Aplicacion de metodos de clasificacion sobre la encuesta a Niños y adolescente que realizan actividad laboral de las gestiones 2016-2017

Primo L. Acho Cachaca

29 de junio de 2024

Resumen

Your abstract.

1. Introduccion

El bienestar y desarrollo de los adolescentes es un tema de vital importancia para cualquier sociedad, ya que estos jóvenes representan el futuro del pais. En Bolivia, la Encuesta a Niños y Adolescentes 2016-2017 realizada por el Instituto Nacional de Estadística (INE) proporciona una valiosa base de datos para el análisis de diversos factores que afectan a este grupo. Este artículo tiene como objetivo aplicar modelos de clasificación avanzados, tales como logit, probit, naive bayes y k-nearest neighbors (KNN), para identificar y entender los determinantes clave del bienestar adolescente en Bolivia.

Los modelos de clasificación elegidos permiten analizar tanto variables categóricas como continuas, proporcionando una comprensión más profunda y detallada de los datos. La elección de estos modelos se basa en su capacidad demostrada para manejar conjuntos de datos complejos y proporcionar predicciones precisas. Este análisis no solo busca contribuir al conocimiento académico, sino también ofrecer recomendaciones prácticas para mejorar las políticas públicas dirigidas a adolescentes.

2. Objetivos

- Identificar los factores y variables determinantes que afectan el bienestar, y desarrollo de los adolescente bolivianos.
- Comparar la eficacia de los diferentes modelos de clasificacion (logit, probit, naive bayes y KNN) en la predicción de resultados basados en los datos de la encuesta.

3. Motivación

La motivacion principal de este tema, radica en la importancia de la situacion de los niños y adolescentes de Bolivia, y asi, a traves del uso de los modelos avanzados, poder brindar informacion util, que pueda servir para su aplicacion en futuras encuestas, futuros trabajos, y poderse usar en el analisis de resultados.

4. Marco teorico

4.1. Encuesta a NNAs que realizan alguna actividad laboral o trabajan 2016 -2017(ENNA2016)

La encuesta presentada por la INE, de la cual este articulo tomo datos, puede obtenerse varios de los datos de la pagina del INE, esta misma proporciona estadisticas e indicadores socioeconomicos y demograficos de la poblacion boliviana en el marco del Modelo, necesarias para su formulacion,

evaluacion, seguimiento.Asi mismo el proposito de la encuesta fue pode cuantificar el numero de niñas, niños y adolescentes de 5 a 14 años de edad y las determinantes que inciden en la actividad laboral o trabajo.

La ENNA 2016 presenta un panorama completo sobre las condiciones de vida de al poblacion boliviana. La unidad de analisis para esta encuesta fueron los hogares de Bolivia, con unidades muestrales como los sectores Censales, Segmentos Censales, Viviendas, y la implementacion de las Unidades Primarias de muestreo (UPM) que concuerdan con los sectores censales o una agrupacion de ellos.

La metodologia aplicada para la recoleccion de la informacion es la Entrevista Directa, conducida por personal debidamente capacitado que visito las viviendas seleccionadas durante el periodo de recoleccion de informacion, utilizando una boleta multitematica que permite el estudio de los hogares.I.N.E. (2017) el cuestionarion en cuestion, toma las siguientes tematicas:

- Formacion educativa
- Condicion de actividad
- Ocupacion y Actividad principal
- Ingresos y Derechos laborales
- Seguridad, dignidad y salud en la ocupacion principal
- Ocupacion y actividad secundaria y derechos laborales
- Tareas domesticas del hogar PARTE B: Seguridad, dignidad y salud en las tareas domesticas del hogar
- Derechos de recreación y asociación

