Universidad Cenfotec

Maestría en Tecnología de Bases de Datos

Documento final de Proyecto de Investigación Aplicada 2

Diseño e implementación de un modelo de aprendizaje automático para la predicción de la probabilidad que tiene un sujeto de conseguir empleo según las variables sociodemográficas con énfasis en la variable educación título

Ing. Esteban Hernández Chavarría

Agosto, 2021

Declaratoria de derechos de autor

Dedicatoria y Agradecimientos

Gracias a Dios, por la salud y vida, por las fuerzas en los momentos duros y por las alegrías de los momentos final de este proyecto.

Gracias a cada uno de los miembros de mi familia, a mi padre y a mi madre por siempre confiar en mi y no permitir que bajara los brazos en ningún momento. A mis dos hermanos, los cuales son mi mayor orgullo por saber esperar y apoyar

Gracias a mi novia, por la paciencia y la tolerancia que me brindo a lo largo de toda la carrera universitaria.

Y por supuesto, gracias al profesor Diego Alonso, por ser un ejemplo a seguir, por su orientación y por sus valiosas enseñanzas.

Universidad Cenfotec
Carrera de Postgrado
Maestría en Tecnología de Bases de Datos

TRIBUNAL EXAMINADOR

Este proyecto fue aprobado por el Tribunal Examinador de la carrera: **Maestría en Tecnología de Bases de Datos,** requisito para optar por el título de grado de **Maestría,** para el estudiante: **Hernández Chavarría Esteban Enrique.**

DIEGO ALONSO ALFARO BERGUEIRO (FIRMA)
Firmado digitalmente por DIEGO ALONSO ALFARO BERGUEIRO (FIRMA)
Fecha: 2021.07.06 08:46:53 -06'00'

MBD Diego Alfaro Bergueiro
Tutor

JOSE IGNACIO BRENES VILLARREAL (FIRMA)
Firmado digitalmente por JOSE IGNACIO BRENES VILLARREAL (FIRMA)
Fecha: 2021.07.07 08:35:01 -06'00'

M. Sc. José Ignacio Brenes Villareal
Lector 1

IGNACIO TREJOS ZELAYA (FIRMA)
Firmado digitalmente por IGNACIO TREJOS ZELAYA (FIRMA)
Fecha: 2021.07.07 10:04:27 -06'00'

M. Sc. Ignacio Trejos Zelaya
Lector 2

San José, Costa Rica, 05 de julio de 2021

Tabla de contenido

Resumen ejecutivo xvi

1. Introducción 1

1.1. Generalidades 1

1.2. Antecedentes del problema 1

1.3. Definición y descripción del problema 2

1.4. Justificación 2

1.5. Viabilidad 3

1.5.1. Punto de vista técnico 3

1.5.2. Punto de vista operativo 4

1.5.3. Punto de vista económico 4

1.6. Objetivos 4

1.6.1. Objetivo general 5

1.6.2. Objetivos específicos 5

1.7. Alcances y limitaciones 5

1.7.1. Alcances 5

1.7.2. Limitaciones 6

1.8. Marco de referencia organizacional y socioeconómico 7

1.8.1. Historia 7

1.8.2. Tipo de negocio y mercado meta 8

1.8.3. Misión, visión y valores 8

1.8.4. Políticas institucionales 9

1.9. Estado de la cuestión 10

1.9.1. Planeación 10

1.9.2. Identificación de los orígenes 13

1.9.3. Selección de fuentes 14
1.9.4. Extracción de la información 16
1.9.5. Análisis de resultados 29
2. Marco conceptual 31
2.1. El desempleo 32
2.2. Sociodemográfica 33
2.3. Modelos estadísticos 34
2.4. CRISP-DM 34
3. Marco metodológico 37
3.1. Tipo de investigación 37
3.2. Alcance investigativo 37
3.3. Enfoque 37
3.4. Diseño 38
3.5. Población y muestreo 38
3.6. Instrumentos de recolección de datos 38
3.7. Técnicas de análisis de la información 38
4. Análisis del diagnóstico 40
4.1. Entendimiento del negocio 40
4.1.1. Objetivos del negocio 40
4.1.2. Evaluación de la situación actual 42
4.1.3. Objetivos de la minería de datos 46
4.1.4. Criterio de éxito de la minería de datos 46
4.1.5. Plan del proyecto 47
4.2. Entendimiento de los datos 47
4.2.1. Recolección de los datos 47

4.2.2. Descripción de los datos 50
4.2.3. Exploración de los datos 54
4.2.4. Verificación de la calidad de los datos 69
4.3. Preparación de los datos 70
4.3.1. Selección de los datos 70
4.3.2. Integración de los datos 74
4.3.3. Construcción de los datos 74
4.3.4. Limpieza de los datos 78
4.4. Modelado 80
4.4.1. Selección de técnicas de modelado 81
4.4.2. Plan de pruebas 81
4.4.3. Construcción del modelo 84
4.5. Evaluación 111
4.5.1. Evaluación de los resultados 111
4.5.2. Modelos aprobados 128
4.6. Reconstrucción del modelo probabilístico 128
4.6.1. Predicción de la probabilidad 128
4.7. Revaluación de los resultados 141
4.7.1. Revisar el proceso 142
4.7.2. Determinar los próximos pasos 143
4.8. Implementación 144
4.8.1. Planeamiento de la implementación 144
4.8.2. Planeamiento del monitoreo y mantenimiento 144
5. Propuesta de solución 146
6. Conclusiones y recomendaciones 147

6.1. Conclusiones 147
6.2. Recomendaciones 149
7. Reflexiones finales 151
8. Trabajos a futuro 152
Referencias 153
Apéndice A 156
Código fuente programado en R 156
Código fuente para el análisis, limpieza, exploración y preparación de los datos 156
Código fuente para el modelado 178
Apéndice B 189
Diccionario de datos 189

Lista de ilustraciones

Ilustración 1: Relación entre el desempleo y los modelos estadísticos 31
Ilustración 2: Nube de palabras 32
Ilustración 3: Metodología CRISP-DM 36
Ilustración 4: Educacion_titulo por cantidad de observaciones 54
Ilustración 5: Cantidad de observaciones por la variable *Trabajo* 55
Ilustración 6: Cantidad de individuos por género 56
Ilustración 7: Histograma de observaciones por edad 57
Ilustración 8: Mapa geográfico de la cantidad de individuos por país de nacimiento 58
Ilustración 9: Cantidad de individuos por país de nacimiento 59
Ilustración 10: Individuos según la zona geográfica de Costa Rica 60
Ilustración 11: Cantidad de observaciones según la región geográfica de Costa Rica 62
Ilustración 12: Gráfica de la cantidad de individuos según la provincia de nacimiento 63
Ilustración 13: Cantidad total de observaciones con título de educación 64
Ilustración 14: Cantidad total de individuos que asisten a un centro educativo 65
Ilustración 15: Ilustración de la cantidad de individuos por nivel educativo 66
Ilustración 16: Relación entre el nivel educativo y el empleo 67
Ilustración 17: Niveles de habla para una lengua no nativa 68
Ilustración 18: Comparativo entre salario bruto y trabajo 72
Ilustración 19: Nombre y tipo del archivo digital 74
Ilustración 20: Sentencia usada para la integración de los datos 74
Ilustración 21: Creación de una nueva variable para la edad 75
Ilustración 22: Visualización de las dos variables por unir 76
Ilustración 23: Datos transformados de la variable *Tipo_seguro* 78
Ilustración 24: Datos transformados de la variable *Trabajo* 78
Ilustración 25: Limpieza de caracteres especiales 79
Ilustración 26: Filtro de edad 80
Ilustración 27: Creación del conjunto de pruebas y aprendizaje 81
Ilustración 28: Matriz de confusión 82
Ilustración 29: Gráfica de la curva ROC 84

Ilustración 30: Resumen de la variable *Permanencia_pais* 86
Ilustración 31: Gráfica de la variable *Permanencia_pais* 86
Ilustración 32: Resumen de la variable *Permanencia_intencion* 87
Ilustración 33: Gráfica de la variable *Permanencia_intencion* 88
Ilustración 34: Resumen de la variable *Permanencia_motivo* 88
Ilustración 35: Gráfica de la variable *Permanencia_motivo* 89
Ilustración 36: Resumen de la variable *Regimen_pension* 90
Ilustración 37: Gráfica de la variable *Regimen_pension* 90
Ilustración 38: Resumen de la variable *Educacion_titulo* 91
Ilustración 39: Gráfica de la variable *Educacion_titulo* 91
Ilustración 40: Conversión de observaciones a ficticias 92
Ilustración 41: Parametrización del modelo K-Vecinos más cercanos 92
Ilustración 42: Parametrización del modelo árboles de decisión 93
Ilustración 43: Parametrización del modelo bosques aleatorios 93
Ilustración 44: Parametrización del modelo máquina de soporte vectorial 94
Ilustración 45: Parametrización del modelo impulso adaptativo 94
Ilustración 46: Parametrización del modelo redes bayesianas 95
Ilustración 47: Parametrización del modelo redes neuronales 95
Ilustración 48: Resumen del modelo árboles de decisión 96
Ilustración 49: Ramas del modelo árboles de decisión 97
Ilustración 50: Gráfico del modelo árboles de decisión 97
Ilustración 51: Resumen del modelo árboles de decisión más *Educacion_titulo* 98
Ilustración 52: Gráfico del modelo árboles de decisión más *Educacion_titulo* 99
Ilustración 53: Selección de las variables del modelo bosques aleatorios 99
Ilustración 54: Resumen del modelo bosques aleatorios 100
Ilustración 55: Selección de variables del modelo bosques aleatorios más *Educacion_titulo* 100
Ilustración 56: Resumen del modelo bosques aleatorios más *Educacion_titulo* 101
Ilustración 57: Gráfica de importancia de variables 102
Ilustración 58: Gráfica de importancia de variables más *Educacion_titulo* 103
Ilustración 59: Resumen del modelo máquina de soporte vectorial 104

Ilustración 60: Resumen del modelo impulso adaptativo 105
Ilustración 61: Gráfico de importancia de las variables del modelo ADA 106
Ilustración 62: Resumen del modelo ADA con la variable *Educacion_titulo*..... 107
Ilustración 63: Gráfico de importancia de variables para el modelo con la variable *Educacion_titulo* 108
Ilustración 64: Resumen del modelo de redes bayesianas 109
Ilustración 65: Gráfico del modelo de redes neuronales de solo un nivel 110
Ilustración 66: Gráfica del modelo de redes neuronales con la variable *Educacion_titulo* 111
Ilustración 67: Código fuente para el modelado de los árboles de decisión..... 112
Ilustración 68: Matriz de confusión del modelo árboles de decisión con la variable *Educacion_titulo* 113
Ilustración 69: Matriz de confusión del modelo árboles de decisión sin la variable *Educacion_titulo* 114
Ilustración 70: Código fuente para el modelado del bosque aleatorio..... 115
Ilustración 71: Matriz de confusión del modelo bosques aleatorios sin la variable *Educacion_titulo* 115
Ilustración 72: Matriz de confusión del modelo bosques aleatorios con la variable *Educacion_titulo* 116
Ilustración 73: Código fuente para el modelado de la máquina de soporte vectorial .. 117
Ilustración 74: Matriz de confusión del modelo máquina de soporte vectorial con la variable *Educacion_titulo* 118
Ilustración 75: Matriz de confusión del modelo máquina de soporte vectorial sin la variable *Educacion_titulo* 119
Ilustración 76: Código fuente para el modelado del impulso adaptativo (ADA)..... 120
Ilustración 77: Matriz de confusión del modelo impulso adaptativo (ADA) con la variable *Educacion_titulo* 120
Ilustración 78: Matriz de confusión del modelo impulso adaptativo (ADA) sin la variable *Educacion_titulo* 121
Ilustración 79: Código fuente para el modelado de la red bayesiana..... 122

Ilustración 80: Matriz de confusión del modelo redes bayesianas con la variable *Educacion_titulo* 123
Ilustración 81: Matriz de confusión del modelo redes bayesianas sin la variable *Educacion_titulo* 124
Ilustración 82: Código fuente para el modelado de la red neuronal 125
Ilustración 83: Matriz de confusión del modelo redes neuronales con la variable *Educacion_titulo* 125
Ilustración 84: Matriz de confusión del modelo redes neuronales sin la variable *Educacion_titulo* 126
Ilustración 85: Código fuente para la generación de las predicciones probabilísticas. 129
Ilustración 86: Distribución de la *Educacion_titulo* por la probabilidad de 1 130
Ilustración 87: Distribución del estado conyugal por la probabilidad de 1 131
Ilustración 88: Distribución de la *Relacion_parentesco* por la probabilidad de conservar el trabajo 132
Ilustración 89: Distribución del *Tipo_seguro* por la probabilidad de conservar el trabajo 133
Ilustración 90: Distribución del *Nivel_educativo* por la probabilidad de conservar el trabajo 134
Ilustración 91: Distribución de *Educacion_asiste* por la probabilidad de conservar el trabajo 135
Ilustración 92: Distribución de un segundo idioma por la probabilidad de conservar el trabajo 136
Ilustración 93: Distribución del *Tipo_seguro* por la probabilidad de conservar el trabajo 137
Ilustración 94: Distribución de la *Provincia_nacimiento* por la probabilidad de conservar el trabajo 139
Ilustración 95: Distribución de la *Zona* por la probabilidad de conservar el trabajo..... 140
Ilustración 96: Distribución de la *Region* por la probabilidad de conservar el trabajo . 141

Lista de tablas

Tabla 1: Selección de estudios 15
Tabla 2: Formulario de extracción de datos 16
Tabla 3: Primera fuente de estudio investigada 17
Tabla 4: Segunda fuente de estudio investigada 18
Tabla 5: Tercera fuente de estudio investigada 20
Tabla 6: Cuarta fuente de estudio investigada 22
Tabla 7: Quinta fuente de estudio investigada 24
Tabla 8: Sexta fuente de estudio investigada 26
Tabla 9: Séptima fuente de estudio investigada 28
Tabla 10: Resumen de resultados de fuentes literarias 30
Tabla 11: Secciones del cuestionario 49
Tabla 12: Descripción de los archivos de datos 50
Tabla 13: Cantidad de observaciones por archivo 50
Tabla 14: Cantidad de variables por sección 51
Tabla 15: Descripción de las variables 52
Tabla 16: Variable con alto número de observaciones sin respuesta 71
Tabla 17: Lista de variables seleccionadas 73
Tabla 18: Diccionario de la variable Tipo_seguro 77
Tabla 19: Diccionario de la variable Trabajo 78
Tabla 20: Lista de modelos de minería de datos 85
Tabla 21: Cálculo de métricas para el modelo de árboles de decisión con la variable Educacion_titulo 113
Tabla 22: Cálculo de métricas para el modelo de árboles de decisión sin la variable Educacion_titulo 114
Tabla 23: Cálculo de métricas para el modelo bosques aleatorios sin la variable Educacion_titulo 116
Tabla 24: Cálculo de métricas para el modelo bosques aleatorios con la variable Educacion_titulo 117
Tabla 25: Cálculo de métricas para el modelo máquina de soporte vectorial con la variable Educacion_titulo 118

Tabla 26: Cálculo de métricas para el modelo máquina de soporte vectorial sin la variable Educacion_titulo 119
Tabla 27: Cálculo de métricas para el modelo impulso adaptativo con la variable Educacion_titulo 121
Tabla 28: Cálculo de métricas para el modelo impulso adaptativo sin la variable Educacion_titulo 122
Tabla 29: Cálculo de métricas para el modelo redes bayesianas con la variable Educacion_titulo 123
Tabla 30: Cálculo de métricas para el modelo redes bayesianas sin la variable Educacion_titulo 124
Tabla 31: Cálculo de métricas para el modelo redes neuronales con la variable Educacion_titulo 126
Tabla 32: Cálculo de métricas para el modelo redes neuronales sin la variable Educacion_titulo 127
Tabla 33: Evaluación de los resultados con la variable Educacion_titulo 127
Tabla 34: Evaluación de los resultados sin la variable Educacion_titulo 127

Resumen ejecutivo

El presente trabajo de tesis se centró en la preocupación que existe a nivel nacional en cuanto a la tasa de desempleo actual, dado que se manifiesta una tendencia al alza y las políticas implementadas por los líderes del país no contrarrestan el crecimiento de esta cifra. Por lo tanto, se buscó proponer mediante el análisis de una serie de variables sociodemográficas obtenidas del Programa Acelerado de Datos de INEC, mediante la Encuesta Continua de Empleo, I, II, III y IV trimestres del 2019, la clasificación de los habitantes costarricenses según el lugar de residencia y la predicción de la probabilidad de que la vivienda individual ocupada pueda variar su estado de ocupación laboral a partir de las diferentes encuestas continuas trimestrales, siguiendo la metodología de CRISP-DM y los diversos modelos de aprendizaje automático supervisado estudiados a lo largo de la carrera de maestría.

El desempleo es una cifra que afecta a toda la población por lo cual entre más herramientas se brinden para evitar el aumento de esta cifra, mejores oportunidades se pueden ofrecer a las familias costarricenses.

Palabras clave

Sociodemográfica, desempleo, subempleo, modelo de aprendizaje automático, CRISP-DM, Instituto Nacional de Estadística y Censos (INEC), encuesta, costarricenses.

1. Introducción

1.1. Generalidades

Actualmente en Costa Rica existe una institución autónoma de derecho público llamada Instituto Nacional de Estadística y Censo (INEC), creada gracias al Decreto n.° 7839 del 15 de octubre de 1998 y con el objetivo de generar la información estadística encomendada por ley y poner a disposición de la sociedad costarricense todos aquellos mecanismos estadísticos que ratifiquen la atención de las necesidades y requerimientos de información.

1.2. Antecedentes del problema

El desempleo es en la actualidad uno de los problemas sociales que más preocupa a los costarricenses al ser la raíz de muchos otros conflictos sociales como la delincuencia, pobreza, desigualdad, deserción escolar, drogadicción, prostitución, entre otros.

Según el Programa Estado de Nación (2019):

> […] en 2018 se agravó la falta de oportunidades para que las personas tengan recursos que les permitan llevar una vida digna. La evidencia disponible muestra una contracción en los ingresos de la mayoría de la población y un mercado laboral que genera empleos escasos y casi siempre de baja calidad, lo cual limita las posibilidades de lograr reducciones en el desempleo, la pobreza y la desigualdad. La situación es crítica para quienes viven fuera del Valle Central, quienes tienen baja escolaridad, las personas jóvenes o con discapacidad y, particularmente, las mujeres.

Los datos muestran que luego de aplicar de forma trimestral la Encuesta Continua de Empleo (ECE) del INEC, se comprueba que entre los meses de abril del 2018 y junio del 2019 las tasas de desempleo y subempleo aumentan casi dos puntos porcentuales, pasando de 10.3 % a 11.9 % en el caso del desempleo, de 7.2 % a 10.3 % para la tasa de subempleo y de 17.6 % a 20.9 % la tasa que suma la población desempleada con las personas que hoy cuentan con empleo, pero que están en busca de cambiar de trabajo.

En relación con lo expuesto, 11.9 % es la tasa de desempleo más alta reportada desde la medición realizada en el año 2010, lo que convierte al desempleo en un problema fundamental para la sociedad costarricense.

1.3. Definición y descripción del problema

El desempleo es una realidad que puede afectar a cualquier persona, ya sea que en la actualidad cuente o no con un puesto laboral, además, hay una serie de problemáticas relacionadas como la inseguridad, la deserción escolar, la prostitución y la drogadicción; por lo tanto, es una situación que debe ser de interés para la sociedad costarricense.

Al respecto, actualmente hay un aumento en la tasa de desempleo y subempleo a nivel nacional, lo que provoca una disminución significativa en los ingresos de la mayoría de la población costarricense.

Como parte del problema, es fundamental indicar que las políticas creadas con la finalidad de revertir el aumento en la tasa de desempleo no funcionan como se esperaba. La desaceleración en las exportaciones genera un desbalance dentro de la economía nacional, ocasionando una reducción en los ingresos de la industria productora y, por consiguiente, en la tasa de contratación de las empresas privadas.

A diario se observa en los noticieros más influyentes de Costa Rica que la cantidad de personas que recurren a las ferias de empleo es cada vez mayor y solo una pequeña fracción es contratada, en especial, dependiendo de las variables sociodemográficas que presenta la persona que solicita el empleo.

Asimismo, en cuanto a esta problemática, se encuentra una falta de herramientas que ayuden a la predicción de comportamientos y probabilidades según el análisis de datos pasados que son obtenidos luego de muchas sesiones de trabajo por cada uno de los encuestadores del INEC.

1.4. Justificación

El INEC ejecuta de forma trimestral la Encuesta Continua de Empleo (ECE) y de manera anual la Encuesta Nacional de Hogares (Enaho), lo que le permite mantener una base de datos sólida y actual del comportamiento de la población costarricense. Sin

embargo, se busca dotar a esta institución de una herramienta que le posibilite clasificar a la población costarricense según variables sociodemográficas y, posteriormente, con un modelo de minería de datos, sea capaz de predecir cuál es la probabilidad de ser desempleado.

El desempleo es un problema que afecta a todos los costarricenses, ya sea de manera directa, al momento de perder el trabajo, o indirecta, cuando la población activa económicamente es afligida por la delincuencia en las calles o por un aumento en la desigualdad, marcando aún más las clases sociales del país.

Por lo anterior, se debe buscar cualquier medio posible para generar empleo, o bien, como se indica en este proyecto de graduación, crear modelos matemáticos y probabilísticos que les permitan a las autoridades competentes presentir la existencia de posibles focos de desempleo antes de que sucedan y originar las políticas exactas y necesarias para evitar el aumento de las tasas de desempleo y subempleo.

A partir de esto, se pretende brindar una mejor calidad de vida a los ciudadanos de este país, pues cada uno tiene diversas necesidades, ya sea de alimentación, higiene, salud, educación, vivienda, recreación, ocio, entre otras, y merece subsanarlas de una forma digna, mediante la ejecución de labores que le retribuyan un pago económico.

1.5. Viabilidad

1.5.1. Punto de vista técnico

El INEC cuenta con un amplio catálogo de productos y servicios, donde se incluyen encuestas continuas y nacionales, estimaciones, proyecciones e indicadores que enriquecen una amplia base de información; además, son herramientas que se ejecutan de manera periódica, algunas de estas de modo trimestral o anual, logrando así mantener sus registros actualizados y apegados a la realidad nacional del momento.

El INEC, como ente productor de información, brinda la posibilidad de acceder a toda la información recopilada y almacenada por cada una de las herramientas antes descritas, para esto crea el Programa Acelerado de Datos, el cual facilita el acceso a la documentación metodológica de las operaciones estadísticas.

1.5.2. Punto de vista operativo

Desde el punto de vista operativo, el presente trabajo de investigación se realiza con la información obtenida de forma libre y gratuita del sitio de bases de datos documentadas llamado Programa Acelerado de Datos del INEC.

Cada una de las bases de datos almacenadas en el sitio del Programa Acelerado de Datos sigue normas de calidad y comparabilidad nacionales e internacionales, no obstante, cada una de las fuentes de datos utilizadas es analizada, depurada y clasificada con el objetivo de determinar la viabilidad de este proyecto.

Dicho lo anterior, se finaliza aclarando que el desarrollo de este trabajo de tesis no afecta el funcionamiento ni la operación diaria del INEC ni representa alguna afectación para sus departamentos internos.

1.5.3. Punto de vista económico

La información obtenida del Programa Acelerado de Datos es de carácter libre y gratuita, cualquier ciudadano está en su libre derecho de acceder, descargar y analizar la misma, por lo que no es necesario efectuar alguna inversión económica para obtener los datos. Así mismo, el presente modelo no es vendido a ninguna institución pública o privada, por lo que no se genera ganancia alguna.

La principal herramienta de trabajo es RStudio, la cual permite analizar e interpretar información en diversos lenguajes de programación, en este caso se utiliza el lenguaje de código abierto R.

1.6. Objetivos

Se selecciona la taxonomía de Bloom, específicamente en dirección a la dimensión cognitiva, debido a la claridad que presentan sus seis niveles de estudio, los cuales ayudan a entender y segmentar los objetivos de este proyecto final de graduación.

Así mismo, el objetivo principal se guía por una de las características más importantes de la ECE, que es la continuidad. La encuesta se desarrolla en más de 9 000 viviendas individuales habitadas y son visitadas cada trimestre, por lo que se registran sus repuestas cuatro veces al año; sin embargo, existe un nivel de rotación de la muestra de un 25 %.

1.6.1. Objetivo general

Implementar un modelo de predicción estadística de aprendizaje automático que permita predecir la probabilidad que tiene un sujeto que habita una vivienda individual ocupada de conservar su empleo según las variables sociodemográficas encontradas en las encuestas continuas trimestrales y la influencia positiva que puede ejercer la variable *Educacion_titulo* en dicha probabilidad.

1.6.2. Objetivos específicos

Definir los posibles modelos de predicción y clasificación estadística aplicables a las fuentes de datos existentes.

Definir cuál debe ser la variable por predecir por parte del modelo.

Distinguir entre las diferentes opciones de modelaje, según sus características y variables, y las que mejor se ajustan a la predicción de la probabilidad de ser un individuo sin empleo.

Analizar cada una de las variables cualitativas y cuantitativas que brindan las encuestas del INEC, escogiendo solo las variables con mayor relevancia y correlación.

Definir cuáles son los valores de cada variable que más importancia tienen en la generación de la probabilidad de encontrar empleo.

Analizar los resultados obtenidos por el modelo y aplicar las correcciones que sean necesarias.

1.7. Alcances y limitaciones

1.7.1. Alcances

El principal alcance de este proyecto de graduación es generar un documento escrito que muestre paso a paso el proceso de análisis de los datos y la búsqueda de las variables con mayor correlación en función de una variable predictora; además, contiene cada uno de los pasos del modelado, analiza cada uno de los resultados obtenidos y explica cuál de todos los modelos probados es el mejor para responder a la necesidad presentada en cada uno de los objetivos.

El alcance secundario es el desarrollo de un modelo de aprendizaje automático, que les permita a los encargados del análisis de datos del INEC crear nuevos mecanismos para interpretar la información y apoyar el diseño de políticas laborales.

La metodología implementada en el ciclo de vida del proyecto de graduación es la llamada CRISP-DM (Cross Industry Standard Process for Data Mining, por su sigla en inglés), se trata de una guía mundialmente utilizada que pretende brindar orden en el proceso de minería de datos, iniciando con la comprensión del negocio, pasando por el importante proceso de la comprensión de los datos y siguiendo por la preparación de los datos, el modelado, la evaluación de los resultados y la implantación del modelo.

Los datos se encuentran dentro de las ECE para los trimestres I, II, III y IV del año 2019, los cuales contienen más de 300 variables, que al final se convierten en 245 por temas de confidencialidad de la información de las personas encuestadas. En cuanto a esto, no todas las variables son posibles candidatas para formar parte del modelo, por lo que se realiza un descarte de aquellas con mucho ruido y con una correlación muy débil.

También uno de los alcances de este proyecto de graduación final es proporcionar un análisis de los pesos que tienen los distintos factores de las variables elegidas en los resultados obtenidos, luego de seleccionar el modelo más adecuado o con mayor porcentaje de exactitud. Se busca observar cómo, por ejemplo, un título de primaria tiene un menor impacto en la probabilidad que un factor como el título universitario, o viceversa, esto si se habla de una variable como la relacionada con la educación titulada.

1.7.2. Limitaciones

La limitación principal está en el mecanismo de recolección de la información, dado que depende estrictamente de los resultados obtenidos en las ECE; de igual modo, queda a total discreción de la persona encargada de este proyecto, el uso o distribución de los datos y resultados obtenidos.

Cada una de las encuestas trimestrales contiene alrededor de 25 000 registros, por lo que el análisis de la información se enfoca en el uso exclusivo de dicha cantidad de registros.

Los resultados obtenidos por medio de los modelos de análisis de datos están apegados única y exclusivamente a los datos aportados por el INEC.

Otra limitación significativa es la cantidad de registros sin respuesta, ya sea porque el encuestado se abstiene de responder o porque el encuestador no le efectúa la pregunta por decisión propia; y que generan mucho ruido a la hora de desarrollar los modelos de predicción y clasificación.

Durante el desarrollo de este proyecto de graduación final, Costa Rica experimenta la pandemia por la COVID-19, lo cual genera una importante afectación en la salud de los ciudadanos, con una tasa de mortalidad del 1 %, y a nivel económico y social, donde existe una parálisis de las actividades monetarias, como la turística en zonas rurales y labores de manufactura e industria en zonas más urbanas; prácticamente no hay actividad económica sin afectación, obteniendo como consecuencia el crecimiento del desempleo en la población nacional, que pasa de un 12 % a un alarmante 25 %.

Por lo tanto, se puede interpretar que la pandemia de la COVID-19 afecta el diseño e implementación de este proyecto; sin embargo, el alcance de este estudio se centra en datos obtenidos en el año natural 2019, mediante la continuidad de una encuesta, por consiguiente, estos datos no se encuentran afectados por la pandemia.

Por el contrario, el beneficio de esta investigación en relación con la capacidad de predecir la posibilidad de un cambio de estado laboral impulsa la creación de nuevas políticas que ataquen de una forma más efectiva el desempleo a nivel nacional.

1.8. Marco de referencia organizacional y socioeconómico

1.8.1. Historia

De acuerdo con el INEC (2017), el primer censo nacional de población se solicita en 1824 por parte del Congreso Constituyente, sin embargo, a pesar de los deseos de realizar dicho censo, no es posible por falta de metodologías.

En el año 1861, se crea la primera Oficina Central de Estadística y se efectúa la promulgación de la Ley de Censos.

Se deben esperar cerca de 40 años para poder ejecutar el Primer Censo General de la Población, considerado como el primer censo oficial.

En 1948, con los decretos n.° 61 y n.° 72 se centraliza la Dirección General de Estadística, la dirección técnica de las estadísticas nacionales, y se crea el Consejo

Nacional de Estadística para orientar técnicamente la estructuración y desarrollo de las estadísticas nacionales.

Ahora bien, en el caso específico del problema de desempleo, las tasas de desempleo y subempleo crecen de modo alarmante; según los valores ofrecidos por el INEC (2019), en el año 2010 la tasa del desempleo ronda el 10 % y en el año 2019 se tiene una tasa que sobrepasa el 12.4 %.

1.8.2. Tipo de negocio y mercado meta

La población que se pretende beneficiar con este proyecto es la relacionada al INEC, al ser la organización gubernamental encargada de generar información y analizar los datos con el objetivo de comprender los comportamientos de la población y brindar los resultados certeros a los poderes del Gobierno responsables de originar políticas de mejora que permitan disminuir el porcentaje de desempleo.

1.8.3. Misión, visión y valores

1.8.3.1. Misión

“Somos responsables de la gestión de las estadísticas nacionales para orientar las decisiones que promuevan el desarrollo del país” (INEC, 2020).

1.8.3.2. Visión

“Seremos líderes en proveer a la sociedad información geoestadística sobre la realidad costarricense” (INEC, 2020).

1.8.3.3. Valores

Integridad: Cada análisis de datos y desarrollo de los modelos debe ser a partir de la información exacta, sin manipular o modificar.

Orientación de servicio: El desarrollo del presente proyecto se planea con la única idea de ayudar a las poblaciones más desprotegidas del país.

Trabajo en equipo: Como sociedad, se debe aprender que se es parte de un gran engranaje y muchas veces se necesita la ayuda de todos.

Calidad: La producción estadística del INEC y sus servicios se sustenta en metodología estadística sólida y procesos estadísticos adecuados; además, es oportuna, confiable y accesible.

Credibilidad: Los análisis presentados y los resultados obtenidos están apegados a la independencia técnica. Asimismo, la credibilidad es objetiva, profesional y transparente de la persona que desarrolla este trabajo final de graduación.

Compromiso: Es una de las ideas que caracteriza al presente trabajo.

1.8.4. Políticas institucionales

El INEC es un ente gubernamental que busca siempre la excelencia en sus procesos, máxime que de estos depende la generación de información, la cual debe ser siempre confiable e íntegra; por tal razón, se crean las siguientes políticas:

1.8.4.1. Política de Control Interno y de Administración de Riesgos

> El Instituto Nacional de Estadística y Censos se compromete a implementar un Sistema de Control Interno y de Valoración de Riesgos, el cual estará implícito en los procesos y procedimientos institucionales. De esta forma cada colaborador(a) ejercerá el autocontrol permanente en el desempeño de sus labores de manera tal que contribuya en la consecución de los objetivos institucionales (INEC, 2018).

1.8.4.2. Política de Desarrollo Organizacional

> El Instituto Nacional de Estadística y Censos se compromete al mejoramiento de su gestión al optimizar el modelo de organización administrativa para responder a la dinámica del entorno y a las demandas de la ciudadanía; además, a los fines por los cuales fue creada, a su misión y a su visión. Esto da como resultado el cumplimiento de los objetivos institucionales (INEC, 2018).

1.8.4.3. Política para la Gestión Misional y de Gobierno

El Instituto Nacional de Estadística y Censos se compromete a desarrollar su gestión con observancia al marco estratégico institucional y al logro de los objetivos de desarrollo nacional. El Instituto Nacional de Estadística Censos se compromete a consolidar el rol rector del Sistema de Estadística Nacional. Esto con el fin de progresar en la producción y difusión de las estadísticas oficiales, desde los principios de confidencialidad estadística, transparencia, especialidad y proporcionalidad (INEC, 2018).

1.9. Estado de la cuestión

En el siguiente proyecto de graduación final se desarrolla el estado de la cuestión siguiendo la plantilla creada por Biolchini, Gomes, Cruz y Horta (2005) en su publicación *Systematic Review in Software Enrineering*, donde se describen cinco etapas: la planeación, la identificación de orígenes, la selección de fuentes, la extracción de la información y el análisis de los resultados.

1.9.1. Planeación

1.9.1.1. Formulación de la pregunta

Determinar los modelos de minería de datos que se utilizan para la predicción del desempleo según la zona geográfica de Costa Rica con ayuda de variables sociodemográficas.

1.9.1.2. Foco de la pregunta

En la presente revisión sistemática, ¿es posible encontrar modelos estadísticos de minería de datos que logren la predicción del desempleo en Costa Rica según una serie de variables sociodemográficas?

1.9.1.3. Amplitud y calidad de la pregunta

1.9.1.3.1. Problema

En la actualidad uno de los principales problemas sociales de Costa Rica es la poca oportunidad de empleo a la que se enfrenta la población. Según el INEC (2019), hoy se está en el 12.4 % de desempleo, uno de los valores porcentuales más altos desde el 2010, cuando el valor se encuentra cerca del 10 %.

Al respecto, analizando la situación desde un punto de vista social, el desempleo es la raíz de muchos otros problemas como la inseguridad, la deserción escolar, la prostitución y la drogadicción, por nombrar algunos. Además, ninguna de las personas que hoy cuenta con un trabajo fijo está exenta de esta problemática, por lo tanto es muy importante medir en qué grado una persona puede llegar a quedarse sin trabajo o continuar con el puesto que desempeña.

En cuanto a esto, existe una organización que se dedica a la recolección de información sociodemográfica, es decir, el INEC, pero hasta el momento no es utilizada la posibilidad de caer en el desempleo para la predicción por medio de modelos estadísticos, según una serie de variables sociodemográficas.

1.9.1.3.2. Pregunta de investigación

¿Es posible determinar la probabilidad que afronta una persona en suelo costarricense de quedarse sin trabajo, al tomar en cuenta datos como la edad, idiomas, sexo, escolaridad, zona de habitación o titulación académica específica dentro del territorio nacional?

1.9.1.3.3. Palabras claves

Data mining, búsqueda de datos, *unemployment*, desempleo, densidad industrial, *industrial density*, subempleo, problemática social, *neuronal net*, *machine learning*.

1.9.1.3.4. Intervención

Predicción de la probabilidad de caer en desempleo.

1.9.1.3.5. Control

Se utiliza una línea de tiempo bastante establecida, dado que se cuenta con los datos de las ECE que se generan de forma trimestral. Por el momento se emplean las generadas en el 2019, pero no se descarta usar las encuestas hechas en el 2018 o información no confidencial originada por la Cámara de Industrias de Costa Rica o la Cámara de Comercio de Costa Rica.

1.9.1.3.6. Efecto

En este apartado se indica cuáles son los posibles resultados obtenidos y lo que se busca en la correcta implementación de un modelo, a saber, que brinde la más alta exactitud y permita mostrar la probabilidad de que la vivienda individual ocupada pueda variar su estado de ocupación laboral a lo largo de las diferentes encuestas continuas trimestrales.

1.9.1.3.7. Medida de salida

Como salida, se busca el modelo estadístico que proporcione la mayor exactitud según las variables sociodemográficas utilizadas.

1.9.1.3.8. Población

La población se refiere a todas aquellas publicaciones indexadas que se asemejen o tengan una fuerte relación con el tema presentado en el trabajo final de graduación.

1.9.1.3.9. Aplicación

Dentro del análisis realizado, se dictamina que los beneficiados directos son todos aquellos costarricenses que cuentan con un trabajo remunerado, los patronos con procesos de contratación actualizados y el INEC por su valiosa labor de recolección de información y la gestión que lleva a cabo para facilitar el acceso a dicha información.

1.9.1.3.10. Diseño

Aún no existe un modelo estadístico que posibilite la generación de un metaanálisis.

1.9.2. Identificación de los orígenes

1.9.2.1. Definición de los criterios de selección de fuentes

Conocimiento de estadistas expertos en el tema, uso de las palabras claves e investigación dentro de los sitios recomendados en la web.

1.9.2.2. Lenguajes estudiados

Español e inglés.

1.9.2.3. Identificación de los orígenes

1.9.2.3.1. Métodos de búsqueda de las fuentes

El principal método de búsqueda es el uso de motores de búsqueda web.

1.9.2.3.2. Cadenas de búsqueda

(“Data Mining” OR “Minería de Datos”) AND (“Machine Learning” OR “Máquina de Aprendizaje”) AND (“Unemployment Prediction” OR “Predicción Desempleo”) AND (“Algorithms” AND “Algoritmo”).

1.9.2.3.3. Lista de fuentes

- ACM Sigsoft Digital Library.
- Google Scholar.
- IEEE Xplore Digital Library.
- Springer Link.
- Ingenta Connect.

1.9.2.4. Selección de fuentes después de la evaluación

Todas aquellas fuentes que cumplan con los criterios de búsqueda, sumado a una lectura de los resúmenes ejecutivos y títulos de cada fuente, muestran una fuerte relación con el problema expuesto en este proyecto de graduación final.

1.9.2.5. Comprobación de las fuentes

Luego de la evaluación del punto anterior, todas las fuentes aquí expuestas son aprobadas.

1.9.3. Selección de fuentes

1.9.3.1. Definición de estudios

1.9.3.1.1. Definición de criterios de inclusión y exclusión

En el estado de la cuestión se incluyen todos aquellos estudios que cumplen con una comparativa efectuada por el encargado de este proyecto de graduación final.

Los estudios deben aportar material vinculado con el desempleo en cualquier región de mundo y con modelos estadísticos de minería de datos que ayuden a la predicción de una variable predictora, no es necesario que sea la misma variable que intenta predecir el presente documento.

De igual modo, se excluyen todos aquellos estudios que no tengan relación con las palabras claves o con el objetivo de este proyecto de graduación final.

1.9.3.1.2. Definición de tipos de estudios

Se incluyen todos aquellos estudios que cumplan con los criterios de búsqueda y, además, se apeguen al procedimiento de selección de estudios, dado que las opciones de incorporación son muchas, pero de inclusión son pocas.

1.9.3.1.3. Procedimiento para la selección de estudios

El procedimiento por seguir consta en primer lugar de la lectura de los títulos de los estudios arrojados por los motores de búsqueda según los criterios empleados; posteriormente, como segundo filtro, se ejecuta la lectura de los resúmenes ejecutivos y de las palabras claves, lo que permite realizar una selección aún más minuciosa de los estudios que se incluyen dentro de este documento.

