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

Ivan Fernando Mujica Mamani

Universidad Católica Boliviana “San Pablo“, BOLIVIA
Maestría en Ciencia de Datos v2.
ifmm87@gmail.com

Resumen. En el presente documento se analizará la clasificación laboral de la población boliviana en base a la Encuesta de Hogares 2019 y tiene por objetivo construir un modelo óptimo de clasificación con diferentes algoritmos para predecir la clasificación laboral de un individuo.

Keywords: Encuesta Hogares, Naive Bayes, Empleabilidad, Condición laboral

1 Introducción

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. Actualmente no se cuentan con estudios públicos abiertos, el Instituto Nacional de Estadística INE bajo tuición del Ministerio de Planificación en el marco de sus competencias realiza estudios que son posteriormente usados para plantear políticas internas.

2 Planteamiento del 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.

3 Metodología y Desarrollo

Actualmente dentro de la minería de datos se cuenta con algoritmos que facilitan en la analítica predictiva. En nuestro caso veremos algoritmos de agrupación (clúster) y algoritmos de clasificación como ser: Árboles de decisión, Naive Bayes, Regresión Logística entre los mas conocidos.

3.1 Caracteristicas de los datos

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 (Plan de Desarrollo Económico y Social para Vivir Bien) 2016 – 2020.

Tabla 1. Presentación de variables.

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

De la tabla anterior nuesta variable objetivo (target) es la variable ***condact*** que indica el grado de empleabilidad que tiene el individuo en su fuente de actividad principal, esta variable tiene 7 categorias:

- p_cesante
- p_ocupado
- p_permanente
- p_aspirante
- p_temporal

3.2 Algoritmos de clasificación

Dentro del área predictiva contamos con algoritmos que nos ayudan con la predicción de nuestras variables objetivo, en nuestro caso la variable ***condición laboral.***

Para ellos se utilizó 3 algoritmos de clasificación que son:

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

4 Experimentos y resultados

4.1 Análisis exploratorio

Implica el uso de gráficos y visualizaciones para explorar y analizar un conjunto de datos. El objetivo es explorar, investigar y aprender, no confirmar hipótesis estadísticas.

