ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ciencias Naturales y Matemáticas

"Aplicación de un modelo para detectar posibles sesgos en una investigación de hogares por parte de encuestadores"

PROYECTO INTEGRADOR

Previo la obtención del Título de:

INGENIERO EN ESTADÍSTICA INFORMÁTICA

Presentado por:
Steeven Fabricio Pérez Jaime
Joan Fernando Mogro Ponce

GUAYAQUIL - ECUADOR
Año: 2021

DEDICATORIA

A Dios, porque aún cuando siento que todo se desmorona me da la claridad y fortaleza para seguir adelante. A mi madre Fátima Yesenia quien es el amor más grande que he tenido en mi vida y a quien se lo debo todo. A mi tío Holger quien me enseñó todo los valores y lo se requiere para ser un hombre de bien. A mi familia por su apoyo incondicional hacia mí, en toda etapa de mi vida.

Steeven Pérez J.

A Dios, porque en los momentos más difíciles me ha brindado sabiduría y fortaleza para continuar. A mi madre Patricia quien me acompañado durante todo el proceso brindándome todo su amor. A mi padre, Carlos Campozano quien me guio en como forjar mi carácter. A Florentino Pérez, quien ha sido una figura determinante en mi vida.

Joan Mogro P.

AGRADECIMIENTOS

Empiezo agradeciendo a Dios, por la resiliencia que me brinda y la fortaleza mental que sin duda son habilidades que me él me ha otorga como herramientas en mi desarrollo en general.

A mi madre, que siempre deseó verme en este momento e hizo todo lo que estuvo en sus manos para verme crecer y que nunca me falte nada en mi vida.

A mi tío, quien fue quien creó mis bases y principios para ser un hombre de bien, nunca olvido cada detalle de sus enseñanzas.

A mi abuela, quien estuve siempre conmigo desde niño, mis tíos, mis hermanos, mi abuelo, quienes nunca dudaron de el alcance que yo podría tener, y estuvieron siempre a mi lado ante cualquier situación.

A la ESPOL, sus docentes, quienes impulsaron mi progreso, mi tutora Mariela González y Sandra García, verdaderas profesionales quienes fueron parte importante en este proyecto.

Steeven Pérez J.

Quiero empezar agradeciendo a Dios, porque gracias a la fortaleza y sabiduría que me brinda he sabido sobrepasar las dificultades día a día.

A mi madre, que me ha apoyado en todo momento de mi vida brindándome su amor incondicional.

A mi padre, que nunca dudó de mis capacidades y desde donde esté siempre me guiará.

A mi abuela, que es parte fundamental de mi vida y seguramente está orgullosa de que haya culminado mis estudios.

A la ESPOL y los docentes que formaron parte del camino por formarme como profesional, especialmente a mi tutora Mariela González y Sandra García por su guía y confianza durante este proyecto.

Joan Mogro P.

DECLARACIÓN EXPRESA

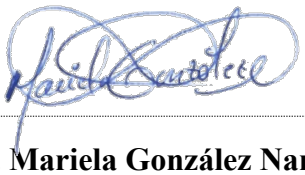
“Los derechos de titularidad y explotación, nos corresponde conforme al reglamento de propiedad intelectual de la institución; Nosotros, Steeven Fabricio Pérez Jaime y Joan Fernando Mogro Ponce damos nuestro consentimiento para que la ESPOL realice la comunicación pública de la obra por cualquier medio con el fin de promover la consulta, difusión y uso público de la producción intelectual."

Steeven Fabricio Pérez Jaime

Joan Fernando Mogro Ponce

EVALUADORES

Ph.D. Sandra García Bustos
PROFESOR DE LA MATERIA



Ph.D. Mariela González Narváez
PROFESOR TUTOR

RESUMEN

Cuando dentro de una población se busca estudiar los diferentes aspectos sociales y económicos además de los factores que los provocan, las encuestas a hogares son la más usuales fuentes de información para la obtención de datos socieconómicos y demográficos. La recopilación de estos datos comúnmente se lo realiza de forma manual y está sujeta a errores en la operación, esto puede ser un generador de sesgos que al final del proceso nos podrían ofrecer una información errónea de lo que queremos investigar. Por esto se ha hace precisa la implementación de técnicas que identifiquen estos sesgos para acciones inmediatas y que no afecten al resultado final del estudio. Una opción para la detección de estos, son métodos manuales y ordinarios como se ha venido haciendo, sin embargo con el fin de mejorar la productividad de esta validación de datos hemos optado por la utilización de técnicas de Machine Learning optimizando estos procesos con la oportunidad de mejorar la detección de estos errores muestrales. En este estudio se ejecutan técnicas de Machine Learning a partir del uso de datos de encuestas realizadas a hogares. Se hicieron pruebas con algoritmos de aprendizaje supervisado. También realizamos un análisis exploratorio con los datos donde incluimos una distribución de variables por encuestador para identificar en cuál de ellos no se sigue el patrón y por lo consiguiente se generaría el sesgo, todo esto previo al modelado. El mejor modelo fue el de Random Forest con una precisión del 0.8579 que quizás no fue la mejor entre todos los modelos, pero demostró una buena proporción entre la sensibilidad de 0.7530 y la especificidad del 0.9628, esto indica una adecuada capacidad de deteccion de verdaderos positivos.

Palabras claves: Población, Datos socieconómicos, Datos demográficos, Sesgo, Machine Learning, Error Muestral, Encuesta de Hogares.

ABSTRACT

When seeking to study the different social and economic aspects of a population, as well as the factors that cause them, household surveys are the most common sources of information for obtaining socioeconomic and demographic data. The collection of this data is commonly done manually and is subject to errors in the operation, this can be a generator of bias that at the end of the process could offer us erroneous information about what we want to investigate. Therefore, it is necessary to implement techniques that identify these biases for immediate actions and that do not affect the final result of the study. One option to detect them are the manual and ordinary methods as has been done, however to improve the productivity of this data validation, the use of Machine Learning techniques has been chosen, optimizing these processes with the opportunity to improve the detection of these sampling errors. In this study, Machine Learning techniques are executed from the use of data from household surveys. Supervised learning algorithms were tested. We also did an exploratory analysis with the data where we included a distribution of variables by interviewer to identify in which of them the pattern is not followed and therefore the bias would be generated, all this prior to modeling. The best model was the Random Forest model with a precision of 0.8579, which was perhaps not the best among all the models, but it showed a good ratio between the sensitivity of 0.7530 and the specificity of 0.9628, this indicates an adequate capacity to detect true positives.

Key words: Population, Socioeconomic Data, Sociodemographic Data, Bias, Machine Learning, Sample Error, Household Surveys.

ÍNDICE GENERAL

ÍNDICE DE GRÁFICOS

ÍNDICE DE TABLAS

ÍNDICE DE CUADROS

CAPÍTULO 1

1. INTRODUCCIÓN

En la actualidad resulta muy demandante la gestión de datos demográficos y socioeconómicos de los hogares e individuos que conforman un país. Este tipo de información se ha convertido de suma importancia para el análisis de políticas tanto económicas como sociales, gestión de programas, planificación de desarrollo y toma de decisiones. Para atender esta demanda, la mejor forma de obtener la información que se considera necesaria es mediante el desarrollo de entrevistas a un determinado número de personas usando cuestionarios cuidadosamente diseñados y validados de tal manera que se obtengan datos idóneos para su posterior análisis (Malhotra, 2004).

Las instituciones encargadas de recopilar este tipo de información frecuentemente recurren a las llamadas encuestas de hogares, por lo que esta práctica se ha convertido en uno de los mecanismos más importante de recolección de información sobre una población (CEPAL, 2021).