1.9.3.2. Ejecución de la revisión

1.9.3.2.1. Selección inicial de estudios

Tabla 1: *Selección de estudios*

Sitio web	Palabra clave	Título del estudio	Autor del estudio	Año
IEEE Xplore Digital Library	*Data* AND *Mining* AND *Unemployment*	"Twitter Data Analysis for Unemployment Crisis"	Nirmala, C., Roopa, G. y Naveen, K.	2015
SpringerLink	*Data* AND *Mining* AND *Unemployment*	"Data Mining for Unemployment Rate Prediction Using Search Engine Query Data"	Xu, W., Li, Z., Cheng, C. y Zheng, T.	2013
SpringerLink	*Unemployment* AND *Neural* AND *Net*	"Forecasting Spanish Unemployment Using Near Neighbour and Neural Net Techniques"	Olmedo, E.	2014
Google Scholar	*Unemployment* AND *Neural Network*	"A Neural Network to Predict Civilian Unemployment Rates"	Aiken, M.	1996
IEEE Xplore Digital Library	*Machine* AND *Learning* AND *Unemployment*	"A Machine Learning Approach for Detecting Unemployment Using the Smart Metering Infrastructure"	Curbelo, C. y Hurst, W.	2020
SpringerLink	*Machine* AND *Learning* AND *Unemployment*	"Prediction of Unemployment Rates with Time Series and Machine Learning Techniques"	Katris, C.	2019
ACM Digital library	*Machine* AND *Learning* AND *Unemployment*	"Could Artificial Intelligence Create an Unemployment Crisis?"	Ford, M.	2013

Fuente: Información obtenida de diferentes fuentes

1.9.3.2.2. Evaluación de la calidad de los estudios

En relación con los estudios obtenidos por medio de los criterios de búsqueda, seis se ajustan completamente a todos los criterios de inclusión definidos con anterioridad.

1.9.3.2.3. Revisión de la selección

Seis estudios son aprobados.

1.9.4. Extracción de la información

1.9.4.1. Definición de criterios para la inclusión o exclusión de la información

Los estudios deben proporcionar técnicas de análisis y minería de datos, como algunos de los mecanismos más comunes de predicción y clasificación de datos, tales como regresiones, líneas de tiempo, árboles de decisión, redes neuronales, entre otras. Otro criterio de inclusión es el uso de términos como desempleo, tasa de empleo y predicción según datos extraídos.

Así mismo, se excluyen todos aquellos estudios que no presenten técnicas de aprendizaje automático supervisado ni contengan dentro de sus respectivos resúmenes ejecutivos conceptos asociados con el desempleo.

1.9.4.2. Formularios de extracción de datos

El formulario que se indica a continuación contiene las pautas resumidas para examinar cada uno de los seis estudios escogidos, el mismo se elabora con ayuda de la publicación de Biolchini et al. (2005).

Tabla 2: *Formulario de extracción de datos*

Extracción de resultados objetivos	
Identificación del estudio	Debe incluir la información más importante del estudio: autores, título, fecha de publicación, entre otra.
Metodología del estudio	Incluye la metodología utilizada para llevar a cabo el estudio.
Resultados del estudio	Muestra cuáles son los resultados obtenidos al final del estudio.
Problemas del estudio	Muestra cuáles son los problemas o las limitaciones experimentados en el desarrollo del estudio.
Extracción de resultados subjetivos	
Información mediante autores	Datos brindados por el autor o autores.
Impresiones generales y abstracciones	Opiniones obtenidas de otros lectores o autores.

Fuente: Elaboración propia con base en Biolchini et al., 2005

1.9.4.3. Ejecución de la extracción

En esta sección se muestra la información de los seis estudios seleccionados a partir de los criterios de búsqueda.

Tabla 3: *Primera fuente de estudio investigada*

Extracción de resultados objetivos	
Identificación del estudio	Nirmala, C., Roopa, G. y Naveen, K. (2015). *Twitter Data Analysis for Unemployment Crisis.* International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT). IEEE, India.
Metodología del estudio	Los autores comienzan el estudio con la extracción de datos del sitio Twitter mediante el uso de la etiqueta (#) relacionada con el desempleo. Los datos son migrados a archivos CSV y luego cargados a R–Studio. La información es limpiada y clasificada según el tipo de Twitter, ya que puede ser utilizada de forma positiva o negativa.
Resultados del estudio	No es posible visualizar los resultados obtenidos dado que el artículo está a la venta, pero no se puede efectuar su compra.
Problemas del estudio	A criterio del investigador de este trabajo final de graduación, el problema es el uso de datos de una fuente no confiable y que puede variar según la tendencia de unas pocas personas influyentes. Así mismo, el fondo del estudio no es similar al presentado en este proyecto de graduación, dado que en este último documento se intenta predecir la posibilidad de caer en el desempleo.
Extracción de resultados subjetivos	
Información mediante autores	Los datos no son solicitados.
Impresiones generales y abstracciones	El estudio cuenta con un gran potencial para la predicción del desempleo, sin embargo, el investigador de este proyecto de graduación final no considera que los datos utilizados para la implementación del modelo estadístico sean de una fuente confiable, en el sentido de que pueden ser manipulables por usuarios influyentes.

Fuente: Nirmala, Roopa y Naveen, 2015

Tabla 4: *Segunda fuente de estudio investigada*

Extracción de resultados objetivos	
Identificación del estudio	Xu, W., Li, Z., Cheng, C. y Zheng, T. (2013). Data Mining for Unemployment Rate Prediction Using Search Engine Guery Data. *SOCA 7,* 33–42.
Metodología del estudio	El estudio inicia con la obtención de datos de investigación por medio de motores de búsqueda; después, se aplica un modelo de selección de variables automatizado que se encarga de indicar cuáles son las características idóneas por utilizar en los modelos. Posteriormente, se utilizan herramientas de minería de datos para describir la relación entre los datos de la tasa de desempleo y los datos consultados en los motores de búsqueda. Luego, se aplican modelos de minería de datos como las redes neuronales y la regresión de vectores de soporte para pronosticar la tendencia del desempleo. Cuando se generan los resultados de cada modelo, se decide cuál es el que presenta mayor exactitud mediante un método de validación cruzada.
Resultados del estudio	El resultado obtenido demuestra que un modelo de minería de datos es eficiente para la predicción de la tasa de desempleo, lo que les permite a los Gobiernos mejorar y fortalecer las políticas que combaten este mal social.
Problemas del estudio	Uno de los problemas más importantes que encuentran los investigadores es decidir cuál de los modelos de minería de datos deben utilizar, dado que unos arrojan mejor rendimiento en sus promedios que otros. Además, otro problema es decidir cuáles de las variables macroeconómicas ajenas a los datos obtenidos por los motores de búsqueda deben usarse para fortalecer el estudio. Por último, se identifica que otro problema es la elección del modelo de minería de datos, dado que se consideran como opciones el uso tradicional de las series de tiempo o si, por el contrario, se emplean modelos poco aplicados a la predicción de la tasa de desempleo, pero que tienen mayor exactitud.

Extracción de resultados subjetivos	
Información mediante autores	Dentro de la investigación, los autores crean un diagrama comparativo entre el valor real de la tasa del desempleo y los valores predichos por el modelo de minería de datos, y visualmente los valores arrojados se acercan bastante a los valores reales siguiendo la tendencia de bajada y subida según el periodo de tiempo.
Impresiones generales y abstracciones	Uno de los puntos más destacables es el uso de datos de muestras cortas, para ser más específicos, muestras trimestrales, tal y como se plantea en el presente proyecto final de graduación. Además, los investigadores utilizan modelos de minería de datos que son vistos a lo largo de la carrera de maestría, por lo que son aplicables a este trabajo. Para terminar, la estructura del trabajo que se emplea se ajusta a la estructura que se usa en el trabajo final de graduación, a saber: 1. Recolección de la información. 2. Selección de las variables. 3. Preparación de conjuntos de datos de prueba y entrenamiento. 4. Aplicación de los modelos. 5. Predicción enfocada en la tasa de desempleo.

Fuente: Xu, Li, Cheng y Zheng, 2013

Tabla 5: *Tercera fuente de estudio investigada*

Extracción de resultados objetivos	
Identificación del estudio	Curbelo, C. y Hurst, W. (2020). A Machine Learning Approach for Detecting Unemployment Using the Smart Metering Infrastructure. *IEEE Access.*
Metodología del estudio	Los autores de esta investigación inician su trabajo luego de entender que la información ofrecida por los medidores inteligentes de electricidad, gas y agua en tiempo real es altamente precisa, lo que abre un abanico de posibilidades para el análisis de las tendencias de cada uno de los hogares, siendo una de estas tendencias el desempleo. Cabe indicar que son datos no tradicionales, lo que brinda una forma totalmente nueva de analizar las tendencias. Los investigadores recolectan esta información para luego crear una serie de modelos de minería de datos no lineales y después comparan los resultados con un modelo de clasificación lineal generalizado, con resultados bastantes prometedores gracias a los valores obtenidos en métricas como el área bajo la curva, la sensibilidad y la especificidad de sus modelos. Es este estudio se utilizan en total seis modelos de clasificación, los cuales son: redes neuronales, regresión logística múltiple, árboles de clasificación y regresión, bosque aleatorio, perceptrón multicapa con caída y discriminación ponderada a distancia.
Resultados del estudio	El estudio obtiene el resultado de seis diferentes mecanismos de minería de datos y aplica un modelo de validación cruzada de desempeño, donde se compara el promedio del área bajo la curva, de la sensibilidad y de la especificidad. Luego de comparar el desempeño de los modelos implementados, los autores indican que los modelos de discriminación ponderada a distancia y perceptrón multicapa con caída son los mejores, porque presentan un 77 % de área bajo la curva, logrando así una mejor predicción de nuevas clases.

Problemas del Estudio	El primer problema que encuentran los autores es cómo obtener la información que alimente el modelo, pues las empresas de servicios públicos que generan los datos son bastante cuidadosas con la seguridad de la información y no brindan los datos con mucha facilidad. Como segundo inconveniente, los autores tienen conocimiento de trabajos realizados con anterioridad donde se utilizan solo los datos ofrecidos por los medidores inteligentes en modelos de redes neuronales y clasificación lineal con resultados poco precisos, por lo que es muy posible que deban combinarse los datos con otros tipos de variables. Por lo tanto, se emplean datos de encuestas a hogares, a partir de los cuales se puede obtener el sexo, la cantidad de ocupantes de la vivienda y el estado de la empleabilidad.
Extracción de resultados subjetivos	
Información mediante autores	Los investigadores efectúan una presentación de los resultados muy visual, lo que facilita la comprensión por parte de las personas interesadas en el estudio. Es una de las formas más fáciles de entender y le brinda al investigador de este trabajo final una muy buena idea para emplearla dentro de los resultados del proyecto.
Impresiones generales y abstracciones	Uno de los puntos que llama la atención es que dentro de la muestra utilizada, los autores ya saben cuáles observaciones son desempleadas y cuáles no, lo que les ayuda a identificar de manera casi única el comportamiento, facilitando la posterior clasificación de las observaciones a partir de los modelos de minería de datos. Al analizar esta investigación, se determina la necesidad que existe de utilizar datos provenientes de censos y cómo estos nutren los estudios relacionados con el desempleo en la población de un país. Para terminar, el uso de múltiples modelos de minería de datos implica que entre más mecanismos se utilicen para predecir, mejores y más exactos van a ser los resultados obtenidos.

Fuente: Curbelo y Hurst, 2020

Tabla 6: *Cuarta fuente de estudio investigada*

Extracción de resultados objetivos	
Identificación del estudio	Katris, C. (2020). Prediction of Unemployment Rates with Time Series and Machine Learning Techniques. *Computational Economics.*
Metodología del estudio	El investigador plantea realizar una comparación de modelos de minería de datos para la predicción de la tasa de desempleo en la región geográfica de Europa, más específicamente en las zonas mediterránea, nórdica, báltica, balcánica y benelux. Todos sus datos son obtenidos de los repositorios de información del Eurostat u Oficina de Estadística de la Unión Europea, ubicada en Luxemburgo, donde los datos son de fácil acceso para todo aquel que desee utilizar la información para cualquier fin que sea en beneficio de la Unión Europea. La información empleada es no lineal y se genera de forma mensual ajustada por temporada de desempleo, para los periodos de tiempo entre el primer mes del año 2000 y el doceavo mes del año 2014. Los modelos de aprendizaje automático que se utilizan son series de tiempo (ARIMA y FARIMA), redes neuronales, regresión de vectores de soporte y *splines* de regresión adaptativa multivariable. Para la comparación del rendimiento de los modelos, se emplean algunas de las métricas más importantes, como el error absoluto medio (MAE) y el error cuadrático medio raíz (RMSE). En cuanto a la comparación de los modelos de minería de datos, se analizan los resultados en distintos horizontes de tiempo y regiones geográficas, para así decidir cuál enfoque es mejor en todos los casos y detectar si el horizonte de tiempo o la zona geográfica afectan el desempeño del modelo.
Resultados del estudio	Con respecto a la efectividad de los modelos de minería de datos, la investigación indica que el mejor mecanismo es la serie de tiempo conocida como media móvil autorregresiva fraccionadamente integrada o FARIMA por su sigla en inglés, dado que presenta el mejor promedio de

	MAE y el mejor promedio de RMSE en comparación con los demás modelos; seguido del diseño FARIMA/GARCH con promedios muy similares a los obtenidos por el modelo FARIMA; luego, en el tercer lugar, están los valores del promedio de las redes neuronales; seguido de cerca por la regresión de vectores de soporte y, para terminar, con los peores promedios de predicción se encuentran los modelos de *splines* de regresión adaptativa multivariable (MARS).
Problemas del estudio	Dentro de los problemas que identifica el autor de este trabajo final de graduación, está el horizonte de tiempo y las ubicaciones geográficas de los distintos países utilizados en los modelos, dado que cuentan con una serie de variables como el clima o la política que afectan el resultado de la investigación; por tal razón, el autor se ve en la necesidad de aplicar los modelos dos veces, la primera vez visualizando las variables geográficas y el horizonte de tiempo como un bloque y, en la segunda ocasión, cada una de las variables anteriores separada y no como un solo bloque.
Extracción de resultados subjetivos	
Información mediante autores	La presentación de los resultados se realiza de una forma muy sencilla y de fácil entendimiento. El autor utiliza una serie de tablas y gráficos que se encargan de mostrar los valores obtenidos en cada modelo. Posterior a cada representación visual, se encuentra una explicación literaria que complementa el contenido gráfico del documento.
Impresiones generales y abstracciones	Cabe destacar que el investigador utiliza información de la entidad encargada de la estadística de la Unión Europea, tal y como se plantea en este proyecto de graduación, donde se consume la información obtenida por el INEC, y con una periodicidad bastante similar, dado que el autor usa datos de un mes de antigüedad y en este documento se emplean datos de censos con tres meses de antigüedad entre cada censo.

Fuente: Katris, 2019

Tabla 7: *Quinta fuente de estudio investigada*

Extracción de resultados objetivos	
Identificación del estudio	Olmedo, E. (2014). Forecasting Spanish Unemployment Using Near Neighbour and Neural Net Techniques. *Computational Economics, 43*(2).
Metodología del estudio	La autora utiliza dos métodos de investigación para la predicción del desempleo en España. El primer mecanismo es un enfoque que usa modelos lineales de predicción como la regresión lineal o la autorregresión vectorial (VAR), aunque más adelante se aprecia que el conjunto de datos empleados es no lineal. Por otra parte, el enfoque que interesa analizar es el segundo, el que se centra en modelos de aprendizaje, donde se encuentran las redes neuronales como el principal y más adecuado mecanismo de predicción, siendo ampliamente utilizado debido a su capacidad para resolver una gran variedad de problemas asociados con la detección y predicción no lineal. Además, es un método que emplea funciones no lineales para hacer sus pronósticos. Esta investigación, tal como la liderada por Katris (2019), usa los datos proporcionados por el Eurostats, de donde toma las tasas de desempleo mensuales ajustadas estacionalmente, desde el periodo comprendido entre enero de 1987 hasta noviembre de 2011, con un total de 298 valores por serie para alrededor de diez regiones de Europa. En la mayoría de los modelos, se toma el conjunto de datos y se divide en dos, el conjunto de pruebas y el conjunto de entrenamiento, pero en esta investigación no es así, la autora divide los datos en tres conjuntos, el primero con 200 observaciones para entrenamiento, el segundo con 49 observaciones para pruebas y el tercero con otras 49 observaciones para ajustar el modelo según se requiera. Para la comparación de los resultados, se utiliza la medida de precisión conocida como error cuadrático medio normalizado (NMSE).
Resultados del estudio	Los resultados se despliegan en dos secciones, la primera se enfoca en el periodo dentro de la muestra, donde los modelos con el promedio más bajo de error cuadrático medio

	normalizado son la autorregresión vectorial y la regresión lineal. En el segundo caso, donde los modelos son analizados fuera del periodo de la muestra, los modelos más exactos son la regresión lineal y las redes neuronales, además muestran que la autorregresión vectorial no es tan exacta como se cree. Por tanto, se resumen los resultados en que el modelo de regresión lineal es el que puede pronosticar con mejor credibilidad el desempleo en los países analizados.
Problemas del estudio	En el documento se detalla que uno de los principales problemas es la elección correcta del modelo de minería para los datos no lineales, dado que otra característica que presentan los datos es que son inestables y dependientes. Sin embargo, se plantea una solución al señalar que se puede usar una técnica no paramétrica y, de este modo, proporcionar la evolución del desempleo en España.
Extracción de resultados subjetivos	
Información mediante autores	La exposición de los resultados se realiza de una forma complicada y con un lenguaje de muy alto nivel estadístico que no es de fácil entendimiento. El autor utiliza una serie de tablas y gráficos que se encargan de mostrar los valores obtenidos en cada modelo; posterior a cada representación visual, se encuentra una explicación técnica de difícil comprensión.
Impresiones generales y abstracciones	La autora comienza su investigación haciendo énfasis en la importancia de utilizar un modelo de predicción no lineal, dado que las técnicas tradicionales emplean mecanismos lineales, lo cual puede explicar el comportamiento diferente e inestable del desempleo en España, país donde intenta predecir la tasa de desempleo. Es también muy valioso para este proyecto de graduación final el uso de los datos generados por instituciones gubernamentales, los cuales son de libre acceso al público en general, son muy ricos a nivel informativo y no son muy usados para originar investigaciones que pueden beneficiar la economía y la política de un país.

Fuente: Olmedo, 2013

Tabla 8: *Sexta fuente de estudio investigada*

Extracción de resultados objetivos	
Identificación del estudio	Ford, M. (2013). Could Artificial Intelligence Create an Unemployment Crisis? *Communications of the ACM, 56* (7), 37-39.
Metodología del estudio	Este artículo no posee una metodología investigativa como los documentos anteriores, pero sí marca una idea muy importante y que puede afectar de forma directa el desempleo en Costa Rica. El autor describe que entre más rutinario sea un puesto de trabajo, más propensa es su automatización por parte de un sistema de información, dado que todas las tareas que ejecuta pueden ser programadas y ejecutadas dentro de una computadora. Señala que en la actualidad los puestos de trabajo que se enfocan en resultados creativos o no rutinarios son el menor porcentaje de la fuerza laboral de la economía. Asimismo, se menciona que el uso de “grandes datos” o *big data,* por su nombre en inglés, es un gran impulsor de esta automatización, porque provee de mucha información histórica y permite la predicción y afinamiento de modelos de minería de datos, aunque originalmente se desarrolla para brindar una ventaja competitiva a las compañías en temas de mercadeo y relación directa con los clientes.
Resultados del estudio	No tanto como resultados, sino como hallazgos, el autor explica cómo, gracias a los modelos de minería de datos, se comienza a transformar la creencia de que los puestos de trabajo creativos o no rutinarios no se pueden automatizar. Las máquinas de aprendizaje automático, como su nombre lo indica, están diseñadas para aprender de datos históricos, tomar tareas no rutinarias y convertirlas en tareas rutinarias fácilmente programables. Por último, cabe resaltar que las máquinas de aprendizaje automático van poco a poco eliminando los trabajos rutinarios y aunque la inteligencia artificial pueda pasar una prueba de Turing, no tiene la destreza que posee un ser humano, no puede igualar su nivel de análisis o innovación.

Problemas del estudio	Uno de los problemas que explica el autor en el artículo es la creciente necesidad de migrar de tareas rutinarias a tareas creativas o no rutinarias, dado que son tareas que no pueden ser automatizadas con facilidad, pero el problema yace justamente ahí, para las personas poco calificadas esto puede presentarse como un reto muy grande, y a la larga, aumentar el índice de la tasa de desempleo en la economía nacional, aun suponiendo que exista la cantidad suficiente de puestos creativos para todos los trabajadores que los necesiten. Es común observar cómo trabajos que antes son desarrollados por profesionales, ahora son ejecutados por máquinas de aprendizaje automático, por ejemplo, algunos puestos en leyes, que antes son ocupados por abogados, ahora son puestos en práctica por sistemas que predicen si los documentos de un juicio son relevantes o no.
Extracción de resultados subjetivos	
Información mediante autores	Los datos no son solicitados.
Impresiones generales y abstracciones	Este artículo se incluye en este proyecto final de graduación por la percepción del autor en relación con la posibilidad de generar desempleo a raíz del desarrollo de máquinas de aprendizaje automático. Es un punto que el investigador del trabajo final de graduación no se formula hasta que lee el artículo, es cómo los especialistas en administración de bases de datos con conocimiento en minería de datos propician que puestos de trabajo de bajo perfil o en los que se requiere baja educación sean sustituidos por computadoras que logren predecir el comportamiento de las personas. Inclusive, este proyecto de graduación está diseñado para intentar predecir una variable de desempleo según una serie de variables encontradas en un conjunto de datos; es decir, intenta predecir si una persona puede quedar sin trabajo, cuando por otro lado los profesionales desarrollan herramientas que buscan automatizar puestos y no depender de personas para ejecutar el trabajo.

Fuente: Ford, 2013

Tabla 9: *Séptima fuente de estudio investigada*

Extracción de resultados objetivos	
Identificación del estudio	Aiken, M. (1996). A Neural Network to Predict Civilian Unemployment Rates. *Journal of International Information Management, 5*(1).
Metodología del estudio	El artículo describe que existen muchas formas de pronosticar variables de tiempo en los modelos económicos de un país o de una región como las técnicas multilíneas y las técnicas relacionadas, pero suelen ser muy inexactas dado que usan indicadores económicos desarrollados por los Gobiernos locales. Los datos son obtenidos del *software* de indicadores del ciclo comercial de la empresa Media Logic Incorporated, que presenta observaciones mensuales y trimestrales con alrededor de 250 series de tiempo macroeconómicas de un periodo de tiempo de 50 años. Las variables de la investigación son seleccionadas mediante una combinación de teoría e inspección subjetiva de las series de tiempo mostradas por el *software*, dentro de las que se encuentran: tasa de desempleo civil, sentimiento del consumidor, expectativas del consumidor, oferta monetaria, entre otras. Siguiendo con la metodología de la investigación, se separan las observaciones en dos subconjuntos, los datos de prueba y los datos de entrenamiento y se utiliza una cantidad considerable de repeticiones para educar al modelo de red neuronal hasta lograr un error de entrenamiento aproximado al 3 %.
Resultados del estudio	Los resultados de la investigación son bastante satisfactorios, luego de ser comparados con otros tres modelos de minería de datos, como lo son regresión lineal, el modelo llamado “no cambio” y con un pronóstico de promedio de tres meses. El porcentaje de error del modelo de regresión logística es 18 veces más alto que el error presentado por la red neuronal. La técnica de promedio de tres meses simplemente brinda el mismo promedio estimado y, por

	último, el modelo llamado “no cambio” es el que más se acerca a los resultados obtenidos en la red neuronal y los más cercanos a las tasas reales.
Problemas del estudio	Uno de los principales problemas de utilizar este tipo de técnicas de predicción es el uso de indicadores económicos elaborados por los Gobiernos de cada país, como son la revisión de los índices, en especial si son compuestos; la revisión de los datos y fuentes, y más aún si la cantidad de la información es demasiado grande; y por último, la definición de los puntos de inflexión, como cuando se debe decidir si el cambio que sufre el indicador en los meses que pasan es de importancia para la predicción o puede llegar a ser una falsa alarma.
Extracción de resultados subjetivos	
Información mediante autores	Los datos no son solicitados.
Impresiones generales y abstracciones	Es un artículo de gran ayuda para esta investigación, que fundamenta aún más usar como modelo central de minería de datos a la red neuronal, al ser un modelo de predicción que no se programa como un *software*, sino que se entrena de forma que entre más se exponga a los datos, mejores van a ser los patrones que diseñe y más exacto el resultado obtenido.

Fuente: Aiken, 1996

1.9.4.4. Resolución de divergencias entre los revisores

No se encuentra divergencia alguna.

1.9.5. Análisis de resultados

1.9.5.1. Resultados del cálculo estadístico

No se aplican cálculos estadísticos.

1.9.5.2. Presentación de resultados

Tabla 10: *Resumen de resultados de fuentes literarias*

Fuente	Estudios	Relevantes	Excluidos	Primarios
IEEE Xplore Digital Library	33	4	2	2
SpringerLink	10	10	7	3
Google Scholar	31	6	4	2
ACM Digital library	2	2	1	1

Fuente: Elaboración propia

1.9.5.3. Análisis de sensibilidad

No aplicable.

1.9.5.4. Comentarios finales

Número de estudios: 76 estudios encontrados, siete estudios seleccionados.

Sesgo de búsqueda, selección y extracción: no se define.

Sesgo de publicación: no se define.

Variación entre revisores: sin variaciones.

Aplicación de resultados: los estudios seleccionados son aquellos que poseen algún valor agregado para este trabajo de graduación final, pues exponen modelos de minería como las redes neuronales que presentan resultados muy valiosos para la predicción de temas relacionados con el desempleo.

Recomendaciones: dentro de las búsquedas es muy valioso utilizar ayudas como *intitle:*, que permite priorizar la indagación de palabras en el título de las publicaciones.

2. Marco conceptual

Antes de crear un marco conceptual, es necesario conocer el problema que se plantea en la investigación y, luego, las averiguaciones desarrolladas con anterioridad acerca de dicho problema.

Al respecto, el problema se establece de forma clara en el primer punto del proyecto de graduación final, donde se aborda la preocupación que existe en Costa Rica en relación con el aumento de la tasa de desempleo y que muy probablemente es producto de las variables sociodemográficas que describen a la población costarricense.

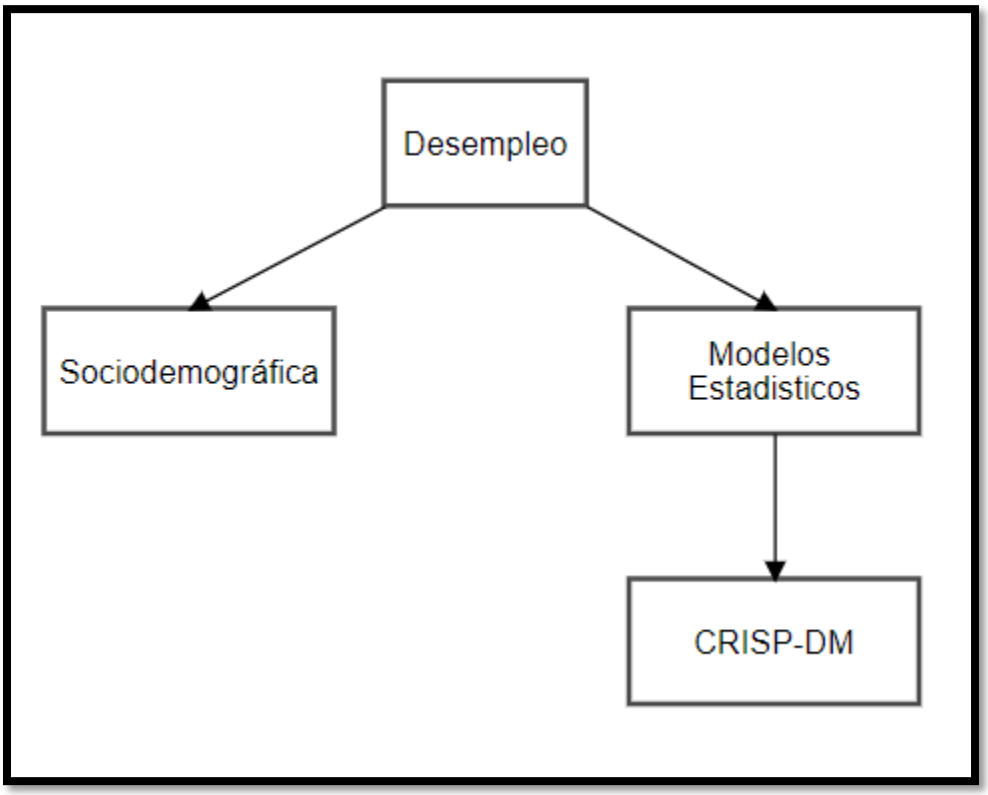


Ilustración 1: Relación entre el desempleo y los modelos estadísticos
Fuente: Elaboración propia

Además, el marco conceptual es necesario para comprender el tipo de investigación y la metodología por implementar; sin embargo, estos puntos se profundizan en el marco metodológico.

Una manera de orientar el desarrollo del marco teórico de este proyecto es la creación de una nube de palabras que permita identificar cuáles son las más usadas en el documento y, así, ampliar los conceptos más relevantes. Para dicho proceso se utiliza por completo el texto de la investigación, el cual se coloca en un generador gratuito de nubes de palabras que da como resultado la siguiente nube:

Ilustración 2: Nube de palabras
Fuente: Elaboración propia

2.1. El desempleo

Son varias las razones que llevan a incluir el concepto de desempleo dentro del marco teórico, la más importante es porque forma parte de la problemática planteada, aspecto que también es central en numerosas investigaciones como se aprecia en el estado de la cuestión, en muchas de las cuales se hace el análisis en territorio europeo, con datos obtenidos de instituciones gubernamentales que ofrecen la información de manera gratuita.

En cuanto a lo expuesto, los investigadores intuyen que la presencia de este concepto en su país es la causa de distintos problemas sociales y económicos como la pobreza, la drogadicción, la violencia social y familiar, la prostitución, la depresión, la ansiedad, la falta de confianza, entre otros problemas sociales, sin olvidar las recesiones económicas, el despilfarro de recursos, la baja producción y la poca exportación de bienes y servicios que pueden afectar en gran medida el desarrollo económico de una nación.

Asimismo, si se observa con detenimiento la nube de palabras generada, se aprecia que la palabra *desempleo* sobresale por la frecuencia con la que aparece en el documento, y junto a esta es posible leer también su respectiva traducción al idioma inglés, a saber *unemployment*, lo que la convierte en un concepto central para esta investigación.

2.2. Sociodemográfica

La sociodemográfica consiste en la noción que pretenden brindar las variables de estudio utilizadas para predecir la posibilidad que tiene un costarricense de caer en desempleo.

Además, la sociodemográfica se encarga de definir las características sociológicas y demográficas de un grupo específico, con el propósito de utilizar dichas características para el estudio de los posibles efectos que pueda tener este grupo en áreas tan significativas como la economía, el comercio, la medicina y el comportamiento social.

Ahora bien, en primer lugar, las variables sociológicas son aquellas que describen al individuo como parte de la sociedad humana, su nivel de pertenencia al hogar, los intereses que persigue el individuo y los grupos sociales que integra.

En segundo lugar, las variables demográficas proporcionan información como el sexo del individuo, edad, lugar de residencia, zona de residencia (si es urbana o rural), estado civil y nivel educativo; variables directamente relacionadas a la probabilidad de que una persona esté desempleada.

2.3. Modelos estadísticos

Los modelos estadísticos son la principal herramienta de investigación en este proyecto de graduación final, pues como se analiza en las averiguaciones del estado de la cuestión, los autores utilizan una serie de modelos como redes neuronales o regresión vectorial de soporte para encontrar la forma más idónea de llegar a los resultados esperados. Sin embargo, estos son solo un par de opciones, pero no las definitivas; es necesario investigar e intentar con otro tipo de modelos estadísticos, muchos de los cuales se estudian durante la maestría y pueden emplearse en este proyecto de graduación.

Los modelos estadísticos, por definición, son aquellos que utilizan una serie de ecuaciones matemáticas que le permiten al investigador analizar y descifrar un conjunto de datos extraídos en un momento en el tiempo y de una sección de la población. Existen muchos tipos diferentes, pero se usan como guía los modelos empleados en investigaciones realizadas en otros países, intentando obtener un mayor margen de éxito.

2.4. CRISP-DM

Esta metodología es la estudiada a lo largo de la maestría universitaria y se considera que la técnica puede soportar la investigación de este proyecto. A continuación, se explica su funcionamiento, mejorando la comprensión de cada una de sus etapas por parte de los lectores.

La técnica CRISP-DM o Cross Industry Standard Process for Data Mining, por su sigla en inglés, es “la metodología que describe en términos de un modelo de proceso jerárquico, que consiste en conjuntos de tareas descritas en cuatro niveles de abstracción” (Smart Vision Europe, 2015).

La técnica utiliza seis fases, las cuales son iterativas y con tareas definidas, que deben dar como resultado un conjunto de documentos e informes. Las fases son las siguientes:

- Entendimiento del negocio: Como fase inicial, es donde se intentan conocer los alcances y objetivos del proyecto. Cada integrante del equipo de trabajo debe saber el problema que se desea solucionar.

- Entendimiento de la información: En esta fase da comienzo la recolección de la información y el entendimiento de los datos, se deben detectar vacíos o problemas de calidad.
- Preparación de los datos: Es donde los investigadores realizan todas las actividades de limpieza y preparación de la información, se eliminan o suprimen datos basura y se escogen las variables más representativas.
- Modelado: Es la etapa donde se implementan los modelos estadísticos, pueden ser uno o varios según la exactitud que se desee lograr; además, los investigadores efectúan una serie de ajustes y configuraciones para su óptimo desempeño.
- Evaluación: En cuanto a esta etapa, el modelo de predicción ya está creado, pero es necesario validar la calidad del análisis mediante una revisión a profundidad.
- Despliegue: No necesariamente es el proceso final del ciclo, a menudo se requiere volver a construir el modelo e implementar los informes de presentación al usuario.

En la ilustración 3 se indica de una forma más clara cómo se desarrolla la técnica CRISP-DM y cómo se relaciona cada una de sus etapas.

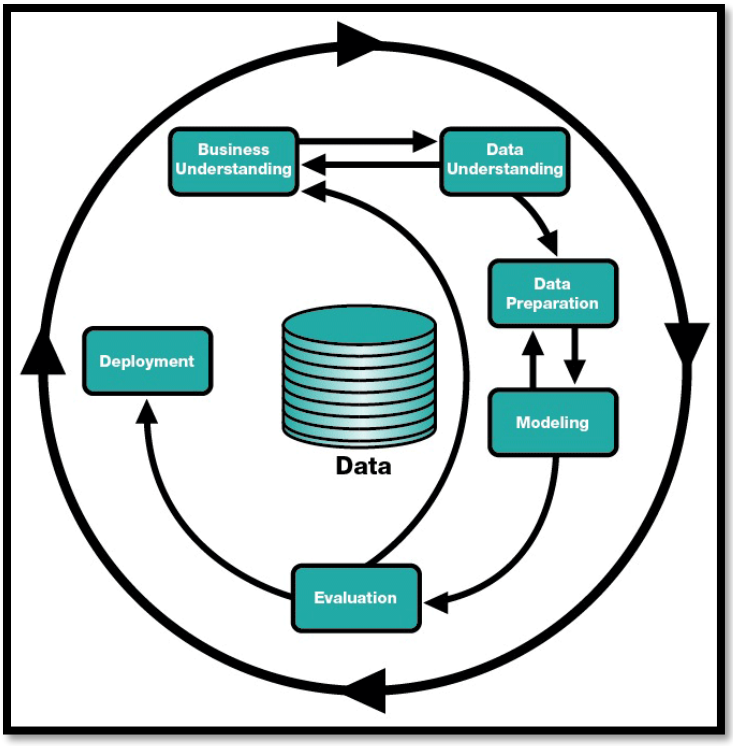


Ilustración 3: Metodología CRISP-DM
Fuente: Smart Vision Europa, 2015

3. Marco metodológico

3.1. Tipo de investigación

En el caso de este proyecto, se utiliza la investigación aplicada porque se trabaja sobre un objetivo en concreto: determinar la probabilidad que existe de que un costarricense pueda caer en desempleo, según un conjunto de variables sociológicas y demográficas, mediante el uso de modelos estadísticos siguiendo la metodología CRISP-DM.

En cuanto a la misma, Lozada (2014) afirma:

> La investigación aplicada tiene por objetivo la generación de conocimiento con aplicación directa y a mediano plazo en la sociedad o en el sector productivo. Este tipo de estudios presenta un gran valor agregado por la utilización del conocimiento que proviene de la investigación básica. De esta manera, se genera riqueza por la diversificación y progreso del sector productivo. Así, la investigación aplicada impacta indirectamente en el aumento del nivel de vida de la población y en la creación de plazas de trabajo (p. 35).

3.2. Alcance investigativo

El alcance definido para este proyecto de graduación final es el exploratorio, al presentarse un trabajo que aún no es desarrollado en Costa Rica, pues luego de la investigación de literatura expuesta en el estado de la cuestión, se determina que en Costa Rica todavía no se efectúa un modelo de predicción que les posibilite a las entidades públicas o privadas conocer la probabilidad que tienen las viviendas individuales ocupadas de cambiar su estado laboral de empleado a desempleado y viceversa.

3.3. Enfoque

El enfoque del trabajo final de graduación es mixto, ya que reúne, investiga y recoge datos de tipo cuantitativo y cualitativo, “así como su integración y discusión conjunta, para realizar inferencias producto de toda la información recabada y lograr un mayor entendimiento del fenómeno bajo estudio” (Hernández y Mendoza, 2009).

De este modo, se utiliza el enfoque cuantitativo para presentar todos los resultados estadísticos y probabilísticos obtenidos gracias a la implementación de los modelos de minería de datos.

Por su parte, el enfoque cualitativo se emplea en cuanto a la posibilidad de tomar decisiones y crear medidas de contingencia con respecto a la probabilidad que existe de caer en desempleo según las características sociológicas y demográficas del conjunto de individuos en estudio.

3.4. Diseño

El diseño adoptado es el exploratorio, aunque el enfoque de la investigación sea mixto, al reconocerse que no existe un proyecto de investigación similar aplicado a la realidad de Costa Rica.

3.5. Población y muestreo

La población que se utiliza en esta investigación es la recopilada por medio de la ECE, la cual se aplica de forma aleatoria y trimestral en el país de Costa Rica por parte del INEC. Adicional, se emplean únicamente los registros del año natural 2019.

El conjunto de datos se puede descargar de manera gratuita desde el sitio web del INEC, cuenta con un instructivo donde se detalla cada una de las más de 300 variables que incluye el conjunto de datos y con cerca de 30 000 observaciones.

3.6. Instrumentos de recolección de datos

La información es recopilada por el INEC mediante el uso de encuestas, posteriormente se descarga de modo gratuito del sitio web del INEC por el investigador a cargo de este proyecto de graduación final.

3.7. Técnicas de análisis de la información

Las técnicas de análisis para el desarrollo del trabajo de investigación son las presentadas dentro de la metodología llamada CRISP-DM, siendo esta una técnica compuesta por seis etapas, las cuales poseen un orden en su ejecución, pero además

se tiene la flexibilidad de regresar a etapas anteriores cuando el investigador lo considere necesario.

El diagrama de la metodología se incluye en el “Marco conceptual”, en la sección 2.4., “CRISP-DM”, ilustración 3. Ahora bien, las etapas que conforman la técnica CRISP-DM son las siguientes:

- Entendimiento del negocio: Como fase inicial, es donde se intentan conocer los alcances y objetivos del proyecto. Cada integrante del equipo de trabajo debe saber el problema que se desea solucionar.
- Entendimiento de la información: En esta etapa se comienza la recolección de la información y el entendimiento de los datos, se deben detectar vacíos o el problema de calidad.
- Preparación de los datos: Es donde los investigadores realizan todas las actividades de limpieza y preparación de la información, se eliminan o suprimen datos basura y se escogen las variables más representativas.
- Modelado: Es la etapa donde se implementan los modelos estadísticos, pueden ser uno o varios según la exactitud que se desee lograr; además, los investigadores efectúan una serie de ajustes y configuraciones para su óptimo desempeño.
- Evaluación: En cuanto a esta etapa, el modelo de predicción ya está creado, pero es necesario validar la calidad del análisis mediante una revisión a profundidad.
- Despliegue: No necesariamente es el proceso final del ciclo, a menudo suele requerirse volver a construir el modelo e implementar los informes de presentación al usuario.