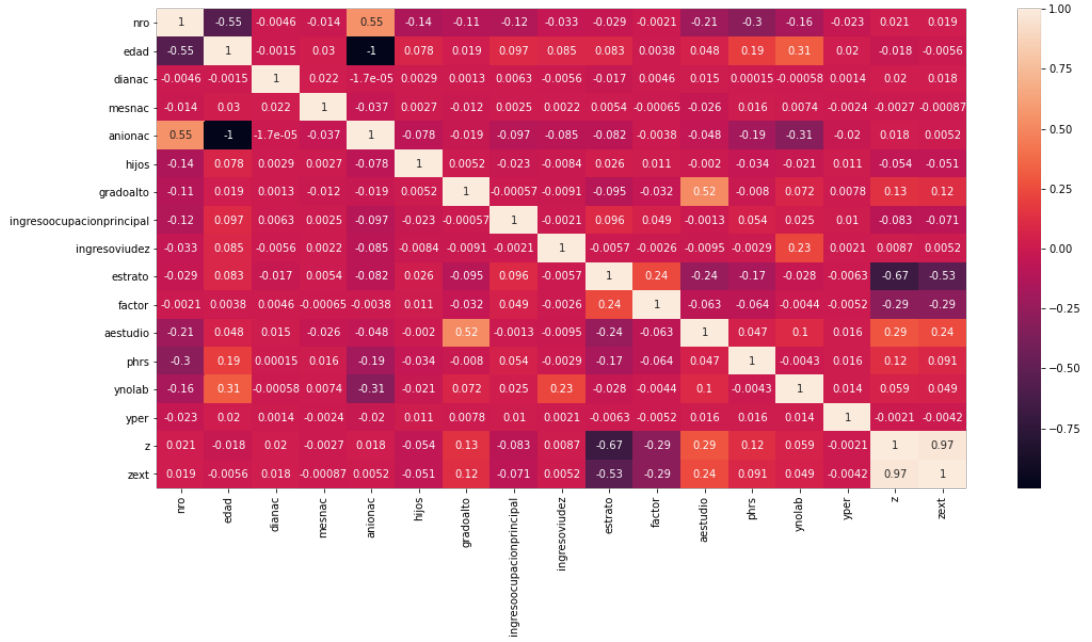

Fig. 1. Heatmap de matriz de correlación del dataset

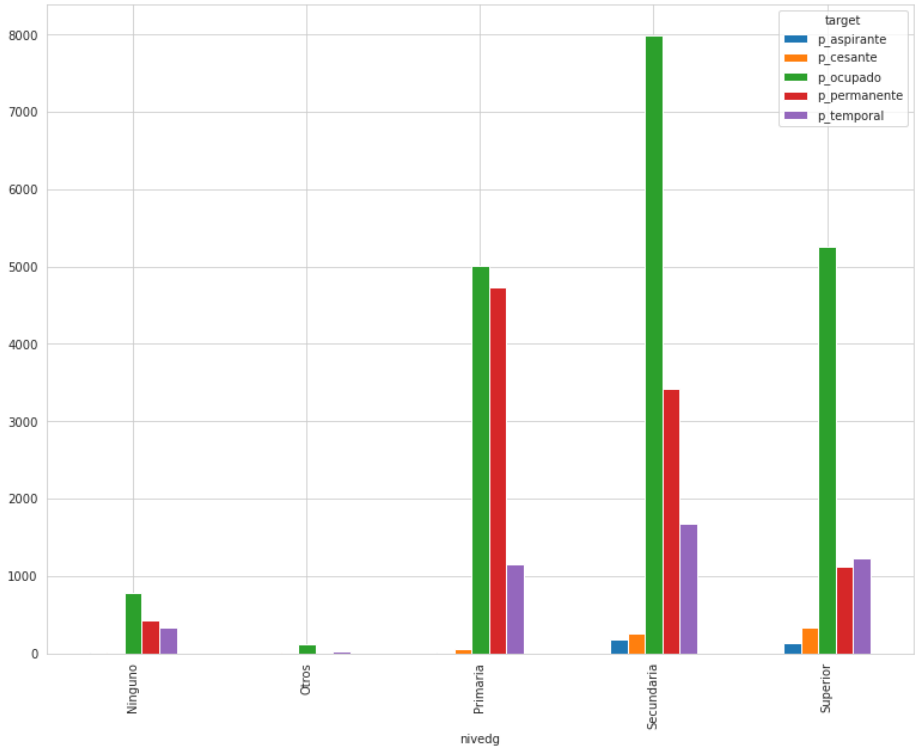

Fig. 2. Histograma del nivel de educación y la condición laboral

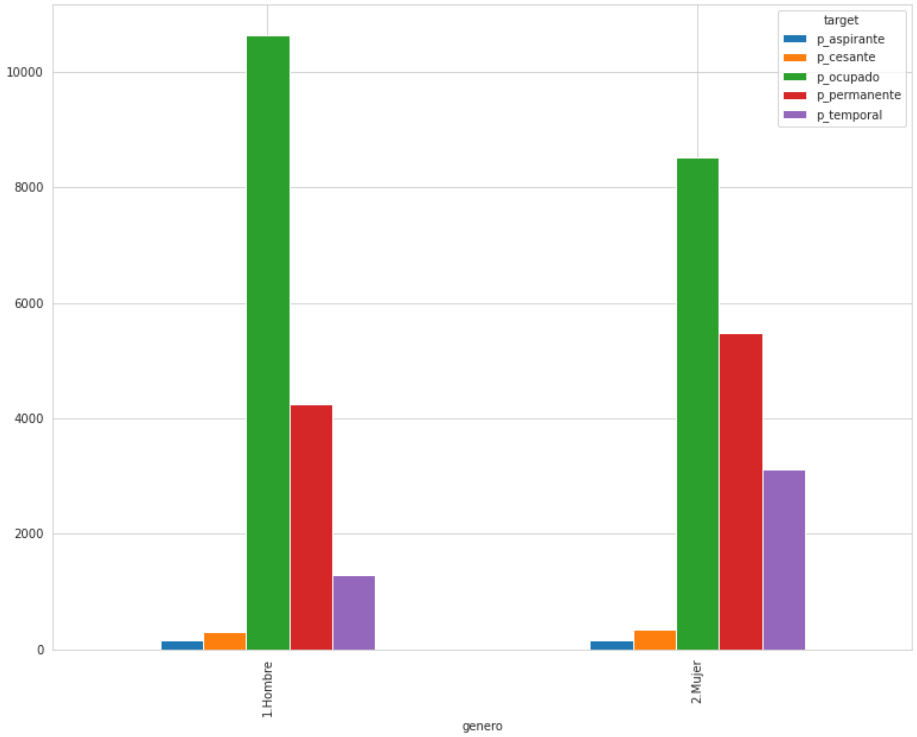

Fig. 3. Histograma sexo y la condición laboral

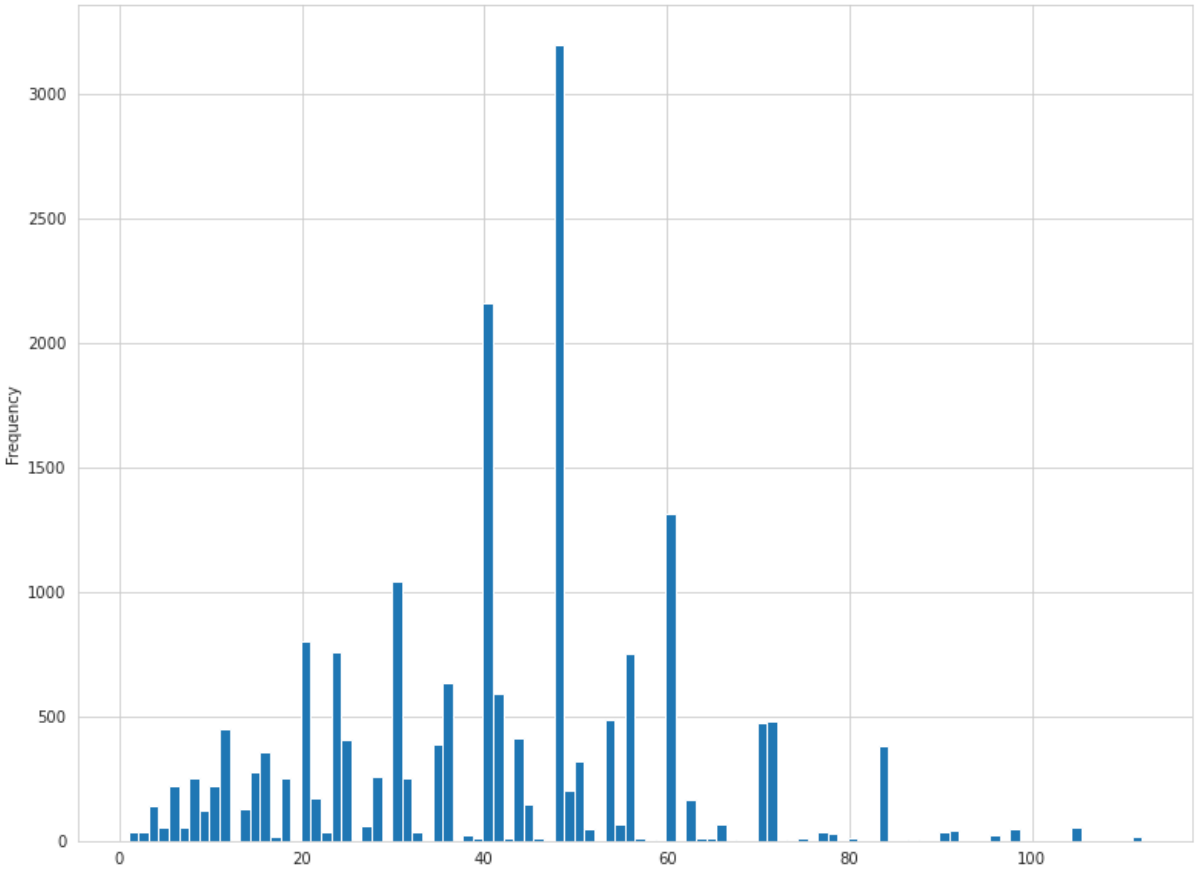

Fig. 4. Histograma horas trabajadas