A pesar de que la implementación de este tipo de encuestas lleva a cabo un minucioso proceso de elaboración y capacitación para los encuestadores, siempre está latente una diferencia entre la información ofrecida por el entrevistado y la información real, lo cual es conocido como error de medición o sesgo (Groves, 1989).

El encuestador resulta ser un factor determinante que puede tener efecto en la respuesta de una pregunta, ya sea porque el encuestador no formule las preguntas tal como están redactadas en el cuestionario, no siga el salto de preguntas correctamente, entre otras cosas. Este tipo de inconvenientes pueden traer consigo

estimaciones erróneas para aquellos indicadores económicos, sociales y de salud que son generados a partir de este tipo de encuesta (Biemer, 2004).

En el presente estudio se desea implementar un modelo para la detección de errores en una investigación de hogares generados por parte de los encuestadores; sin embargo, no se está considerando las características de estos últimos.

1.1 Descripción del problema

Uno de los mayores dilemas en una investigación es la ocurrencia de errores en la toma o en el registro de datos. A estos errores se los denomina sesgos dado que inciden en la veracidad y precisión de los resultados, afectando de cierta forma la validez de la investigación en curso. De esta manera se puede representar la diferencia de lo que se está obteniendo con respecto a lo que se desea obtener (Mateu & Casal, 2003).

A través de toda la recolección y procesado de información se encuentran presentes los conocidos como errores aleatorios y errores sistemáticos, siendo este último incapaz de compensarse incrementando el tamaño de la muestra a diferencia del error aleatorio. Ninguna investigación está exenta de este tipo de errores por lo que es importante conocerlos para intentar evitarlos, minimizarlos o inclusive corregirlos de ser posible (Beaglehole, 2008).

Otro punto para tener claro es que los sesgos pueden producirse en cualquier fase del proceso de investigación; es decir: en la planificación, la conducción, el análisis, la presentación de resultados. Por lo tanto, la finalidad de conocerlos es, en un sentido amplio, poder determinar si influyen por exceso o por defecto en los resultados, y más concretamente tenerlos en cuenta a la hora de interpretarlos (Manterola & Otzen, 2015)

1.2 Justificación del problema

Un gran indicador de la importancia que poseen las encuestas actualmente es la gran cantidad de disciplinas que utilizan habitualmente este método de recogida de información, como podemos evidenciarlo en sociología, psicología, en la salud, estadística, economía, etc. Todo ello se traduce en investigaciones destinadas al conocimiento de los hábitos de los consumidores, el estudio de la personalidad, las habilidades educativas, la preocupación por la salud pública, los hábitos de alimentación, la medición de la económica, las expectativas de los consumidores, la estimación de ventas, conocimiento de la demanda de nuevos productos, etc.

La importancia que lleva la implementación de encuestas es sumamente significativa puesto que nos concede una idea de lo que la gente piensa, la situación en la que se encuentra, entre otras cosas. De esta manera se puede analizar la información recopilada y a partir de ella diseñar un plan estratégico para toma de decisiones o medidas. Por lo tanto, si los datos recopilados se ven sesgados, de cierta manera esto conllevara a un análisis inexacto, erróneo y decisiones equivocadas.

1.3 Objetivos

1.3.1 Objetivo General

Diseñar un modelo estadístico que permita detectar sesgos generados por encuestadores en una investigación de hogares en Ecuador, aplicando técnicas de machine learning.

1.3.2 Objetivos Específicos

- Estimar un modelo que detecte los sesgos automaticamente en casos oportunos por medio de modelos de clasificación en una encuesta en hogares.

- Reducir los errores de información a través del uso de regularizadores que son generados en investigaciones producto de errores sistemáticos.
- Evaluar los resultados conseguidos para la comprobación del modelo esperado.

1.4 Marco Teórico

En esta sección se describen los principales conceptos utilizados en este tipo de encuestas, y así entrar en contexto del estudio que se realiza al procedimiento de levantamiento de información en hogares, además de otras metodologías como son los modelos de clasificación que nos conducen al objetivo de la investigación. Es conveniente que nuestra data este sujeta a más de uno de estos modelos, ya que la idea es comparar y comprobar cual de ellos se ajusta de mejor manera, de tal forma que nos ofrezca la mejor acuracidad entre otros parámetros como la sensibilidad y la especificidad que son variables que tambien incluiremos en esta sección, ya que son indicadores claves en nuestro criterio al momento de seleccionar al mejor modelo aplicado. También se muestra la definicion de cada uno de ellos, tales como Support Vector Machine, Random Forest, Decission Tree, K-nearest Neighbors, Logistic Regression.

En la primera parte se refiere a la variable ingreso, nuestra unidad de observación que son los hogares, y también nuestra unidad de análisis que son las personas quienes cumplen con las categorias de ocupación de Patronos y Trabajadores por Cuenta Propia, en el segundo los modelos de clasificación, mientras que en el tercero el software al que dimos uso.

1.4.1 Formulario

Dentro de la encuesta de Hogares tenemos un cuestionario que se divide en 7 diferentes secciones, donde nos enfocaremos en los datos que nos ofrece la primera sección sobre los miembros del hogar, tales como hogar, sexo, edad, parentesco. Además de esto, nos focalizaremos en las secciones 2 y 3 que

contienen las características ocupacionales y los ingresos mensuales respectivamente, tomando de esta última los ingresos que provengan del trabajo de manera independiente, específicamente las categorías de Patrono y Cuenta propia dado que siendo una encuesta de Hogares, nuestra unidad de análisis son los hogares.

Las demás secciones en donde se incluyen: Aspectos Generales de los Empleados, Datos de la Vivienda, Índice de Confianza como consumidor e indentidad de género u orientación sexual no fueron consideradas para nuestros estudio.

1.4.1.1 Hogar

Se hace referencia a una unidad de convivencia, un grupo de personas que viven juntas, y comparten el domicilio desarrollando un ambiente familiar y privado, además de compartir los recursos económicos. En su mayoria los individuos dentro de un hogar tienen una relación de parentesco o de matrimonio.

1.4.1.2 Patrono

Es considerado Patrono la persona que trabaja sin ningún vinculo de dependencia, son sus propios jefes o únicos socios comerciales participantes de la empresa y tienen bajo su mando al menos una persona asalariada quien se mantiene de forma estable.

1.4.1.3 Trabajador por Cuenta Propia

Se trata de las personas quienes trabajan llevando a cabo su actividad dando uso para ellos únicamente su trabajo propio, no se subordinan a un Patrono ni tampoco se sirven de trabajadores asalariados, sin embargo pueden recibir soporte de familiares con rol de trabajadores pero sin remuneración. Son incluidos también los asociados de cooperativas de trabajo o empresas de gente que no se sirven de asalariados.

1.4.1.4 Ingresos

Los ingresos son utilidades que pueden ser tanto monetarias como no monetarias que se enlazan y producen como resultado un centro de ganancia-consumo. La manera en la que podemos distinguirlos es por el cómo fueron conseguidos, puede ser por un producto o un servicio.

1.4.2 Modelos de Clasificación

En esta parte se definen los modelos de clasificación que implementaremos en nuestro conjuntos de datos.

1.4.2.1 Support Vector Machine

Se trata de modelos que pueden generar sus clasificaciones o las regresiones de datos no lineales en función de la transformación de los datos de entrada a otros campos con mayores dimensiones. Support Vector Machine en cuanto a la clasificación, lo que busca es establecer una frontera de decisión que permita determinar las categorías, separando los datos de entrenamiento, con algunas observaciones del grupo de entrenamiento que son los vectores de soporte, con la idea de que la diferencia sea la más grande posible. Se podría decir que el principal uso de este modelo se da clasificaciones binarias.

