INE), que tiene como objetivo Proporcionar estadísticas e indicadores socioeconómicos y demográficos de la población boliviana, necesarias para la formulación, evaluación, seguimiento de políticas y diseño de programas de acción contenidas en el PDES 2016 - 2020.

El presente estudio tiene la finalidad de realizar un análisis de categorización de la variable ***condición laboral*** en función de las variables sociales y demográficas de la Encuesta de Hogares realizada anualmente por el Instituto Nacional de Estadística de Bolivia desde el 2016 al 2019, con técnicas de Análisis Estadístico y Machine Learning.

El problema

La falta de estudios con técnicas de mineria de datos respecto a la categorización de la condición laboral en función de variables independientes demográficas como ser sexo, edad, nivel de educación, parentezco, pertenencia étnica y sociales como nivel de ingresos, gastos del hogar y otras podrian explicar la categorización de la condición ocupacional del país.

Para ellos se utilizó 3 algoritmos de clasificación que son:

- Árboles de decisión.
- Naive Bayes Gausiano.
- Regresión Logística Multinomial

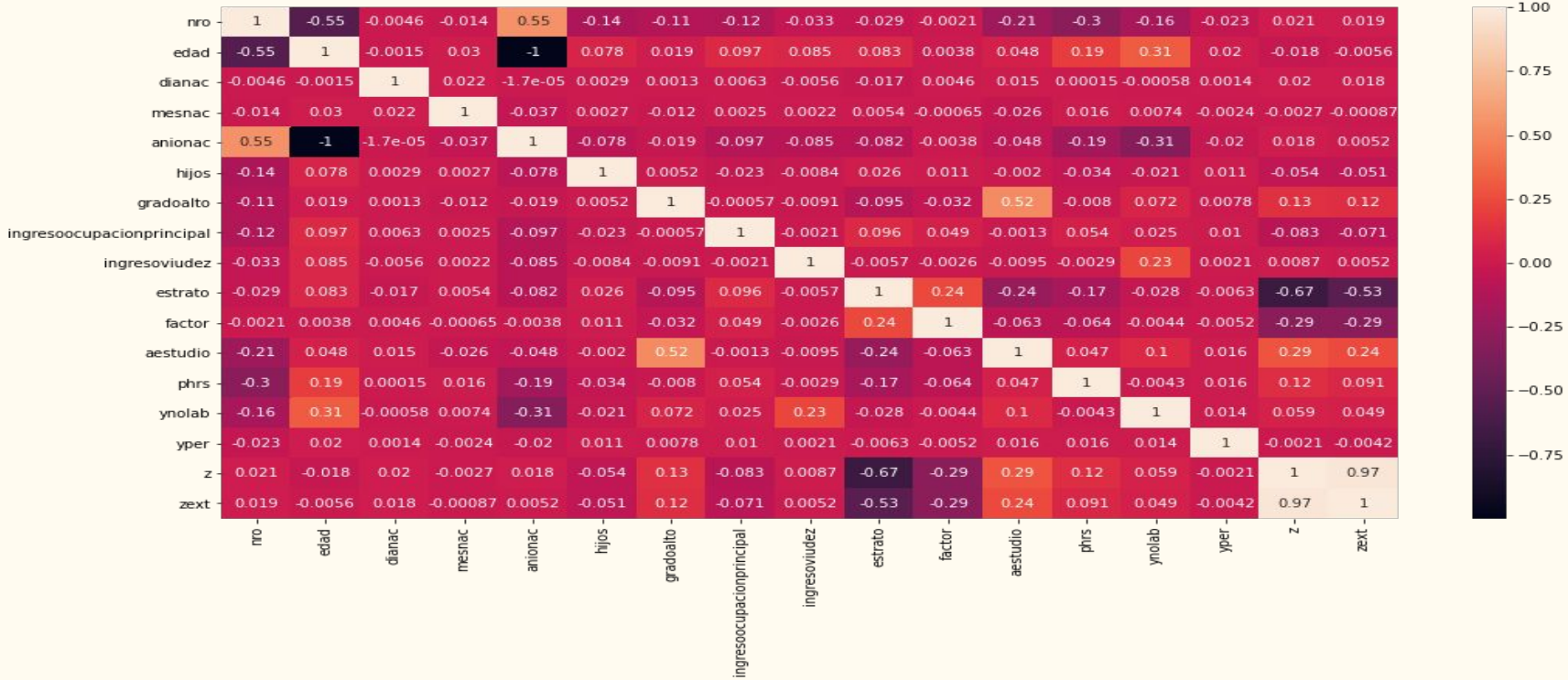
Características de los datos

[illegible]
La variable objetivo (target) es la variable *condact* [illegible] o condicion laboral