4. Análisis del diagnóstico

4.1. Entendimiento del negocio

Es necesario comprender cuáles son los objetivos que tiene el INEC con respecto a las ECE y, de esa forma, entender cuál puede ser la contribución de este proyecto en el análisis de sus datos.

La documentación relacionada con las ECE establece de manera muy clara los objetivos que se buscan satisfacer en el INEC con la información obtenida de las viviendas individuales ocupadas, los cuales se citan a continuación:

- Generar indicadores del mercado laboral trimestralmente.
- Estimar y analizar la dinámica del mercado laboral mediante indicadores de cambio.
- Realizar un operativo de campo periódico que recolecte información estacional sobre la situación de empleo y desempleo en los hogares.

Los objetivos señalados en el párrafo anterior le permiten al investigador marcar las pautas de su proyecto, comprender cuáles son los resultados esperados por el INEC con sus encuestas y ofrecer un modelo de análisis de datos mediante la creación de una máquina de aprendizaje automático que ayude a potenciar los objetivos del INEC y a conducir a una mejor interpretación de la información.

Este proyecto de graduación final tiene como guía la metodología CRISP-DM, y en este apartado se inicia el desarrollo de la primera etapa, llamada “Entendimiento del negocio”, abriendo camino a un plan de proyecto.

4.1.1. Objetivos del negocio

El objetivo principal del negocio es desarrollar un modelo de aprendizaje automático que permita la predicción de la probabilidad que tiene una vivienda individual habitada de variar su estado de empleo actual; por ejemplo, en caso de encontrarse en estado desempleado, el modelo indica la probabilidad de seguir en el desempleo o, por el contrario, la posibilidad de cambiar de estado a empleado.

La veracidad de este modelo busca brindarle a la institución un mejor análisis de la información y, con esto, ofrecer el conocimiento necesario para la mejora de las políticas laborales del país.

4.1.1.1. Contexto

La ECE surge como una necesidad expresa del Banco Central de Costa Rica en el año 2006, dado que para ese momento el INEC aplica de forma anual la Encuesta de Hogares Propósitos Múltiples (EHPM), la cual brinda información de empleo, desempleo, ingresos, gastos, entre otras variables socioeconómicas de las viviendas costarricenses, pero su periodicidad no satisface la precisión y frecuencia de la información que necesita el Banco Central para sus indicadores.

De este modo, en el tercer trimestre del año 2010 se inicia la ECE, la cual genera cada tres meses una base de datos con información que sigue las recomendaciones de la OIT (Organización Internacional del Trabajo), logrando así una homogeneidad entre los datos nacionales e internacionales.

El INEC cuenta con cada base de datos trimestral desde el año 2010, pero el alcance de este proyecto final de graduación se limita a los datos extraídos en el año 2019. No obstante, la información obtenida no es utilizada para el estudio del cambio de estado en los individuos que habitan las viviendas en Costa Rica, siendo este un trabajo que predice la probabilidad de conservar el estado de empleo o desempleo según las variaciones que perduran a lo largo del tiempo, gracias a la continuidad de la encuesta.

4.1.1.2. Objetivos del negocio

La información contenida dentro de cada ECE es muy amplia y diversa, lo que abre un abanico de posibilidades para el investigador, pero en medio de la difícil situación laboral por la que atraviesa el país, se toma la decisión de enfocar los esfuerzos de este proyecto en intentar calcular la probabilidad que puede tener una vivienda individual ocupada en conservar su estado laboral, ya sea que se encuentre desempleada y exista la posibilidad de ubicarse en el mercado laboral o, al contrario, que sea una vivienda donde sus individuos formen parte de la fuerza laboral y haya la posibilidad de perder sus trabajos.

Siguiendo la línea descrita en el párrafo anterior, se establecen los siguientes objetivos del negocio:

- Estimar la posibilidad de cambiar el estado laboral de los integrantes de una vivienda individual ocupada, ayudándose por la continuidad de la encuesta, donde se garantiza una rotación de un 25 % de los hogares encuestados.
- Identificar las variables más representativas, con el objetivo de intensificar su captura en las futuras encuestas continuas.
- Generar una interfaz visual sencilla, pero eficaz que represente los objetivos del negocio desde una forma clara y expedita.

Este análisis de la información pretende enfocar los esfuerzos de las entidades gubernamentales encargadas de diseñar las políticas laborales en las viviendas más vulnerables a cambios de estados laborales, en especial en aquellas que tienen un estado laboral activo y sean más propensas a perderlo, pasando a ser parte de la estadística del desempleo en Costa Rica.

4.1.1.3. Criterio de éxito

Como criterio de éxito para este proyecto, se establece el diseño, desarrollo, ajuste y presentación de un modelo de aprendizaje automático capaz de aprender a interpretar las variables más representativas y predecir si existe riesgo para los habitantes de las viviendas individuales ocupadas en cambiar de estado de empleado a desempleado, o si existe la ganancia de salir del desempleo y formar parte del mercado laboral del país.

Este criterio de éxito pretende que las entidades gubernamentales busquen mejorar las condiciones laborales de los costarricenses mediante la generación de políticas más eficaces en la lucha contra el desempleo que azota al país y, aunque no está dentro del alcance de este proyecto, disminuir el impacto provocado por la pandemia de la COVID-19.

4.1.2. Evaluación de la situación actual

En este momento, el INEC es una organización gubernamental competente y fortalecida. Además, cuenta con una amplia gama de sistemas de información capaces de solventar muchas de sus necesidades tanto en el campo, más precisamente al

momento de recolectar la información que integra a la ECE, como en el tratamiento de la información en las oficinas centrales del INEC.

Ahora bien, siendo más específicos en el tema de la ECE, el INEC es el encargado de recolectar la información mediante dispositivos electrónicos, en su defecto usando tabletas con un sistema operativo Android, lo que permite un mejor control de los errores y la calidad de la información suministrada por los encuestados.

Seguidamente, todas las entrevistas son transferidas a las oficinas centrales mediante un protocolo llamado FTP (Protocolo de Transferencia de Archivos) cada cierto tiempo, puede ser de forma diaria o semanal.

Una de las características más importantes de la ECE es su calidad en los datos, lo cual es debido a la cantidad de filtros que usa el INEC al momento de tratar su información. Uno de estos es el Sistema Administración de Entrevistas para Supervisores (SAES), el cual tiene la función de revisar inconsistencias y favorecer el manejo de reportes de control de calidad.

Después, se inicia el segundo filtro, luego de cargar los datos en los servidores del INEC, donde gracias al Sistema Administrativo de Segmentos o (SASET) se realiza la verificación de la información y, de ser necesario, se corrigen los datos según una serie de parámetros preestablecidos, como la detección de encuestas incompletas o la existencia de inconsistencias en la estructura de las repuestas sugestionadas.

El proceso actual que vive el INEC con respecto a sus datos culmina en la Unidad de Muestreo, la cual con ayuda del sistema estadístico llamado IBM SSPS, busca calcular valores como la estimación, el error típico, el intervalo de confianza, entre otros.

Para finalizar esta evaluación, el INEC está altamente capacitado y la calidad de su información favorece el desarrollo de este proyecto de investigación final, en especial porque dentro de los objetivos clave de la ECE no se encuentra la predicción de ningún tipo de comportamiento en la población meta.

4.1.2.1. Inventario de recursos

Este proyecto de investigación final cuenta para su desarrollo con los siguientes archivos digitales, todos estos extraídos del sitio oficinal de INEC:

- Archivos Microsoft Excel: Único archivo en este formato. Es el encargado de explicarle al interesado en los datos de la encuesta acerca del uso idóneo de la información contenida en los archivos de datos, así como las variables contenidas en las bases de datos, sus características y alcances. Recibe el nombre de “Guía para uso de base de datos ECE PAD.xlsx”.
- Archivos de bases de datos: Son cuatro archivos con extensión *.sav*, cada uno posee la información de un trimestre completo de encuestas para el año 2019.

Definidos los archivos digitales que son usados en este proyecto, se detalla el *software* especializado que emplea el equipo de investigación.

- *Software*: Todo lo asociado con el modelado de la máquina de aprendizaje automático se hace en la herramienta RStudio, teniendo como base el lenguaje R y SQL Server Express Edition para un posible almacenamiento de la información. Para todo lo relacionado con la visualización de la información, se implementa el uso de las librerías *ggplot2*.

4.1.2.2. Requisitos, supuestos y restricciones

A efectos de la investigación, este tema se centra en dos aristas. En primer lugar, se encuentra el uso de la información para el proyecto de investigación final, la cual corresponde a datos de interés público obtenidos por medios informáticos o escritos y de aportación estrictamente voluntaria, siguiendo la Ley n.° 9694, donde se indica: “Se declara de interés público la actividad estadística que permita producir y difundir estadísticas fidedignas y oportunas para el conocimiento veraz e integral de la realidad costarricense, como fundamento para la eficiente gestión administrativa pública y privada” (Ley n.° 9694 de 2019, capítulo II, sección I, art. 4).

Sin embargo, y en segundo lugar, las restricciones a la información para este proyecto se originan en las bases de datos otorgadas por el INEC, dado que se dispone de 307 variables al momento de aplicar la encuesta en las viviendas individuales por trimestre, pero por un tema de confidencialidad estadística, se deben reducir:

> La confidencialidad estadística es la prohibición que tiene el personal de las instituciones del SEN de revelar los datos que se refieran a personas físicas o

> jurídicas determinadas, de los que hayan tenido conocimiento de manera directa o indirecta en el desempeño de sus actividades. Esta prohibición se mantendrá incluso una vez terminado el vínculo con el organismo de que se trate (Ley n.° 9694 de 2019, capítulo II, sección IV, art. 20).

Por lo tanto, cumpliendo la Política para la Divulgación de Estadística, se eliminan 62 variables confidenciales, dentro de las que están el nombre, los apellidos paternos y maternos, los documentos de identificación, las variables geográficas específicas y cualquier otro tipo de información personal que posibilite la identificación de los individuos encuestados.

4.1.2.3. Costos y beneficios

Los datos utilizados en el desarrollo de este proyecto no representan gasto alguno para el investigador, dado que son de acceso público desde el sitio web oficial del INEC, solo es necesario realizar el registro con un correo electrónico personal.

En cuanto a las herramientas de desarrollo que se utilizan, la más importante es RStudio Desktop, la cual cuenta con una licencia de fuente abierta u *open source*, por su nombre en inglés, lo que permite un uso gratuito de la herramienta de forma local, así como de un paquete amplio de librerías de desarrollo.

Cuantificar los beneficios económicos de este proyecto es difícil, en especial porque el objetivo principal está orientado a mejorar la calidad de vida de todos los costarricenses —principalmente de todos aquellos individuos que se encuentran sin un empleo formal o informal que les posibilite llevar el sustento a sus viviendas— mediante la creación de una herramienta que apoye al INEC para ofrecerles análisis a los encargados de crear políticas de generación de empleo.

Los costos asociados a la implementación del proyecto en un ambiente productivo están fuera del alcance de esta investigación, específicamente porque el costo de un ambiente profesional de RStudio puede oscilar entre los $ 995 hasta los $ 11 950 anuales y no es posible determinar si el INEC está interesado en adquirir dichas licencias. Además, no está contemplado el costo de los especialistas o de la infraestructura.

4.1.3. Objetivos de la minería de datos

Dentro del paquete de datos por analizar, se cuenta con 307 variables de acceso público, divididas en diez secciones, lo cual lo convierte en un conjunto de datos complejo. Por esta razón, en los objetivos de la minería de datos de este proyecto se contempla lo siguiente:

- Seleccionar las variables más representativas del conjunto de datos, mediante un proceso de análisis exploratorio de la correlación de las variables independientes con la variable por predecir, el cual le permite al investigador observar las variables con una mayor relación a la variable objetivo.
- Predecir la probabilidad de que un ocupante de una vivienda individual habitada pueda variar su estado de ocupación laboral, pasando del desempleo al empleo o, en caso contrario, de una actividad productiva remunerada a engrosar el porcentaje de personas desempleadas en el país.
- Determinar cuáles son las ocupaciones más representativas al cambio de estado, pero dando prioridad al estudio y análisis de los casos donde el estado laboral es empleado y se da un cambio al estado de desempleado.

4.1.4. Criterio de éxito de la minería de datos

En el presente apartado, y con base en los objetivos descritos en el punto anterior, se divide el criterio de éxito en dos partes. Primero, efectuar un análisis exploratorio detallado que le permita al investigador de este proyecto definir las variables más representativas por utilizar en la generación del modelo de minería de datos principal, teniendo el cuidado necesario de no descartar ninguna variable sin antes verificar su relación con la variable por predecir, porque existe la particularidad de que cada reproducción de la regresión origine un análisis diferente con variables distintas, quedando a discreción del desarrollador del proyecto la decisión de cuál reproducción del modelo utilizar.

El criterio de éxito principal se basa en la creación de un modelo de minería de datos que pueda determinar la probabilidad de que un ocupante de una vivienda individual habitada varíe su estado de ocupación laboral, con la ayuda de las variables con mayor relación encontradas en el criterio de éxito anterior.

4.1.5. Plan del proyecto

El plan de proyecto va de la mano con las etapas de la metodología CRIPS-DM, descrita en el apartado 3.7, “Técnicas de análisis de la información”, donde se indican los seis pasos por seguir, los cuales pueden ser ejecutados de forma lineal o reiterada, repitiendo pasos las veces que se considere necesario. Al respecto, la metodología permite devolverse a pasos anteriores con el objetivo de reforzar el entendimiento de los datos o mejorar el modelo de aprendizaje automático, lo cual es una gran ventaja de la metodología CRISP-DM.

4.2. Entendimiento de los datos

Para iniciar esta etapa del proyecto, es necesario descargar los datos del sitio oficial del INEC (la referencia al sitio web se puede encontrar en la bibliografía de este documento). Los únicos requisitos para la descarga de la información son la creación de una cuenta de registro asociada a un correo electrónico personal y una contraseña segura.

El entendimiento de los datos es fundamental. El encargado del desarrollo del presente proyecto final de graduación debe comenzar con la aclimatación de los datos y a generar sus propias conjeturas y conclusiones.

4.2.1. Recolección de los datos

Los datos son recolectados mediante el modo de encuestas cara a cara con el entrevistado en cerca de 9 024 viviendas individuales ocupadas de forma continua cuatro veces al año, con una rotación del 25 %, lo cual garantiza que un 75 % de las viviendas coincidan entre una encuesta y otra, y en los cuatro trimestres de encuestas, la muestra de viviendas se renueva al 100 %.

El personal responsable de la aplicación de las encuestas está conformado por los siguientes puestos, ordenados por su estructura organizacional: un coordinador, un encargado operativo de campo, supervisores generales, supervisores de zona, encargados de los operadores de equipo móvil y operadores que realizan las encuestas.

La captura de la información se efectúa por medio de dispositivos móviles, en su mayoría tabletas que poseen el sistema operativo Android, lo que permite predisponer

algunas de las repuestas, disminuyendo el error humano al momento de ingresar las respuestas a las preguntas aplicadas.

En los lugares peligrosos o de difícil acceso geográfico donde no es posible utilizar los medios tecnológicos, el personal está capacitado para hacer las preguntas en cuestionarios de papel, pero al final de la semana son transcritos a medios digitales. También, si el entrevistado lo solicita, la encuesta es realizada por teléfono, en caso de que el individuo no pueda atender al personal en un horario diurno normal.

El formulario de preguntas es la parte central en la recolección de los datos. Las preguntas están definidas con cierta correspondencia respecto al cuestionario de la Enaho, en relación con variables demográficas y de empleo. Además, se toman en cuenta las recomendaciones señaladas por la OIT, asegurando una homogeneidad entre los datos obtenidos en el territorio nacional con los datos que obtienen otras entidades a nivel internacional.

El cuestionario está formado por diez secciones, las cuales se detallan a continuación:

Tabla 11: *Secciones del cuestionario*

Sección	Descripción
Identificación hogares	Contiene variables que permiten la identificación de las viviendas individuales ocupadas, como el año, el trimestre, la vivienda y el individuo.
Sección A	Presenta los datos personales, educación e idioma, así como las características sociodemográficas.
Sección B	Son los datos relacionados con el trabajo y la condición de la actividad económica.
Sección C	Información vinculada a la actividad económica que se ejecuta en el trabajo o en la empresa, ligada a los individuos ocupados.
Sección D	Datos del tiempo laborado para las personas ocupadas de forma independiente, como lo son la duración del trabajo, las actividades del negocio, los rebajos sociales, entre otros.
Sección E	Son las variables que comprenden el horario del trabajo de las personas ocupadas asalariadas. Dentro de los valores, se pueden encontrar días de trabajo seguidos, días de descanso seguidos, horario, entre otros.
Sección F	Esta sección se centra en el trabajo secundario, junto con la actividad, horas y forma de pago.
Sección G	Contiene información del empleo inadecuado y del tiempo de más que se gasta ejerciéndolo, como las horas extraordinarias.
Sección H	Datos de los individuos desempleados, como su último trabajo, el tiempo que tiene sin poder desempeñar un empleo y si actualmente está en busca de una actividad que le genere una entrada económica.
Sección I	Es la última de las secciones del cuestionario y centra su atención en las labores agrícolas o de producción de productos de autoconsumo del hogar.

Fuente: Elaboración propia

Los archivos que contienen los datos por utilizar en este proyecto se encuentran distribuidos por trimestre, como se aprecia en la tabla 12:

Tabla 12: *Descripción de los archivos de datos*

Nombre de archivo	Peso	Descripción
I Trimestre 2019.sav	9.842 KB	Información de las encuestas aplicadas en el primer trimestre del 2019.
II Trimestre 2019.sav	9.792 KB	Información de las encuestas aplicadas en el segundo trimestre del 2019.
III Trimestre 2019.sav	9.688 KB	Información de las encuestas aplicadas en el tercer trimestre del 2019.
IV Trimestre 2019.sav	9.649 KB	Información de las encuestas aplicadas en el cuarto trimestre del 2019.

Fuente: Elaboración propia

4.2.2. Descripción de los datos

Los archivos están en formato *.sav*, que es una extensión genérica para almacenar datos, y cada uno de los archivos contiene la siguiente cantidad de observaciones:

Tabla 13: *Cantidad de observaciones por archivo*

Nombre del archivo	Cantidad de observaciones
I Trimestre 2019.sav	25 530
II Trimestre 2019.sav	25 695
III Trimestre 2019.sav	25 397
IV Trimestre 2019.sav	25 342

Fuente: Elaboración propia

El siguiente paso es la descripción de cada una de las variables que conforman los archivos de datos; no obstante, entablar una lista con todas las 307 variables de acceso público es poco eficiente e innecesario.

En cambio, se realiza un análisis preliminar de la cantidad de variables que componen las diferentes secciones del conjunto de datos, lo que facilita mucho el trabajo de descripción y escogencia de las variables más importantes para el desarrollo del modelo.

Tabla 14: *Cantidad de variables por sección*

Sección	Cantidad de variables
Identificación Hogares	6
Sección A	21
Sección B	8
Sección C	16
Sección D	40
Sección E	35
Sección F	26
Sección G	7
Sección H	10
Sección I	11

Fuente: Elaboración propia

Se toma la decisión de enseñar en el presente proyecto de graduación final únicamente las variables de la sección A, porque es la sección que cuenta con las variables más objetivas, con mejor relación entre las mismas y más nutridas con respecto a valores no vacíos, lo que le posibilita al modelo aprender con información real y no con datos no contestados.

Tabla 15: *Descripción de las variables*

Variable	Descripción	Tipo de dato	Sección
Relacion_parentesco	Relación de parentesco que tienen los miembros de la vivienda individual ocupada.	Factor de 20 niveles	Sección A
Sexo	Género del individuo que contesta la encuesta.	Factor de 2 niveles	Sección A
Edad	Edad de la persona que habita en la vivienda.	Numérico	Sección A
Estado_conyugal	Relación sentimental del individuo.	Factor de 9 niveles	Sección A
Lugar_nacimiento	Cuando el individuo nace, ¿dónde vive su madre?	Factor de 4 niveles	Sección A
Permanencia_pais	¿Cuánto tiempo tiene el individuo de vivir en Costa Rica?	Factor de 3 niveles	Sección A
Permanencia_intencion	¿Tiene el individuo intenciones de seguir viviendo o establecerse en el país?	Factor de 3 niveles	Sección A
Permanencia_motivo	¿Por qué motivo se encuentra en este país?	Factor de 4 niveles	Sección A
Seguro	¿Tiene el individuo seguro social o lo cubre el seguro de algún familiar?	Factor de 3 niveles	Sección A
Tipo_seguro	¿Qué tipo de seguro tiene el individuo?	Factor de 10 niveles	Sección A
Regimen_pension	¿Para cuál régimen de pensiones cotiza el individuo?	Factor de 3 niveles	Sección A
Plan_voluntario	¿Está el individuo afiliado a algún plan voluntario de pensiones complementarias?	Factor de 3 niveles	Sección A
Educacion_asiste	Indica si el individuo asiste a un centro educativo.	Factor de 8 niveles	Sección A
Educacion_nivel_grado	El valor de esta variable muestra el último grado aprobado por el individuo.	Factor de 39 niveles	Sección A
Educacion_titulo	Variable que señala el último título recibido por el individuo.	Factor de 10 niveles	Sección A
Educacion_codigotitulo	Despliega el código de la carrera.	Factor de 49 niveles	Sección A
EducacionNoregular_asiste	Aparte de la educación regular, ¿el individuo recibe algún curso u otro tipo de formación del que tenga título o certificación?	Factor de 3 niveles	Sección A

EducacionNoregular_codigo	Código del curso de la educación no regular al que asiste el individuo.	Factor de 49 niveles	Sección A
EducacionNoregular_institucion	Nombre del instituto o centro donde cursa la educación no regular.	Factor de 10 niveles	Sección A
Idioma	Indica si el individuo habla, lee y escribe fluidamente algún otro idioma aparte de su lengua materna.	Factor de 3 niveles	Sección A
Idioma_cual	Muestra cuáles son las lenguas extranjeras que habla, lee y escribe el individuo.	Factor de 6 niveles	Sección A
Tipo_poblacion	Variable que señala si la población es joven o adulta.	Factor de 2 niveles	Sin sección
Pais_nacimiento	Despliega el país de nacimiento de los individuos entrevistados.	Factor de 11 niveles	Sin sección
Provincia_nacimiento	Indica la provincia donde nace el individuo, basándose en la división territorial administrativa de Costa Rica disponible en el Manual de Clasificación Geográfica con Fines Estadísticos de Costa Rica.	Factor de 8 niveles	Sin sección
Region	Se refiere a las regiones establecidas por el MIDEPLAN. Se mencionan seis regiones: Central, Chorotega, Pacífico Central, Brunca, Huetar Norte y Huetar Caribe.	Factor de 6 niveles	Sin sección
Zona	Variable constituida por la zona de residencia donde se ubica el hogar individual encuestado, los niveles son determinados por la Unidad de Cartografía del INEC.	Factor de 2 niveles	Sin sección
Nivel_educativo	Corresponde al grado más avanzado de estudios aprobados dentro del ciclo de educación regular.	Factor de 8 niveles	Sin sección

Fuente: Elaboración propia

Es muy importante acotar que las variables descritas en la tabla anterior no son las definitivas, es posible excluir o agregar más variables si el desarrollo del presente proyecto lo requiere. No obstante, si es necesario agregar variables de otras secciones, el investigador cuenta con la obligación de realizar una breve explicación de la variable, su alcance, el tipo de dato y la razón por la cual se incluye en el análisis.

En el resumen anterior se indican ciertas variables que pueden ser de mucha utilidad. Conforme avance el proyecto de investigación, el encargado del estudio profundiza en el análisis de los datos, donde es posible observar comportamientos significativos en las variables, como en las correspondientes a la edad y la cantidad de individuos con algún título en educación.

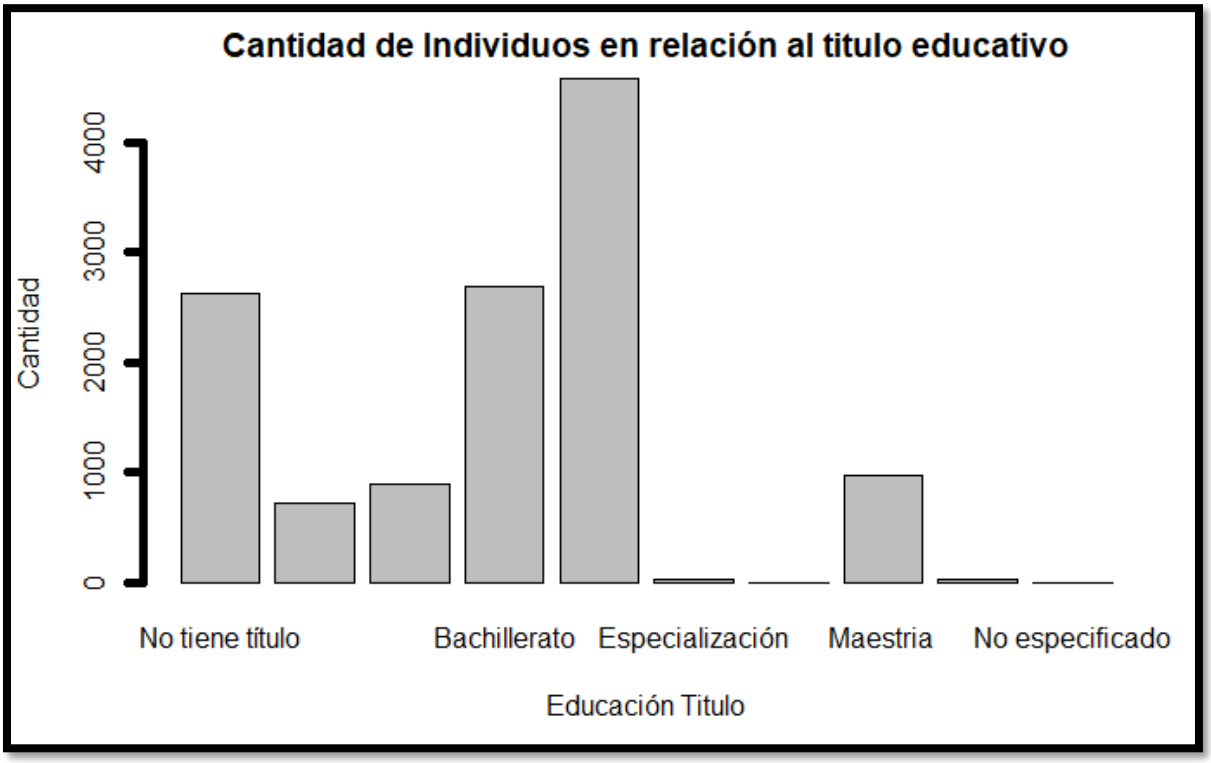


Ilustración 4: Educacion_titulo por cantidad de observaciones
Fuente: Elaboración propia

La ilustración 4 contiene la cantidad de observaciones según la variable llamada *Educacion_titulo*, donde se resalta la gran cantidad de individuos que no tienen un título, siendo casi igual a las observaciones de individuos con bachillerato. Estos dos factores pueden ser muy relevantes en el aprendizaje del modelo.

4.2.3. Exploración de los datos

La presente exploración de datos se realiza con la totalidad de las observaciones. Luego de agrupar cada una de las encuestas trimestrales, se contabilizan 101 964

muestras. Se espera que cada una de las visualizaciones expuestas dentro de este apartado sirva como guía para escoger el modelo más adecuado en función al objetivo principal de este proyecto.

En primer lugar, se inicia con la visión global de la variable por predecir, de forma tal que marque las pautas para las siguientes visualizaciones.

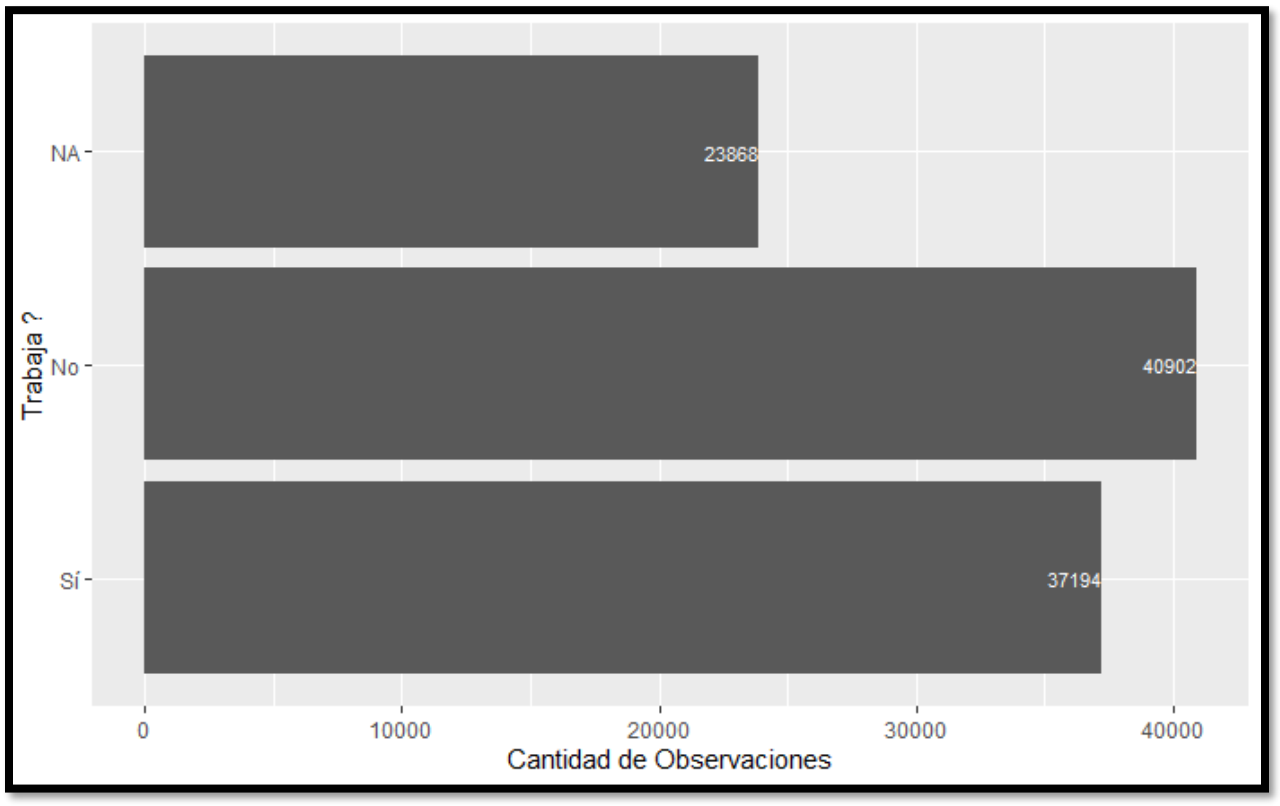


Ilustración 5: Cantidad de observaciones por la variable *Trabajo*
Fuente: Elaboración propia

En la ilustración 5 se destaca la cantidad de individuos sin trabajo, a saber, casi 41 mil personas, 4 mil más que los individuos que sí cuentan con un trabajo y, por ende, con un ingreso para sus hogares. Con relación a lo expuesto, una mayor cantidad de individuos desempleados enciende las alarmas de una sociedad que no sabe cómo atacar este problema social.

Se debe tener cuidado con el número de observaciones sin respuesta (cerca de 24 mil), pues hace pensar que muchas de las personas encuestadas tienen algún tipo

de timidez en decir que no cuentan con un trabajo estable, por lo que prefieren no contestar la pregunta. Estos valores no aportan al aprendizaje del modelo y, por consiguiente, en el apartado de “Selección de los datos” son separados del conjunto de trabajo.

4.2.3.1. Datos básicos de los individuos

El perfil de los individuos que conforman la encuesta es analizado en las siguientes visualizaciones:

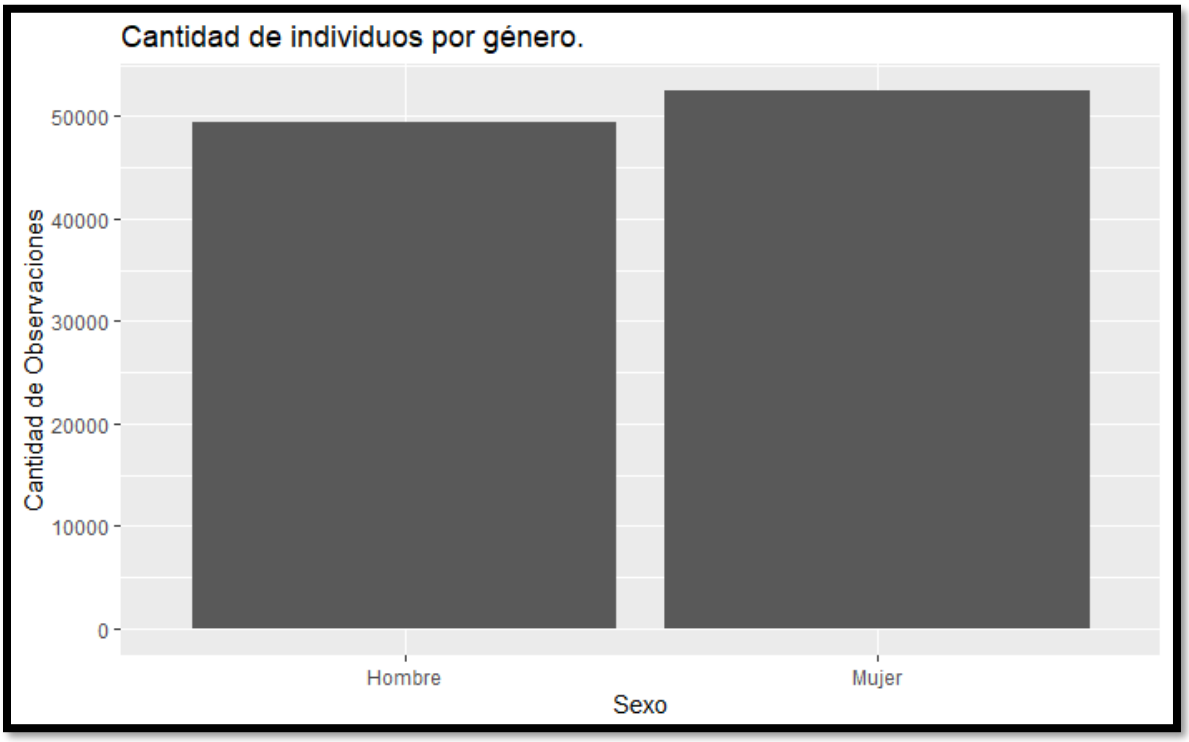


Ilustración 6: Cantidad de individuos por género
Fuente: Elaboración propia

La primera exploración muestra un número ligeramente mayor de mujeres en el set de datos (cerca de 3 000 observaciones), lo cual es muy aceptable para esta investigación y genera confianza por ser un conjunto balanceado, donde el 51.47 % son mujeres y un 48.53 % son hombres.

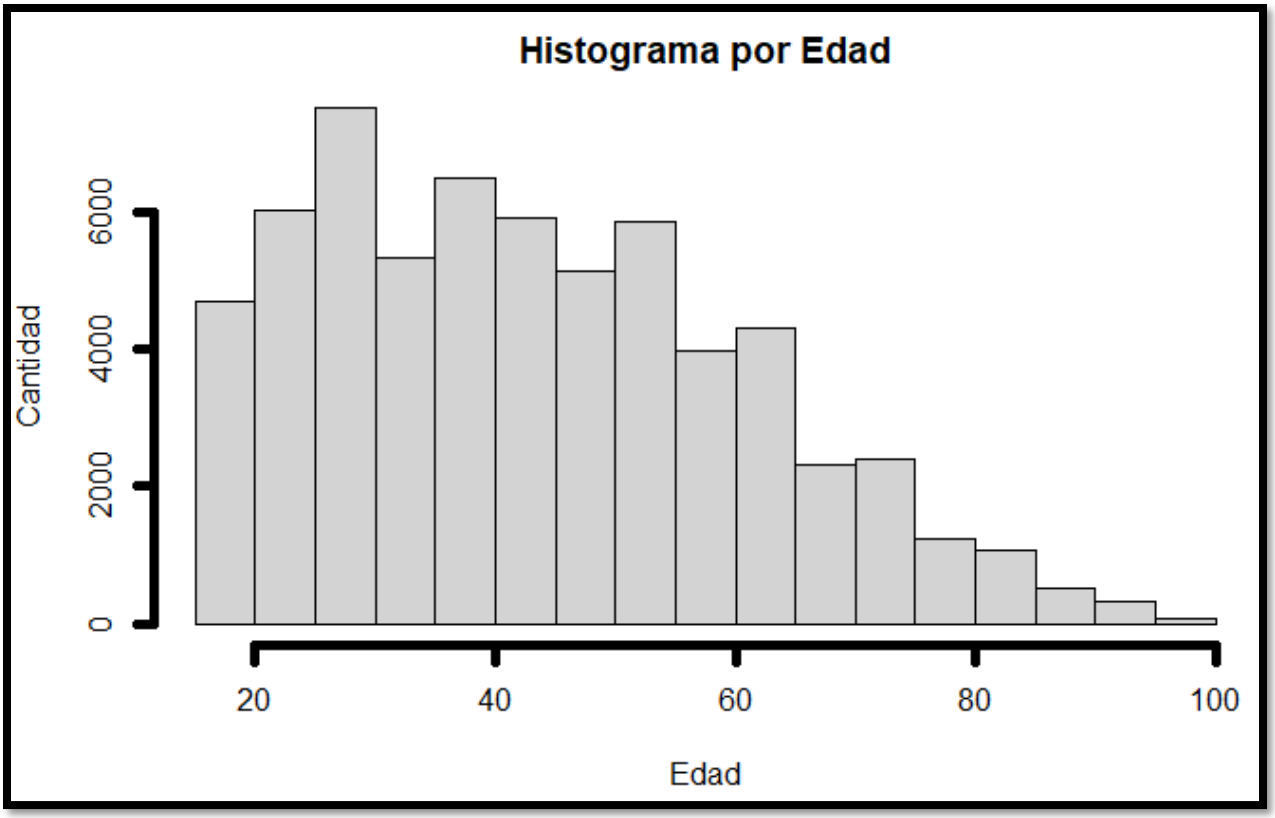


Ilustración 7: Histograma de observaciones por edad
Fuente: Elaboración propia

La segunda exploración despliega un histograma de la edad de los individuos, donde el grueso de las observaciones se encuentra entre los 20 y 60 años. Este análisis ayuda en especial al momento de limpiar los datos, dado que los valores por debajo de los 18 años no le son útiles al modelo, por un tema de legalidad laboral.