A la frontera de desición podemos llamarla Hiperplano, ésta dependiendo de sus variables, puede tomar diferentes dimensiones. Si analizamos la función del hiperplano podremos notar que cualquier punto que pertenezca a ella tiene una característica importante, y es que al reemplazarlo en la ecuación tomará el valor de cero, mientras que si reemplazamos en la ecuación del hiperplano algunos de los puntos dentro de la categoria que representa una clase tomaría valores mayores a cero, y si realizamos la misma práctica con la otra clase, al reemplazar los puntos en la ecuacion del hiperplano, en este caso nos resultará un valor menor que cero.

Partiendo de esto, definimos el algoritmo para la clasificación de los datos. Primero, obtenemos la ecuacion del hiperplano calculando lo coeficientes, y luego ya con esta ecuación podemos reemplazar el dato que queremos clasificar en dicha ecuación y según el dato que queramos clasificar el signo del valor resultando del reemplazo nos indicará si es de una clase o de la otra.
Es posible indicar que la complejidad de este modelo, radica en encontrar el Hiperplano Óptimo, que se lo encuentra midiendo la distancia de los vectores mas cercanos y trazando una linea entre ellos, estos ultimos son los llamados vectores de soporte ya que ayudan a definir la curva que divide una clase de la otra.

En la actualidad Support Vector Machine tiene un uso extendido en combinacion con estructuras más complejas como redes neuronales o redes convolucionales.

1.4.2.2 Decission Tree

El modelo de Decission Tree o Árboles de Decisión, son modelos predictivos que son usados no solo en clasificación sino también en regresión, y cuya actividad se basa en la elaboración de reglas lógicas (partición de los datos en condiciones o en rangos) partiendo de los datos de entrada.
Al entrenar árboles de decisión, la idea es buscar acrecentar al máximo la ganancia de información cuando se crean las reglas lógicas que componen un árbol.
Se busca entrenar un modelo que sea capaz de determinar la categoria a la que pertenece un dato en particular, la principal idea es que el modelo aprenda a calcular una frontera de decisión que permita asignar un dato a una u otra categoría. Hoy en día el algoritmo mas usado para la clasificación por Decission Tree es CART (Classification And Regression Trees), este algoritmo, en el caso de clasificación, utiliza el criterio del índice de Gini para la división de un nodo de un subnodo. El índice de Gini se define de la siguiente manera:

$$GI = \sum_{i=0}^{c} Pi(1 - P_i)$$

Que también puede ser escrito como:

$$GI = 1 - \sum_{i=0}^{c} P_i^2$$

Donde sabemos que ***P*** es la probabilidad de la clase ***i*** en un total ***c*** de clases.

Representados todos los puntos u observaciones en un plano cada una con sus características, donde lo mas común de ver en los árboles de decisión son sets de datos con 3 o más características. La idea principal es encontrar una forma de separar los datos de una clase de la otra, es decir, calcular unas fronteras de decisión que permitan posteriormente clasificar nuevos datos en alguna de las categorías. La idea de la clasificación es que iterativamente se irán generando particiones binarias sobre la región de interés buscando que cada nueva partición genere un nuevo subgrupo de datos lo mas homogéneo posible, en este con el algoritmo de CART se apoyan estas subdivisiones bajo el criterio de Gini.

Todas las particiones binarias que se van dando, pueden verse también como un árbol de decisión en donde el nodo principal de donde parte, es llamado La Raíz, y es quien contiene la primera partición indicando si se cumple o no la condición, esto se replica generando otros nodos internos, y finalizando en las hojas que corresponde a regiones donde finalizarían las subdivisiones.

1.4.2.3 Random Forest

Este modelo de clasificación involucra la elaboración aleatoria de un gran número de árboles de decisión en un mismo conjunto de datos, y la decisión que se hace luego de la clasificación es considerada en función del cálculo del voto de la mayor parte de las predicciones dadas por cada árbol que forma parte del bosque.

Los bosques aleatorios, son usados normalmente para resolver los problemas que nos pueden dejar los arboles de decisión como el overfitting que demuestran un buen funcionamiento en el entrenamiento, mientras que cuando queremos hacer predicciones con nuevos datos no se ajusta de la misma manera.

Los bosques aleatorias preservan el bajo sesgo de los arboles de decisión pero además de eso, reducen su varianza, esto se logra al agregarlo 2 variantes a los árboles de decisión que son:

En primer lugar el entrenamiento ya no de uno, sino de varias arboles de decisión, puede ser entre decenas o cientos, y la aleatoriedad nos permite entrenar cada árbol con un subset diferente. Por lo tanto si se generase algún ruido, podría afectar a algunos árboles pero no a la totalidad del bosque.

En segundo lugar, al agregar los resultados para generar la predicción los árboles que no funcionen muy bien, no representarán un impacto significativo en el resultado final. Por lo que al combinar estos 2 elementos que son la aleatoriedad y la agregación se logra reducir la varianza de los árboles individuales.

1.4.2.4 K-Nearest Neighbors

Para este modelo de clasificación su entrenamiento consiste en el cálculo de la distancia que existen entre un dato y uno nuevo, ó también puede tratarse de un dato que se quiere destinar a una clase, con la mayor parte de las clases a las que corresponden sus k vecinos mas cercanos. K es considerado un parámetro dentro del algoritmo.

Este modelo memoriza las instancias de formación que luego se utilizarán como conocimiento para el momento de la predicción. K-NN demanda un costo elevado de memoria por el almacenamiento de grandes grupos de datos, resulta un coste computacional en el periodo de prueba debido a que una observación determinada requiere un agotamiento de todos los datos.

Suponiendo que queremos predecir un punto en un plano, procedemos a buscar punto K mas cercano, y luego debemos clasificar los otros puntos entorno a al que queremos predecir, para el voto mayoritario de sus vecinos K. Cada objeto vota por su clase y la clase que contenga mas votos es la que se tomara como la predicción.

Para poder encontrar los puntos similares mas cercanos, se encuentra la distancia entre puntos utilizando las medidas de distancia. La opción mas común es la distancia euclidiana definida de la siguiente manera:

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Sin embargo, existen otras medidas que dependiendo del caso podrían llegar a ser mas adecuadas como:

Distancia Manhattan:

$$\sum_{i=1}^{k} |x_i - y_i|$$

Distancia Minkowski:

$$[\sum_{i=1}^{k} (|x_i - y_i|)^4]^{\frac{1}{4}}$$

En general, el algoritmo consiste en calcular las distancias, encontrar sus vecinos mas cercanos y votar por las etiquetas, ya que esto último es lo que define el resultado de la predicción.

1.4.2.5 Logistic Regression

En este modelo de clasificación que además es un modelo de regresión, la funcionalidad que tiene es la de predecir variables categóricas. En otras palabras, lo que busca es determinar una relación entre una variable dependiente definida como un valor entero y determinado de clases, con respecto a otras variables independientes.
En general, dentro de la regresión logística se realiza el cálculo de probabilidades de acontecimientos de alguna de las clases que son parte del modelo tras la utilización de la función logística.

La Regresión Logística es modelo que nos ayuda a predecir clases binarias, por lo que nuestra variable objetivo será de naturaleza dicotómica. Este modelo escribe y estima la relación entre una variable binaria dependiente y las otras variables independiente.