- p_cesante
- p_ocupado
- p_permanente
- p_aspirante
- p_temporal

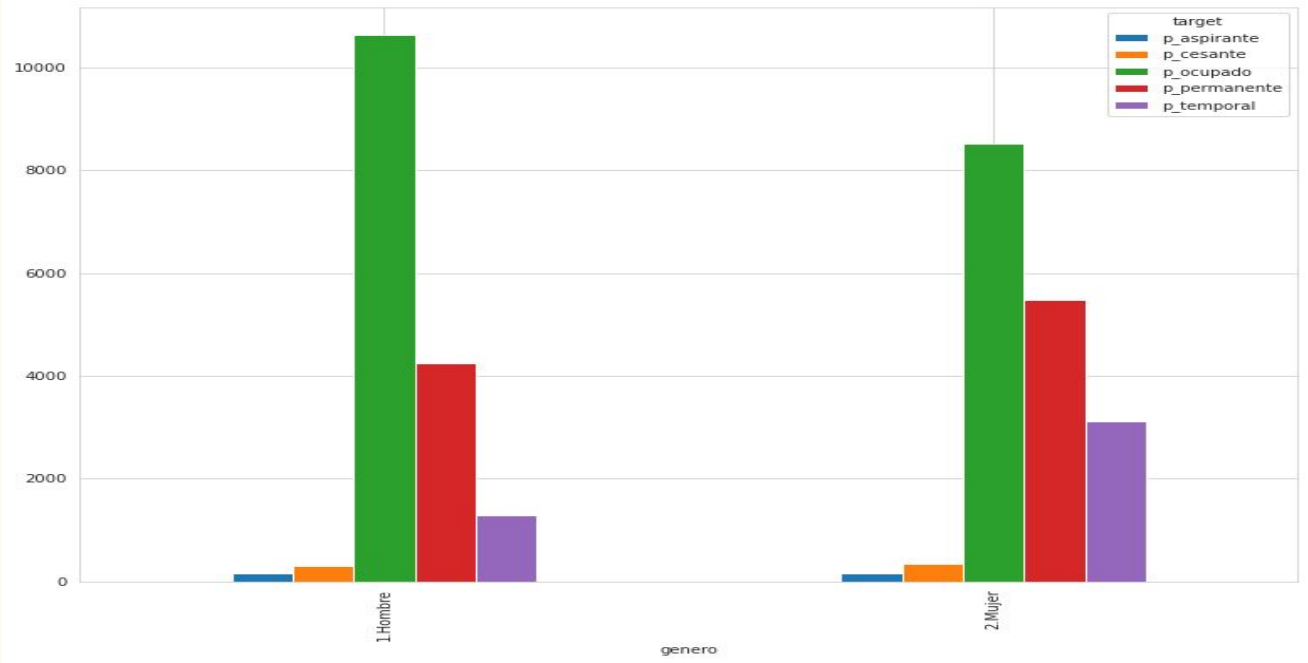
Variable	Tipo	Descripción	Variable	Tipo	Descripción
genero	Numeric	¿Es hombre o mujer?	hnv_ult_a	Numeric	Hijos nacidos vivos en el último año
edad	Numeric	¿Cuántos años cumplidos tiene?	quienatenparto	Numeric	Personal de atención del parto
dia_nac	Numeric	¿Cuál es la fecha de su nacimiento?(día)	dondeatenparto	Numeric	Lugar de atención del parto
mes_nac	Numeric	¿Cuál es la fecha de su nacimiento?(mes)	educ_prev	Numeric	Años de estudio previos
anio_nac	Numeric	¿Cuál es la fecha de su nacimiento?(año)	aestudio	Numeric	Años de estudio
relacion_jefe_hogar	Numeric	¿Qué relación o parentesco tiene con el jefe o jefa del hogar?	cob_op	Numeric	Grupo Ocupacional ocupación principal
estado_civil	Numeric	¿Cuál es su estado civil o conyugal actual?	caeb_op	Numeric	Clasificación de Actividad Económica de Bolivi Ocupacion principal
fuma	Numeric	¿Durante los últimos 12 meses () ha fumado cigarrillos?	pet	Numeric	Poblacion en edad de trabajar
bebe	Numeric	¿Durante los últimos 12 meses () ha consumido bebidas alcohólicas?	ocupado	Numeric	Poblacion Ocupada
frecuencia_bebe	Numeric	¿Con que frecuencia ha consumido bebidas alcohólicas ?	cesante	Numeric	Poblacion Desocupada Cesante
grado_alto	Numeric	Ingrese el Curso o Grado	aspirante	Numeric	Poblacion Desocupada Aspirante
ocupacion	Numeric	¿Esta ocupacion usted la realiza?	desocupado	Numeric	Poblacion Desocupada
tiene_seguro	Numeric	¿En su actual ocupación Ud. recibe o recibirá los siguientes beneficios: Seguro de salud	pea	Numeric	Poblacion Activa
ingreso_ocupacion_principal	Numeric	¿Cuánto es su ingreso total en su ocupación principal? Monto Bs	temporal	Numeric	Poblacion Inactiva Temporal
estrato	String	Estrato	permanente	Numeric	Poblacion Inactiva Temporal
factor	Numeric	Factor de expansión	pei	Numeric	Poblacion Inactiva
cobersalud	Numeric	Cobertura de Seguro de Salud	condact	Numeric	Condicion de Actividad Ocupacion Principal
phrs	Numeric	Horas trabajadas a la semana Ocupación Principal	ynolab	Numeric	Ingreso no laboral (Bs/Mes)
shrs	Numeric	Horas trabajadas a la semana Ocupacion Secundaria	yper	Numeric	Ingreso Personal (Bs/Mes)
ylab	Numeric	Ingreso laboral (Bs/Mes)	yhog	Numeric	Ingreso del Hogar (Bs/Mes)