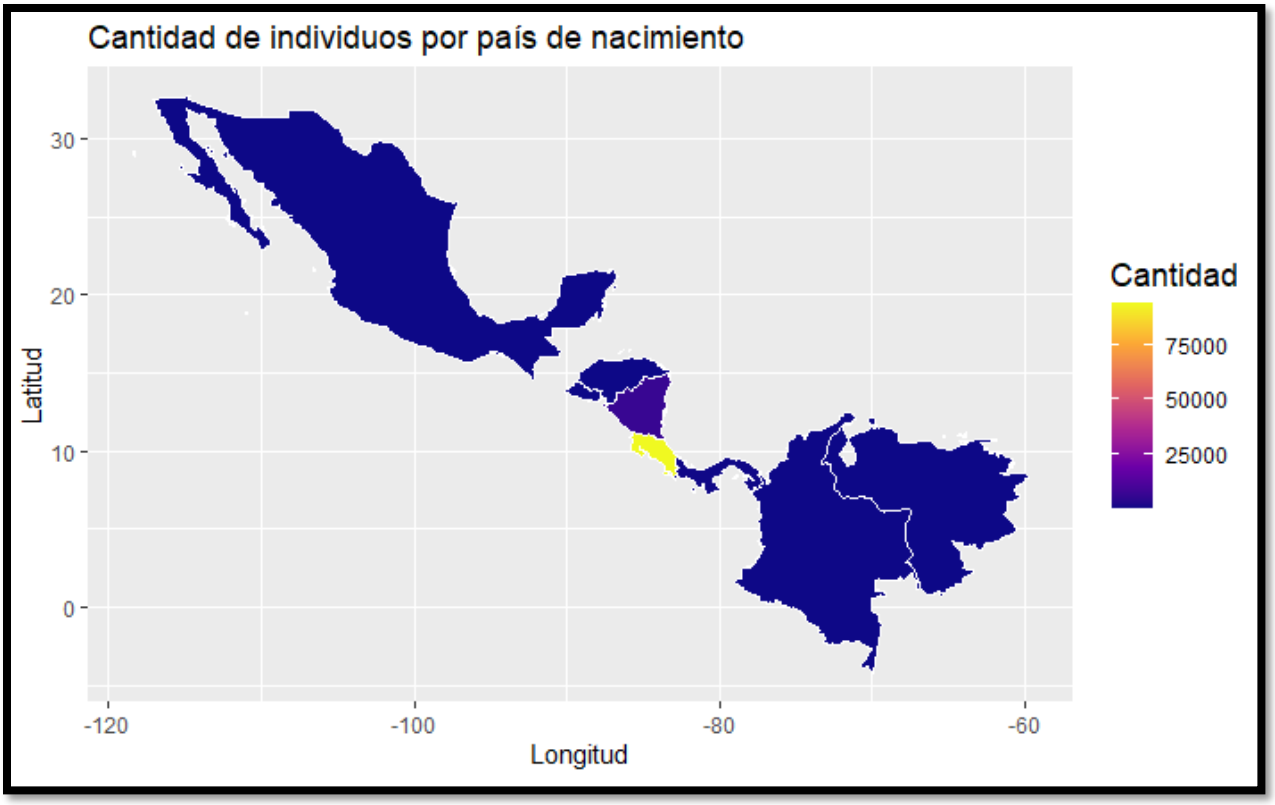


Ilustración 8: Mapa geográfico de la cantidad de individuos por país de nacimiento
Fuente: Elaboración propia

Las personas encuestadas son en su gran mayoría nacidas en Costa Rica, sin embargo, hay individuos de países como Venezuela, México, Colombia, Panamá, Nicaragua, Honduras y El Salvador, aunque en menor medida, pero no por esto dejan de ser significativos para el aprendizaje del modelo. En la ilustración 9 se indica la cantidad respectiva de observaciones por país de nacimiento.

4.2.3.2. Datos geográficos de los individuos

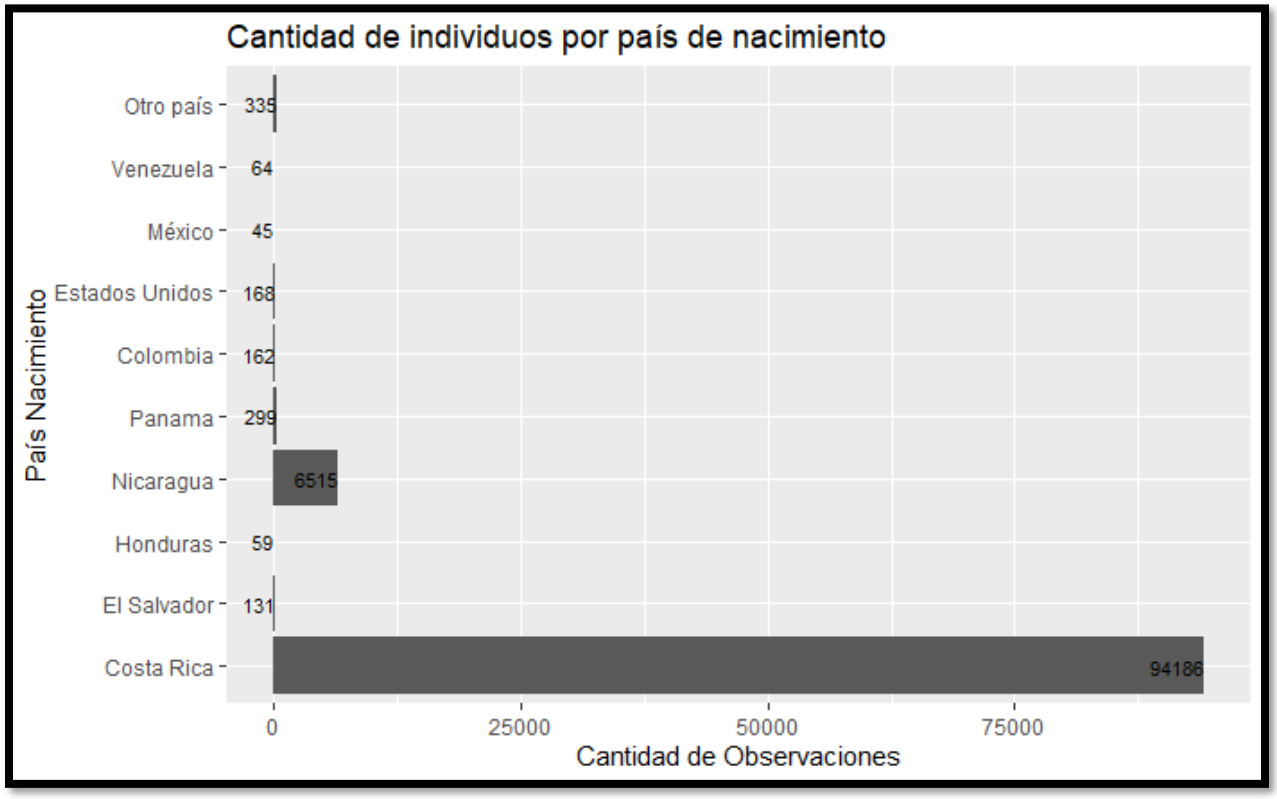


Ilustración 9: Cantidad de individuos por país de nacimiento
Fuente: Elaboración propia

El 92.37 % de las observaciones son costarricenses, seguido por el 6.38 % de individuos nacidos en Nicaragua y el restante 1.25 % de otras nacionalidades. Esta variable permite identificar plenamente el desempleo en los costarricenses.

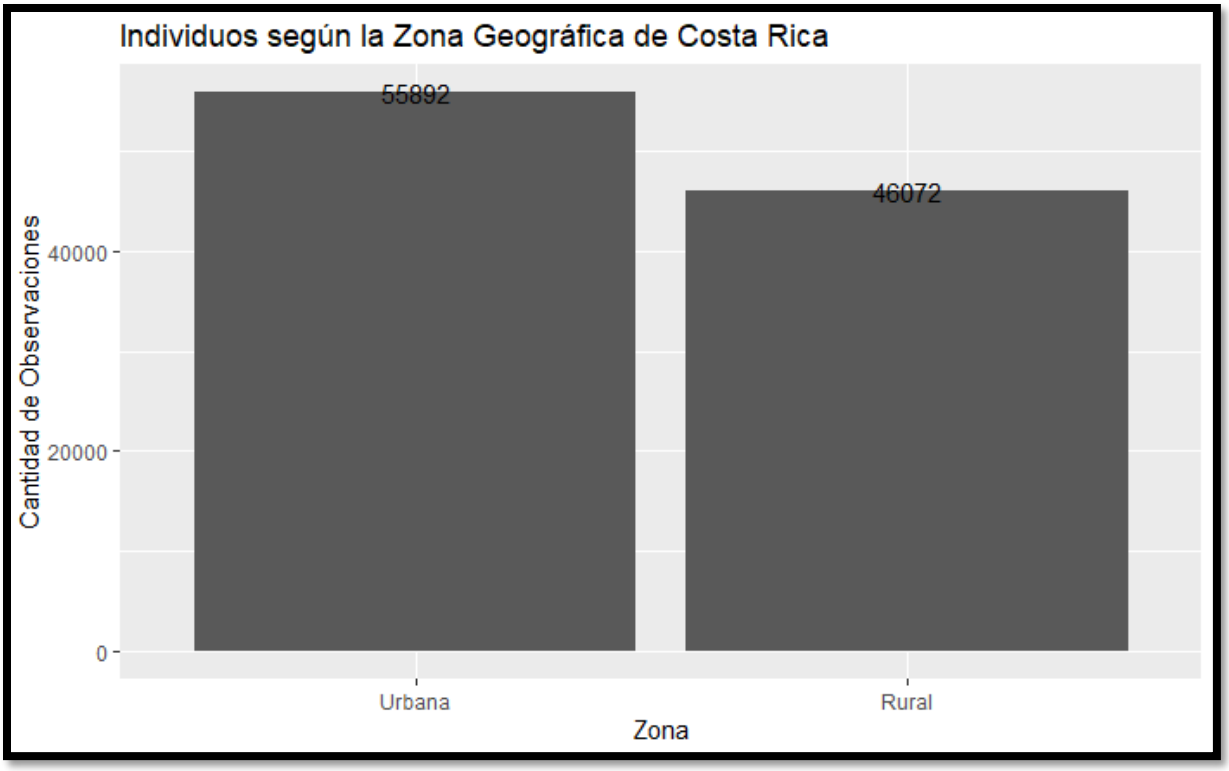


Ilustración 10: Individuos según la zona geográfica de Costa Rica
Fuente: Elaboración propia

La clasificación de zonas nace en el Censo Nacional de Población y Vivienda del año 2011 por la necesidad de crear una Unidad Geoestadística Mínima; así, se divide en zona urbana y zona rural. Al respecto, una zona urbana se define como:

> […] áreas que se delimitaron a 'priori' para el Censo Nacional de Población del 2011, con criterio físico y funcional, tomando en cuenta elementos tales como: cuadrantes claramente definidos, calles, aceras, servicios urbanos (recolección de basura, alumbrado público) y actividades económicas como: industria, grandes comercios y servicios diversos. La delimitación geográfica se realizó a partir de los centros administrativos de cada cantón o distrito y se amplió de manera compacta en función de las características antes señaladas; además se consideraron como urbanos otros conglomerados de viviendas ubicados fuera se ese compacto (barrios, condominios y otros sentamientos), que poseen características como las descritas para las zonas urbanas (INEC, 2016, p. 13).

Por otro lado, la zona rural se entiende del siguiente modo:

> […] el resto de áreas del país no ubicadas en el área urbana, que reúnen características, tales como: un predominio de actividades agropecuaria, pecuarias, silvícola y turísticas; dentro de esas zonas se pueden encontrar conglomerados de viviendas y viviendas dispersas; así como centros poblados, con disposición de servicios de infraestructura como electricidad, agua potable y teléfono; cuentan con servicios como escuela, iglesia, parque o plaza de esparcimiento, centro de salud, guardia rural, etc.; pequeños o medianos comercios relacionados algunos con el suministro de bienes para la producción agrícola; y un nombre determinado que distintivo que los distingue de otros poblados o caseríos (INEC, 2016, p. 14).

En relación con lo expuesto, la cantidad de observaciones es más alta en la zona urbana, con 55 892 registros (54.82 %), lo que concuerda con la densidad de la población, la cual se ubica en su mayoría en el área metropolitana; no obstante, la zona rural no se queda atrás y posee en la actualidad 46 072 registros para un 45.18 % del total de observaciones, un muy buen indicador de lo balanceado que se encuentra el conjunto de datos.

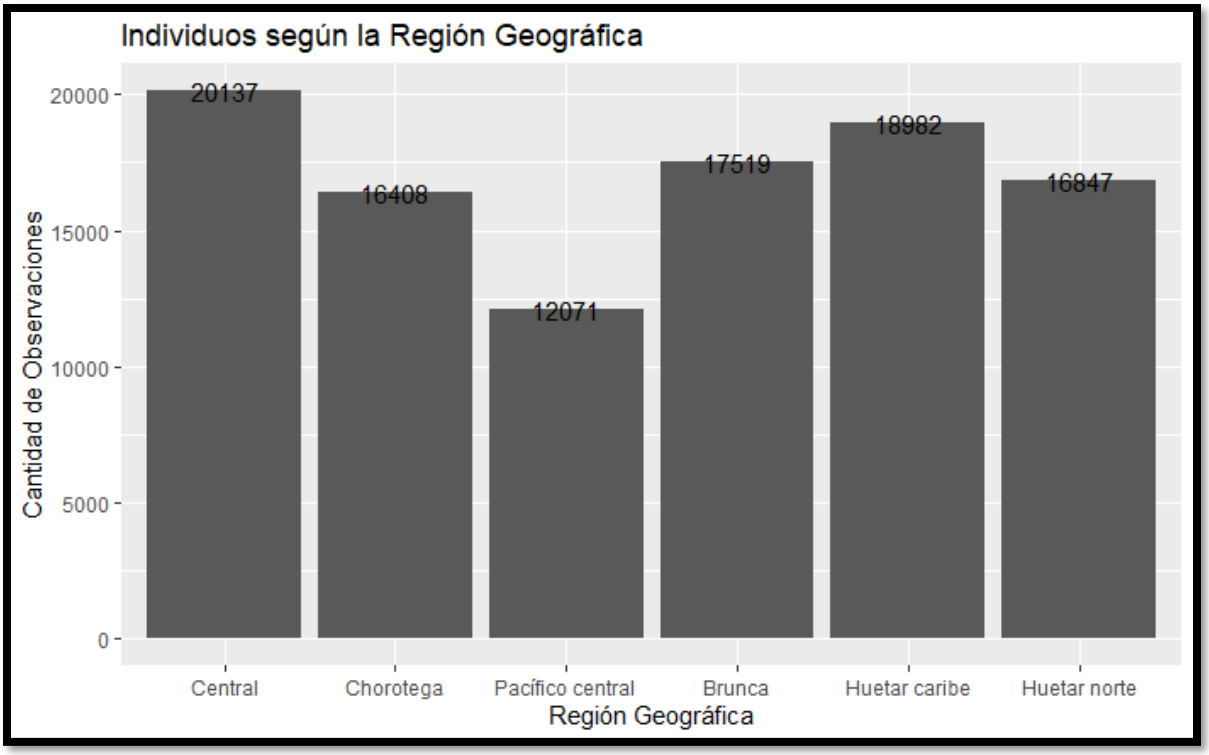


Ilustración 11: Cantidad de observaciones según la región geográfica de Costa Rica
Fuente: Elaboración propia

La definición de las regiones se encuentra a cargo del Sistema Nacional de Planificación (SNP), que a su vez forma parte del Ministerio de Planificación Nacional y Política Económica (MIDEPLAN) de Costa Rica.

En cuanto a esto, se establecen seis grandes regiones: la primera es la llamada Central, compuesta por los principales cantones de San José, Alajuela, Heredia y Cartago; luego la región Chorotega, donde se incluyen los principales cantones de la provincia de Guanacaste; seguidamente la región del Pacífico Central, donde se abarcan los cantones de Alajuela y Puntarenas; después, la región Brunca, configurada por los cantones de Puntarenas Sur y Pérez Zeledón de San José; por último, las regiones Huetar Norte y Caribe, que se dividen los cantones de la provincia de Limón, los cantones al norte de Alajuela y Heredia.

A nivel de observaciones, es muy importante que dicha variable no cuenta con observaciones vacías y que las cantidades se encuentran relativamente cercanas una

de la otra, teniendo mayor cantidad la región Central, con 19.75 %, y menor cantidad el Pacífico Central, con 11.84 %.

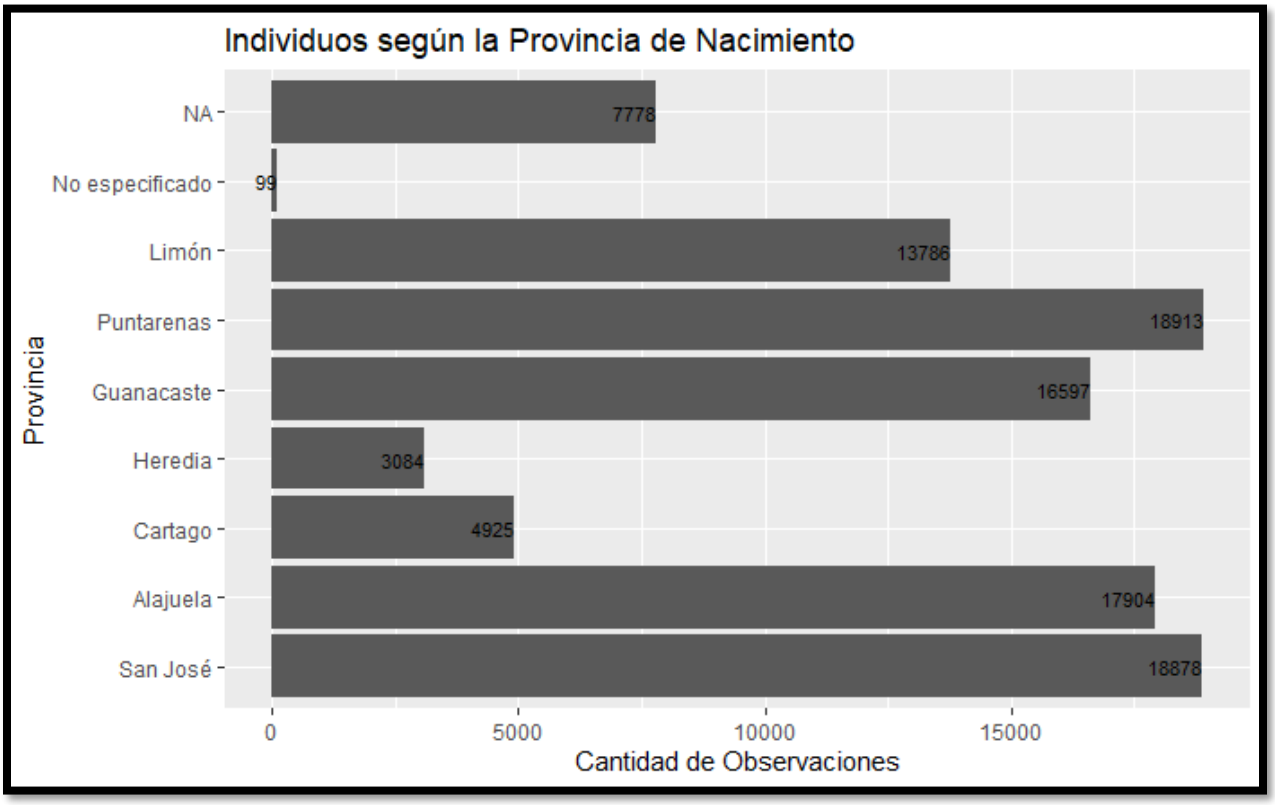


Ilustración 12: Gráfica de la cantidad de individuos según la provincia de nacimiento
Fuente: Elaboración propia

Cada ECE ofrece una división territorial administrativa elaborada por el mismo INEC, que define el concepto de provincia como:

> […] territorio geográfico con una población no menor al diez por ciento de la población total del país, y que por ley deben conservar al menos ese mismo porcentaje de población. Los límites geográficos serán acordados por el Instituto Geográfico Nacional (IGN) (INEC, 2016, p. 9).

Una vez entendido el término de provincia, se analizan los datos obtenidos en dicha variable. Afortunadamente la cantidad de observaciones sin respuesta no supera

el 7.62 % de la muestra, lo cual es un buen indicador si se desea utilizar esta variable para enseñarle al modelo a predecir. Otro aspecto positivo es la gran cantidad de individuos encuestados fuera de la gran área metropolitana. Como se observa, Puntarenas lidera con el 18.55 % de las observaciones, seguido muy de cerca por Guanacaste con el 16.28 % y Limón con el 13.52 %.

Sin embargo, no deja de preocupar la cantidad de observaciones generadas en Heredia y Cartago, con el 3.02 % y el 4.83 % respectivamente; en especial porque son provincias que cumplen con la función de hospedaje, dado que la mayoría de los trabajadores sale muy temprano a ejercer sus labores en otras provincias como San José y regresan en las noches a descansar y dormir.

4.2.3.3. Datos educativos de los individuos

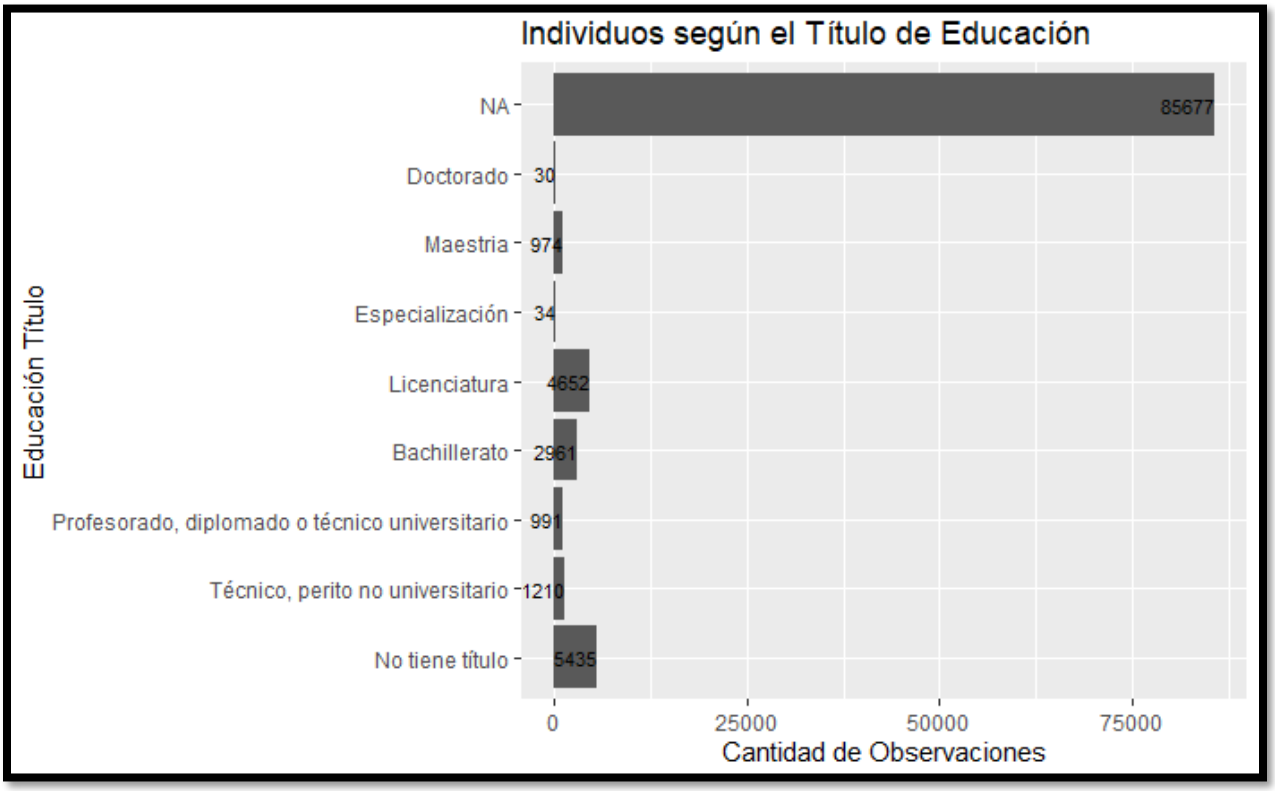


Ilustración 13: Cantidad total de observaciones con título de educación
Fuente: Elaboración propia

La variable denominada *Educacion_titulo* tiene un porcentaje muy alto de observaciones sin respuesta (84 %), lo cual no es positivo para el modelo, porque no puede aprender de valores vacíos, pero es una variable que contribuye mucho al modelo si se limpian los datos, porque se puede determinar si el valor del empleo se encuentra estrechamente ligado a un título en educación o si, por el contrario, no tener un título en educación va de la mano con el desempleo existente en el país.

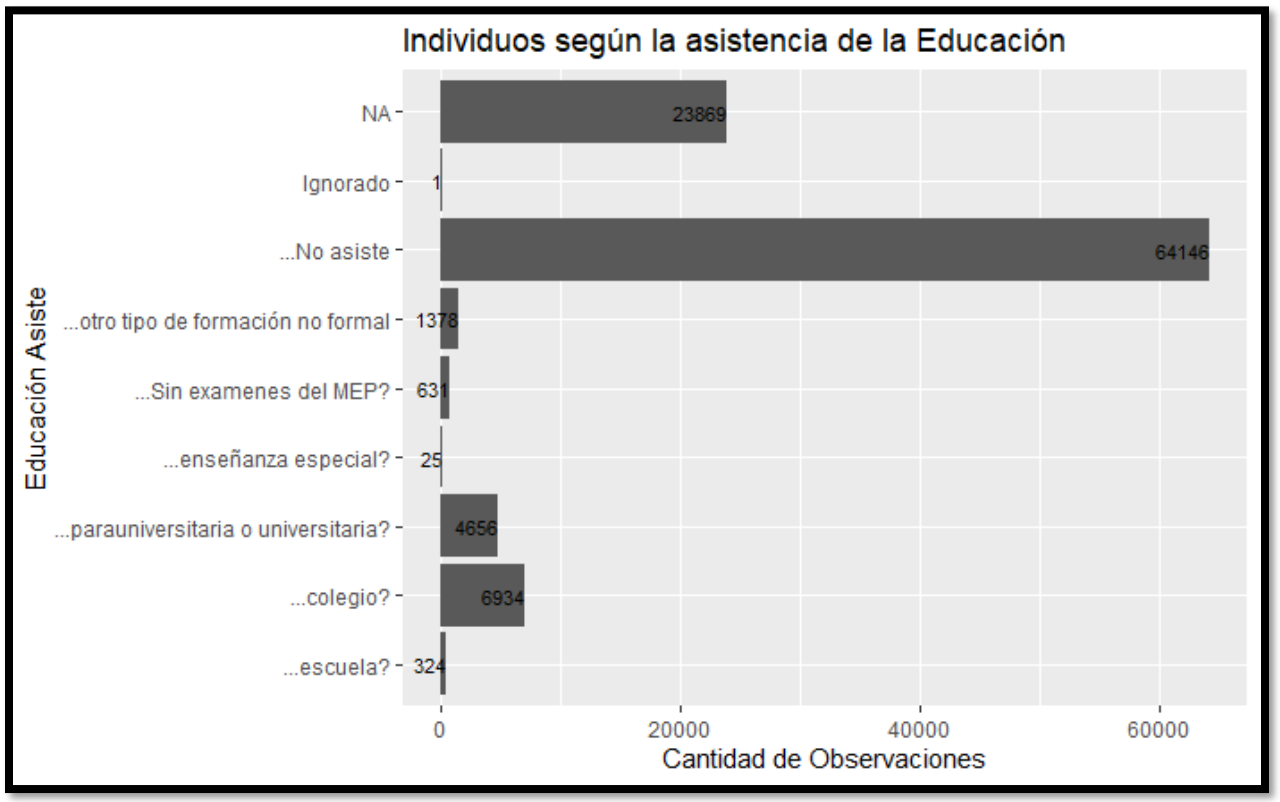


Ilustración 14: Cantidad total de individuos que asisten a un centro educativo
Fuente: Elaboración propia

La asistencia a un centro educativo es de gran aporte para el crecimiento humano. El conocimiento se encuentra directamente vinculado a la superación de las personas, así como les ofrece mejores puestos de trabajo y mayores ingresos económicos. Visualizando los datos de la ilustración 14, es muy preocupante que el 62.91 % de los individuos no asista a ningún centro educativo del país, factor que puede tener una alta relación con el desempleo.

Al igual que ocurre con la variable *Educacion_titulo*, la variable de asistencia a centros educativos tiene un 23.40 % de observaciones sin respuesta, pero es considerada como valiosa para el modelo, por lo que en la sección de “Limpieza de datos” se valora quitar o no los valores vacíos.

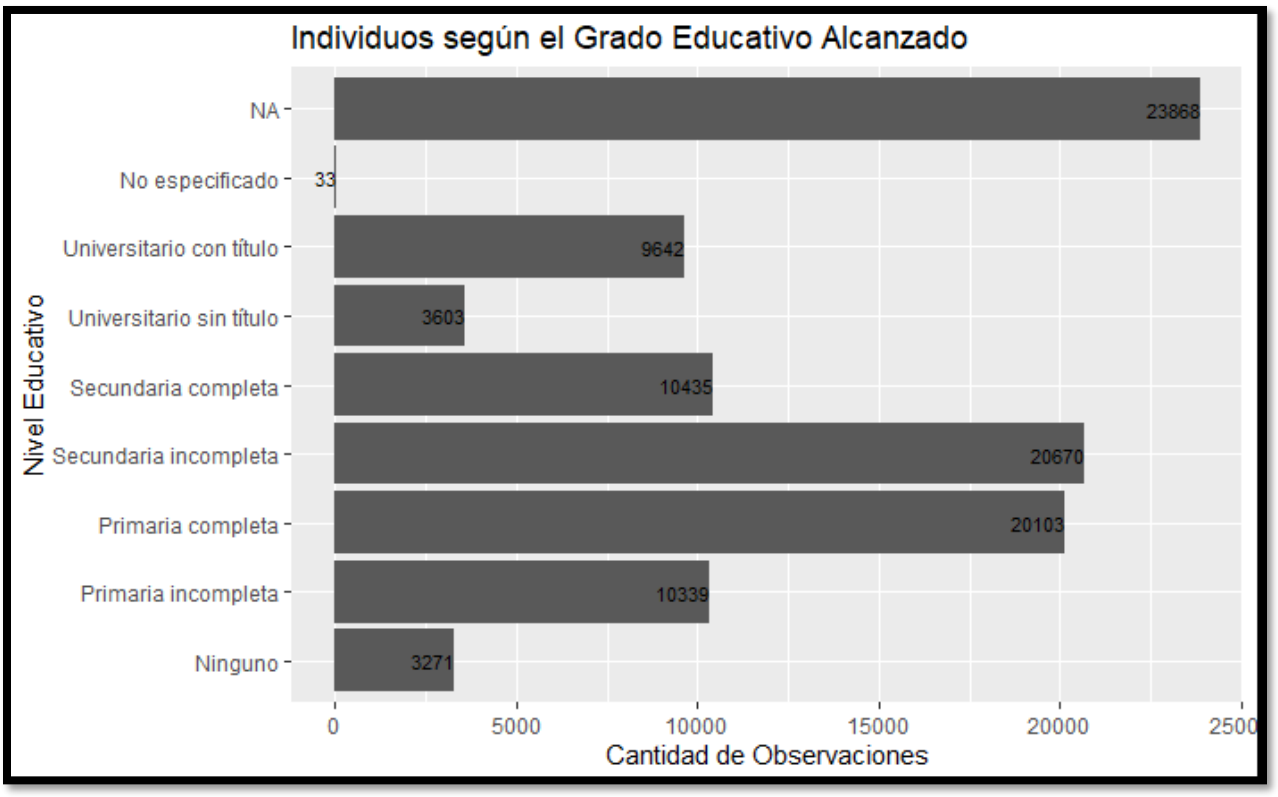


Ilustración 15: Ilustración de la cantidad de individuos por nivel educativo
Fuente: Elaboración propia

En la ilustración 15 se aprecia cómo la mayoría de los participantes sigue estando en la población que no responde la pregunta, pero dicha cantidad no supera el 25 % de las observaciones por lo que la variable puede ser utilizada en el proceso de aprendizaje del modelo.

La gráfica también muestra cómo el grueso de los individuos entrevistados se ubica entre las etapas de primaria completa y secundaria incompleta, con el 39.98 %, un valor que da a entender que los individuos buscan la forma de combatir el desempleo con el conocimiento.

Otros dos valores importantes son la secundaria completa con un 10.23 % de la cantidad total de individuos y el 9.45 % que representa el tamaño de los individuos que terminan sus estudios universitarios y obtienen un título, dando a entender que existe una buena porción de observaciones donde se busca una especialización académica superior que les permita acceder a mejores condiciones laborales.

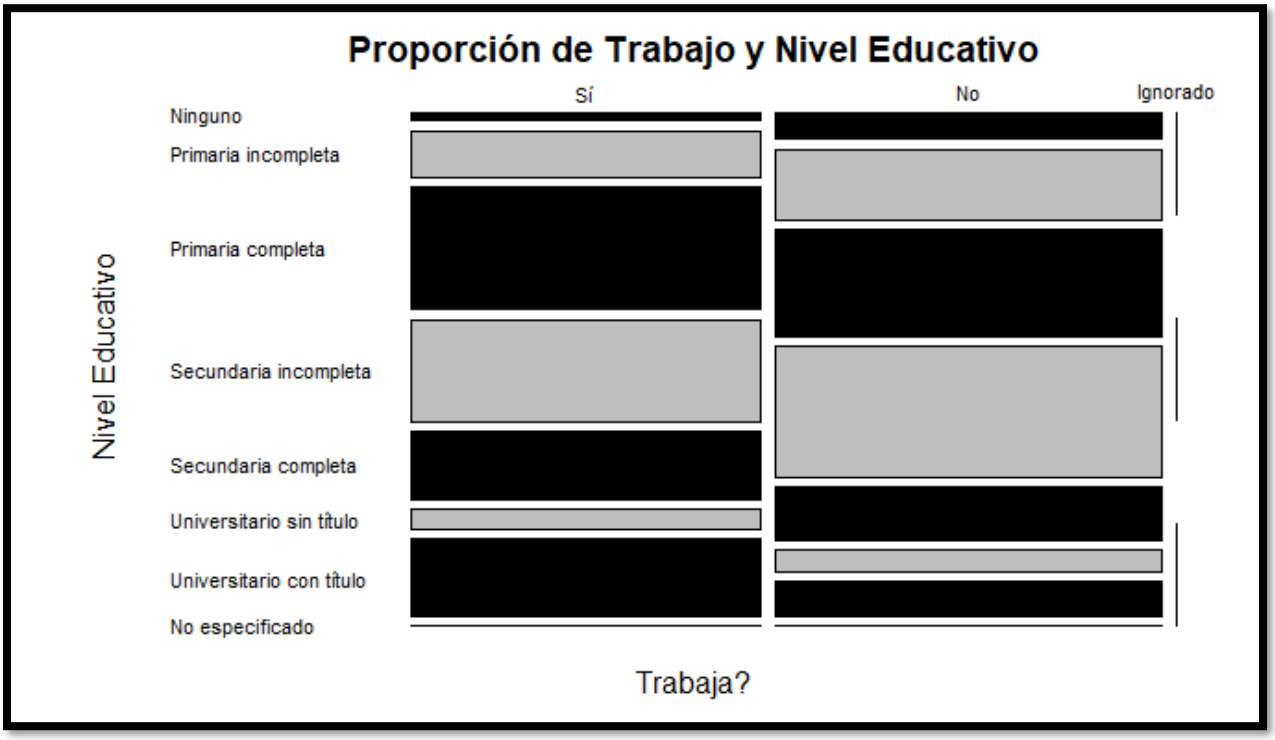


Ilustración 16: Relación entre el nivel educativo y el empleo
Fuente: Elaboración propia

Luego del análisis de la variable de *Nivel_educativo*, surge la duda de cómo se puede comportar la misma en función de la variable por predecir, llamada *Trabajo*; por lo tanto, en la ilustración 16 se observa un gráfico de mosaico, donde se relacionan estas dos variables. En el eje X se encuentran dos columnas, la primera indica si el individuo trabaja y la segunda columna muestra si, por el contrario, el individuo no trabaja. Por otra parte, en el eje Y se despliegan los ocho niveles que conforman la variable de nivel educativo.

Ahora bien, los valores contenidos en la columna de “no trabaja” son mayores que los valores de la columna de “sí trabaja”, a excepción de tres niveles, a saber, primaria completa, secundaria completa y universitario con título. En cada uno de los tres casos anteriores, la cantidad de individuos que sí laboran es mayor a la cantidad de personas desempleadas, lo cual hace pensar que los patronos de este país tienen como prioridad contratar a personas que logran terminar sus ciclos educativos.

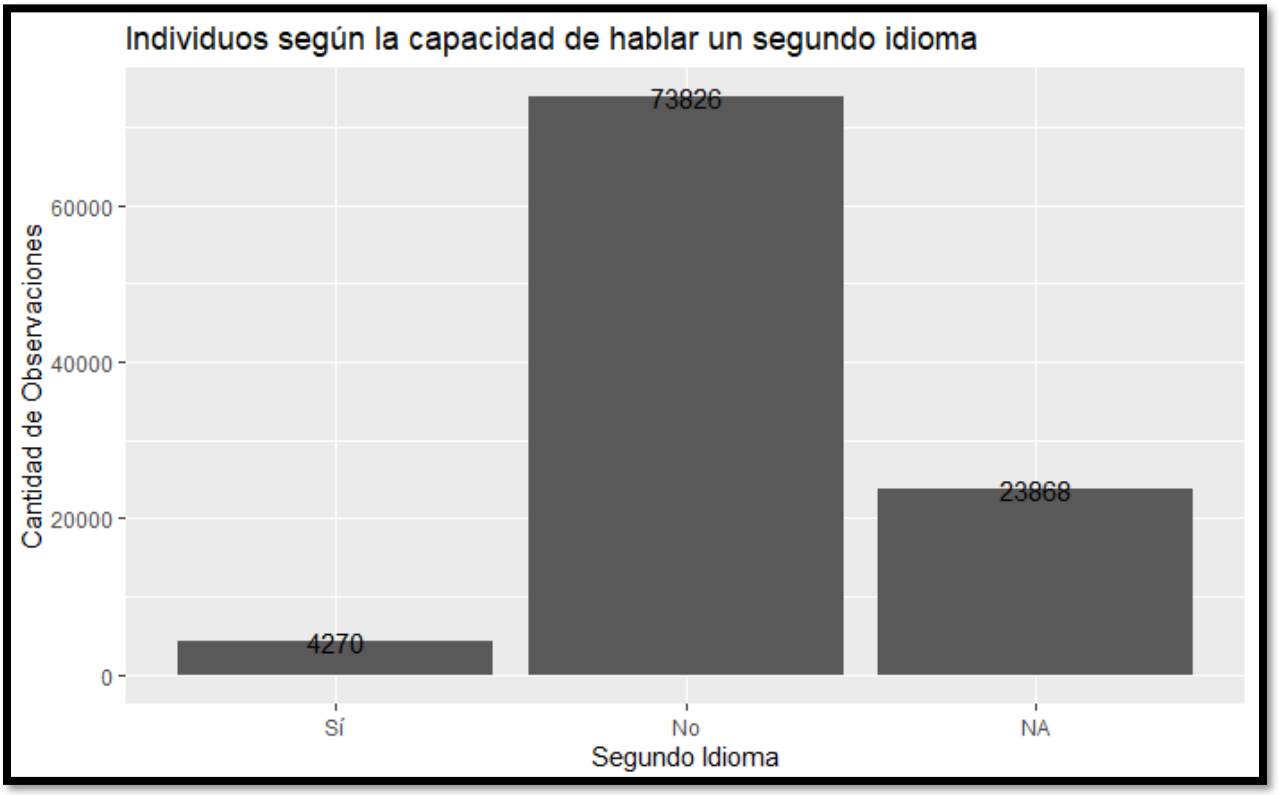


Ilustración 17: Niveles de habla para una lengua no nativa
Fuente: Elaboración propia

La capacidad de hablar otra lengua no nativa es un gran impulso a la hora de conseguir una oportunidad laboral, sin embargo, la ilustración 17 muestra que el 71.42 % de los individuos no domina otro idioma diferente a su lengua materna, lo que limita mucho su posibilidad de salir del desempleo si así fuera el caso.

Solo el 4.19 % de las observaciones sabe una segunda lengua, es un valor sumamente bajo y desesperanzador para el desarrollo personal y social de las personas

que habitan este país; pero es útil conocer este tipo de información, porque el modelo puede aprender mucho de este tipo de variables.

4.2.4. Verificación de la calidad de los datos

La calidad de los datos utilizados en este proyecto de investigación final es muy buena, incluso son datos obtenidos en su mayoría mediante medios digitales, lo que disminuye considerablemente el factor de error humano ante la posible duplicidad de registros o ingreso de valores no permitidos en campos abiertos. Cabe recalcar que los datos no son recopilados ni preparados para el desarrollo de esta investigación, por lo que muchos de los puntos expuestos e investigados en este proyecto se deben ajustar a las variables presentadas en el conjunto de datos.