4.2. Modelos de Elección discreta Binaria

Muchas veces en el análisis econométrico se suelen usar variables binarias, un caso particular de estas variables, es cuando estas son las dependientes, es decir, el modelo tratará de explicar la probabilidad de ocurrencia de cierto evento, o bien, la probabilidad de la ausencia o presencia de cierta característica en las observaciones. Cuando se habla de este tipo de modelo se está interesado en predecir la probabilidad de ocurrencia de cierto evento en base a variables explicativas. Se puede definir la ocurrencia de un evento mediante un indicador de estructura binaria con valor igual a 1 cuando el evento ocurre (éxito) y 0 cuando no ocurre (fracaso). Por ejemplo:

$$\begin{aligned} &y = Situacion\ Laboral \\ &1\ si\ el\ individuo\ trabaja \\ &0\ en\ otro\ caso \end{aligned} \tag{1}$$

Los modelos de eleccion discreta binaria son aquellos modelos que explican la probabilidad de ocurrencia para el evento en la variable y, condicionado por un conjunto de variables explicativas, denotado de la siguiente manera:

$$p_i = Prob(y_i = 1|x_i) i = 1, ..., n. \tag{2}$$

Debe notarse que como ɛ"solo toma valores 0 y 1, la disstribucion de ɛçondicional en "x.es la de bernoulli. Por lo tanto, si se denota $Prob(y_i = 1|x_i) = p_i$, entonces: $Pr(v_i = 0|x_i) = p_i$, y por lo tanto:

$$E(y_i|x_i) = 1 * p_i + 0 * (1 - p_i) = p_i : esperanzacondicionaldey_i \tag{3}$$

$$V(y_i|x_i) = p_i(1 - p_i) :' Varianza\ condicionada\ de\ y_i \tag{4}$$

4.2.1. Modelo lineal de probabilidad

Si tenemos el siguiente modelo:

$$y_i = \beta_0 + \beta_1 x_i + e \tag{5}$$

Donde x_i es una variable explicativa numérica cualquiera,y_i es una variable dependiente tipo binario. Como la variable y_i es binaria, este modelo se denomina Modelo Lineal de Probabilidad(MLP). La esperanza o el valor viene dado por la siguiente expresión:

$$E(y_i|xi) = \beta_0 + \beta_1 * x_i \tag{6}$$

Las probabilidades se distribuirán de la siguiente manera:

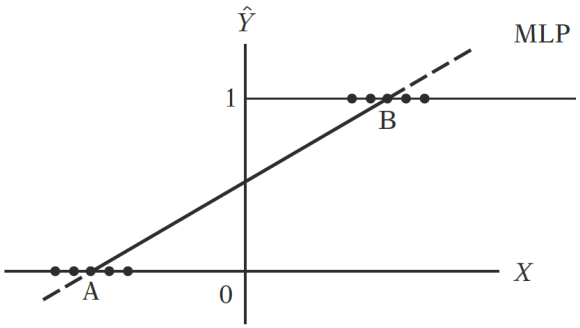

Figura 1: Modelo lineal de probabilidad

LA varianza como ya se explico tiene el siguiente valor $p_i : (1 - p_i)$, y los valores pronosticados de cada una de las probabilidades no saguran que la probabilidad estimada este comprendida entre 0 y 1, por lo que el modelo tendra valores absurdos. Con el avance de las herramientas informaticas en la econometria, este modelo es obsoleto en uso y los modelos que se usan son los logit y probit.Hosmer (2000)

4.2.2. Modelo LOGIT

Si tenemos el siguiente modelo donde la variable dependiente es binaria:

$$y_i = \beta_0 + \beta_1 x_1 + e \tag{7}$$

Como el modelo lineal de probabilidad no proporciona buenos estimadores, ecisten alternativas, una de estas es el modelo Logit, este tipo de estimacion usa a la funcion de distribucion logistica:

$$Prob(y_i = 1) = p_i = \frac{e^{z_i}}{1 + e^{-z_i}} = \Lambda(z) \tag{8}$$

Donde : $z_i = \beta_0 + \beta_1 x_i$ La funcion $\Lambda(z)$ es la distribucion logistica, esta funcion de distribucion servira para la estimacion de los paramteros en un modelo con variable de eleccion discreta binaria. Se puede notar que p_i no esta linealmente relacionado con z_i, esto quiere decir que no se puede estimar los parametros por MCO, pero se puede linealizar, haciendo lo siguiente, si p_i es la probabilidad de exito, la probabilidad de fracaso es :