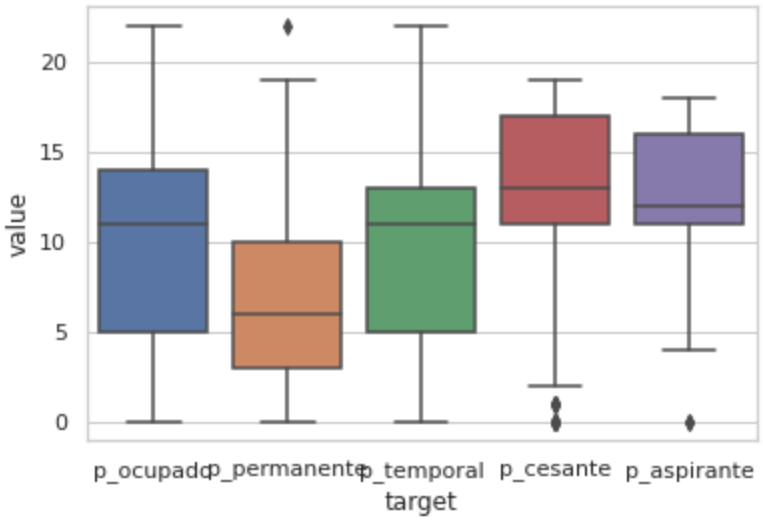

Fig. 5. Boxplot años de estudio y condicion laboral

Diagrama de cajas, muestra la variabilidad la mediana y los cuartiles y las medidas de asimetria. de los años de estudio por cada nivel de la variable target y muestra 4 outliers

4.2 Árboles de Decisión

Como primer parso del estudio aplicaremos Arboles de Decisión que nos ayudara a