Cada uno de los sets de información obtenidos de forma trimestral cumple con la Ley de la República n.° 7839 (1998), artículo 4, sección D, donde se vela por la calidad de los datos divulgados:

> Evaluar la calidad de sus estadísticas y las del SEN, promover la investigación, el desarrollo, el perfeccionamiento y la aplicación de la metodología estadística en los entes que generan estadística básica o de síntesis, así como apoyar y brindar asistencia técnica a los servicios estadísticos del Estado y a otros usuarios, mediante convenios de cooperación mutua.

Posterior a la Ley n.° 7839, surge la nueva ley de la República n.° 9694 en el año 2019, derogando muchas de las normas establecidas en la Ley n.° 7839, no obstante, también cuenta con diversas secciones donde se obliga a salvaguardar la calidad de los datos divulgados, como el artículo 33, sección I: “Establecer la política y el marco de calidad que regirá la producción y divulgación de las estadísticas oficiales; promover su adopción en las instituciones del SEN, y evaluar la calidad de las estadísticas del SEN”.

El mecanismo de recolección de datos es mediante la aplicación de una serie de preguntas contenidas en una encuesta, las cuales siguen las recomendaciones hechas por el OIT, logrando así una homogeneidad nacional e internacional con la manera de estructurar cada pregunta.

Como se detalla en el apartado de la “Exploración de los datos”, las variables elegidas cuentan, dependiendo del caso, con cantidades significativas de observaciones sin respuestas, también llamadas *nulas*, esto puede o no ser un problema, dado que la relevancia de las variables sin respuesta está definida por el alcance y el objetivo principal de este proyecto investigativo. No obstante, los modelos de aprendizaje automático necesitan aprender y de valores nulos o sin respuesta no se aprende nada útil, tampoco es posible utilizar a plenitud el conjunto de datos, puesto que es sometido a una limpieza de valores sin respuesta, pero dependiendo del objetivo que se desea solucionar, se adecúa.

4.3. Preparación de los datos

Posterior a la exploración inicial, se preparan los datos con el fin de acomodar y ajustar cada una de las variables para su análisis en el modelado de las técnicas de minería de datos.

La exploración hecha con anterioridad posibilita elegir y seleccionar las variables por utilizar. Seguidamente, y solo si el modelado lo permite, se ensaya la integración de variables y la construcción de otras nuevas variables, como por ejemplo, la nueva clasificación de intervalos de edades. Para concluir, se aplica una limpieza a las observaciones que aumenta la calidad de los datos y optimiza el aprendizaje de los modelos.

4.3.1. Selección de los datos

Los datos empleados corresponden a los obtenidos en la ECE en el año 2019, descargados del sitio web del Programa Acelerado de Datos del INEC, siendo este un sitio de acceso público.

Los registros se encuentran distribuidos en cuatro archivos digitales, esto se debe a que el INEC genera un archivo de datos por cada trimestre en el cual aplica una encuesta. Además, contienen 307 variables agrupadas en diez secciones (247 variables) y otras 60 variables sin agrupación. No todas las variables son elegibles, dentro de las razones que se pueden comentar está, primero, que la calidad de las observaciones sin

respuesta o nulas (NA) es muy alta y hace poco funcional un aprendizaje efectivo, como por ejemplo:

Tabla 16: *Variable con alto número de observaciones sin respuesta*

Variable	Cantidad NA	Porcentaje
EducacionNoregular_codigo	95 625	93.78 %
EducacionNoregular_institucion	95 624	93.78 %
Educacion_codigotitulo	85 677	84.02 %
Joven_nini	98 605	96.70 %

Fuente: Elaboración propia.

Segundo, existen muchos datos que no aportan al aprendizaje del modelo; en cuanto a esto, al intentar predecir la probabilidad de que un individuo pueda cambiar su estado laboral mediante el análisis de la variable *Trabajo*, las variables que no aportan son dependientes de la variable predictora, por ejemplo, las variables salariales o de aguinaldo del individuo presentan un valor monetario si la variable *Trabajo* es afirmativa, en cambio, si el individuo no cuenta con un valor monetario, la variable predictora es negativa o sin respuesta, tal y como se puede observar en la ilustración 18:

Salario_bruto	Trabajo
175000	Sí
120000	Sí
NA	*NA*
NA	*NA*
170000	Sí
170000	Sí
NA	*NA*
NA	No
NA	No
NA	No
NA	*NA*

Ilustración 18: Comparativo entre salario bruto y trabajo
Fuente: Elaboración propia

Los datos seleccionados para el aprendizaje de los modelos son los siguientes:

Tabla 17: *Lista de variables seleccionadas*

Variable	Tipo de dato	Sección
Relacion_parentesco	Factor de 20 niveles	Sección A
Sexo	Factor de 2 niveles	Sección A
Edad	Numérico	Sección A
Estado_conyugal	Factor de 9 niveles	Sección A
Lugar_nacimiento	Factor de 4 niveles	Sección A
Permanencia_pais	Factor de 3 niveles	Sección A
Permanencia_intesion	Factor de 3 niveles	Sección A
Permanencia_motivo	Factor de 4 niveles	Sección A
Seguro	Factor de 3 niveles	Sección A
Tipo_seguro	Factor de 10 niveles	Sección A
Regimen_pension	Factor de 3 niveles	Sección A
Plan_voluntario	Factor de 3 niveles	Sección A
Educacion_asiste	Factor de 8 niveles	Sección A
Educacion_titulo	Factor de 10 niveles	Sección A
EducacionNoregular_asiste	Factor de 3 niveles	Sección A
Idioma	Factor de 3 niveles	Sección A
Idioma_cual	Factor de 6 niveles	Sección A
Tipo_poblacion	Factor de 2 niveles	Sin sección
Pais_nacimiento	Factor de 11 niveles	Sin sección
Provincia_nacimiento	Factor de 8 niveles	Sin sección
Region	Factor de 6 niveles	Sin sección
Zona	Factor de 2 niveles	Sin sección
Nivel_educativo	Factor de 8 niveles	Sin sección

Fuente: Elaboración propia

Las variables nombradas en la tabla 17 cuentan con un potencial importante de aprendizaje, pero eso no es garantía de su uso, dado que al inicio del modelado hay una etapa más de exploración estadística, donde existe la viabilidad de encontrar que no aportan al modelo y se descartan del análisis.

4.3.2. Integración de los datos

Los datos se descargan en cuatro archivos *.sav* y se almacenan en forma digital en la computadora del investigador:

Ilustración 19: Nombre y tipo del archivo digital
Fuente: Elaboración propia

Es necesaria una integración de los cuatro archivos digitales mediante la combinación de los cuatro conjuntos de datos, con el objetivo de facilitar la manipulación de la información de los individuos. La misma se lleva a cabo con ayuda de la herramienta denominada RStudio, con participación de la librería *dplyr* y del comando *bind_rows()*.

```
library(dplyr)
library(foreign)
#Lectura de las 4 encuestas
Trimestre_1 <- read.spss("I_Trimestre_2019.sav", to.data.frame = TRUE, reencode = NA, use.missings = FALSE)
Trimestre_2 <- read.spss("II_Trimestre_2019.sav", to.data.frame = TRUE, reencode = NA, use.missings = FALSE)
Trimestre_3 <- read.spss("III_Trimestre_2019.sav", to.data.frame = TRUE, reencode = NA, use.missings = FALSE)
Trimestre_4 <- read.spss("IV_Trimestre_2019.sav", to.data.frame = TRUE, reencode = NA, use.missings = FALSE)

#Unión de 4 datasets

Encuesta_2019 <- bind_rows(Trimestre_1, Trimestre_2, Trimestre_3, Trimestre_4)
```

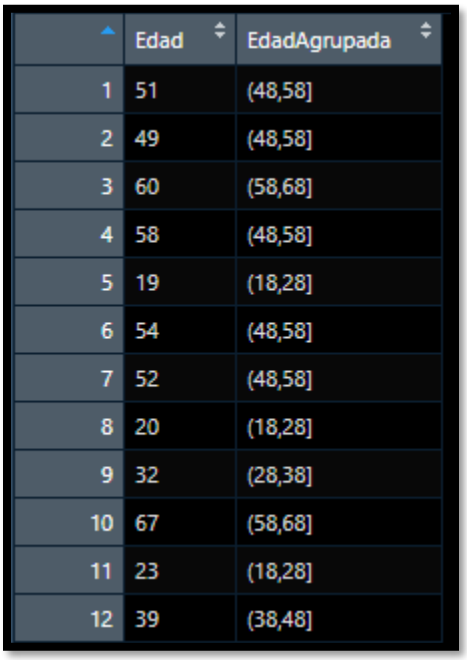
Ilustración 20: Sentencia usada para la integración de los datos
Fuente: Elaboración propia

4.3.3. Construcción de los datos

En este rubro del presente proyecto se explican algunas de las modificaciones que se aplican al conjunto de datos, principalmente se atribuyen a las variables de tipo factor con muchos niveles y a la unión de variables que poseen una relación estrecha.

4.3.3.1. Atributos generados

Se toma la variable edad, la cual es de tipo factor de 100 niveles, y se convierte a tipo numérico, con el fin de poder manipular más fácil los datos del conjunto. Además, dentro del alcance se visualizan las edades de las personas afectadas por el desempleo, con el propósito de proponer acciones de acuerdo con la edad del individuo y no acciones generalizadas que no aplican para todos los individuos. Se determina que la visualización es más efectiva si se agrupan las edades por segmentos de diez años cada una, por lo que se crea una nueva variable llamada *EdadAgrupada*.



	Edad	EdadAgrupada
1	51	(48,58]
2	49	(48,58]
3	60	(58,68]
4	58	(48,58]
5	19	(18,28]
6	54	(48,58]
7	52	(48,58]
8	20	(18,28]
9	32	(28,38]
10	67	(58,68]
11	23	(18,28]
12	39	(38,48]

Ilustración 21: Creación de una nueva variable para la edad
Fuente: Elaboración propia

4.3.3.2. Atributos unificados

Dos variables muy importantes son *Poblacion_joven* y *Poblacion_adulto*, sin embargo, existe una relación tal entre estas que cuando una observación es *Poblacion_joven*, se coloca un valor nulo en la variable *Poblacion_adulto*, y cuando una observación tiene valor en la variable *Poblacion_adulto*, se le asigna un valor nulo a la variable *Poblacion_joven*.

Por esta razón, se crea una variable nueva a partir de la unión de estas dos variables, para de este modo lograr un mayor aprovechamiento de los datos al momento de enseñarle al modelo. En la ilustración 22 se establece cómo están distribuidas las columnas que se pretenden unir:

	Poblacion_joven	Poblacion_adulto
1	*NA*	Población adulta
2	*NA*	Población adulta
3	*NA*	Población adulta
4	*NA*	Población adulta
5	*NA*	Población adulta
6	Población joven	*NA*
7	*NA*	Población adulta
8	*NA*	Población adulta
9	*NA*	Población adulta
10	Población joven	*NA*
11	*NA*	Población adulta
12	*NA*	*NA*
13	*NA*	Población adulta
14	*NA*	Población adulta
15	*NA*	Población adulta
16	*NA*	Población adulta
17	*NA*	Población adulta

Ilustración 22: Visualización de las dos variables por unir
Fuente: Elaboración propia

4.3.3.3. Diccionario de datos

El diccionario de datos nace con el propósito de evitar confusión en los modelos estadísticos, en especial porque existen descripciones muy extensas en la mayoría de los niveles de cada factor y eso perjudica la comprensión de las variables y el desempeño del análisis de la variable por predecir.

Uno de los ejemplos más claros es el caso de la variable llamada *Tipo_Seguro*, donde se encuentran niveles denominados como "¿[...] mediante convenio como asociaciones, sindicatos, cooperativas, etc.?", lo cual provoca que los análisis estadísticos de las variables sean engorrosos y poco exactos.

A continuación, se muestra un par de ejemplos de lo realizado con los niveles de las variables. El diccionario de datos completo se encuentra en el apéndice B de este documento de investigación.

4.3.3.3.1. Variable *Tipo_seguro*

Tabla 18: *Diccionario de la variable* Tipo_seguro

Variable	Valor original	Valor asignado
Tipo_seguro	...asalariado?	1
	...mediante convenio como asociaciones, sindicatos, cooperativas, etc.?	2
	...cuenta propia o voluntario?	3
	...pensionado de la CCSS, Magisterio u otro?	4
	...familiar de asegurado directo o pensionado?	5
	…asegurado por el Estado, incluye familiar de asegurado por el Estado?	6
	…pensionado del régimen no contributivo monto básico, gracia o guerra?	7
	…seguro privado o del extranjero?	8
	…otras formas como seguro de estudiante, de refugiado u otros?	9

Fuente: Elaboración propia

Como se observa en la tabla 18, las descripciones de la variable son muy extensas y el aprendizaje del modelo se ve afectado, por lo que se asigna un valor numérico a cada nivel, comenzado por el número 1 y terminando en el número 9.

Siempre se respeta la cantidad de niveles, solo se cambia la descripción. En la ilustración 23 se aprecia el resultado obtenido luego de la conversión:

```
> summary(Enc2019_SeccionAnalisis$Tipo_seguro)
    1     2     3     4     5     6     7     8     9  NA's
18863   850  6915  6303 12634  3823  3340    93   123 10191
> str(Enc2019_SeccionAnalisis$Tipo_seguro)
 Factor w/ 9 levels "1","2","3","4",..: 8 5 3 1 1 1 1 1 1 4 ...
```

Ilustración 23: Datos transformados de la variable *Tipo_seguro*
Fuente: Elaboración propia

4.3.3.3.2. Variable por predecir *Trabajo*

La variable por predecir es de tipo dicotómica, en otras palabras, puede adoptar solo dos valores, en este caso, un sí o un no, pero es más sencillo si se trabaja únicamente con una variable de tipo booleana, en este caso 1 o 0.

Tabla 19: *Diccionario de la variable* Trabajo

Variable	Valor original	Valor asignado
Trabajo	Sí	1
	No	0

Fuente: Elaboración propia

Como en el caso anterior, los niveles del factor que conforman la variable no son afectados, lo único que se modifica es la descripción, con el objetivo de facilitar su interpretación.

```
> summary(Enc2019_SeccionAnalisis$Trabajo)
    1     0
33656 29479
> str(Enc2019_SeccionAnalisis$Trabajo)
 Factor w/ 2 levels "1","0": 2 2 1 2 1 1 1 1 1 2 ...
```

Ilustración 24: Datos transformados de la variable *Trabajo*
Fuente: Elaboración propia

4.3.4. Limpieza de los datos

La limpieza de los datos es una etapa necesaria en todo proyecto de minería de datos, en especial porque los valores con los cuales se pretende investigar en el actual

proyecto no se obtienen en función a los objetivos planteados en este documento de graduación final.

Asimismo, la limpieza de los datos en este proyecto se centra en tres ejes: la exactitud, la variabilidad y el balanceo de cada una de las variables por modelar.

4.3.4.1. Caracteres especiales

Al momento de generar algunas de las visualizaciones utilizadas en la exploración de los datos, es necesario un reemplazo de caracteres especiales por otros de fácil comprensión para las librerías que se utilizan en RStudio; por ejemplo, para la generación de la ilustración 8 (“Mapa geográfico de la cantidad de individuos por país de nacimiento”), se requiere la ayuda de una librería llamada *maps* y al ser desarrollada en el idioma inglés, ninguno de los nombres de las regiones cuenta con tildes, sin embargo, los niveles de los factores de la variable geográfica utilizada sí tienen tildes, tal es el caso de México, por lo que se aplica un reemplazo del nombre del factor de la variable de la siguiente forma:

```
levels(Encuesta_2019.region$region)[levels(Encuesta_2019.region$region)=="Estados Unidos"] <- "USA"
levels(Encuesta_2019.region$region)[levels(Encuesta_2019.region$region)=="México"] <- "Mexico"
```

Ilustración 25: Limpieza de caracteres especiales
Fuente: Elaboración propia

4.3.4.2. Valores sin respuesta

Con respecto a los valores sin respuesta o valores nulos, dentro de la exploración de los datos se aprecia cómo dos variables de gran importancia para el aprendizaje del modelo, denominadas *Educacion_asiste* y *Nivel_educativo*, presentan un comportamiento muy similar, ambas cuentan con un 23.41 % de observaciones sin respuesta y como se comenta en secciones anteriores, de los valores vacíos los modelos no aprenden, por lo que se eliminan dichos valores y luego se mide el comportamiento de las demás variables.

4.3.4.3. Valores no numéricos

La variable *Edad* originalmente es un factor de 100 niveles, pero no es práctico para el presente estudio analizar la variable como factor, por lo que es preciso convertir dicho factor en un dato numérico.

Como medida de limpieza, se deben quitar tres niveles dentro de la variable *Edad*, la primera llamada “Menor de un año”, la segunda “97 años y más, menor de 15 años con edad ignorada” y “Mayor de 15 años con edad ignorada”, con el fin de no afectar la conversión de tipo factor a tipo numérico.

Continuando con la limpieza de la variable *Edad*, esta contiene registros con edades menores o iguales a los 18 años, por lo que a efectos de esta investigación, las edades menores a los 18 años no aplican dentro del alcance, al no ser es necesario predecir si un menor de edad puede cambiar su estado laboral.

```
47
48  ##### Se filtran las edades que son menores a 18 años.
49
50  Enc2019_SeccionAnalisis <- filter(Enc2019_SeccionAnalisis, Edad >= 18)
51
```

Ilustración 26: Filtro de edad
Fuente: Elaboración propia

Por último, el conjunto de datos luego de la limpieza está comprendido por 63 136 observaciones y 25 variables, con estos datos se comienza con el modelado.

4.4. Modelado

En el actual apartado se identifica la mejor técnica de modelado, iniciando con el uso de un conjunto de datos de entrenamiento, donde se incluye la variable por predecir, llamada *Trabajo* (Sí -> 1, No -> 0), y se aplican algoritmos estadísticos para comprobar si las 25 variables seleccionadas son de utilidad para el aprendizaje del modelo o si, por el contrario, generan ruido y el modelo no logra aprender lo requerido, caso en el cual son retiradas del conjunto de datos de entrenamiento.

4.4.1. Selección de técnicas de modelado

Uno de los pasos más relevantes para la selección del modelo de minería de datos de este proyecto de investigación es el desarrollo del estado de la cuestión, pues en los diversos artículos científicos donde se analizan los datos vinculados con la problemática del desempleo se aprecia cómo los principales algoritmos, o al menos los que más exactitud logran, son las redes neuronales y las máquinas de soporte vectorial. Sin embargo, no se excluyen de este trabajo final otros algoritmos supervisados y no supervisados como lo son K-Vecinos más cercanos, árboles de decisión, bosques aleatorios, red bayesiana o potenciación.

Al final, sin importar cuál modelo de minería de datos se debe seleccionar, se predice una variable dicotómica, donde sus únicos valores son sí o no (1 o 0), apegado a un umbral de discriminación para el cálculo de la probabilidad de 0.5.

4.4.2. Plan de pruebas

El plan de pruebas determina la calidad de un modelo de minería de datos y, como se explica en las clases de la maestría, las pruebas se trabajan con dos conjuntos de datos, uno para el aprendizaje del modelo y el segundo para las pruebas. Estos conjuntos de datos se obtienen gracias a la función *Sample()* de R, la cual permite tomar una muestra de tamaño específico del conjunto de datos original. La proporción utilizada en esta investigación es del 70 % para la colección de datos de aprendizaje y de un 30 % de datos para el set de datos de pruebas. En la ilustración 27 se muestra el procedimiento para la separación de los datos:

```
datos <- Enc2019_SeccionAnalisis[, c(1:2, 4:5, 9:10, 12:13, 15:16, 18:23)]

muestra <- sample(1:nrow(datos),floor(nrow(datos)*0.30))
ttesting <- datos[muestra,]
taprendizaje <- datos[-muestra,]
```

Ilustración 27: Creación del conjunto de pruebas y aprendizaje
Fuente: Elaboración propia

En la ilustración 27 se observa cómo son incluidas las variables seleccionadas como resultado de la exploración de datos y posteriormente, gracias a la función *Sample()*, se almacena el 30 % de las observaciones en un *dataset* llamado *ttesting* y el restante 70 % de las observaciones en el *dataset* denominado *taprendizaje*.

Las vías utilizadas para medir la calidad de los modelos son la matriz de confusión, seguida del cálculo de la sensibilidad, especificidad, exactitud y curva ROC o característica operativa del receptor.

La matriz de confusión es el mecanismo visual encargado de mostrar la información vinculada con los datos reales y los datos previstos arrojados por un algoritmo de clasificación. En la ilustración 28 se aprecia la estructura de una matriz de confusión:

		Predicted	
		Negative	Positive
Actual	Negative	a	b
	Positive	c	d

Ilustración 28: Matriz de confusión
Fuente: Santra y Christy, 2012

La matriz de confusión está formada por una serie de valores; en el caso de la ilustración 29, los valores son a, b, c y d, y su respectiva explicación es la siguiente:

A es el número de predicciones correctas en una instancia negativa.

B es el número de predicciones incorrectas en una instancia positiva.

C es el número de predicciones incorrectas en una instancia negativa.

D es el número de predicciones correctas en una instancia positiva.

Una vez que se conocen los valores de la matriz de confusión, es posible calcular la sensibilidad, la especificidad y la exactitud del modelo mediante la aplicación de unas sencillas fórmulas matemáticas, las cuales se describen a continuación:

$$\text{Exactitud} = \frac{A+D}{Total\ observaciones\ prueba}$$

$$\text{Sensibilidad} = \frac{D}{Total\ observaciones\ positivas}$$

$$\text{Especificidad} = \frac{A}{Total\ observaciones\ negativas}$$

Las métricas más importantes para este proyecto son la exactitud y la sensibilidad, las cuales se relacionan con la probabilidad de conservar el empleo.

En primer lugar, la sensibilidad permite identificar todos los casos que son verdaderos positivos; en otras palabras, se aprueba el porcentaje de individuos que son empleados y que el modelo determina que efectivamente son empleados según el análisis de sus características o variables independientes.

Así mismo, la exactitud también es una valiosa métrica para esta investigación al indicarle al investigador el porcentaje de cada predicción favorable, o sea, revela si la predicción de individuos con empleo es verdadera o si la predicción de individuos desempleados de igual modo es verdadera y sigue engrosando la estadística de desempleo de este país.

Finalmente, la curva ROC es otro mecanismo que posibilita medir el desempeño de los modelos de minería de datos de forma visual mediante la generación de un gráfico en dos dimensiones, donde el eje X es la tasa de predicciones incorrectas en una instancia positiva y el eje Y representa la tasa de predicciones correctas en una instancia positiva. El modelo idóneo es el que presenta el punto (0,1), donde se clasifican correctamente todos los casos positivos y negativos.

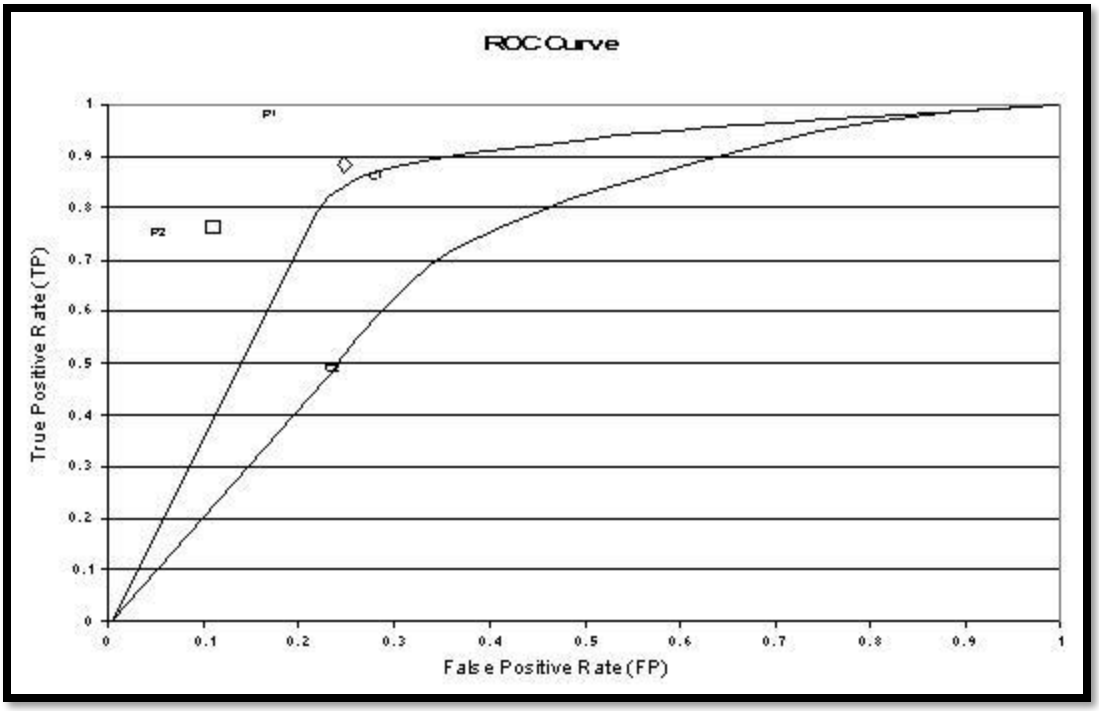


Ilustración 29: Gráfica de la curva ROC
Fuente: Hamilton, 2018

4.4.3. Construcción del modelo

El proyecto final de graduación utiliza los siguientes algoritmos de modelado, la decisión de emplearlos radica en las lecturas del estado de la cuestión, donde ciertas investigaciones encuentran a las redes neuronales y a las máquinas de soporte vectorial como los modelos más exactos, pero además se incluyen otros algoritmos vistos en clase que pueden ayudar con el alcance de los objetivos de este trabajo.

Seguidamente, se expone la lista completa de algoritmos:

Tabla 20: *Lista de modelos de minería de datos*

Modelos
Árboles de decisión
Bosques aleatorios
Máquinas de soporte vectorial
Potenciación
Red bayesiana
Redes neuronales

Fuente: Elaboración propia

4.4.3.1. Ajuste de parámetros

El ajuste de los parámetros para cada uno de los modelos varía según el tipo, no obstante, todos los algoritmos reciben un parámetro en común: el conjunto de datos. Previamente, en el punto de exploración de los datos, se define una serie de variables por utilizar en el modelado, pero no todas las variables son estadísticamente significativas, por lo que se efectúa otro análisis descriptivo de las variables seleccionadas y así se determina si el aporte al modelo es significativo o no.

Este análisis arroja resultados negativos para las tres variables relacionadas con la permanencia: permanencia país, permanencia intención y permanencia motivo. A continuación, se analiza cada una de estas variables.

4.4.3.1.1. *Permanencia_pais*

La variable cuenta con un gran número de observaciones sin respuesta, exactamente 56 882, lo que representa el 90.10 % de la cantidad total de observaciones; en cuanto a esto, un modelo de minería de datos no puede aprender de una variable así.

```
> summary(Enc2019_SeccionAnalisis$Permanencia_pais)
Menos de un año    Un año o más        Ignorado            NA's
            109            6145               0           56882
> tabla <- table(Enc2019_SeccionAnalisis$Permanencia_pais)
> prop.table(tabla)*100

Menos de un año    Un año o más        Ignorado
       1.742885       98.257115       0.000000
>
```

Ilustración 30: Resumen de la variable *Permanencia_pais*
Fuente: Elaboración propia

La ilustración 31 expone de una forma más sencilla la cantidad de observaciones de la variable *Permanencia_pais*:

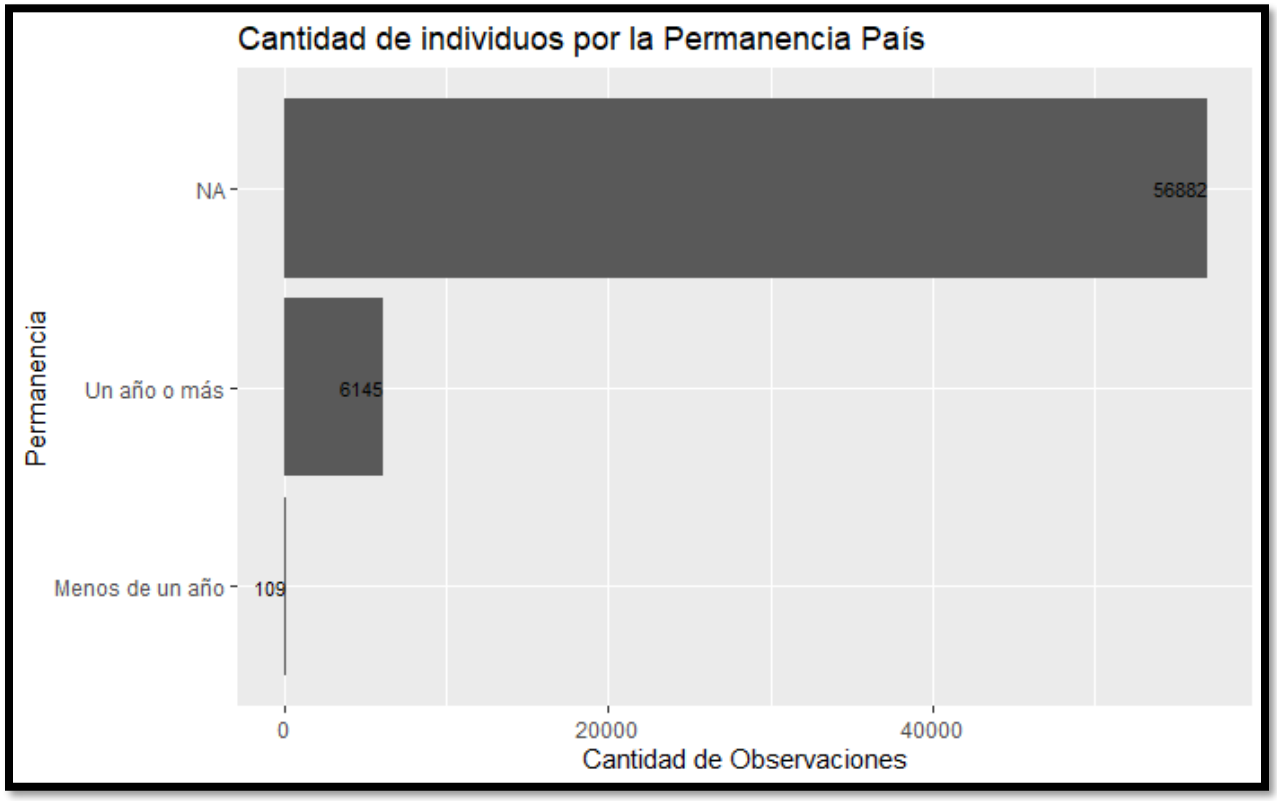


Ilustración 31: Gráfica de la variable *Permanencia_pais*
Fuente: Elaboración propia

4.4.3.1.2. *Permanencia_intención*

Esta variable pertenece al grupo de variables de permanencia, además contiene 63 027 observaciones sin respuesta, lo cual significa que el 99.82 % de sus datos no sirven para que los modelos aprendan.

```
> summary(Enc2019_SeccionAnalisis$Permanencia_intension)
...Costa Rica?  ...otro país?       Ignorado           NA's
           100              9              0          63027
> tabla <- table(Enc2019_SeccionAnalisis$Permanencia_intension)
> prop.table(tabla)*100

...Costa Rica?  ...otro país?       Ignorado
     91.743119       8.256881       0.000000
>
```

Ilustración 32: Resumen de la variable *Permanencia_intencion*
Fuente: Elaboración propia

En el siguiente diseño se establece que solo hay 109 observaciones con respuesta. Dentro de las recomendaciones finales de la investigación, se hace referencia a la necesidad de mejorar la manera de recolectar la información en esta variable.

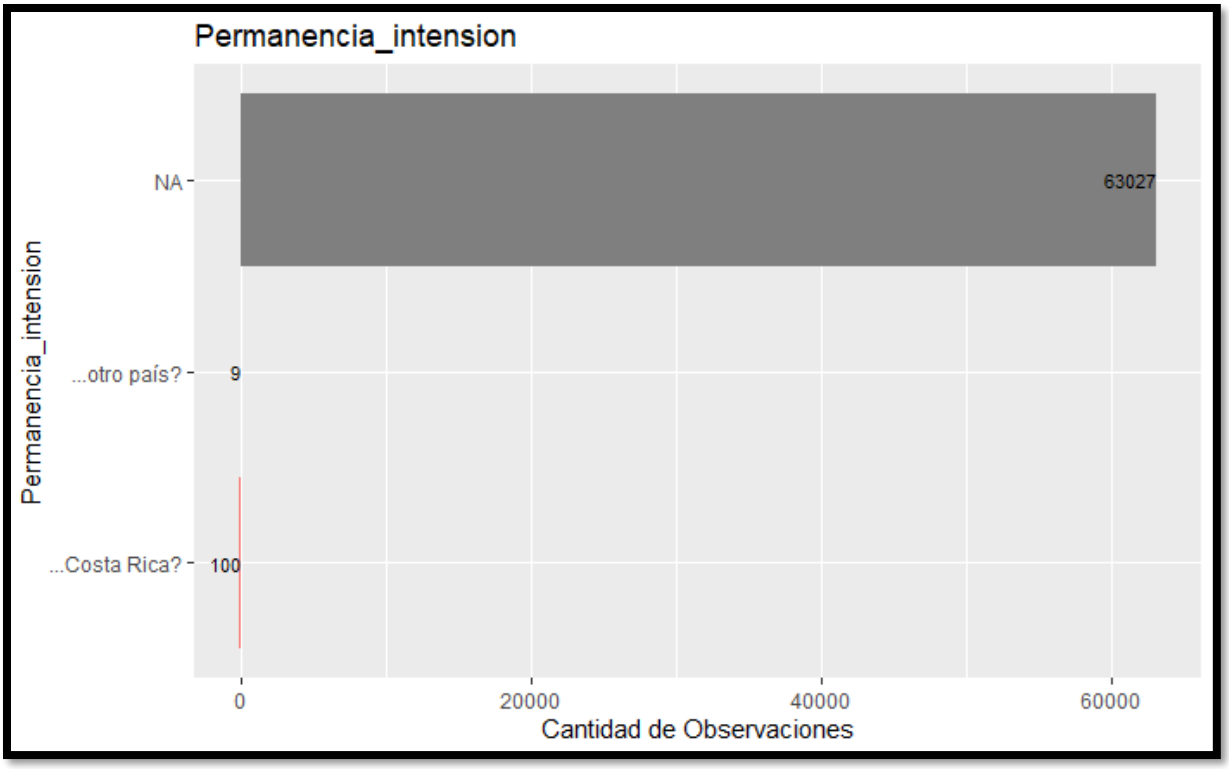


Ilustración 33: Gráfica de la variable *Permanencia_intencion*
Fuente: Elaboración propia

4.4.3.1.3. *Permanencia_motivo*

Es la última de las variables del grupo de permanencia y sigue el patrón de las dos variables anteriores, ya que tiene 63 127 observaciones nulas o sin respuesta y únicamente cuenta con nueve variables con valores que pueden ser usados para el aprendizaje.

```
> summary(Enc2019_SeccionAnalisis$Permanencia_motivo)
 Estudio  Trabajo     Otro Ignorado     NA's
       0        7        2        0    63127
> tabla <- table(Enc2019_SeccionAnalisis$Permanencia_motivo)
> prop.table(tabla)*100

 Estudio  Trabajo     Otro Ignorado
 0.00000 77.77778 22.22222  0.00000
```

Ilustración 34: Resumen de la variable *Permanencia_motivo*
Fuente: Elaboración propia

En relación con la ilustración 35, en definitiva se determina que esta variable no se utilice en ninguno de los modelos.

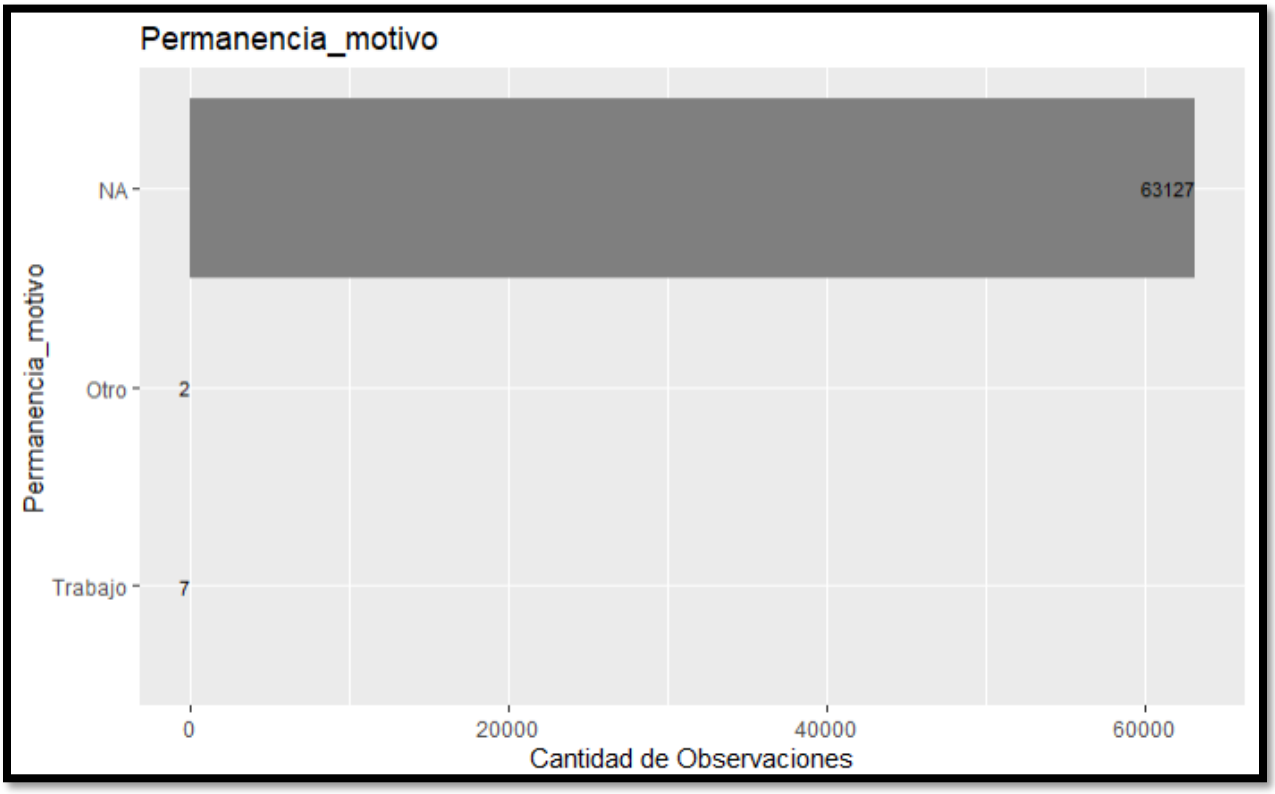


Ilustración 35: Gráfica de la variable *Permanencia_motivo*
Fuente: Elaboración propia

El análisis descriptivo también identifica un par de variables fuera del grupo de permanencia con muy poco valor estadístico, a saber *Regimen_pension* y *Educacion_titulo.* A continuación, se describen los resultados del análisis.

4.4.3.1.4. *Regimen_pension*

Es una variable de tipo factor con únicamente tres niveles, pero con 36 507 observaciones sin respuesta, lo que representa el 57.82 % del total de los registros contenidos en el conjunto de datos. Cuando se tiene una variable con más del 25 % de las observaciones nulas, se puede decir que no es representativa para el aprendizaje del modelo.

```
> summary(Enc2019_SeccionAnalisis$Regimen_pension)
                                          Ninguno                                Régimen de IVM de la CCSS?
                                             1771                                                     23208
Otro régimen (Magisterio, Poder Judicial, Hacienda, etc.)?                                           NA's
                                             1650                                                     36507
```

Ilustración 36: Resumen de la variable *Regimen_pension*
Fuente: Elaboración propia

Luego de visualizar el gráfico, se determina que la variable no se tome en cuenta dentro del análisis.