$$1 - p_i = \frac{1}{1 + e^{z_i}} \tag{9}$$

Por consiguiente, se tiene:

$$\frac{p_i}{1 - p_i} = \frac{1 - e^{z_i}}{1 + e^{z_i}} \tag{10}$$

La ultima expresion es llamada la razon de las probabilidades en favor de tener exito, es decir, la razon de probabilidad para que y_i sea igual a uno o tambien llamada **Ratio de Odds**. Por ejemplo, si este

valor es 0.5 esto nos dira que la razon de la probabilidad de tener exito respecto de la probabilidad de fracasar es igual a $\frac{1}{2}$. si se toma el logaritmo natural de esta expresion se tendra:

$$\begin{aligned} L_i &= ln(\frac{p_i}{1-p_i}) = z_i \\ L_i &= ln(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_1 \end{aligned} \tag{11}$$

Caracteristicas del Modelo Logit

- Aunque las probabilidades se encuentren entre 0 y 1, esto no quiere decir que los logit estaran acotados de esta forma.
- Aunque L es lineal en x, las probabilidades en si mismas no lo son, a diferencia del MLP, donde las probabilidades aumentan linealmente con x.
- Si bien se ha indicado un modelo con una sola variable explicativa, el logit admite tantos regresores como lo requiera el modelo a estimarse.
- Si L, el logit, es positivio, significa que cuando se incrementa el valor de la(s) regresora(s), aumentan las posibilidades de que la regresada sea igual a 1 (lo cual indica que sucedera algo de interes). Si L es negativo, las probabilidades de que la regresada iguale a 1 disminuye conforme se incrementa el valor de x. Para expresarlo de otra forma, el logit se convierte en negativo y se incrementa en gran medida conforme la razon de las probabilidades disminuye de 1 a 0; ademas se incrementa en gran medida y se vuelve positivo conforme la razon de las probabilidades aumenta de 1 a infinito.Hosmer (2000)

4.2.3. Modelo PROBIT

En el modelo probit se especifica a traves de la siguiente funcion de distribucion acomulada normal:

$$F(z) = \Psi(z) = \int_{-\infty}^{z} \Psi(v)dv \tag{12}$$

Dpmde $\Psi(z)$ es la distribucion normal estandar:

$$\phi(v) = (2\pi)^{-\frac{1}{2}} exp(-\frac{z^2}{2}) \tag{13}$$

Por lo cual el modelo quedaria especificado de la siguiente manera:

$$y_i = \int_{-\infty}^{z} (2\pi)^{-\frac{1}{2}} exp(-\frac{z^2}{2})dv + e_i \tag{14}$$

4.2.4. Modelo Naive Bayes

El algoritmo Naive Bayes es una tecnica de clasificacion basada en el teorema de Bayes, con una suposicion de independencias entre predictores. Es simple pero eficas para grandes volumenes de datos. Naice Bayes es utilizado en tareas de filtrado de spam, diagnostico medico, y analisis de sentimientos. A pesar de su suposicion simplificadora, funciona bien incluso cuando la independencia entre caracteristicas no se mantiene completamente.

Naive Bayes para clasificacion Naive Bayes calcula la probabilidad de cada clase bajo la suposicion de independencia de variables, y clasifica una nueva observacion en la clase con la mayor probabilidad posterior. Matematicamente, la probabilidad de una clase dado un vector de caracteristicas (x) se calcula como:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{15}$$

donde:

- $P(A|B)$ es la probabilidad de A dado B,
- $P(B|A)$ es la probabilidad de B dado A,
- $P(A)$ y $P(B)$ son las probabilidades de A y B independientemente de cada uno

Para la clasificacion, Naive Bayes calcula la probabilidad de que un dato pertenezca a cada posible clase, dadas sus caracteristicas, y asinga el dato a la clase con la mayor probabilidad. La suposicion de que las caracteristicas son independientes entre si simplifica los calculos pero no se ajusta a la realidad.Manning (2008)

Ventajas de Naive Bayes

- **Eficiencia:** Rapido en terminos de tiempo de entrenamiento y prediccion.
- **Escalabilidad:** Maneja bien grandes volumenes de datos.
- **Simplicidad:** Daciel de implementar y entender. menos exigente en preprocesamiento

Desventajas de Naive Bayes

- **Suposicion de independencia:** La suposicion de independencia entre caracteristicas no siempre es valida.
- **Rendimiento**: Puede ser superado por modelos mas complejos en tareas con relaciones complejas entre caracteristicas.