Por otro lado, la regresión logística tiene el nombre de la función que se utiliza en el núcleo del método, la Función Sigmoide, que se estima de la siguiente manera:

$$p = \frac{1}{1 + e^{-y}}$$

Que es una curva que puede tomar valores entre 0 y 1, si la curva tienda a infinito positivo, se convertirá en 1, pero si pasa hacia el infinito negativo se convertirá en 0.

Por lo tanto, si la salida de la función sigmoide supera el 0,5 podríamos clasificar el resultado como SI, mientras que si es menor a 0,5 como NO. Esto para aplicación de este logaritmo como en detección de enfermedades graves en pacientes, o la clasificación entre si un correo es spam o no.

CAPÍTULO 2

2. METODOLOGÍA Y DATOS

Para este capítulo se indican los modelos que implementamos, como se definen y su aplicación en este estudio, además, del desempeño de cada uno a partir de nuestro conjunto de datos en una encuesta a hogares elaborada en el año 2021.

Previo al modelado, se contaba con una base con carencias de orden, y poco entendible, sin embargo, se trabajó en el tratamiento de datos y se dio uso de tres técnicas de balanceo para óptimos resultados de nuestros modelos, tales como Under-Sampling, Both-Sampling, y una relación de 80/20 aplicados a la data.

Luego de la aplicación de estos algoritmos, se procedió a considerar 4 principales valores asociados a los resultados del modelo, que son: la acuracidad, el intervalo de confianza, la sensibilidad y la especificidad. Estos valores son indicadores para nuestro criterio de selección del mejor modelo.

2.1 Datos y Variables

Para la realización de los cálculos pertinentes, se validaron datos provenientes de un total de diez bases de datos anonimizadas de una encuesta a hogares sobre el año 2021. Las cuales contienen 200 variables que se dividen en tres secciones: datos demográficos de la población encuestada, características ocupacionales e ingresos del hogar. Cabe mencionar que el conjunto total de datos nos proporciona un total de 30.568 observaciones de personas de 5 años y más.

2.1.1 Variables Discretas

Cuando nos referimos a variables discretas, hablamos de variables cuantitativas para las cuales es posible tomar un número finito o contable de valores. Para este estudio, en la determinación de la variable respuesta no fue considerada ningún tipo de variable discreta.

2.1.2 Variables Continuas

Se trata de variables cuantitativas que pueden expresar una conjunto infinito de valores o puede tomar cualquier valor dentro de un intervalo definido. Dentro de este estudio se adoptaron las siguientes variables de este tipo:

- Edad (variable medida en años).
- Ingreso (variable medida en USD americanos).
- Costos de Materiales (variable medida en USD americanos).
- Gastos de Operación (variable medida en USD americanos).
- Autoconsumo de las Personas (variable medida en USD americanos).

Para la variable Edad, dentro de nuestras observaciones se observó un mínimo de edad de 12 años, mientras que, el máximo de edad observado fue 94. Sin embargo, fueron consideradas únicamente personas con rango de edad de 15 años en adelante, ya que estas son consideradas Personas Económicamente Activas (PEA).

Las variables: Costos de Materiales, Gastos de Operación, Autoconsumo fueron consideradas para el cálculo del Ingreso de las personas, mientras que a partir de esta último, se realizó la estimación de la variable respuesta.

2.1.3 Variables Categóricas

Cuando hablamos de variables categóricas, nos referimos a variables cualitativas que constan de un número finito de propiedades o de categorías, y para esto no se requiere tener un orden. A continuación, las variables categóricas utilizadas:

- Provincia (Bolívar, Guayas, Los Ríos, Manabí, Santo Domingo, Santa Elena, Galápagos)

- Tipo de Vivienda (Casa o Villa, Departamento, Cuartos en casa de inquilinato, Mediagua, Rancho/Covacha, Choza, Otra)
- Género (M: Masculinio, F: Femenino)
- Parentesco (Yerno o Nuera, Padres o Suegros, Otros No Parientes, Otro Pariente, Nieto o Nieta, Jefe, Hijo/a, Cónyuge)
- Nivel de Instrucción (Superior Universitario, Superior No Universitario, Secundaria, Primaria, Postgrado, Ninguno, Educación Básica, Analfabetismo, Bachillerato)
- Ocupación (Cuenta Propia, Patrono)
- ID (Encuestadores) (con variable alfanumérica)

Para este estudio nuestra cobertura geográfica como la acabamos de indicar en nuestras variables categóricas, fue constituida por 7 provincias del Ecuador.

El tipo de vivienda nos especifica el lugar de residencia del hogar, y su situación respecto al lugar físico donde se encuentra.

Se consideró también el parentesco que se refiere al vínculo entre miembros dentro de un hogar. Se observó el nivel de instrucción de cada uno para conocer si esta variable influye en la variable respuesta. Su ocupación, categoría a tener en cuenta entre Cuenta Propia o Patrono.

La variable de ID son los códigos alfanuméricos que generan una identificación a cada encuestador, esto es muy importante puesto que es la idea base es evaluar el desempeño de los mismos.

2.2 Variable de Respuesta

La variable de respuesta consiste en el resultado obtenido de nuestro modelo,

según este estudio, la respuesta a la pregunta sobre si existe inconsistencias o no en el levantamiento de informacion por parte de los encuestadores, agrupación por clase, sea la mayoritaria o la minoritaria.

2.2.1 Clase Mayoritaria

Clasificación de la variable de respuesta con mayor cantidad de observaciones. Para este estudio corresponde a las observaciones que no presentaron ningún tipo de error.

2.2.2 Clase Minoritaria

Clasificación de la variable de respuesta con menor cantidad de observaciones. Para este estudio corresponde a las observaciones que presentaron error.

2.3 Métodos de Balanceo

2.3.1 Undersampling

Método de balanceo de datos que reduce la cantidad de observaciones de la clase mayoritaria, igualando el número de observaciones con respecto a la clase minorita.

2.3.2 Oversampling

Método de balanceo de datos que aumenta la cantidad de observaciones de la clase minoritaria, igualando el número de observaciones con respecto a la clase mayoritaria.

2.3.3 Bothsampling

Método de balanceo de datos resultante de una combinación entre los métodos de undersampling y oversampling. Reproduce las observaciones de la clase minoritaria y reduce las observaciones de la clase mayoritaria hasta llegar a cierto equilibrio.

2.3.4 Balanceo de Datos 80/20

Método de balanceo de datos que toma toda la clase minoritaria y se selecciona una muestra aleatoria de la clase mayoritaria de tal forma que la proporción de datos resulta en un 80% de las observaciones sin error y un 20% de las observaciones que presentaron error.

2.4 Criterios de Clasificación

2.4.1.1 Precisión

Nos referimos explícitamente a la precisión de la clasificación en sí. Se considera el número de predicciones correctas entre el número total de muestras de entrada.

$$Precisión = \frac{Número\ de\ Predicciones\ Correctas}{Número\ total\ de\ predicciones\ realizadas}$$

2.4.1.2 Sensibilidad

Este término es utilizado para la tasa de verdaderos positivos, correspondiente a la proporción de puntos de puntos de datos positivos que efectivamente son considerados como positivos, en relación a los puntos de datos positivos en su totalidad.

$$Tasa\ de\ Verdaderos\ Positivos = \frac{Verdaderos\ Positivos}{Falsos\ Negativos + Verdaderos\ Positivos}$$

2.4.1.3 Especificidad

Es la que nos indica la tasa negativa verdadera. Es la tasa de verdaderos negativos correspondiente a la relación de puntos de datos negativos que efectivamente son considerados como negativos, en relación a los puntos de datos negativos en su totalidad.

$$Tasa\ Negativa\ Verdadera = \frac{Verdaderos\ Negativos}{Verdaderos\ Negativos + Falsos\ Positivos}$$

2.4.2 Software

El software que usamos en el estudio es R, que se trata de una herramienta open source para manipulación de datos, elaboración de gráficos en base a los mismos, diferentes cálculos, entre otros. Todo esto aplicando un lenguaje de programación. La versión de R que usamos fue la 4.1.1. Los paquetes de los cuales dimos uso son:

Paquete caret

Incluye una serie de funciones unificadas provenientes de distintos paquetes, facilitando el uso de decenas de métodos complejos de clasificación y regresión. (Kuhn, 2019)

Una de sus funciones principales es la función train() que contiene todos los algoritmos de clasificación además de permitir la partición de datos de entrenamiento y datos de test.