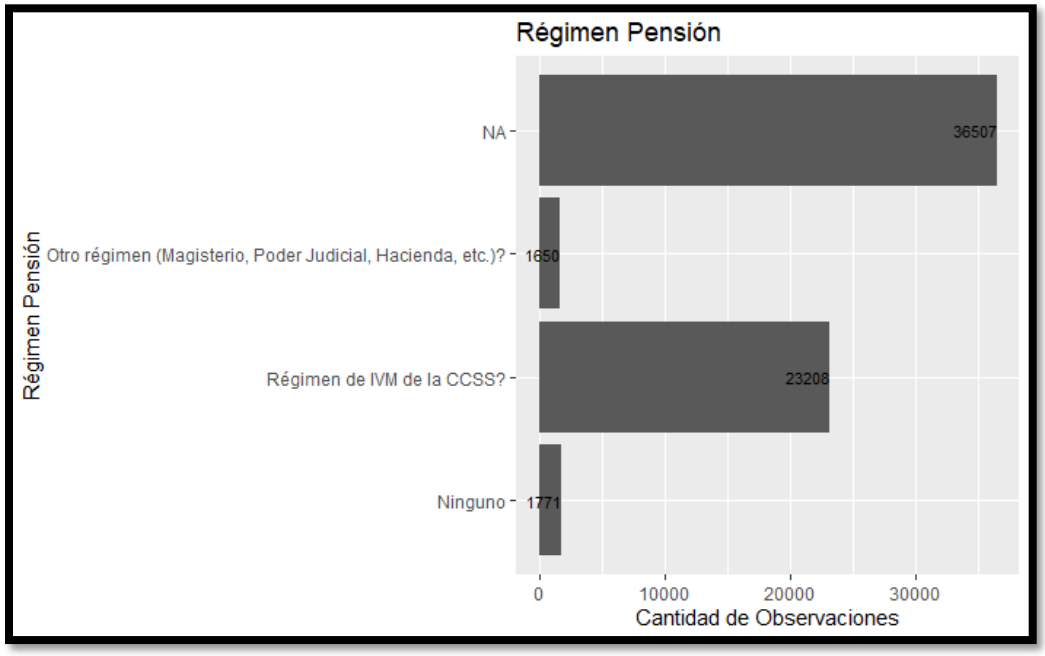


Ilustración 37: Gráfica de la variable *Regimen_pension*
Fuente: Elaboración propia

4.4.3.1.5. *Educacion_titulo*

Para terminar con el análisis descriptivo, se presenta la variable llamada *Educacion_titulo*, la cual contiene 50 569 registros nulos. Si se aplican las reglas que se usan para descartar las variables anteriores, la variable no es de importancia para el aprendizaje; no obstante, a nivel social se considera que sí tiene valor y puede aportar al modelo de minería, por lo que antes de descartarla, se utiliza y se mide su aporte final, de no obtener resultados positivos, se prosigue con la eliminación del conjunto de datos.

```
> summary(Enc2019_SeccionAnalisis$Educacion_titulo)
                          No tiene título                  Técnico, perito no universitario Profesorado, diplomado o técnico universitario
                                     2633                                              724                                              890
                               Bachillerato                                      Licenciatura                                    Especialización
                                     2697                                             4585                                               34
                  Maestría y doctorado                                         Maestria                                        Doctorado
                                        0                                              974                                               30
                       No especificado                                             NA's
                                        0                                            50569
```

Ilustración 38: Resumen de la variable *Educacion_titulo*
Fuente: Elaboración propia

A nivel social, lo que hace suponer que la variable aporta al conocimiento del modelo son sus niveles, porque a una persona que no cuenta con un título en educación le es más difícil encontrar un puesto de trabajo.

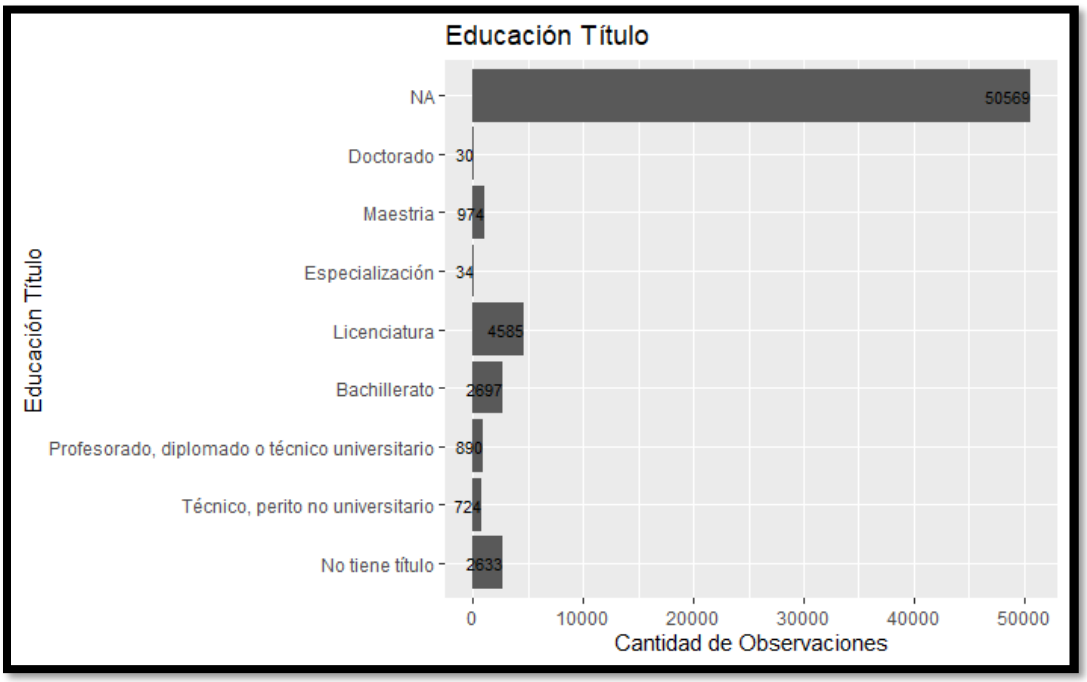


Ilustración 39: Gráfica de la variable *Educacion_titulo*
Fuente: Elaboración propia

La parametrización del algoritmo de K-Vecinos más cercanos comienza con la conversión del conjunto de datos en variables ficticias eficientes y flexibles, mejor conocidas como variables *dummys*, a excepción de la variable de respuesta. La conversión se realiza con la función *dummy.data.frame(datos, sep = ".")*, donde se envía

el conjunto de datos y un punto *(.)*, el cual se utiliza para nombrar las variables ficticias, de forma tal que se encarga de separar el nombre de la variable y el valor asignado.

```
### Conversión de las variables Factor a Dummy, con excepción de la variable respuesta.

datos <- Enc2019_SeccionAnalisis[, c(1:2, 4:5, 9:10, 12:13, 15:16, 19:23)]
datos2 <- dummy.data.frame(datos, sep = ".")
datos2$Trabajo <- Enc2019_SeccionAnalisis[, "Trabajo"]#datos[, "Trabajo"]
```

Ilustración 40: Conversión de observaciones a ficticias
Fuente: Elaboración propia

El modelo es configurado con tres parámetros, el primero espera recibir la fórmula y es donde se indica cuál es la variable que se desea predecir, en este caso la llamada *Trabajo*; el segundo parámetro es el conjunto de datos con las variables ficticias y, por último, el tercer parámetro es la cantidad de vecinos cercanos que se desea que el algoritmo tome en cuenta.

El valor de retorno predictivo, en este caso es una probabilidad.

```
modelo<-train.knn(Trabajo~., data=taprendizaje, kmax=9)
modelo

prediccion <- predict(modelo, ttesting, type = "prob") # Para que me retorne la probabilidad
prediccion
```

Ilustración 41: Parametrización del modelo K-Vecinos más cercanos
Fuente: Elaboración propia

El siguiente modelo por implementar es el llamado árboles de decisión, para esto es necesario definir la fórmula o el marco de datos junto con la variable respuesta como primer parámetro; después, se introduce el conjunto de datos ficticios, pueden ser los mismos utilizados en el modelo de K-Vecinos más cercanos; para terminar, se indica el número mínimo de observaciones que deben existir en un nodo para intentar una división.

```
############################## Arboles de decision

modelo <- train.rpart(formula = Trabajo~.,
                      data = taprendizaje,
                      minsplit = 2)

prediccion <- predict(modelo,
                      ttesting,
                      type = "prob")
```

Ilustración 42: Parametrización del modelo árboles de decisión
Fuente: Elaboración propia

El tercer modelo implementado es el llamado bosques aleatorios. Este algoritmo no se trabaja con variables ficticias, sino con las variables tal y como están en el conjunto de datos. Se divide en dos subconjuntos, uno de prueba y el otro de aprendizaje. Para la parametrización del modelo, se ingresa la fórmula y se señala cuál es la variable de respuesta que se busca predecir. Como segundo parámetro, se define el conjunto de datos, luego la importancia, que como valor por defecto es falso, pero en este caso se establece verdadero (True o T), indicándole al modelo que debe calcular la importancia de la característica para un análisis más detallado. Por último, se señala que se desea omitir las observaciones sin respuesta.

El tipo de predicción que se busca es de tipo probabilidad.

```
############################################## Bosques aleatorios

datos <- Enc2019_SeccionAnalisis[, c(1:2, 4:5, 9:10, 12:13, 15:16, 18:23)]

muestra <- sample(1:nrow(datos),floor(nrow(datos)*0.30))
ttesting <- datos[muestra,]
taprendizaje <- datos[-muestra,]

RF.modelo <- train.randomForest(formula = Trabajo~., data = taprendizaje, importance = T, na.action=na.omit)
prediccion <- predict(RF.modelo, ttesting, type = "prob", na.action = na.omit) # Para que me retorne la probabilidad
```

Ilustración 43: Parametrización del modelo bosques aleatorios
Fuente: Elaboración propia

Continuando con la parametrización de los modelos, se encuentra la máquina de soporte vectorial, en la cual el primer parámetro configurado es la fórmula, donde se indica el marco de datos y la variable por predecir. Por su parte, el segundo parámetro es el conjunto de datos de aprendizaje.

Para la predicción es necesario ingresar el modelo generado, el conjunto de datos de prueba y el tipo de predicción que se busca, el cual es de tipo probabilidad.

```
################################################## maquina de soporte vectorial

modelo <- train.svm(Trabajo ~ .,data = taprendizaje)
prediccion <- predict(modelo, ttesting, type = "prob") # Para que me retorne la probabilidad
```

Ilustración 44: Parametrización del modelo máquina de soporte vectorial
Fuente: Elaboración propia

El impulso adaptativo es un algoritmo de aprendizaje automático que pretende mejorar el rendimiento. En este caso la parametrización es bastante sencilla, primero se establece cuál es la fórmula sobre la que se desea predecir y, como segundo argumento, se indica el conjunto de datos de aprendizaje.

Una vez más, se ajusta la predicción con el modelo generado con anterioridad, el conjunto de datos de prueba y el tipo de predicción; en este caso se busca que retorne la probabilidad.

```
################################################## #Impulso Adaptativo

modelo <- train.ada(formula = Trabajo~., data = taprendizaje)
prediccion <- predict(modelo, ttesting, type = "prob") # Para que me retorne la probabilidad
```

Ilustración 45: Parametrización del modelo impulso adaptativo
Fuente: Elaboración propia

El sexto modelo en parametrizar es un modelo de redes bayesianas llamado clasificador de Bayes, y a pesar de su simpleza, puede lograr niveles muy altos de precisión.

El primer parámetro que se utiliza es la fórmula, la cual —como en los casos anteriores— se define en función de la variable por predecir, y el segundo argumento corresponde a enviar el conjunto de datos de aprendizaje.

La predicción se realiza según el modelo, el conjunto de datos de prueba y un tipo de predicción probabilística, como en todas las predicciones anteriores.

```
########################################### bayes

modelo <- train.bayes(Trabajo~.,data=taprendizaje)
prediccion <- predict(modelo, ttesting, type = "prob") # Para que me retorne la probabilidad
```

Ilustración 46: Parametrización del modelo redes bayesianas
Fuente: Elaboración propia

Para finalizar, se configura el modelo de aprendizaje automático denominado redes neuronales, uno de los más utilizados en los estudios analizados en el estado de la cuestión. La red neuronal recibe los mismos argumentos que los modelos anteriores, siendo la fórmula el primero de estos y luego el conjunto de datos de aprendizaje, pero se debe sumar un argumento más, el *hidden* o tamaño, el cual permite indicar los niveles de la capa intermedia de la red.

```
#############################################################################
library(neuralnet)

datos.aprendizaje.red <- model.matrix(~.,
                                      data = taprendizaje)

datos.test.red <- model.matrix(~.,
                               data = ttesting)

colnames(datos.aprendizaje.red) <- make.names(colnames(datos.aprendizaje.red))
colnames(datos.test.red) <- make.names(colnames(datos.test.red))

#Construir MOdelo

modelo.nnet2 <- neuralnet(Trabajo0~.,
                          data = datos.aprendizaje.red, hidden = 2)

summary(modelo.nnet2)

# ver detalles del modelo
print(modelo.nnet2)

neuralnet::gwplot(modelo.nnet2)

plot(modelo.nnet2)
```

Ilustración 47: Parametrización del modelo redes neuronales
Fuente: Elaboración propia

4.4.3.2. Resultado del modelado

Los resultados expuestos en la siguiente sección se obtienen con ayuda de la herramienta de programación RStudio, con una división de datos del 30 % (18 940 observaciones) para las pruebas y el 70 % (44 195 observaciones) para el aprendizaje del modelo.

4.4.3.2.1. Árboles de decisión

Según los parámetros introducidos en la función de modelado de árboles de decisión *train.rpart()*, estos son los resultados arrojados:

```
> summary(modelo.rpart)
Call:
rpart::rpart(formula = Trabajo ~ ., data = taprendizaje, minsplit = 2)
  n= 44195

          CP nsplit rel error    xerror        xstd
1 0.37775829      0 1.0000000 1.0000000 0.005107307
2 0.07813337      1 0.6222417 0.6222417 0.004642149
3 0.07759754      2 0.5441083 0.5814214 0.004546744
4 0.03838472      3 0.4665108 0.4665108 0.004219018
5 0.01000000      4 0.4281261 0.4281261 0.004087463

Variable importance
       Tipo_seguro.1              Sexo.0              Sexo.1        Tipo_seguro.4        Tipo_seguro.7    Nivel_educativo.7    Plan_voluntario.0
                  43                  12                  12                   7                    6                    5                    3
   Plan_voluntario.1 Relacion_parentesco.1 Relacion_parentesco.2        Tipo_seguro.5        Tipo_seguro.3            Idioma.0            Idioma.1
                   3                   2                   2                   1                    1                    1                    1
```

Ilustración 48: Resumen del modelo árboles de decisión
Fuente: Elaboración propia

La ilustración 48 muestra las variables independientes con mayor importancia para el modelo: *Tipo_seguro.Asalariado*, *Sexo.Mujer*, *Sexo.Hombre*, *Tipo_seguro.PensionadoCCSS*, *Tipo_seguro.PensionadoRegimenNoContributivo* y *Nivel_educativo.UniversitarioConTitulo*. (Ver el apéndice B, “Diccionario de datos”, para comprender el significado de las variables más significativas).

De la ilustración 49 se deduce que las observaciones personas asalariadas (*Tipo_seguro.1 = Asalariado?*), de sexo es masculino (*Sexo.1 = Masculino*) y, por último, aseguradas (*Tipo_seguro.4 = PensionadoCCSS*) corresponden a quienes tienen una mayor probabilidad de conservar su trabajo o, en otras palabras, en quienes es menos probable que cambie su estado laboral.

```
> modelo.rpart
n= 44195

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 44195 20529 1 (0.53549044 0.46450956)
   2) Tipo_seguro.1>=0.5 13244  1176 1 (0.91120507 0.08879493) *
   3) Tipo_seguro.1< 0.5 30951 11598 0 (0.37472133 0.62527867)
     6) Sexo.1>=0.5 12506  5451 1 (0.56412922 0.43587078)
      12) Tipo_seguro.4< 0.5 10043  3423 1 (0.65916559 0.34083441)
        24) Tipo_seguro.7< 0.5 9031  2523 1 (0.72062894 0.27937106) *
        25) Tipo_seguro.7>=0.5 1012   112 0 (0.11067194 0.88932806) *
      13) Tipo_seguro.4>=0.5 2463   435 0 (0.17661389 0.82338611) *
     7) Sexo.1< 0.5 18445  4543 0 (0.24629981 0.75370019) *
```

Ilustración 49: Ramas del modelo árboles de decisión
Fuente: Elaboración propia

A continuación, se indican las ramas del árbol de forma gráfica para entender más fácil la predicción resultante:

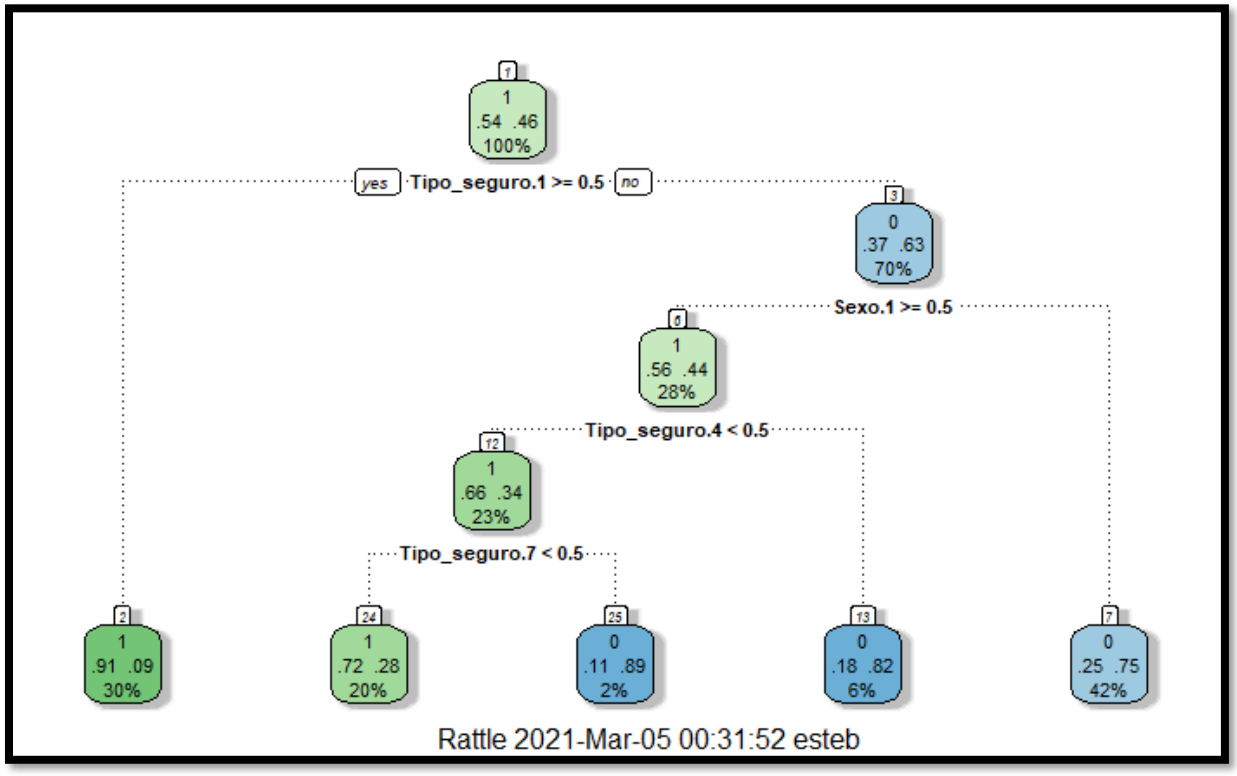
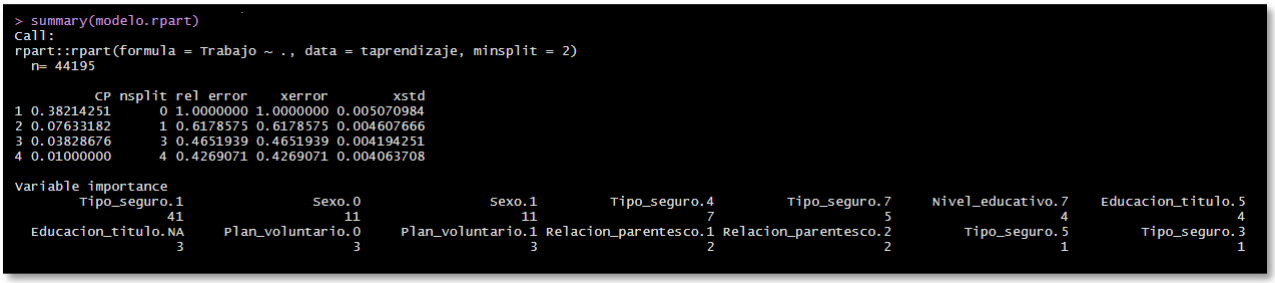


Ilustración 50: Gráfico del modelo árboles de decisión
Fuente: Elaboración propia

Sin embargo, cabe recordar que no se está incluyendo dentro del modelo la variable *Educacion_titulo*, la cual contiene una gran cantidad de observaciones nulas o sin respuesta, pero se considera que tiene mucho potencial de aprendizaje, por lo que con el propósito de comparar resultados, se ejecuta el análisis, pero con la variable *Educacion_titulo*.



```
> summary(modelo.rpart)
Call:
rpart::rpart(formula = Trabajo ~ ., data = taprendizaje, minsplit = 2)
  n= 44195

          CP nsplit rel error    xerror        xstd
1 0.38214251      0 1.0000000 1.0000000 0.005070984
2 0.07633182      1 0.6178575 0.6178575 0.004607666
3 0.03828676      3 0.4651939 0.4651939 0.004194251
4 0.01000000      4 0.4269071 0.4269071 0.004063708

Variable importance
       Tipo_seguro.1              Sexo.0              Sexo.1        Tipo_seguro.4        Tipo_seguro.7    Nivel_educativo.7   Educacion_titulo.5
                  41                  11                  11                   7                    5                    4                    4
  Educacion_titulo.NA    Plan_voluntario.0    Plan_voluntario.1 Relacion_parentesco.1 Relacion_parentesco.2        Tipo_seguro.5        Tipo_seguro.3
                   3                   3                   3                   2                    2                    1                    1
```

Ilustración 51: Resumen del modelo árboles de decisión más *Educacion_titulo*
Fuente: Elaboración propia

En la ilustración 51 se observa cómo el modelo incluye dentro de sus variables independientes más importantes los valores de la variable *Educacion_titulo.5 (Educacion_titulo.5 = Licenciatura)* en la posición 7 y conserva las mismas seis variables del modelo que no tiene la variable *Educacion_titulo*. Por lo anterior el investigador cree que efectivamente la variable es significativa para el aprendizaje del modelo, aunque tenga una gran cantidad de observaciones sin respuesta. En la ilustración 52 se visualiza el gráfico del modelo:

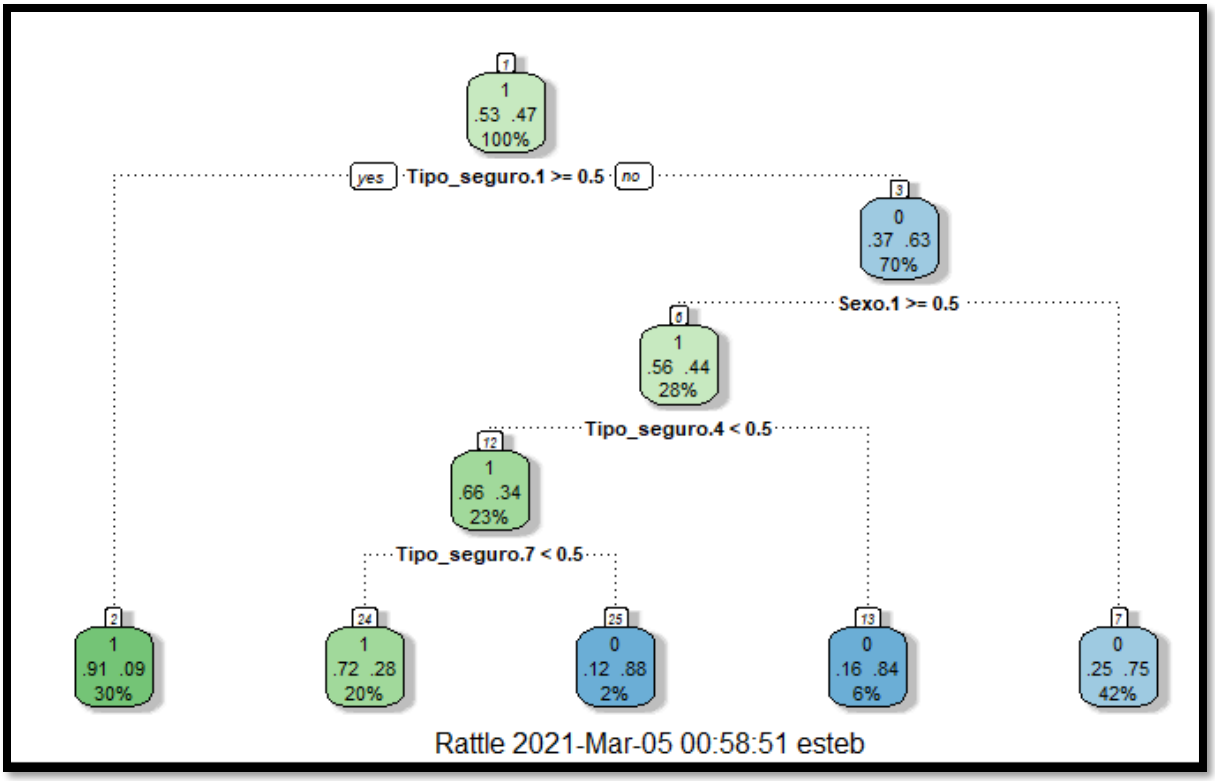


Ilustración 52: Gráfico del modelo árboles de decisión más *Educacion_titulo*
Fuente: Elaboración propia

4.4.3.2.2. Bosques aleatorios

Los resultados obtenidos del modelo de bosques aleatorios con ayuda de la función *train.randomForest()* y sin la inclusión de la variable de *Educacion_titulo,* son los siguientes:

```
datos <- Enc2019_SeccionAnalisis[, c(1:2, 4:5, 9:10, 12:13, 15:16, 18:23)]# Sin incluir la variable de Educación Titulo

#datos <- Enc2019_SeccionAnalisis[, c(1:2, 4:5, 9:10, 12:16, 18:23)]# Incluyendo la variable Educación Titulo

muestra <- sample(1:nrow(datos),floor(nrow(datos)*0.30))
ttesting <- datos[muestra,]
taprendizaje <- datos[-muestra,]
```

Ilustración 53: Selección de las variables del modelo bosques aleatorios
Fuente: Elaboración propia

El resumen del modelo muestra cómo se generan 500 árboles para todas las variables menos *Educacion_titulo*, obteniendo un error estimado de predicción del 15.58 %.

```
> modelo.RF

Call:
 randomForest(formula = Trabajo ~ ., data = taprendizaje, importance = T,      na.action = na.omit)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 15.58%
Confusion matrix:
      1     0 class.error
1 14839  3113   0.1734069
0  2214 14028   0.1363133
```

Ilustración 54: Resumen del modelo bosques aleatorios
Fuente: Elaboración propia

Continuando con el ejercicio, se genera otro modelo más, pero esta vez con la variable *Educacion_titulo*:

```
#datos <- Enc2019_SeccionAnalisis[, c(1:2, 4:5, 9:10, 12:13, 15:16, 18:23)]# Sin incluir la variable de Educación Titulo

datos <- Enc2019_SeccionAnalisis[, c(1:2, 4:5, 9:10, 12:16, 18:23)]# Incluyendo la variable Educación Titulo

muestra <- sample(1:nrow(datos),floor(nrow(datos)*0.30))
ttesting <- datos[muestra,]
taprendizaje <- datos[-muestra,]
```

Ilustración 55: Selección de variables del modelo bosques aleatorios más *Educacion_titulo*
Fuente: Elaboración propia

En este caso, el resumen del modelo despliega el uso de también 500 árboles, pero se mejora en la tasa del error estimado de predicción, bajando a un 14.02 %.

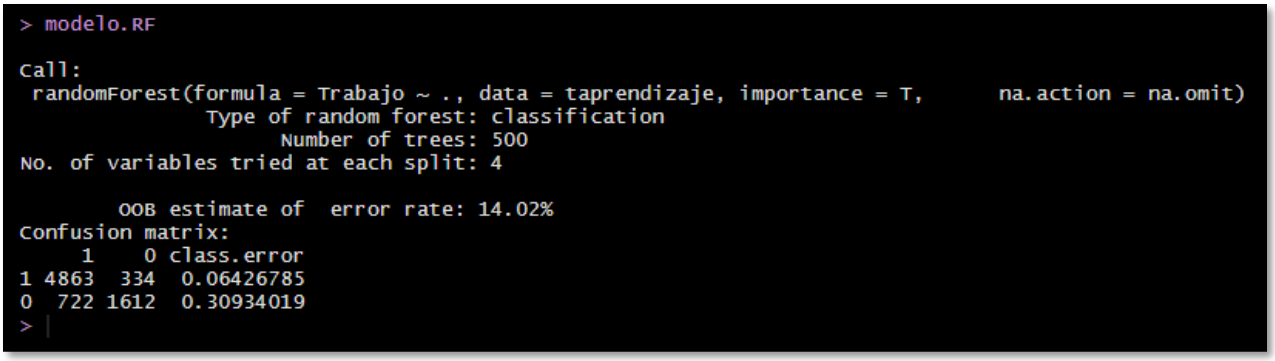


Ilustración 56: Resumen del modelo bosques aleatorios más *Educacion_titulo*
Fuente: Elaboración propia

Asimismo, se generan los gráficos de importancia de variables para el modelo, uno sin contemplar la variable *Educacion_titulo* y el segundo con la variable incluida dentro del modelo. A continuación, se exponen los resultados.

En la columna de la derecha es posible observar la relevancia de la variable con respecto a su contribución a la precisión y al grado de clasificación errónea.

Para la clasificación realizada, sin la variable *Educacion_titulo,* se obtiene que las variables en orden de importancia son: *Tipo_seguro, Nivel_educativo, Sexo, Estado_conyugal, Relacion_parentesco, Provincia_nacimiento*, entre otras.

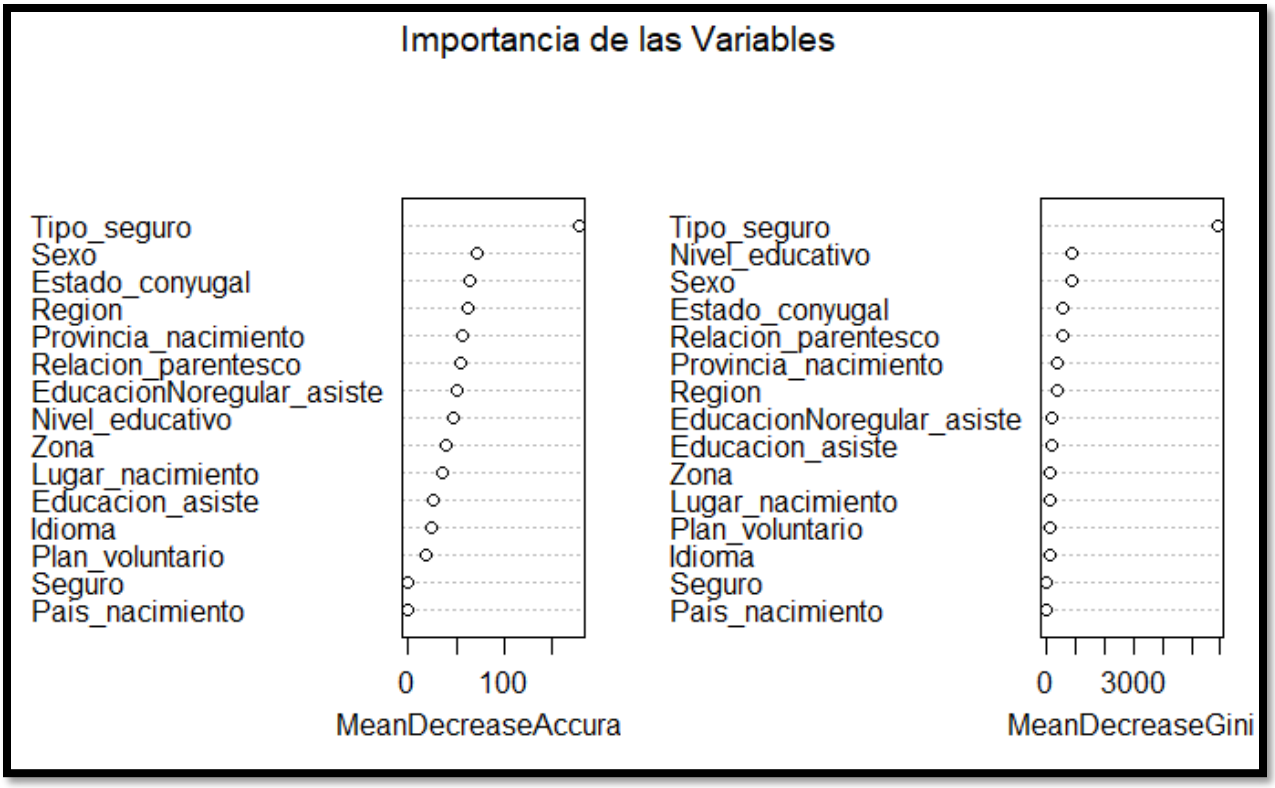


Ilustración 57: Gráfica de importancia de variables
Fuente: Elaboración propia

En cuanto al segundo modelo de bosques aleatorios, la *Educacion_titulo* se ubica en el cuarto lugar de importancia de variables para el modelo, desplazando a la variable *Sexo* a un séptimo lugar, por lo que la variable *Educacion_titulo*, a pesar de su gran cantidad de observaciones sin respuesta, aporta mucho al aprendizaje del modelo.

En la ilustración 58, la columna de la derecha muestra la lista de las variables con mayor relevancia:

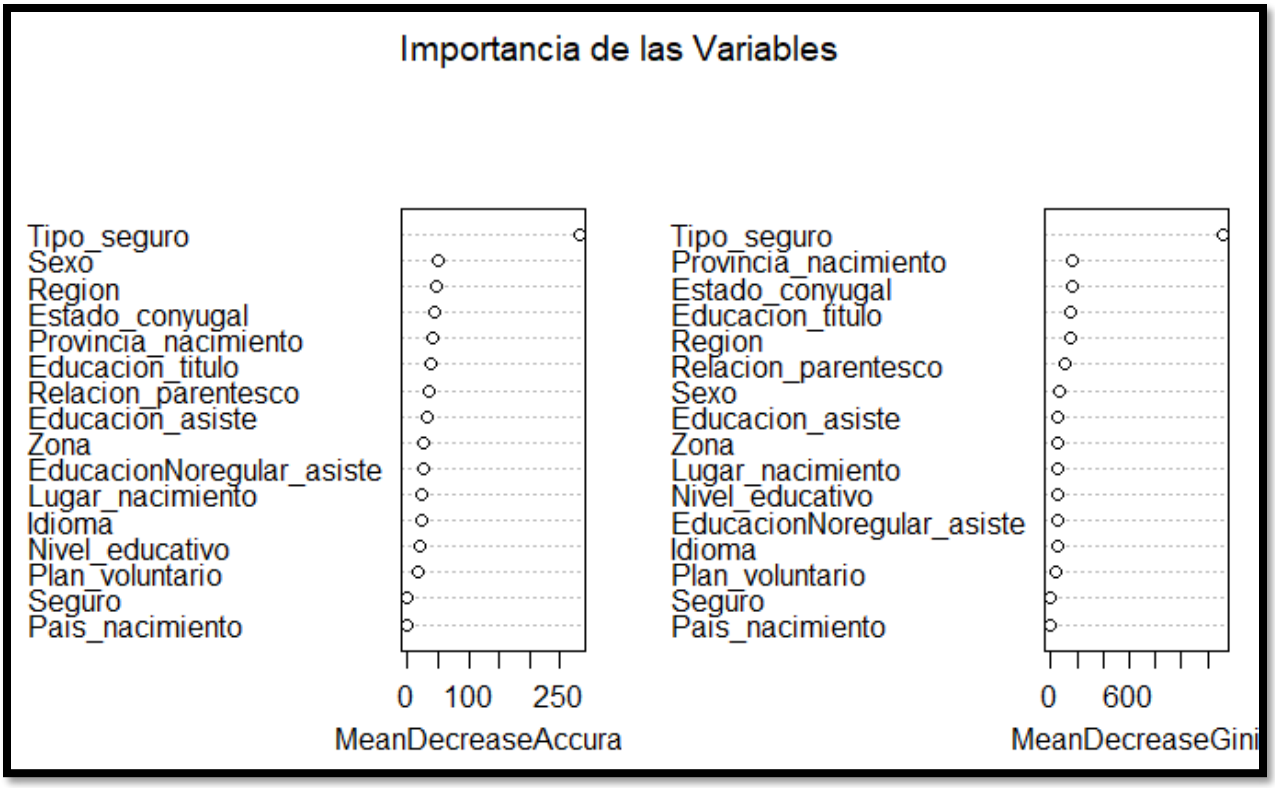


Ilustración 58: Gráfica de importancia de variables más *Educacion_titulo*
Fuente: Elaboración propia

4.4.3.2.3. Máquina de soporte vectorial

Los resultados del modelo de máquinas de soporte vectorial son los siguientes:

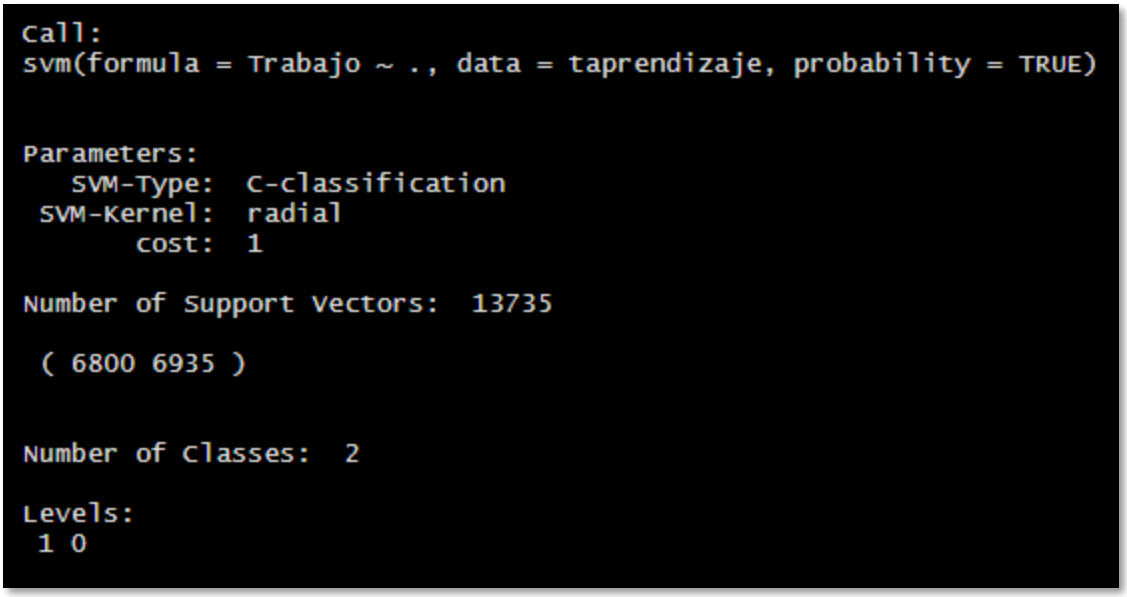


Ilustración 59: Resumen del modelo máquina de soporte vectorial
Fuente: Elaboración propia

Como se aprecia en el resumen, el costo de la generación del modelo es de 1, para un total de 13 735 vectores de soporte generados.