4.2.5. Algoritmo KNN

El algoritmo de las K vecinas mas cercanas K-nearest neighbors(KNN) es un algoritmo de Machine Learning que pertenece a los algoritmos de aprendizaje supervisado simples y fáciles de aplicar que pueden ser utilizados para resolver problemas de clasificación y de regresión. La logica detras del algoritmo de las K vecinas mas cercanas es una de las mas sencillas de todos los algoritmos de Machine Learning supervisados:Cover (1967)

- **Etapa 1:** Seleccionar el numero de K vecinas
- **Etapa 2:** Calcular la distancia

$$\sum_{i=1}^{n} |x_i - y_i| \ \textit{distancia euclideana} \tag{16}$$

$$\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \ \textit{distancia de manhattan} \tag{17}$$

- **Etapa 3:** Tomar las K vecinas mas cercanas segun la distancia calculada
- **Etapa 4:** entre las vecinas, contar el numero de puntos en cada categoria.
- **Etapa 5:** Atribuir un nuevo punto a la categoria mas presente entre las k vecinas
- **Etapa 6:** El modelo esta listo.

5. Descripción de la base de datos

Para la aplicacion de los modelos de clasificacion anteriormente descritos, se hara uso de la encuesta ninios ninias y adolescentes que realizan actividad laboral o trabajan de la gestion 2016-2017. La base de datos proporcionada por el INE cuenta con una muestra de 10.488 registros, y 212 variables. dichas varias toman en cuenta datos como formacion educativa, datos basicos de la persona, y varias referente al estado actual de la empleabilidad respecto a ingresos, seguridad y dignidad, ocupacion y actividades en la poblabion mencionada(ninios ninias y adolescentes).
Como bien lo menciona el INE en los reportes correspondientes a este cuestionario, la muestra fue elegida de acuerdo a las unidades muestras tomadas del censo, es decir los sectores censales, segmentos censales. y los datos obtenidos fueron realizados de una encuesta multitematica, que permite asi mismo la correcta extraccion de la informacion

6. Metodologia

El lenguaje utilizado para el estudio de esta base de datos es el lenguaje R, que nos brindara paquetes y herramientas que hara util la extraccion de informacion, y la aplicacion de los modelos de clasificacion mencionados en la teoria.
En la practica para poder utilizar los modelos de clasificacion *Logit, Probit, NaiveBayes*, debemos contar con los siguientes criterios minimos:

- Definir una variable dependiente y binomial (de dos estapas i.e. 1 y 0)
- Tener un conjunto de covariable que nos ayudan a poder clasificar los datos respecto de y
- Contar con dos conjuntos de datos, uno de entrenamiento y otro de prueba que nos ayudara a ver que tan efectivo es el modelo

Para nuestro caso definiremos la variable y de la seccion de preguntas de la encuestas *ENNA 2016* que nos ayudan a verificar si el ninio ninia o adolescente alguna vez a trabajado, tomando en cuenta los siguientes criterios: Si en NNA ha trabajado alguna vez, si el NNA trabaja actualmente, y por ultimo si este trabajo es remunerado o no. dichas preguntas son tomadas de las secciones B y C del cuestionario y se toman las principales del conjunto de variables.
La definicion de covariables se las tomara del siguiente conjunto de datos tomados de la encuestas:

- Departamento
- Area: Rural o Urbana
- Edad: el rango de edad en el que se encuentra
- Si el NNA sabe leer y escribir
- Ultimo curos vencido del NNA
- Si esta cursando algun curso el anio de la encuesta
- tiene tiempo libre para dedicarse a otras actividades
- Si tiene impedimentos para realizar sus actividades cotidianas
- Que tipos de impedimentos no le dejan realizar sus actividades