Análisis Exploratorio



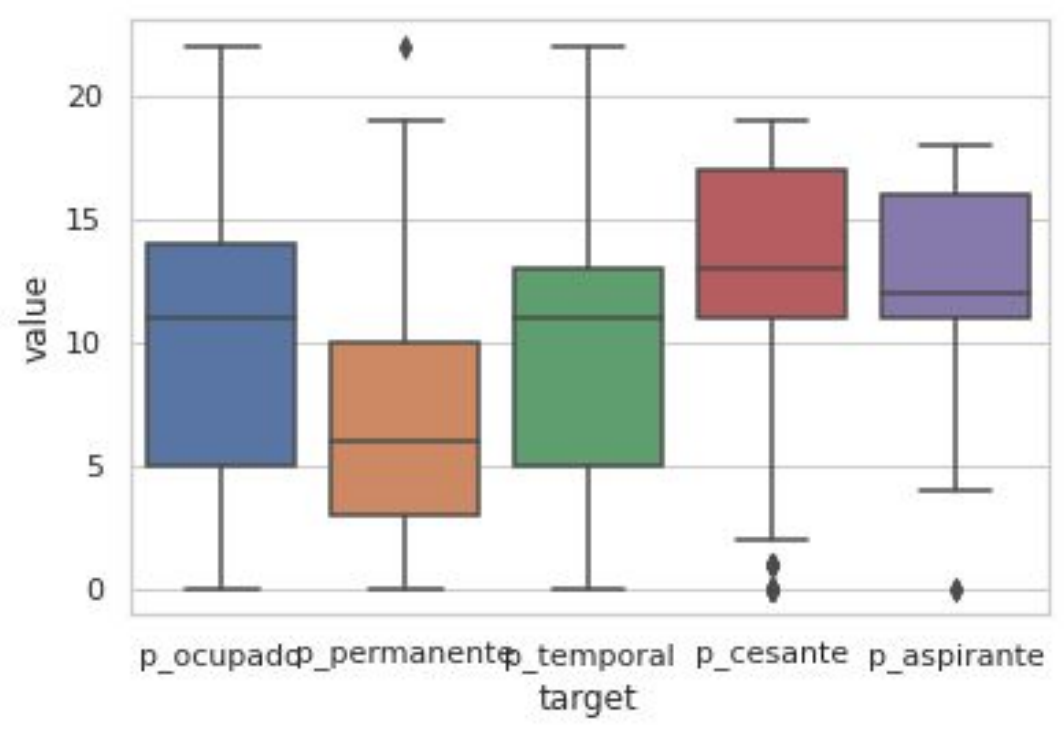
Heatmap de matriz de correlación del dataset