4.4.3.2.4. Impulso adaptativo

En la ilustración 60 se exponen los resultados del modelo ADA para el impulso adaptativo creado con el conjunto de datos de aprendizaje:

```
> summary(modelo.ada)
Call:
ada(Trabajo ~ ., data = taprendizaje)

Loss: exponential Method: discrete   Iteration: 50

Training Results

Accuracy: 0.812 Kappa: 0.623

> print(modelo.ada)
Call:
ada(Trabajo ~ ., data = taprendizaje)

Loss: exponential Method: discrete   Iteration: 50

Final Confusion Matrix for Data:
          Final Prediction
True value     0     1
         0 16769  3828
         1  4480 19118

Train Error: 0.188

Out-Of-Bag Error:  0.19  iteration= 50

Additional Estimates of number of iterations:

train.err1 train.kap1
        49         49
```

Ilustración 60: Resumen del modelo impulso adaptativo
Fuente: Elaboración propia

Los resultados anteriores se obtienen sin la variable *Educacion_titulo* y se logra una precisión del 81.2 % y una medida de concordancia entre la clasificación observada y la clasificación predicha (Kappa) de un 62.3 %, lo cual sugiere que puede ponérsele más iteraciones para que mejore la precisión. El valor del error de entrenamiento es bastante bajo, rondando el 19 %.

En la ilustración 61 se describen las variables más importantes para el modelo, a saber, *Pais_nacimiento, Plan_voluntario, Relacion_parentesco, Educacion_asiste, Idioma*, entre otros. Sin embargo, llama mucho la atención cómo la variable *Tipo_seguro*, que es tan relevante para los modelos anteriores, se encuentra en los últimos lugares.

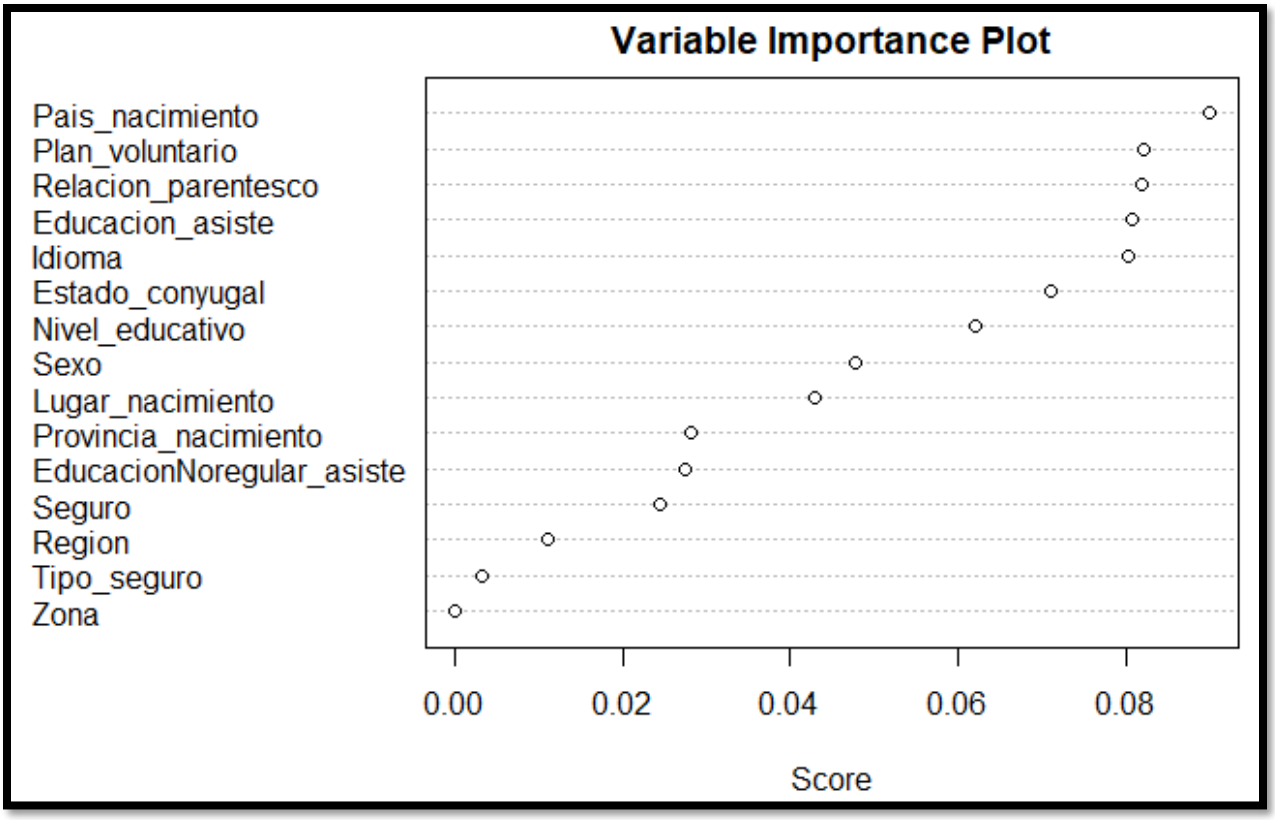


Ilustración 61: Gráfico de importancia de las variables del modelo ADA
Fuente: Elaboración propia

Seguidamente, se crea de nuevo el modelo tomando en cuenta la variable *Educacion_titulo*, dado que el investigador de este proyecto de graduación cree que a pesar de la gran cantidad de observaciones sin respuesta, es una variable fundamental.

```
> summary(modelo.ada)
Call:
ada(Trabajo ~ ., data = taprendizaje)

Loss: exponential Method: discrete   Iteration: 50

Training Results

Accuracy: 0.813 Kappa: 0.624

> print(modelo.ada)
Call:
ada(Trabajo ~ ., data = taprendizaje)

Loss: exponential Method: discrete   Iteration: 50

Final Confusion Matrix for Data:
          Final Prediction
True value     0     1
         0 16828  3778
         1  4507 19082

Train Error: 0.187

Out-Of-Bag Error:  0.19  iteration= 50

Additional Estimates of number of iterations:

train.err1 train.kap1
        48         48
```

Ilustración 62: Resumen del modelo ADA con la variable *Educacion_titulo*
Fuente: Elaboración propia

Los valores de precisión y Kappa cambian de forma mínima, además el error de entrenamiento tiene una reducción muy leve, por lo cual se deduce que la variable no aporta al aprendizaje del modelo y es preferible no utilizarla.

Por otra parte, en el gráfico de importancia de las variables se observa que la variable *Educacion_titulo* queda al final de lista, respaldando la decisión de no incluirla en el aprendizaje de este modelo.

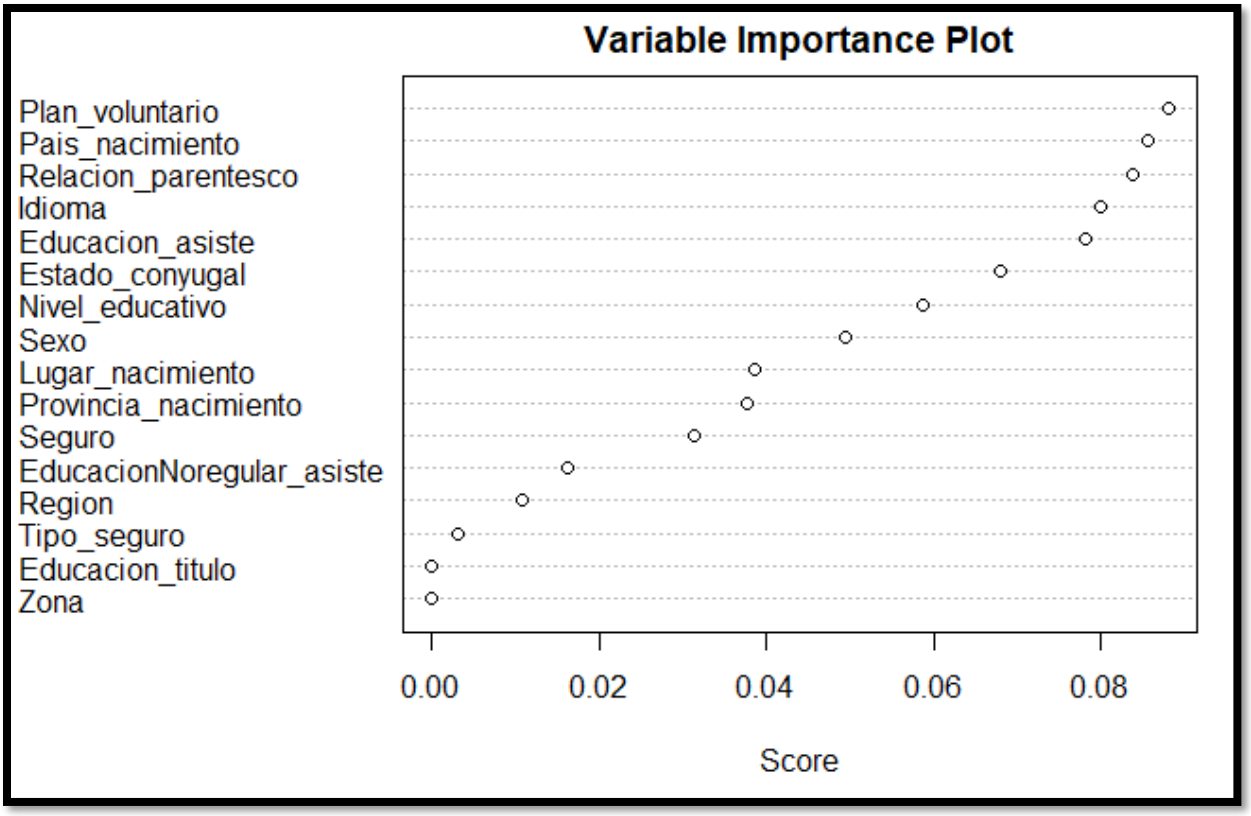


Ilustración 63: Gráfico de importancia de variables para el modelo con la variable *Educacion_titulo*
Fuente: Elaboración propia

4.4.3.2.5. Redes bayesianas

Se parte de un primer modelo de redes bayesianas elaborado a partir de las probabilidades *a priori* y se calcula con base en eso las *a posteriori*. En el siguiente resumen se indica la probabilidad de la variable *Trabajo a priori*, donde el 53 % es para el valor 1 o trabaja y el 47 % para un valor 0 o desempleado.

No se muestran las probabilidades *a posteriori* porque son muchas según la cantidad de variables que se usan en la creación del modelo.

```
> print(modelo.bayes)

Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
        1         0
0.5320964 0.4679036
```

Ilustración 64: Resumen del modelo de redes bayesianas
Fuente: Elaboración propia

4.4.3.2.6. Redes neuronales

El proceso de aprendizaje del modelo de redes neuronales es muy completo. Al respecto, el tiempo de ejecución que demora el modelo en entrenarse depende de la cantidad de niveles de capa intermedia.

Para esta investigación, se crean dos modelos, ambos con un solo nivel, como se aprecia en la ilustración 65:

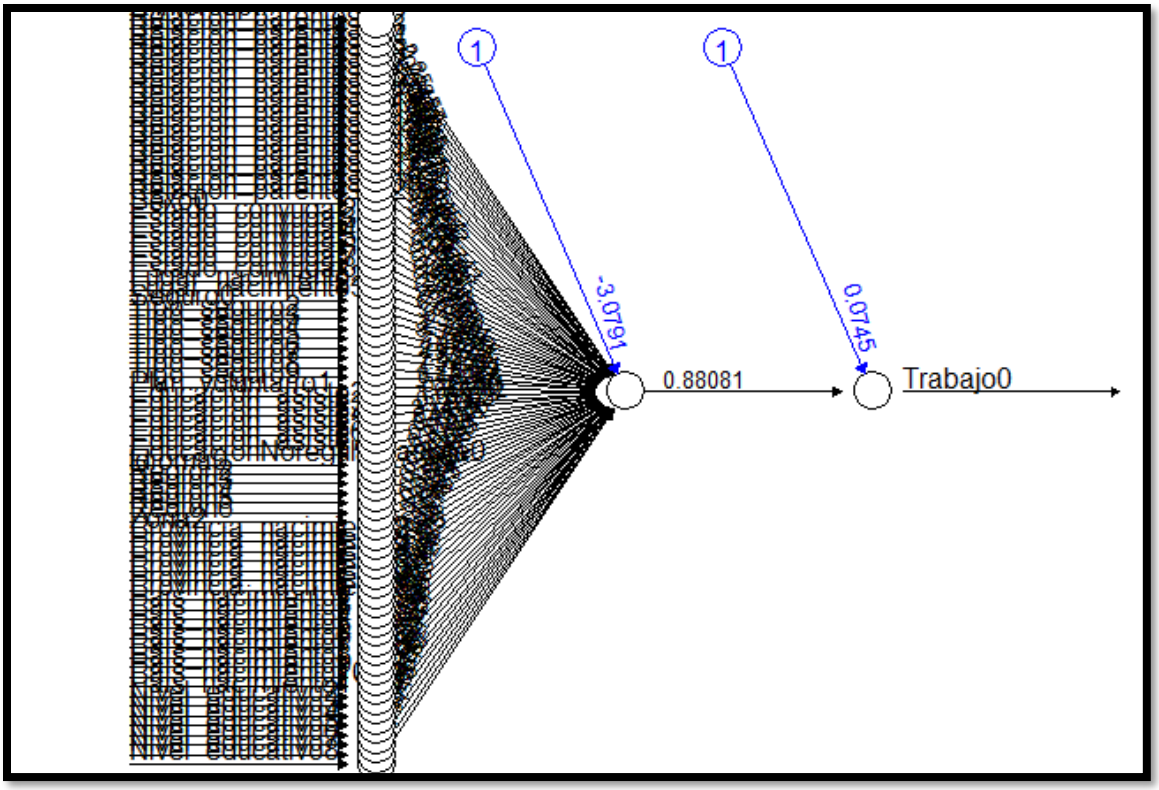


Ilustración 65: Gráfico del modelo de redes neuronales de solo un nivel
Fuente: Elaboración propia

Sin embargo, parte importante del modelo de esta investigación gira en torno a la variable *Educacion_titulo* y su relevancia en el aprendizaje del modelo, por lo que se incluye en el modelo de red neuronal y se analiza su comportamiento y exactitud en el apartado de “Evaluación”.

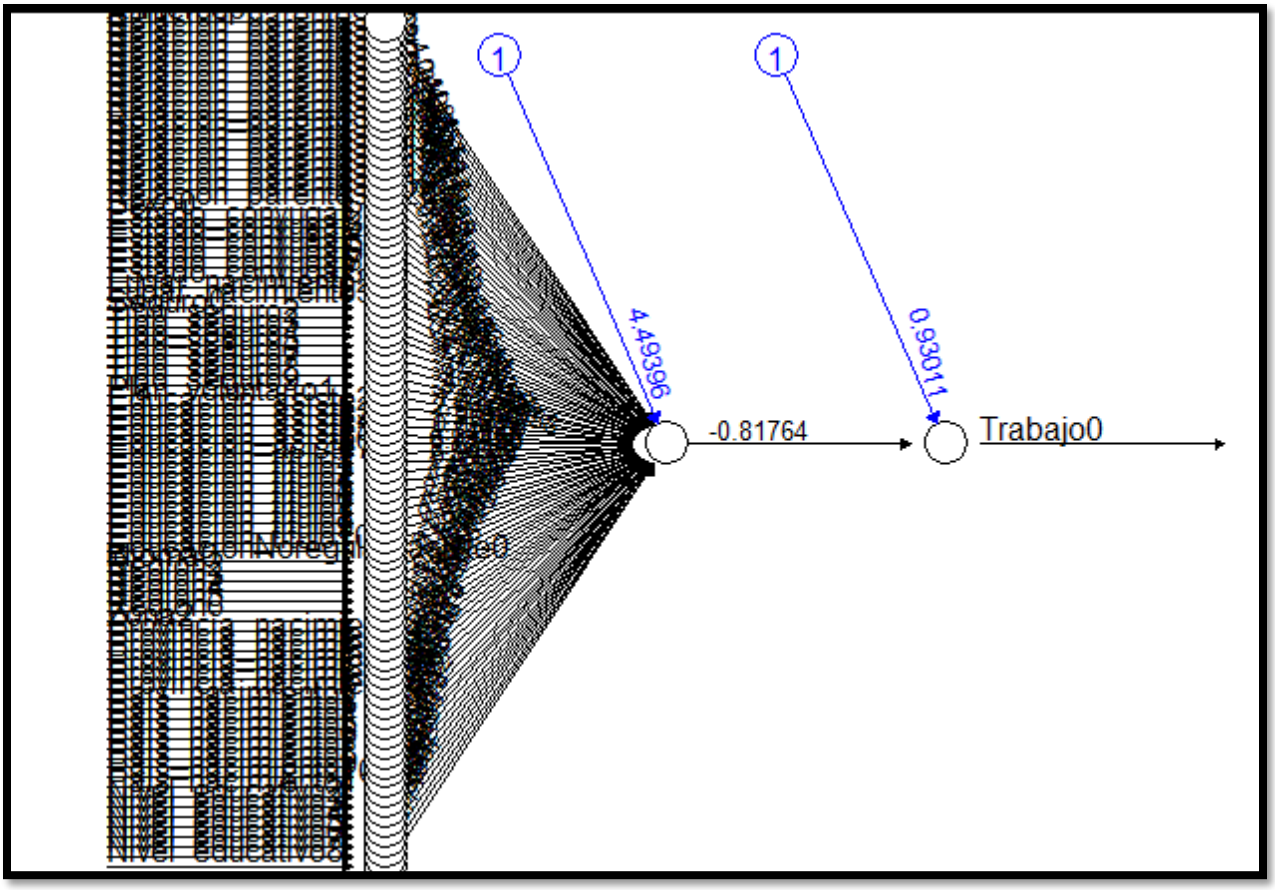


Ilustración 66: Gráfica del modelo de redes neuronales con la variable *Educacion_titulo*
Fuente: Elaboración propia

4.5. Evaluación

En esta sección se pretende determinar cuál de los seis modelos y sus variantes desarrollados en la etapa anterior logra el porcentaje de exactitud más adecuado para la predicción de la variable *Trabajo*, así como si el modelo es capaz de establecer si el objetivo principal puede ser solventado.

4.5.1. Evaluación de los resultados

Esta investigación centra su estudio en estimar la probabilidad que tiene un individuo de cambiar su estado laboral según sus variables sociodemográficas; por lo tanto, se busca predecir con la mayor exactitud posible la variable *Trabajo*, permitiendo a las autoridades pertinentes del INEC generar mejoras en la recolección de la

información y políticas que faciliten atacar de una forma más efectiva el desempleo en la población costarricense.

Los algoritmos de minería de datos aplicados en la sección anterior cuentan con una matriz de confusión, de la cual es posible obtener la exactitud, entre otros valores.

4.5.1.1. Árboles de decisión

Se realizan dos modelos de árboles de decisión, uno con la variable *Educacion_titulo* y otro que no incluye dicha variable.

En primer lugar, se presenta la evaluación del modelo que sí la incluye; de este modo, mediante la función *table()* se obtienen los valores de la matriz de confusión:

```
prediccion.rpart <- predict(modelo.rpart, ttesting, type = "prob") # Para que me retorne la probabilidad

#Ver las predicciones
head(prediccion.rpart)
summary(prediccion.rpart)

Clase2 <- ttesting[, "Trabajo"]
head(Clase2)
Score2 <- prediccion.rpart$prediction[,2]
head(Score2)
Corte <- 0.5
Prediccion <- ifelse(Score2 > Corte, "0", "1")
MC2 <- table(Clase2, Pred = factor(Prediccion, levels = c("1", "0")))
general.indexes(mc=MC2)
plotROC(Score2,Clase2)
```

Ilustración 67: Código fuente para el modelado de los árboles de decisión
Fuente: Elaboración propia

Los resultados del modelo con la variable *Educacion_titulo* son los siguientes:

```
> MC2 <- table(Clase2, Pred = factor(Prediccion, levels = c("1", "0")))
> general.indexes(mc=MC2)

Confusion Matrix:
      Pred
Clase2    1    0
     1 8100 2021
     0 1602 7217

Overall Accuracy: 0.8087
Overall Error:    0.1913

Category Accuracy:

           1           0
    0.800316    0.818347
```

Ilustración 68: Matriz de confusión del modelo árboles de decisión con la variable *Educacion_titulo*
Fuente: Elaboración propia

En la tabla 21 se encuentra el cálculo de métricas con la variable *Educacion_titulo*:

Tabla 21: *Cálculo de métricas para el modelo de árboles de decisión con la variable* Educacion_titulo

Métrica	Fórmula	Resultado
Exactitud	$\frac{8100 + 7217}{18940}$	0.8087
Sensibilidad	$\frac{8100}{9702}$	0.8348
Especificidad	$\frac{7217}{9238}$	0.7812
Tasa error	$\frac{7915 + 7274}{18940}$	0.1913

Fuente: Elaboración propia.

A continuación, se exponen los resultados obtenidos sin la variable de *Educacion_titulo*:

```
> MC2 <- table(Clase2, Pred = factor(Prediccion, levels = c("1", "0")))
> general.indexes(mc=MC2)

Confusion Matrix:
      Pred
Clase2    1    0
     1 8072 2068
     0 1695 7105

Overall Accuracy: 0.8013
Overall Error:    0.1987

Category Accuracy:

           1           0
    0.796055    0.807386
```

Ilustración 69: Matriz de confusión del modelo árboles de decisión sin la variable *Educacion_titulo*
Fuente: Elaboración propia

En la tabla 22 se encuentra el cálculo de métricas sin la variable *Educacion_titulo*:

Tabla 22: *Cálculo de métricas para el modelo de árboles de decisión sin la variable* Educacion_titulo

Métrica	Fórmula	Resultado
Exactitud	$\frac{8072 + 7105}{18940}$	0.8013
Sensibilidad	$\frac{8072}{9767}$	0.8264
Especificidad	$\frac{7105}{9173}$	0.7745
Tasa error	$\frac{1695 + 2068}{18940}$	0.1987

Fuente: Elaboración propia

Las evaluaciones son positivas y aunque no sea mucho el incremento, la variable *Educacion_titulo* sí contribuye con el entrenamiento del modelo.

4.5.1.2. Bosques aleatorios

Siguiendo la mecánica anterior, se crean dos modelos de bosques aleatorios, uno con la variable *Educacion_titulo* y otro que no incluye dicha variable.

En primer lugar, se presenta la evaluación del modelo que sí la incluye; así, mediante la función *table(),* se obtienen los valores de la matriz de confusión:

```
prediccion.RF <- predict(modelo.RF, ttesting, type = "prob", na.action = na.omit) # Para que me retorne la probabilidad.

Clase3 <- ttesting[, "Trabajo"]
head(Clase3)
Score3 <- prediccion.RF$prediction[,2]
head(Score3)
Corte <- 0.5
Prediccion <- ifelse(Score3 > Corte, "0", "1")
MC3 <- table(Clase3, Pred = factor(Prediccion, levels = c("1", "0")))
general.indexes(mc=MC3)
plotROC(Score3,Clase3)
```

Ilustración 70: Código fuente para el modelado del bosque aleatorio
Fuente: Elaboración propia

Los resultados del modelo sin la variable *Educacion_titulo* son los siguientes:

```
> MC3 <- table(Clase3, Pred = factor(Prediccion, levels = c("1", "0")))
> general.indexes(mc=MC3)

Confusion Matrix:
      Pred
Clase3    1    0
     1 6382 1299
     0  941 5943

Overall Accuracy: 0.8462
Overall Error:    0.1538

Category Accuracy:

           1           0
    0.830881    0.863306
```

Ilustración 71: Matriz de confusión del modelo bosques aleatorios sin la variable *Educacion_titulo*
Fuente: Elaboración propia

A continuación, se expone el cálculo de métricas para el modelo bosques aleatorios sin la variable *Educacion_titulo*:

Tabla 23: *Cálculo de métricas para el modelo bosques aleatorios sin la variable* Educacion_titulo

Métrica	Fórmula	Resultado
Exactitud	$\frac{6382 + 5943}{14565}$	0.8462
Sensibilidad	$\frac{6382}{7323}$	0.8715
Especificidad	$\frac{5943}{7242}$	0.8206
Tasa error	$\frac{1299 + 941}{14565}$	0.1538

Fuente: Elaboración propia

Como se puede observar, la variable cuenta con una gran cantidad de observaciones sin respuesta y no son tomadas en cuenta en el modelo.

```
> MC3 <- table(Clase3, Pred = factor(Prediccion, levels = c("1", "0")))
> general.indexes(mc=MC3)

Confusion Matrix:
      Pred
Clase3    1    0
     1 2066  175
     0  304  681

Overall Accuracy: 0.8515
Overall Error:    0.1485

Category Accuracy:

           1           0
    0.921910    0.691371
```

Ilustración 72: Matriz de confusión del modelo bosques aleatorios con la variable *Educacion_titulo*
Fuente: Elaboración propia

En la tabla 24 se indica el cálculo de métricas para el modelo bosques aleatorios con la variable *Educacion_titulo*:

Tabla 24: *Cálculo de métricas para el modelo bosques aleatorios con la variable* Educacion_titulo

Métrica	Fórmula	Resultado
Exactitud	$\frac{2066 + 681}{3226}$	0.8515
Sensibilidad	$\frac{2066}{2370}$	0.8717
Especificidad	$\frac{681}{856}$	0.7955
Tasa error	$\frac{175 + 304}{3226}$	0.1485

Fuente: Elaboración propia

Por la cantidad tan baja de observaciones útiles, es mejor no usar la variable *Educacion_titulo* por su poco aporte al aprendizaje del modelo.

4.5.1.3. Máquina de soporte vectorial

Para la evaluación de este modelo, también se consideran dos aristas, una con la variable *Educacion_titulo* y otra que no la toma en cuenta.

En primer lugar, se detalla la evaluación que sí considera la variable; en cuanto a esto, con ayuda de la función *table()* se genera la matriz de confusión:

```
prediccion.msv <- predict(modelo.msv, ttesting, type = "prob") # Para que me retorne la probabilidad

length(Clase4)
length(Score4)
length(Prediccion)
length(na.omit(ttesting))

summary(modelo.msv)

ttesting_NaOmit <- na.omit(ttesting)
Clase4 <- ttesting_NaOmit[, "Trabajo"]

head(Clase4)
Score4 <- prediccion.msv$prediction[,2]
head(Score4)
Corte <- 0.5
Prediccion <- ifelse(Score4 > Corte, "0", "1")
MC4 <- table(Clase4, Pred = factor(Prediccion, levels = c("1", "0")))
general.indexes(mc=MC4)
plotROC(Score4,Clase4)
```

Ilustración 73: Código fuente para el modelado de la máquina de soporte vectorial
Fuente: Elaboración propia

Los resultados del modelo con la variable *Educacion_titulo* son los siguientes:

```
> MC4 <- table(Clase4, Pred = factor(Prediccion, levels = c("1", "0")))
> general.indexes(mc=MC4)

Confusion Matrix:
      Pred
Clase4    1    0
     1 2049  125
     0  337  607

Overall Accuracy: 0.8518
Overall Error:    0.1482

Category Accuracy:

           1           0
    0.942502    0.643008
```

Ilustración 74: Matriz de confusión del modelo máquina de soporte vectorial con la variable *Educacion_titulo*
Fuente: Elaboración propia

En la tabla 25 se describe el cálculo de métricas para el modelo máquina de soporte vectorial con la variable *Educacion_titulo*:

Tabla 25: *Cálculo de métricas para el modelo máquina de soporte vectorial con la variable* Educacion_titulo

Métrica	Fórmula	Resultado
Exactitud	$\frac{2049 + 607}{3118}$	0.8518
Sensibilidad	$\frac{2049}{2386}$	0.8587
Especificidad	$\frac{607}{732}$	0.8292
Tasa error	$\frac{125 + 337}{3118}$	0.1482

Fuente: Elaboración propia

A continuación, se muestran los resultados obtenidos sin la variable de *Educacion_titulo*:

```
> MC4 <- table(Clase4, Pred = factor(Prediccion, levels = c("1", "0")))
> general.indexes(mc=MC4)

Confusion Matrix:
      Pred
Clase4    1    0
     1 6144 1454
     0 1028 5949

Overall Accuracy: 0.8297
Overall Error:    0.1703

Category Accuracy:

           1           0
    0.808634    0.852659
```

Ilustración 75: Matriz de confusión del modelo máquina de soporte vectorial sin la variable *Educacion_titulo*
Fuente: Elaboración propia

En la tabla 26 se indica el cálculo de métricas para el modelo máquina de soporte vectorial sin la variable *Educacion_titulo*:

Tabla 26: *Cálculo de métricas para el modelo máquina de soporte vectorial sin la variable* Educacion_titulo

Métrica	Fórmula	Resultado
Exactitud	$\frac{6144 + 5949}{14575}$	0.8297
Sensibilidad	$\frac{6144}{7172}$	0.8566
Especificidad	$\frac{5949}{7403}$	0.8035
Tasa error	$\frac{1028 + 1454}{14575}$	0.1703

Fuente: Elaboración propia

Se determina que el modelo mejora su exactitud si toma en cuenta la variable *Educacion_titulo*, por lo cual es un dato valioso para el aprendizaje.

4.5.1.4. Impulso adaptativo

A partir de los datos de prueba y de la predicción del modelo, se obtiene la matriz de confusión, con ayuda de la función *table()*.

```
prediccion.ada <- predict(modelo.ada, ttesting, type = "prob") # Para que me retorne la probabilidad

summary(modelo.ada)
print(modelo.ada)

varplot(modelo.ada, plot.it = TRUE, type = c("none","scores"),max.var.show=30)

table(ttesting$Trabajo, prediccion.ada$prediction[, 2] >= 0.5)

Clase5 <- ttesting[, "Trabajo"]
head(Clase5)
Score5 <- prediccion.ada$prediction[,2]
head(Score5)
Corte <- 0.5
Prediccion <- ifelse(Score5 > Corte, "1", "0")
MC5 <- table(Clase5, Pred = factor(Prediccion, levels = c("1", "0")))
general.indexes(mc=MC5)
```

Ilustración 76: Código fuente para el modelado del impulso adaptativo (ADA)
Fuente: Elaboración propia

Los resultados del modelo con la variable *Educacion_titulo* son los siguientes:

```
> MC5 <- table(Clase5, Pred = factor(Prediccion, levels = c("1", "0")))
> general.indexes(mc=MC5)

Confusion Matrix:
      Pred
Clase5    1    0
     1 8186 1952
     0 1615 7187

Overall Accuracy: 0.8117
Overall Error:    0.1883

Category Accuracy:

           1           0
    0.807457    0.816519
```

Ilustración 77: Matriz de confusión del modelo impulso adaptativo (ADA) con la variable *Educacion_titulo*
Fuente: Elaboración propia

En la tabla 27 se indica el cálculo de métricas para el modelo impulso adaptativo con la variable *Educacion_titulo*:

Tabla 27: *Cálculo de métricas para el modelo impulso adaptativo con la variable* Educacion_titulo

Métrica	Fórmula	Resultado
Exactitud	$\frac{8186 + 7187}{18940}$	0.8117
Sensibilidad	$\frac{8186}{9801}$	0.8352
Especificidad	$\frac{7187}{9139}$	0.7864
Tasa error	$\frac{1615 + 1952}{18940}$	0.1883

Fuente: Elaboración propia

A continuación, se muestran los resultados obtenidos sin la variable de *Educacion_titulo*:

```
> MC5 <- table(Clase5, Pred = factor(Prediccion, levels = c("1", "0")))
> general.indexes(mc=MC5)

Confusion Matrix:
      Pred
Clase5    1    0
     1 8090 1946
     0 1620 7284

Overall Accuracy: 0.8117
Overall Error:    0.1883

Category Accuracy:

           1           0
    0.806098    0.818059
```

Ilustración 78: Matriz de confusión del modelo impulso adaptativo (ADA) sin la variable *Educacion_titulo*
Fuente: Elaboración propia

En la tabla 28 se aprecia el cálculo de métricas para el modelo impulso adaptativo sin la variable *Educacion_titulo*:

Tabla 28: *Cálculo de métricas para el modelo impulso adaptativo sin la variable* Educacion_titulo

Métrica	Fórmula	Resultado
Exactitud	$\frac{8090 + 7284}{18940}$	0.8117
Sensibilidad	$\frac{8090}{9710}$	0.8332
Especificidad	$\frac{7284}{9230}$	0.7892
Tasa error	$\frac{1615 + 1952}{18940}$	0.1883

Fuente: Elaboración propia

Cabe resaltar que la exactitud y la tasa de error son prácticamente las mismas para ambas predicciones; por consiguiente, en este caso la variable *Educacion_titulo* no agrega valor al modelo.

4.5.1.5. Redes bayesianas

De igual modo, en cuanto al modelo de redes bayesianas, se efectúan dos evaluaciones, una con la variable *Educacion_titulo* y la segunda sin dicha variable.

La sentencia utilizada para obtener las matrices de confusión es la siguiente:

```
prediccion.bayes <- predict(modelo.bayes, ttesting, type = "prob") # Para que me retorne la probabilidad

table(ttesting$Trabajo, prediccion.bayes$prediction[, 2] >= 0.5)

summary(modelo.bayes)
print(modelo.bayes)
modelo.bayes

Clase7 <- ttesting[, "Trabajo"]
head(Clase7)
Score7 <- prediccion.bayes$prediction[,2]
head(Score7)
Corte <- 0.5
Prediccion <- ifelse(Score7 > Corte, "1", "0")
MC7 <- table(Clase7, Pred = factor(Prediccion, levels = c("1", "0")))
general.indexes(mc=MC7)
```

Ilustración 79: Código fuente para el modelado de la red bayesiana
Fuente: Elaboración propia

Los resultados del modelo con la variable *Educacion_titulo* son:

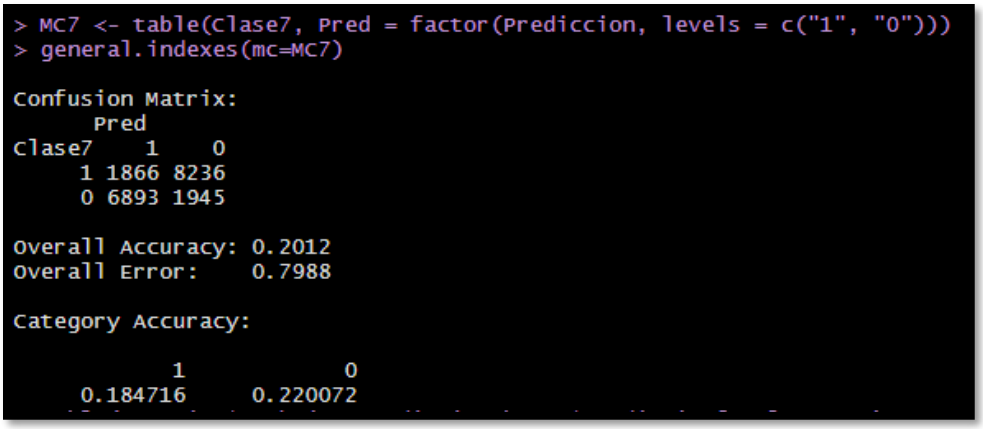


Ilustración 80: Matriz de confusión del modelo redes bayesianas con la variable *Educacion_titulo*
Fuente: Elaboración propia

En la tabla 29 se expone el cálculo de métricas para el modelo redes bayesianas con la variable *Educacion_titulo*:

Tabla 29: *Cálculo de métricas para el modelo redes bayesianas con la variable* Educacion_titulo

Métrica	Fórmula	Resultado
Exactitud	$\frac{1866 + 1945}{18940}$	0.2012
Sensibilidad	$\frac{1866}{8759}$	0.2130
Especificidad	$\frac{1945}{10181}$	0.1910
Tasa error	$\frac{6893 + 8236}{18940}$	0.7987

Fuente: Elaboración propia

A continuación, se describen los resultados obtenidos sin la variable *Educacion_titulo*:

```
> MC7 <- table(Clase7, Pred = factor(Prediccion, levels = c("1", "0")))
> general.indexes(mc=MC7)

Confusion Matrix:
      Pred
Clase7    1    0
     1 1975 8123
     0 7022 1820

Overall Accuracy: 0.2004
Overall Error:    0.7996

Category Accuracy:

           1           0
    0.195583    0.205836
```

Ilustración 81: Matriz de confusión del modelo redes bayesianas sin la variable *Educacion_titulo*
Fuente: Elaboración propia

En la tabla 30 se detalla el cálculo de métricas para el modelo redes bayesianas sin la variable *Educacion_titulo*:

Tabla 30: *Cálculo de métricas para el modelo redes bayesianas sin la variable* Educacion_titulo

Métrica	Fórmula	Resultado
Exactitud	$\frac{1975 + 1820}{18940}$	0.2004
Sensibilidad	$\frac{1975}{8997}$	0.2195
Especificidad	$\frac{1820}{9943}$	0.1830
Tasa error	$\frac{7022 + 8123}{18940}$	0.7996

Fuente: Elaboración propia

Las evaluaciones del modelo de redes bayesianas son muy decepcionantes, pues la variable *Educacion_titulo* no aporta mejoría alguna.

4.5.1.6. Redes neuronales

Por último, se evalúan los resultados de los modelos de redes neuronales creados en el apartado anterior, para lo cual se crean las matrices de confusión con ayuda de la función *table()*:

```
modelo.nnet1 <- neuralnet(Trabajo0~.,
                          data = datos.aprendizaje.red, hidden = 1)

predicciones.nnet1 <- compute(modelo.nnet1, datos.test.red)

table(datos.test.red[, 'Trabajo0'], predicciones.nnet1$net.result >= 0.5)
```

Ilustración 82: Código fuente para el modelado de la red neuronal
Fuente: Elaboración propia

Los resultados del modelo con la variable *Educacion_titulo* son los siguientes; al respecto, es posible observar que el modelo omite las observaciones sin respuesta o en NA:

```
> predicciones.nnet1 <- compute(modelo.nnet1, datos.test.red)
> table(datos.test.red[, 'Trabajo0'], predicciones.nnet1$net.result >= 0.5)

    FALSE TRUE
  0  2041  175
  1   327  681
```

Ilustración 83: Matriz de confusión del modelo redes neuronales con la variable *Educacion_titulo*
Fuente: Elaboración propia

En la tabla 31 se encuentra el cálculo de métricas para el modelo redes neuronales con la variable *Educacion_titulo*:

Tabla 31: *Cálculo de métricas para el modelo redes neuronales con la variable* Educacion_titulo

Métrica	Fórmula	Resultado
Exactitud	$\frac{681 + 2041}{3224}$	0.8443
Sensibilidad	$\frac{681}{856}$	0.7956
Especificidad	$\frac{2041}{2368}$	0.8619
Tasa error	$\frac{327 + 175}{3224}$	0.1557

Fuente: Elaboración propia

A continuación, se indican los resultados obtenidos sin la variable *Educacion_titulo*:

```
> predicciones.nnet1 <- compute(modelo.nnet1, datos.test.red)
>
> table(datos.test.red[, 'Trabajo0'], predicciones.nnet1$net.result >= 0.5)

    FALSE TRUE
  0  6282 1389
  1  1004 5938
```

Ilustración 84: Matriz de confusión del modelo redes neuronales sin la variable *Educacion_titulo*
Fuente: Elaboración propia

En la tabla 32 se detalla el cálculo de métricas para el modelo redes neuronales con la variable *Educacion_titulo*:

Tabla 32: *Cálculo de métricas para el modelo redes neuronales sin la variable* Educacion_titulo

Métrica	Fórmula	Resultado
Exactitud	$\frac{5938 + 6282}{14613}$	0.8363
Sensibilidad	$\frac{5938}{7327}$	0.8104
Especificidad	$\frac{6282}{7286}$	0.8622
Tasa error	$\frac{1004 + 1389}{14613}$	0.1637

Fuente: Elaboración propia

Siguiendo el patrón de los modelos evaluados con anterioridad, la variable *Educacion_titulo* presenta una mejoría en la exactitud de la predicción, pero no tan importante.