Paquete Rose

Incluye funciones que permiten realizar diversos balanceos de datos, mediante la creación de muestras sintéticas (Menardi & Torelli, 2013).

Una de sus funciones principales es la función ovun.sample() que permite especificar el método de balanceo de datos que se quiera realizar y el tamaño de muestra del mismo.

2.5 Elaboración del Modelo

La sección estudiada del cuestionario presentó un total de 14531 observaciones que representaron a las personas que fueron catalogadas como cuenta propia o patronos. Para la obtención de la variable de respuesta, que determina si los datos que fueron tomados de una persona presentan algún tipo de inconsistencia, se utilizó los datos de ingresos, costos de materiales, gastos de operación y autoconsumo de

las personas. Las variables tomadas para formar parte de los distintos algoritmos de clasificación fueron variables sociodemográficas como: la provincia de residencia, ciudad en la que habita, tipo de vivienda en el que reside, género, edad, nivel de instrucción, parentesco o papel en la vivienda y ocupación de la persona.

Dado que la proporción de datos que presentaron algún tipo de inconsistencia fue muy reducida en comparación con los datos que están bien tomados, se realizaron distintos métodos de balanceo de datos para ser probados en los algoritmos de clasificación seleccionados.

2.5.1 Contribución de las variables para los modelos de clasificación

Dado el alto costo computacional que puede tomar la implementación de algoritmos de machine learning y sobre todo ante un volumen elevado de datos, la selección de variables que forman parte de los modelos resulta ser una parte importante del proceso. Ahorrando tiempo de ejecución para el programa, disminuyendo la varianza interna de los modelos, o sobreajustes en los mismos.

Por medio de la función VarImp() en R se puede obtener un valor de entre 0 y 100 que representa el aporte de la variable en cuestión para los modelos de clasificación. Siendo 0 la cota inferior que representa que la variable no tiene aporte alguno para el modelo y 100 la cota superior que representa una alta contribución de una variable para dicho modelo.

Cada modelo de clasificación recibe un aporte diferente por parte de las distintas variables, por lo cual no hay un “grupo” exacto de variables que puedan ser escogidas; sin embargo se puede llegar a notar un conjunto de variables que a través de los distintos métodos de clasificación resulten aportar de gran manera para la implementación de los modelos.

CAPÍTULO 3

3. RESULTADOS Y ANÁLISIS

En esta sección se realiza un análisis univariante exploratorio de las variables consideradas en el estudio, para luego hacer uno bivariante finalmente, se presentan los resultados obtenidos de los modelos estadísticos empleados en el estudio de errores en la encuesta a hogares.

3.1 Análisis Univariante

En la tabla 3.1 se observa que para la sección 3 del cuestionario analizado, se cuenta con un total de 7 provincias de las cuales se obtuvieron los datos. Siendo, la provincia de Guayas el lugar donde más datos se registran con un 50.35%, seguida por la provincia de Santa Elena con un 15.73%. Luego, las provincias de Santo Domingo de los Tsáchilas, Manabí, Los Ríos y Bolívar con unos porcentajes similares. Cuyos valores son 8.49%, 8.73%, 6.48% y 6.86% respectivamente.

Tabla 3.1 Provincia de residencia de las personas encuestadas.

Provincias	n	%
Bolivar	997	6.86%
Guayas	7316	50.35%
Los Rios	941	6.48%
Manabí	1268	8.73%
Galapagos	489	3.37%
Santo Domingo de los Tsáchilas	1234	8.49%
Santa Elena	2286	15.73%
Total	**14531**	**100%**

De la tabla 3.3 se concluye que hay más hombres registrados en la sección de cuenta propia o patronos, correspondiendo a un 59.78%. A diferencia de las mujeres que constituyen el 40.22% restante.

Tabla 3.2 Género de las personas encuestadas.

Género	n	%
Hombre	8687	59.78%
Mujer	5844	40.22%
Total	**14531**	**100%**

De la tabla 3.4 se observa que de todas las personas encuestadas para esta sección, aquellas que son consideradas como jefes del hogar representan un 57.42%, seguido por el cónyuge con 18.55% y el hijo/a con un 16.62%.Con porcentajes más reducidos se encuentran las personas con un parentesco de otro pariente, yerno o nuera, nieto o nieta, otros no parientes y padres o suegros. Con valores de 2.99%, 1.78%, 0.94%, 0.88% y 0.82%.

Tabla 3.3 Parentesco de las personas encuestadas.

Parentesco	n	%
Cónyuge	2695	18.55%
Hijo/a	2415	16.62%
Jefe	8344	57.42%
Nieto o Nieta	137	0.94%
Otro Pariente	434	2.99%
Otros no parientes	128	0.88%
Padres o Suegros	119	0.82%
Yerno o Nuera	259	1.78%
Total	**14531**	**100%**

En la tabla 3.5 se observa que el 36.69% de los cuenta propia o patronos son personas con una educación secundaria, valor similar de personas con una educación primaria de 32.88%, luego se encuentran las personas con un nivel superior universitario con un 12.98% y las personas con bachillerato con 6.08%. Las personas con una educación de centro de alfabetización, educación básica, superior

no universitario, postgrado o simplemente no recibieron ningún nivel de estudio, representan una parte pequeña de la población.

Tabla 3.4 Nivel de Instrucción alcanzado de las personas encuestadas.

Nivel de instrucción	n	%
Bachillerato	884	6.08%
Centro de alfabetización	6	0.04%
Educación Básica	224	1.54%
Primaria	415	32.88%
Secundaria	164	36.69%
Superior No Universitario	4778	1.79%
Superior Universitario	5331	12.98%
Postgrado	261	1.13%
Ninguno	2468	2.86%
Total	**14531**	**100%**

Con la tabla 3.6 se concluye que las personas que representan al grupo de cuenta propia tienen un 92.63% de la población con respecto a los patronos con un 7.37%.

Tabla 3.5 Ocupación de las personas encuestadas.

Ocupación	n	%
Cuenta Propia	13460	92.63%
Patrono	1071	7.37%
Total	**14531**	**100%**

De la tabla 3.7 se puede concluir que los errores para esta sección son mínimos con un 3.07%, mientras que las observaciones que han sido tomadas correctamente representan un 96.94%

Tabla 3.6 Errores generados en la toma de datos de las personas encuestadas.