clasificar la condicion laboral de las personas en funcion a los predictores, mediante metodos euristicos, segun la informacion obtenida sin utilizar metodos estadisticos
se tomo en cuenta tres variables predictoras: ***edad, hijos, años de estudio.***

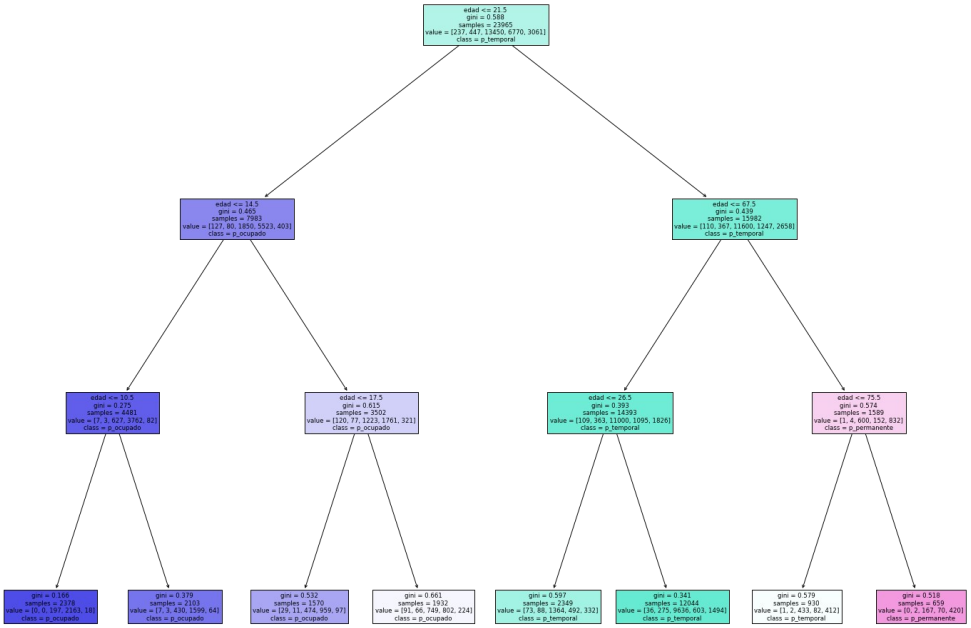

Fig. 6. Arbol de decisión

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.