Por ultimo como contamos con una extensa base de datos, se optara por usar la mitad de los datos, para usarlos como datos de entrenamiento y la otra mitad como datos de prueba. tambien dado que muchas de las preguntas no fueron contestadas, hay varias tuplas que incompletas, para lo cual se hara uso del metodo PPM(Predicted Mean Matching) para completar los valores faltantes.

modelo	accuracy Efectividad	Sensitivity Sensibilidad	Especificity Especificdad
Logit	0.988	0.99976	0.9259
Probit	0.988	0.99976	0.9259
Naive Bayes	0.9638	1.000	0.39
KNN	0.9623	0.9945	0.4893

7. Resultados y analisis

7.1. Resultados

Habiendo seguido los pasos detallados en la metodologia, *R* tiene distintas funciones capaces de aplicar los algoritmos antes mecionados, para el caso el modelo Logit, y Probit se uso la funcion **glm** del paquete *Caret y e1071*, para el modelo de *Naive bayes*, y para el modelo KNN se hace uso del paquete *class*.
Haciendo uso tanto de la base de datos de entrenamiento y de prueba, cuyo codigo se adjunta como enlace en este documento. de cada modelo, son dos datos que nos ayudaran a ver que tan efectivos fueron cada modelos ante la prediccion de los datos de prueba. estos son la accuracy(efectividad), sensitivity(capacidad de prediccion de casos positivos) y specifity(capacidad de prediccion de los valores negativos)

7.2. Analisis

De los datos obtenidos de analisis, claramente vemos que los algoritmos Logit y Probit son los mas efectivos, prediciendo los datos de prueba proporcionados, igualmente extraidos de la encuesta. Ambos algoritmos tienen resultados iguales dentro de los datos proporcionados en la matriz de confusion, por tanto podriamos decir que para este conjunto de datos, y las covariables definidas ambos modelos son igual de efectivos.
En la practica podriamos deducir que mediante los valores seleccionados,es decir las covariables, y la variables y, son suficientes para de terminar si un NNA es propenso a tener o ha tenido un trabajo remunerado o no, durante su infancia. lo cual es util para determinar la situacion de cada NNA.

8. Conclusiones y Recomendaciones

8.1. Conclusiones

Del pequenio trabajo que se presento, se puede evidenciar, que es y varias encuestas son bastante utiles para poder determinar la situacion actual de varios sectores, como se realizo en este caso, de los NNAs, sin embargo el INE ha abarcado bastantes otras areas, que pueden ser igualmente explotadas, para hacer uso de metodos modernos de analisis de datos, tal cual ahora lo hemos realizado.
El uso de estos algoritmos, y otros metodos de analisis de datos son utiles y necesarios para poder realizar asi conclusiones mas acertadas, ya que como se realizo en este trabajo, se pueden determinar casos positivos y negativos, con un minimo de variables, que si bien tienen un pequenio margen de error, nos sirven para poder mejorarlok y obetener mejores resultados en un futuro

8.2. Recomendaciones

Para el desarrollo de este trabajo, no se cuenta con el conocimiento en pleno de cada pregunta que forma parte del cuestionario que brindo la base de datos final, por lo cual se recomienda en futuros trabajos, indagar un poco mas en ellos, para tener una clara perspectiva, de como realizar o determinar las covariables correctas, y asi mismo una correcta definicion de la variable y.
Tambien es menester seguir utilizando este y varios otro metodos para el analisis de datos de las futuras encuestas y otros programas realizados por el INE.

Referencias

Cover, . H. P. E., T. M. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory.*.

Hosmer, . L. S., D. W. (2000). Applied logistic regression. *Willey*.

I.N.E. (2017). Encuesta a nnas que realizan alguna actividad laboral o trabajan 2016 -2017(enna2016. *Reporte base*.

Manning, R. P. . S. H., C. D. (2008). Introduction to information retrieval. *Cambridge university Press*.