Errores	n	%
Si	446	3.07%

No	14085	96.93%
Total	**14531**	**100%**

3.2 Análisis Descriptivo

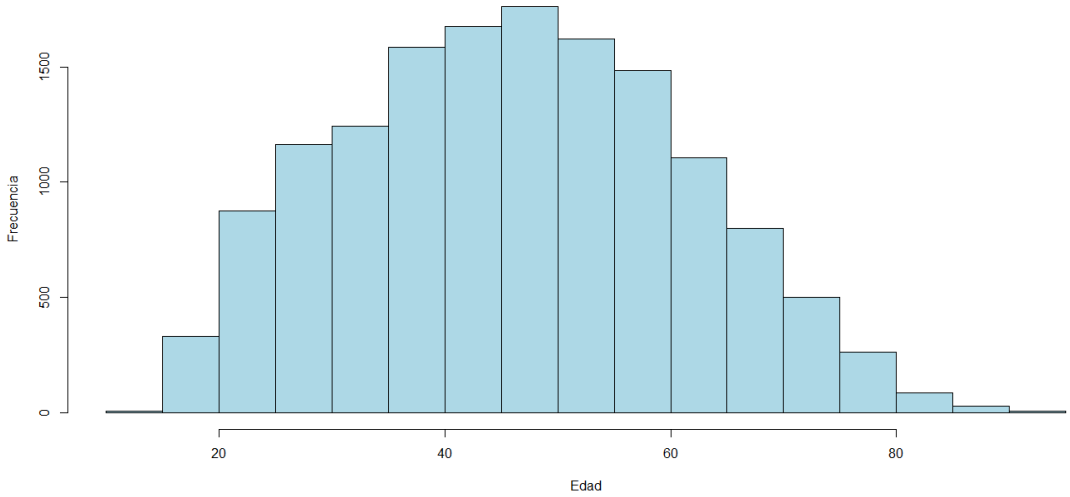
La tabla 3.8 muestra que el rango de edad definido para este estudio es mínimo de 12 años y máximo de 94 años. La mediana indica que el 50% de los datos encuentra como límite los 47 años. A su vez, la media representa que la edad promedio de las personas que pertenecieron al grupo de cuenta propia o patronos mostraron una edad de 46.82 años.

Tabla 3.7 Estadísticas descriptivas de la edad de las personas encuestadas.

Mínimo	Q1	Mediana	Q3	Máximo	Media	Varianza	Desviación
12.00	36.00	47.00	57.50	94.00	46.82	221.35	14.88

El grafico 3.1 muestra un histograma de frecuencias de la edad de las personas encuestadas con un ligero sesgo positivo.

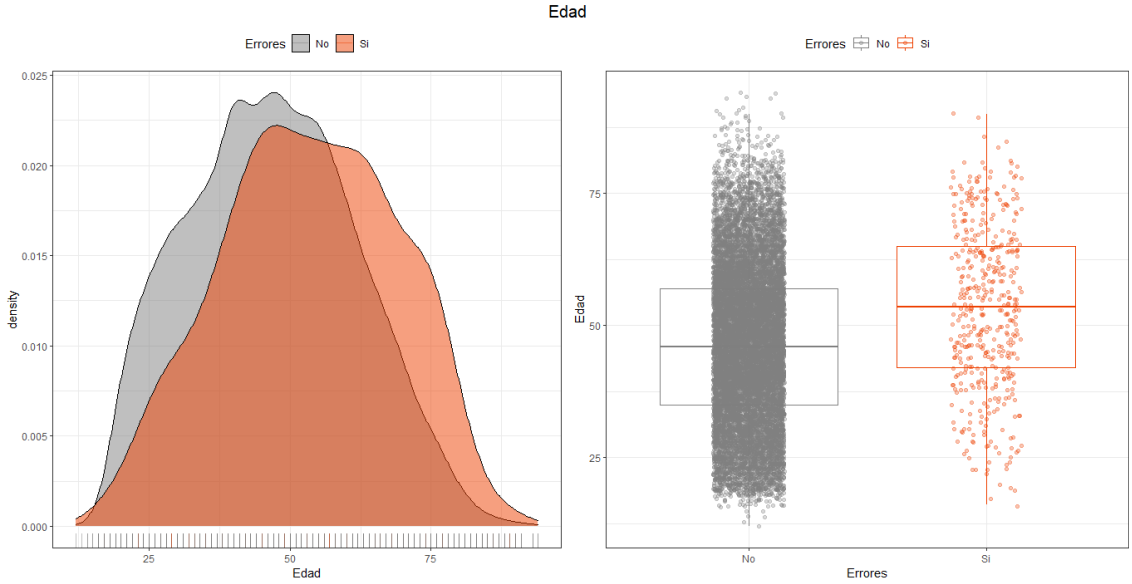
Gráfico 3.1 Histograma de frecuencias de la edad de las personas encuestadas.



3.3 Análisis bivariante

El grafico 3.2 muestra que la distribución de la variable edad de las personas para las que se cometió un error en la entrevista, es muy similar para las personas que no tuvieron error alguno en la entrevista, con una pequeña diferencia de rango para las personas de mayor edad. Además, se observa que el valor de las medianas según se haya encontrado un error en la entrevista es diferente para las personas con una entrevista correcta.

Gráfico 3.2 Gráfico de densidad y diagrama de cajas de la edad de los encuestados vs errores



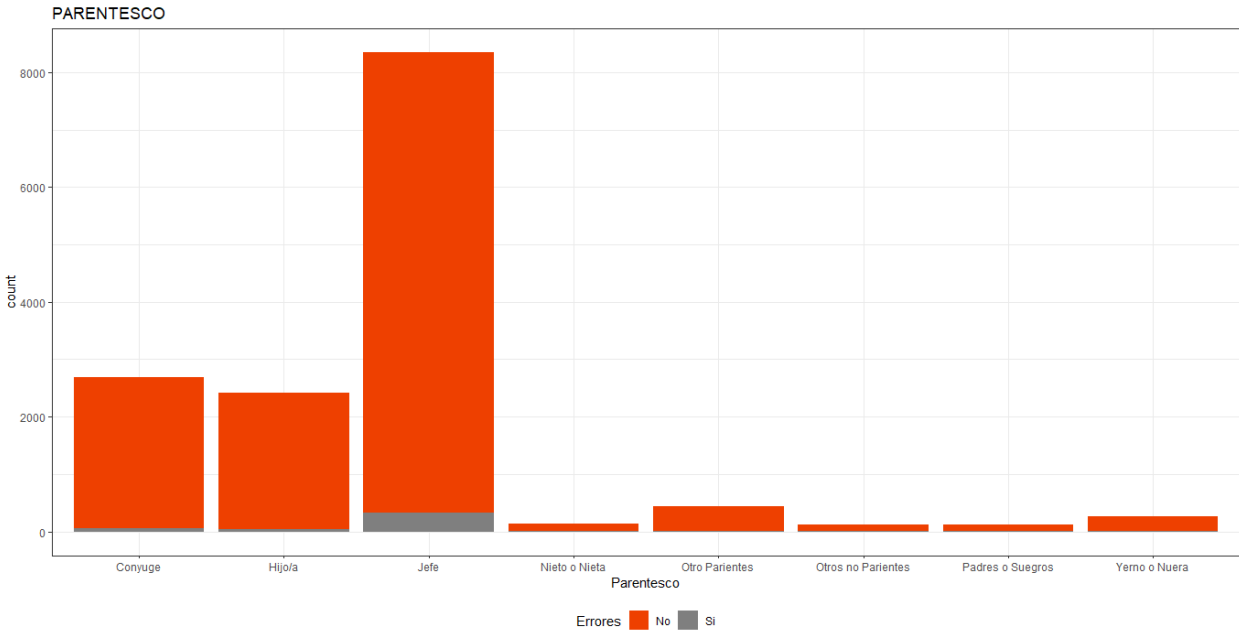
De la tabla 3.9 se puede observar que las personas con las cuales se comete algún error en la entrevista presentaron una media de edad más elevada con 53.39 años, en comparación con las personas que tuvieron una entrevista correcta y que presentaron una edad promedio de 46.61 años.