4.2 Naive Bayes

Otra alternativa popular es el algoritmo Naive Bayes, se deberá usar el Gausiano pues este permite trabajar con mas de dos clases en las variables target.

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)***

Tabla 2. Dataset categorizado Naive Bayes

	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, eso signitica que de cada 100 persona el modelo predice 90 a una categoria de condicion laboral. A diferencia del algoritmo Naive Bayes Multinomial que solo nos dio un una precisión de 62%.

4.2 Regresión Logística Multinomial

La regresión logística multinomial se utiliza cuando la variable dependiente en cuestión es nominal (equivalente categórica, lo que significa que puede incluirse en una de un conjunto de categorías que se excluyen) y para los cuales hay más de dos categorías.
Las variables consideradas para este modelo fueron ***edad, genero, hijos, tipohogar, cobersalud, razontrabaja, relacionjefehogar, ocupacion, ingresoocupacionprincipal, aestudio y la condicion laboral.***

Tabla 3. Dataset categorizado Naive Bayes

	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

Los resultados fueron los siguientes:

Coef de determinación: 0.910
Cross-validation R^2: 0.912
Precision: 0.910
Proporción de verdaderos positivos : 0.910
Contribución de la precision ponderada(F1 score): 0.910

El modelo predice en todos los casos al 91, es decir de cada 100 personas podemos categorizar a 91.
El cálculo de la matriz de confusión se puede ver en las filas los datos observados y en las columnas las predichas, también se puede ver en la diagonal principal esta bien clasificado.

```
[[   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]]
```

Fig. 7. Matriz de confusión

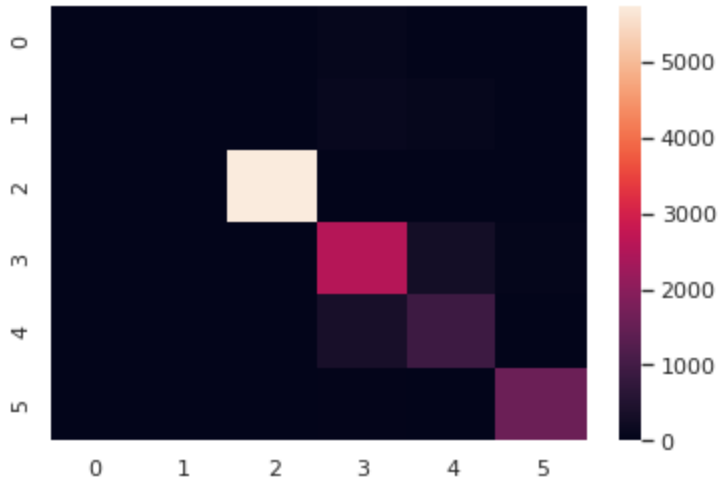

Fig. 7. Heatmap de la matriz de confusión

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

5 Conclusiones y trabajo futuro

Los resultados del modelo fueron buenos alcanzando 90 % de precisión ,es decir, de cada 100 persona el modelo predice 90 a una categoria de condicion laboral. Pero a diferencia con el algoritmo Naive Bayes Multinomial con el mismo dataset solo obtuvo 62 % de precisión.

La regresión logistica también dio buenos resultados alcanzando una precision de 91%, siendo también una buena alternativa para clasificar la condición laboral de las personas.

Antes de implementar algún algoritmo de clasificación para predicciones, se debe probar y ensayar con varios algoritmos evaluando el desempeño de cada uno de ellos sobre los datos de interés. Tomar decisiones basados en evidencia debe ser prioridad para las instancias privadas y gubernamentales, para ello, el uso de analítica predictiva es de mucha importancia ya que con el uso de datos, algoritmos estadísticos y técnicas de machine learning, se podría identificar probabilidades de resultados futuros basados en datos históricos.

ANEXOS