Análisis Exploratorio



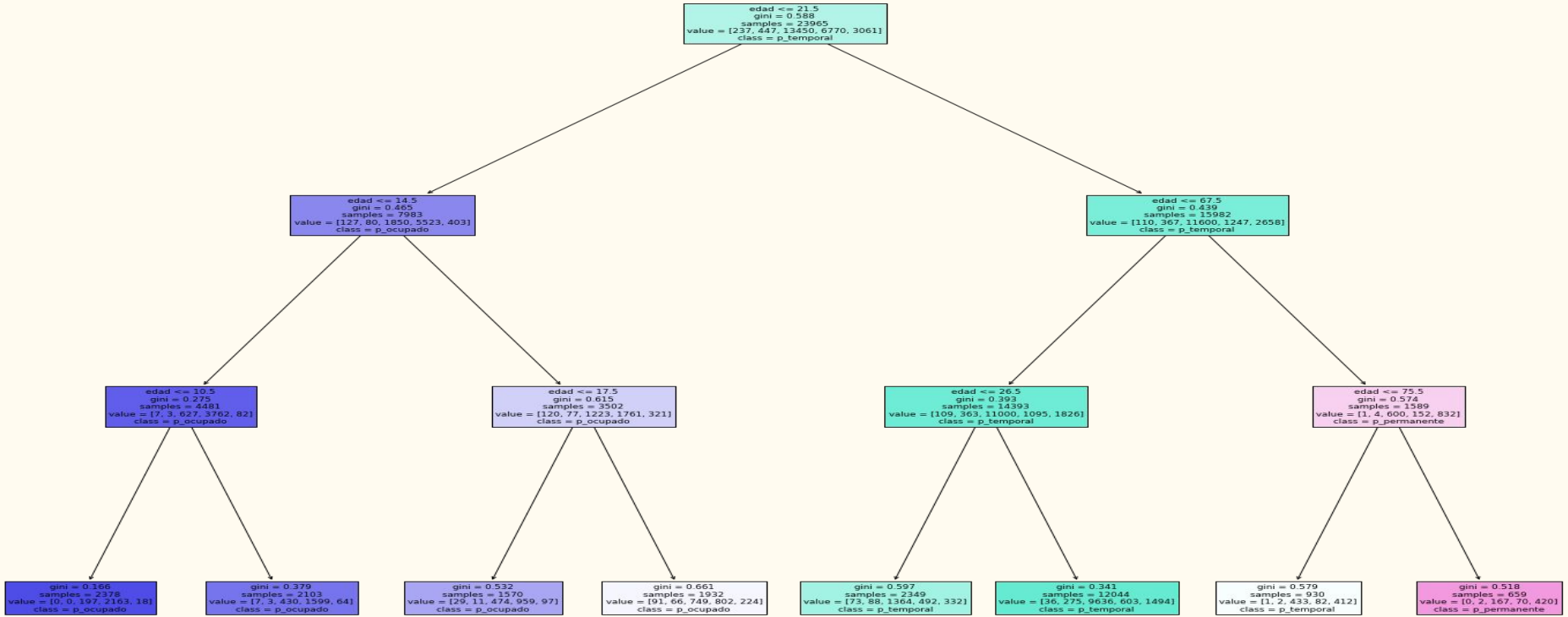
Histograma genero y la condición laboral

Análisis Exploratorio



Boxplot años de estudio y condicion laboral

Árboles de Decisión



Segun el grafico, la clasificacion de una persona cualquiera a una de las 6 categorias de condicion laboral, segun el modelo de arbol de decision, prima mas la edad, el numero de hijos, y los anios de estudio con 3 niveles de profundidad en cuanto al modelo. La precisión del modelo alcanza al 72 % de precisión, lo cual indica un nivel aceptable de exactitud.