Tabla 3.8 Estadísticas descriptivas de la edad de las personas encuestadas según los errores.

Errores	Mínimo	Mediana	Máximo	Media
No	12	46.00	94	46.61
Si	16	53.59	90	53.39

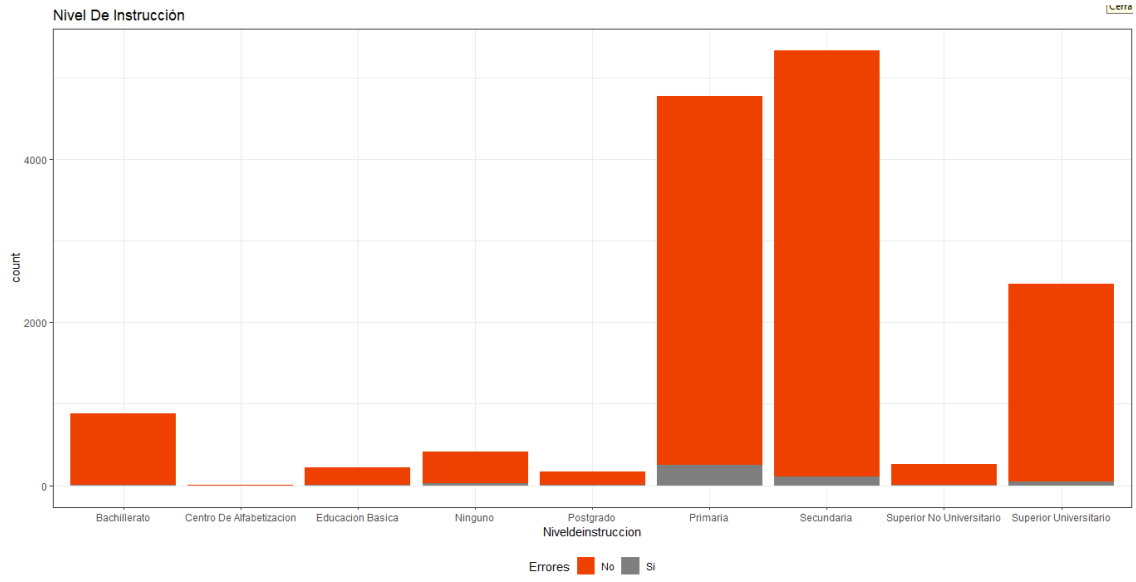
En el grafico 3.3 se observa que el tipo de parentesco registrado como jefe de hogar, es el que registra un mayor número de inconsistencias en las entrevistas.

Gráfico 3.3 Gráfico de barras de parentesco según los errores



En el grafico 3.4 se observa que aquellas personas que alcanzaron un nivel de estudio primario, son quienes registraron mayor tipo de errores en sus entrevistas, seguido por las personas con una educación secundaria.

Gráfico 3.4 Gráfico de barras de Nivel de instrucción según los errores



3.4 Implementación de algoritmos según balanceo

3.4.1 Algoritmos de clasificación con Undersampling

En la tabla 3.9 se puede observar que el modelo con mejor precisión y sensitividad de clasificación es el random forest, mientras que el método de Logistic Regression es el que presentó una especificidad más alta que los demás modelos.

Tabla 3.9 Comparación de los modelos según técnica de balanceo Undersampling

Algoritmos de clasificación					
	Support Vector Machine	**Random Forest**	**Decision Tree**	**KNN**	**Logistic Regression**
Precisión	0.706	0.8196	0.7327	0.6526	0.7973
95% CI	(0.6615, 0.7478)	(0.7808, 0.8541)	(0.6892, 0.7731)	(0.6065, 0.6966)	(0.7571, 0.8336)
Sensitividad	0.6854	0.84270	0.4944	0.68540	0.6742
Especificidad	0.7111	0.81390	0.79170	0.6444	0.8278

Para un balanceo de datos por medio de método undersampling el mejor modelo resultó ser el random forest con una precisión de 0.8196, sensitividad de 0.9326 y una especificidad de 0.7917

En el cuadro 3.1 se puede observar resultados adicionales del modelo escogido.

Cuadro 3.1 Resultados Completos del mejor modelo por medio de Undersampling (Random Forest)

```
               Accuracy : 0.8196
                 95% CI : (0.7808, 0.8541)
    No Information Rate : 0.8018
    P-Value [Acc > NIR] : 0.188

                  Kappa : 0.5606

 Mcnemar's Test P-Value : 4.171e-14

            Sensitivity : 0.9326
            Specificity : 0.7917
         Pos Pred Value : 0.5253
         Neg Pred Value : 0.9794
             Prevalence : 0.1982
         Detection Rate : 0.1849
   Detection Prevalence : 0.3519
      Balanced Accuracy : 0.8621

       'Positive' Class : NO
```

En el siguiente cuadro se puede observar las 11 variables que generan mayor aporte para el modelo escogido por medio de undersampling, siendo la ciudad con una codificación de 090159 la variable de mayor importancia, seguida por la edad del entrevistado y la provincia de residencia santa elena. También se puede notar la influencia del nivel de instrucción primaria y el género mujer para la implementación del modelo.

Cuadro 3.2 Variables con mayor contribución al modelo por técnica de balanceo Undersampling

Aporte	variable	Ranking
100.000000	Ciudad090150	1
76.140649	Edad	2
31.017776	Provincias24	3
26.286406	NiveldeinstruccionPrimaria	4
20.114046	Provincias23	5
15.824546	Ciudad240152	6
13.761029	GeneroMujer	7
9.146911	Viviendas09	8
9.078666	Ciudad230150	9
8.686215	ParentezcoConyuge	10
8.573208	Viviendas03	11

3.4.2 Algoritmos de clasificación con bothsampling

En la tabla 3.10 se puede observar que el modelo con mejor precisión y sensitividad de clasificación es el random forest, mientras que el método de Vector support machine es el que presenta una especificidad más alta en comparación al resto de modelos.

Tabla 3.10 Comparación de los modelos según técnica de balanceo Bothsampling

Algoritmos de clasificación					
	Support Vector Machine	**Random Forest**	**Decision Tree**	**KNN**	**Logistic Regression**
Precisión	0.8018	0.8263	0.6771	0.7461	0.7973
95% CI	(0.7619, 0.8377)	(0.788, 0.8602)	(0.6316, 0.7201)	(0.7032, 0.7857)	(0.7571, 0.8336)
Sensitividad	0.44944	0.92130	0.8202	0.55060	0.8018
Especificidad	0.88889	0.80280	0.64170	0.7944	0.6204

Para un balanceo de datos por medio del método bothsampling el mejor modelo resulta ser el random forest con una precisión de 0.8263, sensitividad de 0.9213 y una especificidad de 0.8028

Cuadro 3.3 Resultados Completos del mejor modelo por medio de Bothsampling (Random Forest)

```
                Accuracy : 0.8263
                  95% CI : (0.788, 0.8602)
    No Information Rate : 0.8018
    P-Value [Acc > NIR] : 0.1056

                   Kappa : 0.5699

 Mcnemar's Test P-Value : 9.796e-13

               Sensitivity : 0.9213
               Specificity : 0.8028
            Pos Pred Value : 0.5359
            Neg Pred Value : 0.9764
                Prevalence : 0.1982
            Detection Rate : 0.1826
      Detection Prevalence : 0.3408
         Balanced Accuracy : 0.8621

          'Positive' Class : NO
```

En el siguiente cuadro se puede observar las 11 variables que generan mayor aporte para el modelo escogido por medio de bothsampling, siendo las primeras 5 variables las mismas que generan mayor aporte para el modelo de undersampling aunque con otro valor de aporte. También se puede notar el aporte de una nueva variable como el nivel de instrucción secundario.

Cuadro 3.4 Variables con mayor contribución al modelo por técnica de balanceo Bothsampling

Aporte	variable	Ranking
100.000000	Ciudad090150	1
85.289096	Edad	2
23.461652	Provincias24	3
21.675895	NiveldeinstruccionPrimaria	4
17.832523	Provincias23	5
15.423675	Provincias09	6
14.982223	GeneroMujer	7
13.323883	ParentezcoConyuge	8
12.770430	Ciudad240152	9
11.445626	Ciudad240250	10
9.050558	NiveldeinstruccionSecundaria	11

3.4.3 Algoritmos de clasificación con balanceo 80/20

En la tabla 3.11 se puede observar que el modelo con mejor precisión fue el logistic regression con 0.8731, el random forest presentó una mejor sensitividad con 0.7530 y finalmente el método de vector support machine mostró una especificidad más alta con 0.997222

Tabla 3.11 Comparación de los modelos según técnica de balanceo 80/20

Algoritmos de clasificación					
	Support Vector Machine	**Random Forest**	**Decision Tree**	**KNN**	**Logistic Regression**
Precisión	0.8018	0.8579	0.8441	0.8151	0.8731
95% CI	(0.7619, 0.8377)	(0.7914, 0.8951)	(0.8072, 0.8764)	(0.7761, 0.8500)	(0.8387, 0.9024)
Sensitividad	0.0112	0.7530	0.4157	0.2023	0.573
Especificidad	0.9972	0.9628	0.9500	0.9667	0.9472

Para un balanceo de datos con estas proporciones de 80% para la clase mayoritaria y 20% para la clase minoritaria, el modelo escogido es el random forest debido a su alta capacidad para detectar los verdaderos negativos.

Cuadro 3.5 Resultados Completos del mejor modelo por medio del balanceo 80/20 (Random Forest)

```
               Accuracy : 0.8579
                 95% CI : (0.7914, 0.8951)
    No Information Rate : 0.8018
    P-Value [Acc > NIR] : 0.1056

                  Kappa : 0.5699

 Mcnemar's Test P-Value : 9.796e-13

            Sensitivity : 0.7530
            Specificity : 0.9628
         Pos Pred Value : 0.5359
         Neg Pred Value : 0.9764
             Prevalence : 0.1982
         Detection Rate : 0.1826
   Detection Prevalence : 0.3408
      Balanced Accuracy : 0.8621

       'Positive' Class : NO
```

En el siguiente cuadro se puede observar las 11 variables que generan mayor aporte para el modelo escogido por medio del balanceo 80/20, siendo la variable de edad la que genera mayor aporte para este modelo, seguido de variables como la ciudad, nivel de instrucción primera y el genero mujer. También se puede notar que toma en cuenta la participación de ciertos encuestadores y el tipo de vivienda que habitan los entrevistados.

Cuadro 3.4 Variables con mayor contribución al modelo por técnica de balanceo 80/20

Aporte	variable	Ranking
100.000000	Edad	1
56.524883	Ciudad090150	2
20.719997	NiveldeinstruccionPrimaria	3
20.158385	Provincias24	4
19.983624	GeneroMujer	5
13.646391	Ciudad240152	6
12.351192	Viviendas05	7
11.963944	ID4ER40BA	8
11.950469	ParentezcoConyuge	9
11.382700	ID14TH18RU	10
11.255820	Provincias09	11

CAPÍTULO 4

4. CONCLUSIONES Y RECOMENDACIONES

La metodología explicada en el capítulo 3 permitió la construcción de algoritmos de clasificación por distintos métodos de balanceo de datos. Por consiguiente, en este capítulo se exhiben las conclusiones y recomendaciones de los resultados obtenidos sobre los modelos de clasificación para detección de errores en encuesta hogares.

4.1 Conclusiones

- La aplicación de diferentes métodos de balanceo de datos permitió una comparación entre los algoritmos de clasificación a probar y permitió una mejor interpretación de los mismos teniendo en cuenta las ventajas y desventajas según el balanceo de datos aplicado.
- Los algoritmos de clasificación por medio de undersampling pueden sufrir una gran pérdida de información al reducir las observaciones de la clase mayoritaria e igualarlas a la clase minoritaria, provocando que los modelos quizá no llegan a una generalización de la información.
- Los algoritmos de clasificación por medio de bothsampling pueden presentar tanta pérdida de información al reducir la clase mayoritaria y un sobreajuste de datos al aumentar la clase minoritaria, presentando modelos aparentemente buenos pero que deben ser analizados minuciosamente.
- Los algoritmos de clasificación con una distribución de 80% para la clase mayoritaria y 20% para la clase minoritaria presentaron una estructura más realista de los datos a trabajar.
- El algoritmo random forest con el método de distribución 80/20 es el seleccionado de entre todos los modelos, tomando en cuenta criterios como el balanceo de los datos, precisión, especificidad y principalmente la sensitividad.

- Las observaciones que presentaron algún tipo de error, ocurrió principalmente en aquellas personas que tuvieron como ocupación cuenta propia y un nivel de instrucción primario
- La provincia del Guayas es aquella que registró el mayor número de personas clasificadas como cuenta propia y patronos, y a su vez presentó un mayor número de errores en sus observaciones en comparación con el resto de las provincias analizadas.

4.2 Recomendaciones

- Se debe conocer y seguir correctamente el flujo de preguntas que conforman el cuestionario, para poder analizar de forma idónea las distintas secciones que forman parte de él.
- Conocer la proporcionalidad de la variable de respuesta es de suma importancia para poder lidiar con datos desbalanceados y así evitar problemas al entrenar algoritmos de clasificación.
- Para evitar inconvenientes como el sobreajuste es necesario utilizar variables que sean realmente necesarias e importantes para los modelos a emplear.
- La selección de un modelo de clasificación no puede depender simplemente de la precisión del mismo ya que según la situación, puede resultar más importante el nivel de verdaderos negativos clasificados.

BIBLIOGRAFÍA

Beaglehole. (2008). *Epimediologia Basica* (Segunda ed.). Washington: Organizacion Panamericana de la Salud.

Biemer. (2004). *Measurement Errors in Surveys.* New York: John Wiley & Sons.

CEPAL. (2021). *Encuestas de ingresos y gastos de los hogares: experiencias en America Latina y el Caribe.* Santiago: CEPAL.

Groves, R. (1989). *Survey Errors and Survey Costs.* New York: John Wiley & Sons.

Kuhn. (27 de Marzo de 2019). *Github*. Obtenido de Github: https://topepo.github.io/caret/

Malhotra, N. (2004). *Investigacion de mercados Un enfoque aplicado* (Cuarta ed.). Mexico.

Manterola, & Otzen. (17 de Julio de 2015). *Scielo Analytics.* Obtenido de Scielo Analytics: https://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0717-95022015000300056

Mateu, & Casal. (2003). Los sesos y su control. *Revista de epidemiología y medicina preventiva*, 15-22.

Menardi, & Torelli. (2013). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, 92-122. Obtenido de ROSE: Generation of synthetic data by Randomly Over Sampling.