Naive Bayes Gausiano

Para la aplicación del algoritmo necesitamos categorizar las variables feature, las seleccionadas son: g***enero, tipohogar, razontrabaja, cobersalud,hijos, ocupacion, relacionjefehogar*** y la variable ***target (condicion laboral)***

	edad	genero_e	hijos_e	tipohogar_e	cobersalud_e	razontrabaja_e	relacionjefehogar_e	ocupacion_e	ingresoocupacionprincipal	aestudio	target
0	42	0	0	1	0	0	0	2	0	17	2
1	44	0	0	4	3	0	0	2	0	16	2
2	4	0	0	1	3	0	6	0	0	0	5
3	41	0	0	4	3	0	0	1	900	6	2
4	31	1	8	4	3	0	5	2	0	4	2

Existe una precisión del 90, esto significa que de cada 100 personas el modelo predice 90 a una categoría de condicion laboral. A diferencia del algoritmo Naive Bayes Multinomial que solo nos daba una precisión de 62%.

Regresión Logística Multinomial

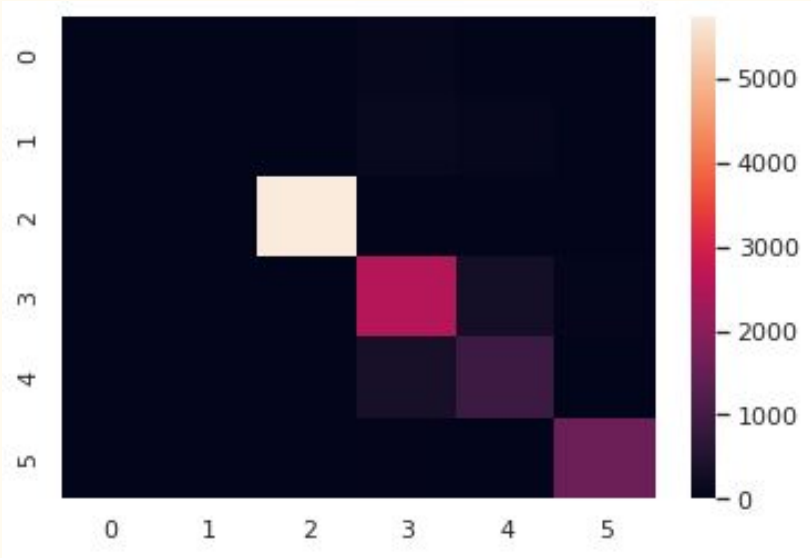
Para la aplicación del algoritmo necesitamos categorizar las variables feature, las seleccionadas son: g***enero, tipohogar, razontrabaja, cobersalud,hijos, ocupacion, relacionjefehogar*** y la variable ***target (condicion laboral)***

	edad	genero_e	hijos_e	tipohogar_e	cobersalud_e	razontrabaja_e	relacionjefehogar_e	ocupacion_e	ingresoocupacionprincipal	aestudio	target
0	42	0	0	1	0	0	0	2	0	17	2
1	44	0	0	4	3	0	0	2	0	16	2
2	4	0	0	1	3	0	6	0	0	0	5
3	41	0	0	4	3	0	0	1	900	6	2
4	31	1	8	4	3	0	5	2	0	4	2

Regresión Logística Multinomial

Los resultados fueron los siguientes:
Coef de determinación: 0.910
Cross-validation R^2: 0.912
Precisión: 0.910
Proporción de verdaderos positivos : 0.910
Contribución de la precisión ponderada(F1 score): 0.910

```
[[   0    1    0   89   13    0]
 [   0    7    0  115   79    0]
 [   0    0 5738    1    1    0]
 [   0   11    0 2558  299   54]
 [   0    4    0  362  934    0]
 [   0    0    0   42    0 1574]]
```



Se aprecia que los colores oscuros muestran errores muy bajos en cuanto a los falsos positivos y verdaderos negativos.

Gracias.

Proyecto disponible en:

https://github.com/ifmm87/mineria-datos-2