El resumen de las evaluaciones calculadas de cada uno de los modelos se aprecia en las siguientes tablas, la primera contiene los valores relacionados con la variable *Educacion_titulo* y la segunda omite dicha variable:

Tabla 33: *Evaluación de los resultados con la variable* Educacion_titulo

	Modelos de Mineria de Datos					
Metrica	**Arbolers de decisión**	**Bosques aleatorios**	**Máquinas de soporte vectorial**	**Impulso Adaptativo**	**Redes bayesianas**	**Redes Neuronales**
Exactitud	0.8087	0.8515	0.8518	0.8117	0.2012	0.8443
Sensibilidad	0.8348	0.8717	0.8587	0.8352	0.2130	0.7956
Especificidad	0.7812	0.7955	0.8292	0.7864	0.1910	0.8619
Tasa de Error	0.1913	0.1485	0.1482	0.1883	0.7987	0.1557

Fuente: Elaboración propia

Tabla 34: *Evaluación de los resultados sin la variable* Educacion_titulo

	Modelos de Mineria de Datos					
Metrica	**Arbolers de decisión**	**Bosques aleatorios**	**Máquinas de soporte vectorial**	**Impulso Adaptativo**	**Redes bayesianas**	**Redes Neuronales**
Exactitud	0.8013	0.8462	0.8297	0.8117	0.2004	0.8363
Sensibilidad	0.8264	0.8715	0.8566	0.8332	0.2195	0.8104
Especificidad	0.7745	0.8206	0.8035	0.7892	0.1830	0.8622
Tasa de Error	0.1987	0.1538	0.1703	0.1883	0.7996	0.1637

Fuente: Elaboración propia

4.5.2. Modelos aprobados

Todos los modelos estudiados en esta investigación final de graduación presentan una exactitud superior al 80 % y una tasa de error inferior al 20 %, a excepción del modelo de redes bayesianas, considerando tanto el estudio donde se adjunta la variable *Educacion_titulo* como el otro en el que no se toma en cuenta esta variable.

En el caso de las redes bayesianas, la exactitud es del 20 %, un valor que no cumple con los objetivos establecidos en este documento, por tal razón es descartado de las siguientes etapas de esta investigación.

Ahora bien, uno de los objetivos es predecir la probabilidad de que los sujetos que habitan en viviendas individuales conserven su estado laboral activo o sigan en un estado laboral de desempleo, por consiguiente los valores obtenidos en la sensibilidad como en la especificidad son igual de importantes y entre más alto sea su valor porcentual, mejor para el estudio.

Para los análisis con la variable *Educacion_titulo*, el modelo con la exactitud más alta es máquina de soporte vectorial (85.18 %) y en cuanto a los modelos que no incluyen dicha variable, el que tiene la exactitud más alta es bosques aleatorios (84.62 %).

4.6. Reconstrucción del modelo probabilístico

Tanto el modelo de máquinas de soporte vectorial como el de bosques aleatorios son los modelos con los criterios de evaluación más altos; son algoritmos de aprendizaje automático supervisado, simples de entrenar y fáciles de ajustar.

Ambos algoritmos permiten realizar la clasificación a partir de un conjunto de datos, donde se indica cuál es la variable cualitativa que se desea estimar según su probabilidad, comenzando con una serie de variables cualitativas y cuantitativas independientes.

4.6.1. Predicción de la probabilidad

Respecto al cálculo de la predicción de la probabilidad en ambos modelos, se utiliza la función *predict*(). Así, es necesario ingresar como parámetro de esta función el modelo generado con los datos de aprendizaje, seguidamente del conjunto de datos de prueba y, por último, el tipo de predicción que obtiene el investigador. Para el alcance

del presente proyecto final de graduación, se busca la predicción de la probabilidad de una variable binomial (1 o 0).

Sin embargo, la predicción de la probabilidad del modelo de bosques aleatorios no emplea observaciones sin respuesta, por lo que se agrega un cuarto parámetro de acción en relación con los NA: a efectos de este estudio, se deben omitir, como se observa en la ilustración 85:

```
prediccion.msv <- predict(modelo.msv, ttesting, type = "prob") # Para que me retorne la probabilidad
prediccion.RF <- predict(modelo.RF, ttesting, type = "prob", na.action = na.omit) # Para que me retorne la probabilidad.
```

Ilustración 85: Código fuente para la generación de las predicciones probabilísticas
Fuente: Elaboración propia

Como resultado se obtiene una matriz de probabilidades de clase, donde hay una columna para cada clase (0 y 1) y una fila para cada observación. Finalmente, con ayuda de la función *boxplot*(), se visualiza de una forma gráfica la distribución de las probabilidades según las variables más relevantes para cada modelo.

4.6.1.1. Distribución de las probabilidades generadas en el modelo de máquina de soporte vectorial

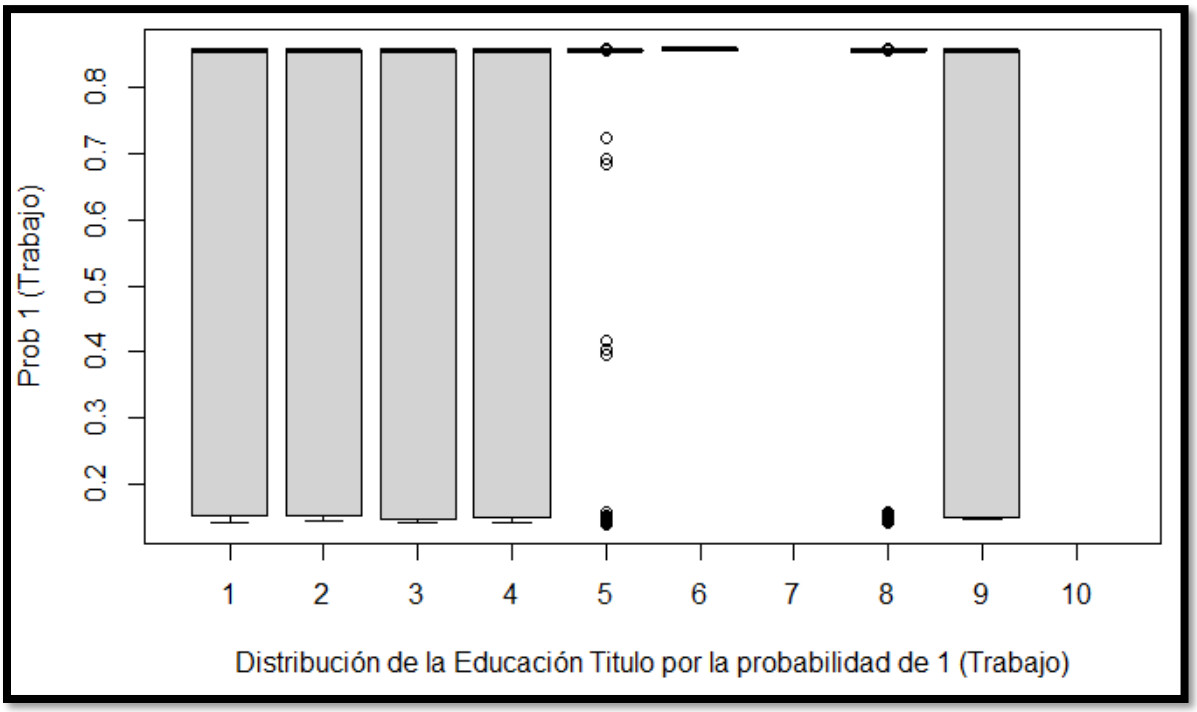


Ilustración 86: Distribución de la *Educacion_titulo* por la probabilidad de 1
Fuente: Elaboración propia

La variable *Educacion_titulo* provoca que el estudio se divida en dos modelos, uno que contempla dicha variable y otro que no. En la ilustración 86 se aprecia cómo la variable no dice mucho, ocasionado por la gran cantidad de observaciones sin respuesta y por los niveles que cubren todo el rango de probabilidad, donde su primer cuartil comienza en 0.1 y su media está en lo más cercano a 1, como lo son el nivel 1 (no tiene título), 2 (técnico, pero no universitario), 3 (profesorado, diplomado o técnico universitario), 4 (bachillerato) y 9 (doctorado).

No obstante, es de interés para la investigación determinar cómo se polariza la probabilidad de conservar o conseguir empleo en este país. De este modo, se visualiza que las observaciones dentro de los niveles 5 (licenciaturas), 6 (especialización) y 8 (maestrías) tienen más del 95 % de probabilidad de trabajar; en otras palabras, si un

ciudadano cuenta con un título en una de estas áreas de estudio, tiene un empleo garantizado.

Lo anterior origina la siguiente pregunta: ¿Es esta la realidad que se vive actualmente en Costa Rica? Y es así como el objetivo de este proyecto se comienza a cumplir, al mostrar el comportamiento de los individuos a los líderes de Costa Rica, con el fin de diseñar y desarrollar nuevas políticas que faciliten e incentiven a los habitantes de hogares individuales a efectuar estudios superiores, con becas o costos más bajos y tal vez hasta horarios más flexibles.

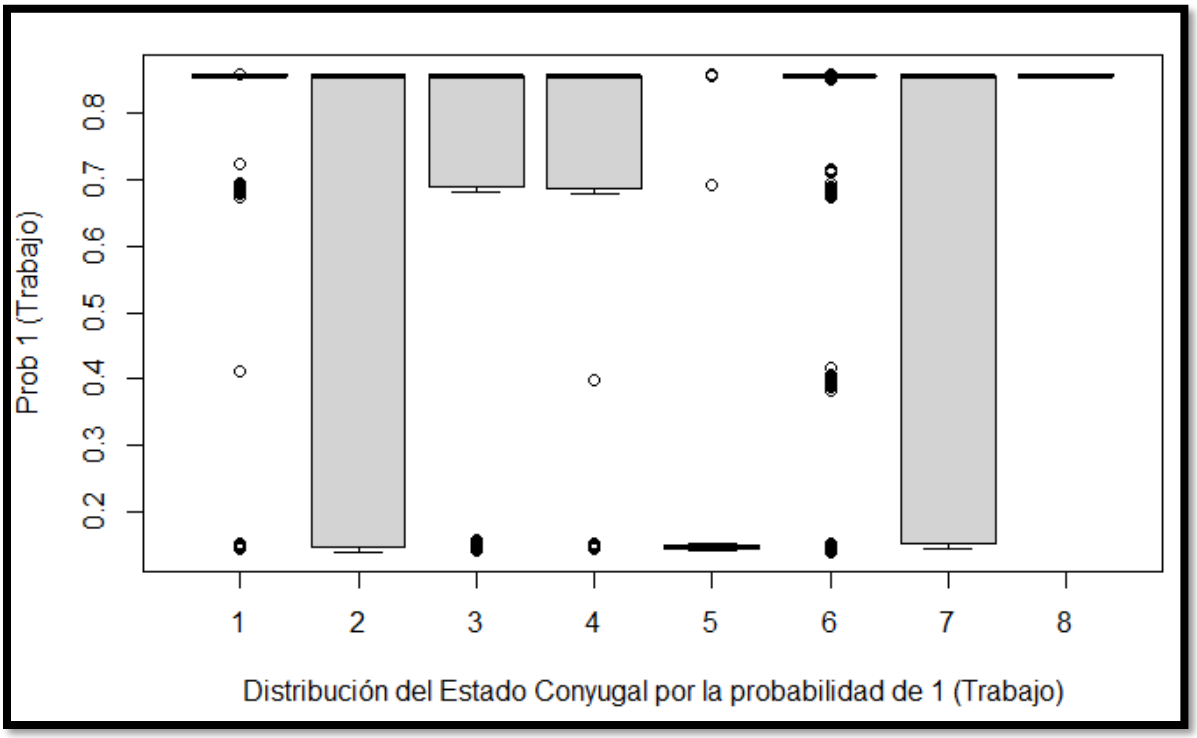


Ilustración 87: Distribución del estado conyugal por la probabilidad de 1
Fuente: Elaboración propia

El estado conyugal es la variable que posee la distribución con mayor cantidad de observaciones. Sin embargo, los niveles 2 (casado) y 7 (casado con cónyuge no residente) presentan una distribución que abarca todo el rango de probabilidades, desde 0.1 en su primer cuartil hasta 0.9 en su tercer cuartil y media, lo que no contribuye con la estimación de la variable *Trabajo*.

En cuanto a los niveles 3 (divorciado) y 4 (separado), su comportamiento es muy diferente al brindar un mayor aporte a la probabilidad, pues su primer cuartil está cerca del 0.7 y su tercer cuartil y media están más arriba del 0.9 %, por ende son los niveles más probables en conservar su trabajo.

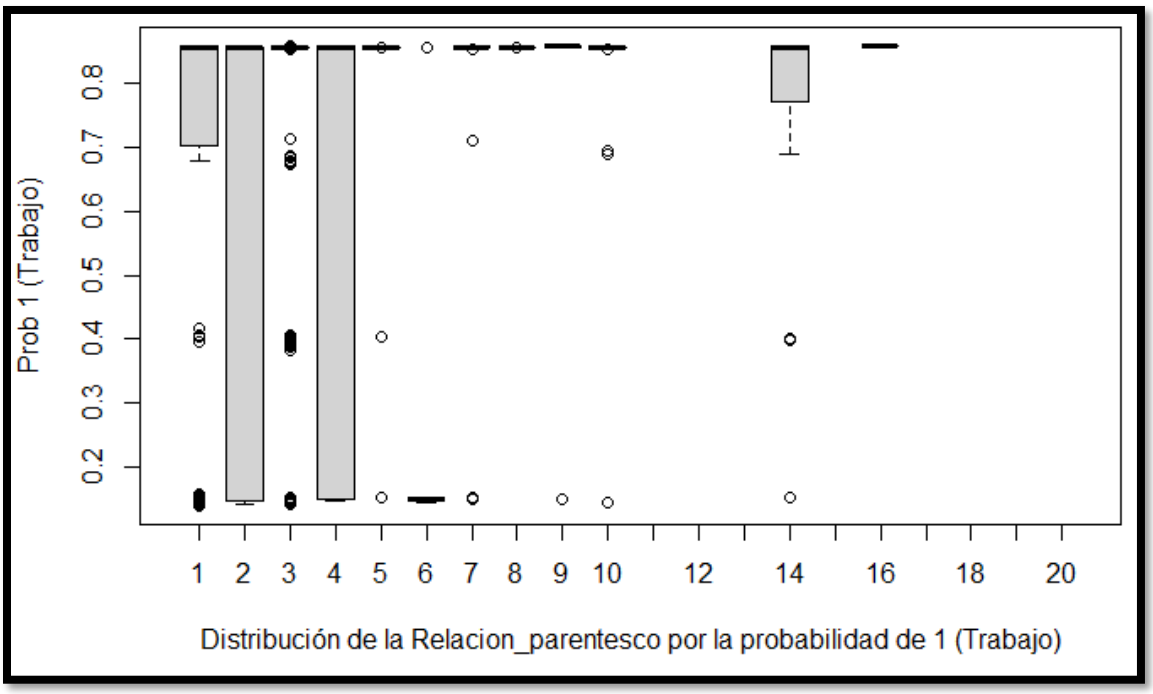


Ilustración 88: Distribución de la *Relacion_parentesco* por la probabilidad de conservar el trabajo
Fuente: Elaboración propia

En la ilustración 88 se aprecia que los niveles 5 (nieto), 7 (hermano), 8 (cuñado), 9 (otro familiar), 10 (otro no familiar) y 16 (unión libre con personas del mismo sexo) tienen una alta correlación con la variable por predecir, aunque posean algunas observaciones fuera del valor mínimo.

Por su parte, los niveles 2 (esposo o compañero) y 4 (yerno o nuera) cuentan con el rango máximo, ya que su primer cuartil está cerca del 0.1 y su tercer cuartil y media se ubican aproximadamente en el valor 1. Esto los convierte en niveles que no aportan a la predicción.

Para terminar, los niveles 1 (jefe) y 14 (hijastro) cuentan con el 0.7 y 0.75 de probabilidad de conservar su trabajo; así, estas dos relaciones de parentesco se vinculan con la variable por predecir.

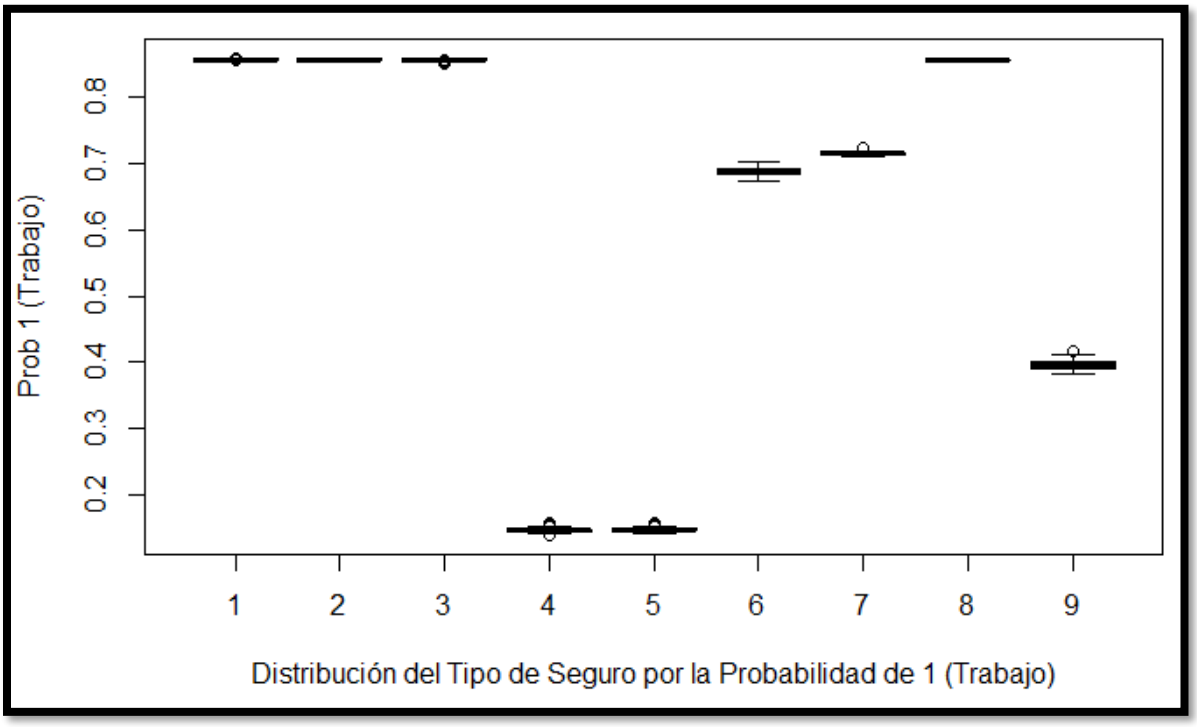


Ilustración 89: Distribución del *Tipo_seguro* por la probabilidad de conservar el trabajo
Fuente: Elaboración propia

Es necesario destacar que la variable *Tipo_seguro* es altamente correlacional con la variable por predecir, en cuanto a los niveles 1 (asalariado), 2 (mediante convenio como asociaciones, sindicatos, cooperativas, etc.), 3 (cuenta propia o voluntario) y 8 (seguro privado o extranjero).

Después se aprecia que los niveles 4 (pensionado de la CCSS, Magisterio u otro) y 5 (familiar de asegurado directo o pensionado) cuentan con una probabilidad baja de tener trabajo o de cambiar su estado de desempleado a trabajador.

Luego, dos niveles con cerca del 70 % de probabilidad de cambiar su estado a trabajador son el nivel 6 (asegurado por el Estado, incluye familiar de asegurado por el Estado) y el nivel 7 (pensionado del régimen no contributivo de monto básico, gracia o

guerra). El último nivel por comentar es el código 9 (otras formas como seguro de estudiante, de refugiado y otros), el cual ronda el 40 % de probabilidad de trabajo.

4.6.1.2. Distribución de las probabilidades generadas en el modelo de bosques aleatorios

En este apartado se analizan algunas de las variables más llamativas del modelo de minería de datos de bosques aleatorios, siendo este el modelo con los valores de exactitud más altos conseguidos sin la variable *Educacion_titulo*.

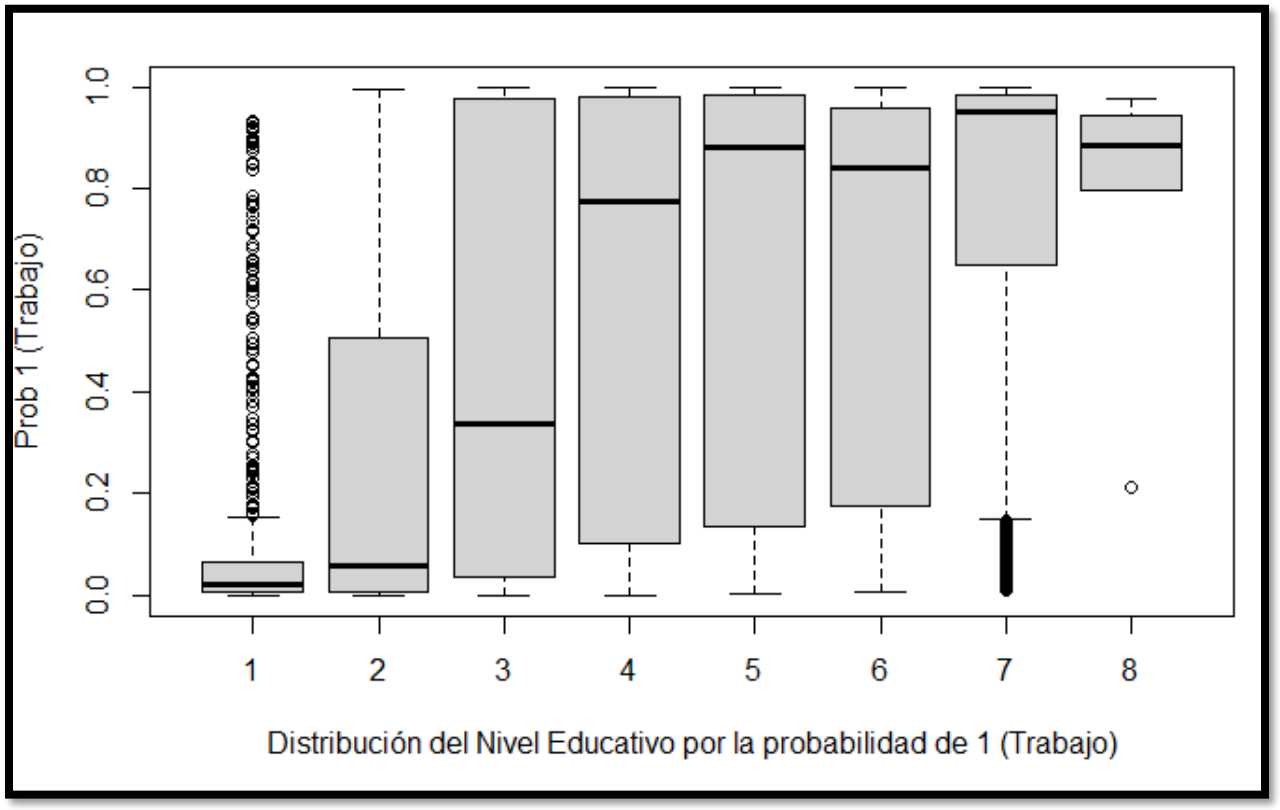


Ilustración 90: Distribución del *Nivel_educativo* por la probabilidad de conservar el trabajo
Fuente: Elaboración propia

El *Nivel_educativo* posee una distribución diversa, donde los dos primeros códigos 1 (ninguno) y 2 (primaria incompleta) tienen la media de la probabilidad más baja de salir del desempleo o de encontrar trabajo, lo cual se asocia con la realidad social que vive el

país, donde entre menor preparación académica, mayor es la probabilidad de estar desempleado.

Siguiendo con esta tendencia, entre más completo es el nivel educativo de las observaciones, mayor es la probabilidad de salir del desempleo o de cambiar su estado a trabajador. Visualmente es posible apreciar esta predisposición a partir del nivel 3 (primaria completa), que tiene una probabilidad media cercana al 40 %; luego está el nivel 4 (secundaria incompleta), donde se dispara el ascenso de la probabilidad con una media que ronda el 80 %; continuando con el nivel 5 (secundaria completa), el cual tiene una media del 90 %, muy similar al nivel 6 (universitario incompleto).

Destaca que la media de la probabilidad del código 7 (universitario completo) se ubica bastante cerca del 100%, por lo cual se puede concluir que un grado académico universitario completo es garantía de estabilidad laboral en el país, aunque algunas de las observaciones se encuentran por debajo del primer cuartil.

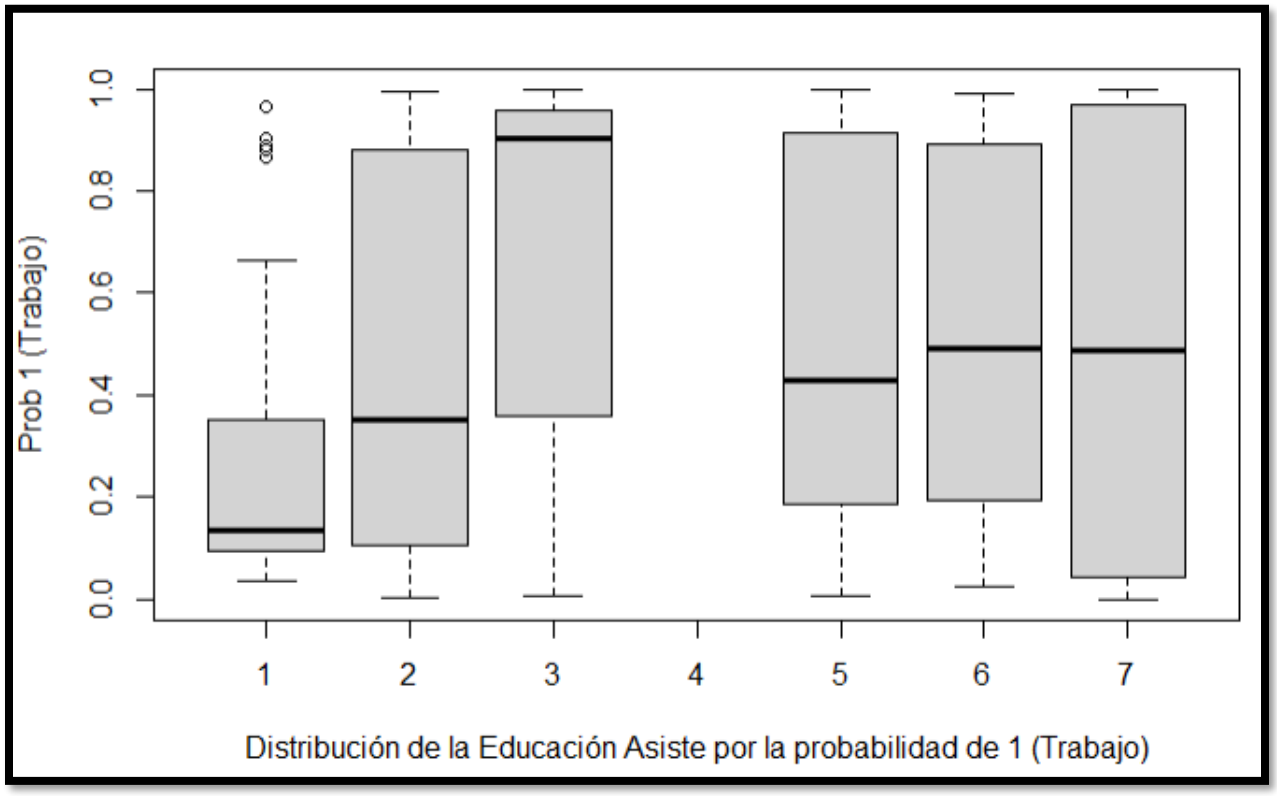


Ilustración 91: Distribución de *Educacion_asiste* por la probabilidad de conservar el trabajo
Fuente: Elaboración propia

La variable *Educacion_asiste* se encarga de recopilar la información de los encuestados que al momento de responder las preguntas de la encuesta se encuentran cursando algún grado educativo.

Si las observaciones están dentro de los niveles 1 (escuela) y 2 (secundaria), las medias de la probabilidad no son muy altas, para el nivel 1 es de menos del 20 % y para el nivel 2 se ubica por debajo del 40 %. No obstante, en el nivel 3 (parauniversitaria o universitario), como pasa con la variable *Nivel_educativo*, entre mayor es el grado académico, mayor es la probabilidad y en el caso del nivel parauniversitario o universitario, la media se dispara hasta alcanzar una probabilidad del 0.9, lo que es una muy buena señal para todos los individuos que habitan en el país.

En relación con la enseñanza especial (código 4), no se registran observaciones y para los niveles 5 (educación abierta), 6 (otro tipo de formación) y 7 (no asiste), las medias oscilan entren el 0.4 y el 0.5 de probabilidad de no salir del estado de desempleo. Por lo expuesto, una educación superior es vital para lograr una estabilidad laboral.

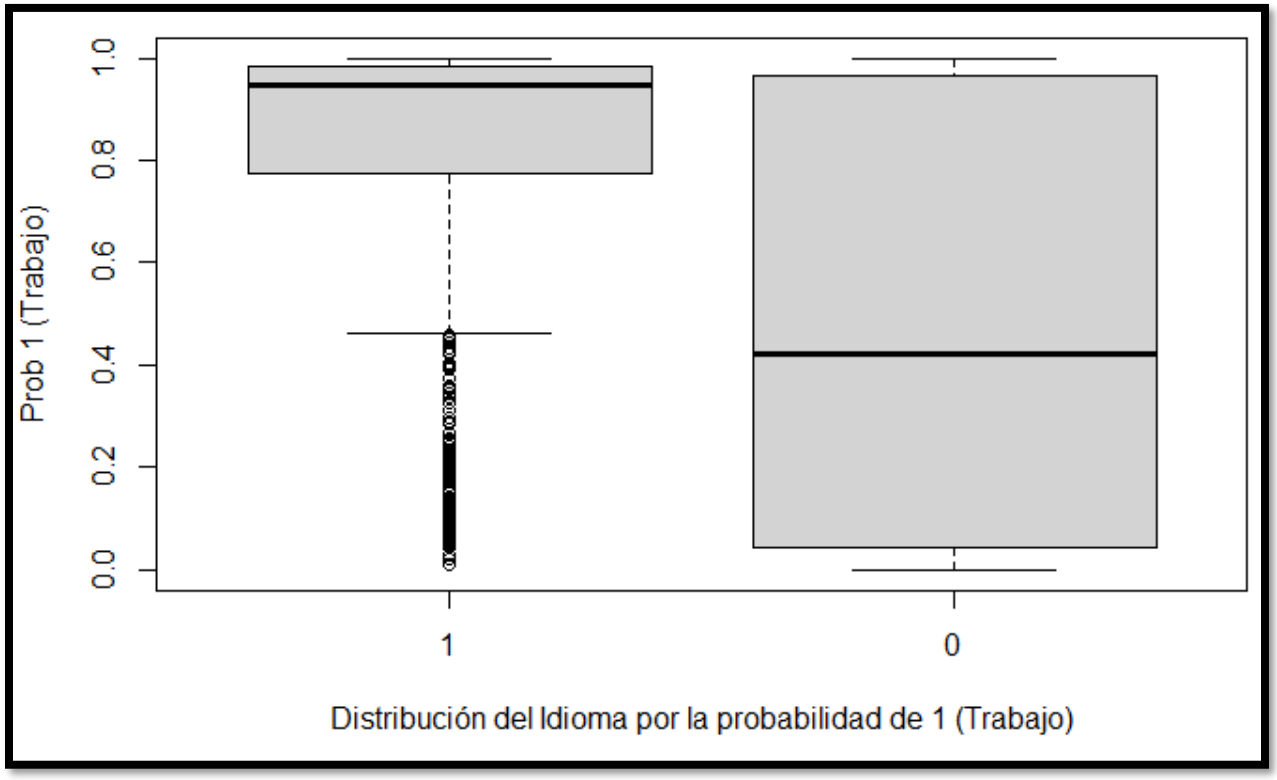


Ilustración 92: Distribución de un segundo idioma por la probabilidad de conservar el trabajo
Fuente: Elaboración propia

Hablar una segunda lengua es el complemento idóneo para todos aquellos individuos que desean salir del desempleo en Costa Rica. En la gráfica anterior se aprecia cómo las observaciones que sí hablan una segunda lengua (nivel 1) tienen una media de probabilidad ubicada cerca del 100 %, aunque se presenta una cantidad menor de observaciones por debajo del primer cuartil.

Para todas aquellas observaciones que no hablan un segundo idioma (nivel 0), la situación no mejora, cuentan con una media de la probabilidad del 40 %, lo cual es un diferencia muy marcada con los individuos que sí conversan en otro idioma distinto a su lengua natal.

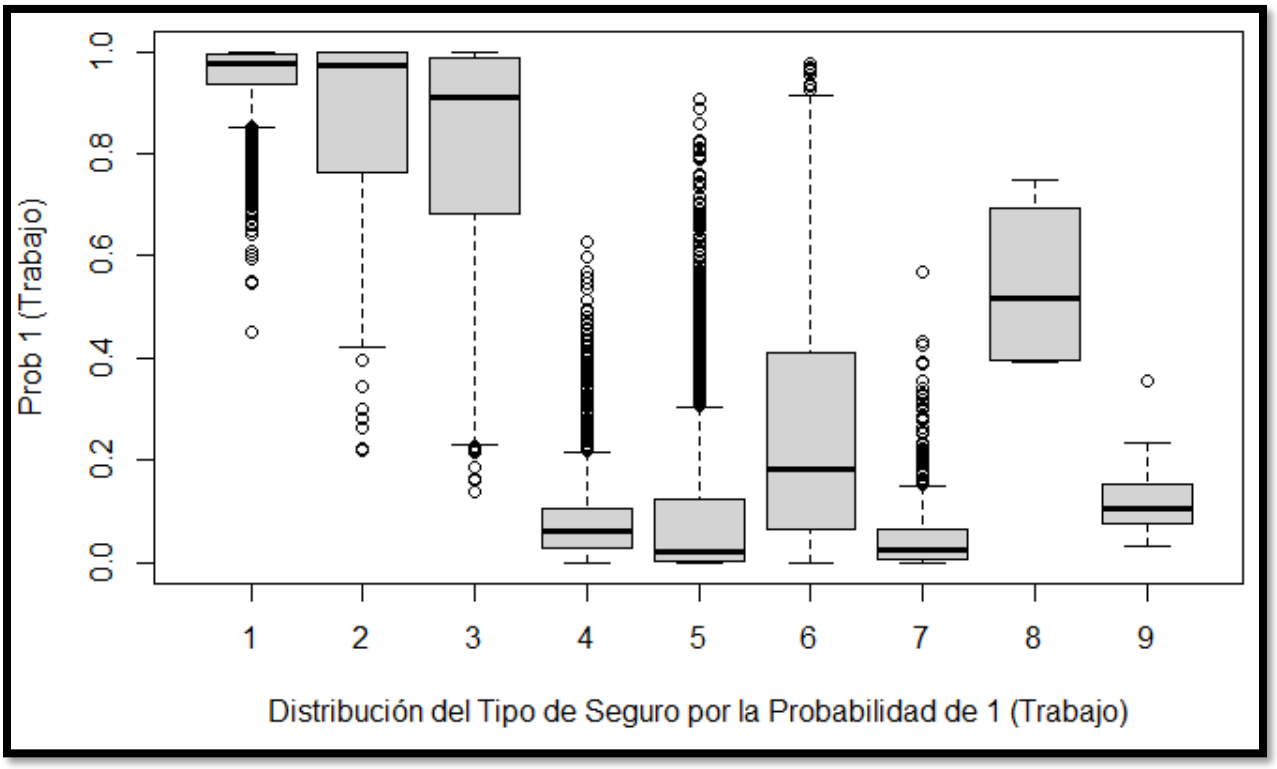


Ilustración 93: Distribución del *Tipo_seguro* por la probabilidad de conservar el trabajo
Fuente: Elaboración propia

En el análisis de las probabilidades del modelo de máquinas de soporte vectorial, la variable *Tipo_seguro* muestra una correlación sumamente alta. Lo anterior provoca que la misma variable sea analizada en la predicción de las probabilidades del modelo de bosques aleatorios.

La correlación de los tres primeros niveles no es tan alta como la correlación de las probabilidades de la predicción de las máquinas de soporte vectorial, pero sí es superior al 90 %. En cuanto a los valores de la media de los niveles 4 (pensionado de la CCSS, Magisterio u otro) y 5 (familiar de asegurado directo o pensionado), son bajos, no llegan al 0.1.

Seguidamente, se nota un cambio en las observaciones llamadas “asegurado por el Estado, incluye familiar de asegurado por el Estado” (nivel 6), donde en el caso anterior ronda en el 0.7 y en la media de la probabilidad actual se ubica en el 0.2 y “pensionado del régimen no contributivo monto básico, gracia o guerra” (nivel 7), que presenta para las predicciones del modelo de máquinas de soporte vectorial un 0.75 y en la media de la probabilidad para el modelo de bosques aleatorios se localiza en cerca del 0.0.

Para los niveles 8 (seguro privado o extranjero) y 9 (otras formas como seguro de estudiante, de refugiado y otros), la media disminuye y se sitúa en el 0.50 y 0.1 respectivamente.

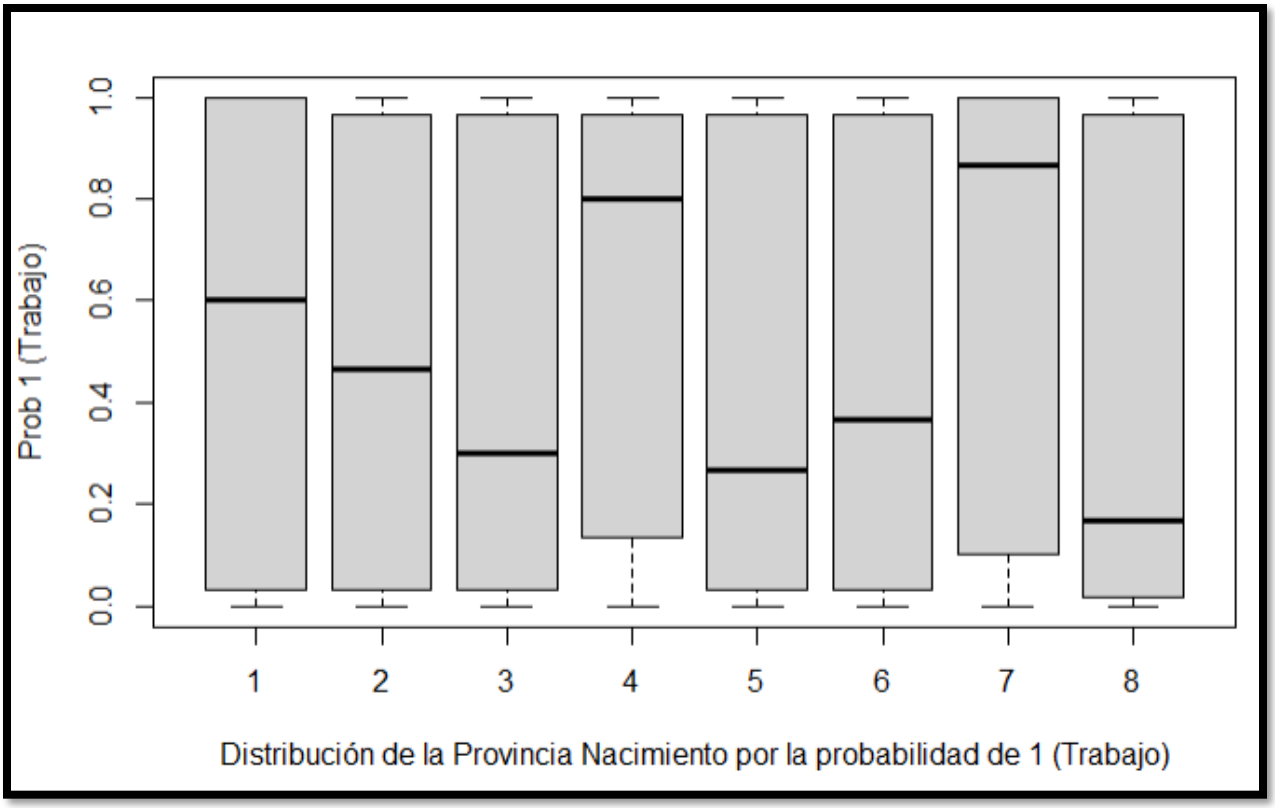


Ilustración 94: Distribución de la *Provincia_nacimiento* por la probabilidad de conservar el trabajo
Fuente: Elaboración propia

El análisis de las variables geográficas es igual de importante que el estudio de las variables sociales de este trabajo de graduación, en este caso se comienza con la variable *Provincia_nacimiento*, donde dos provincias sobresalen con sus altas probabilidades, las cuales son Heredia (nivel 4), por encima del 80 %, y Limón (nivel 7), la cual se encuentra por encima de la provincia de Heredia, llegando a casi 90 %.

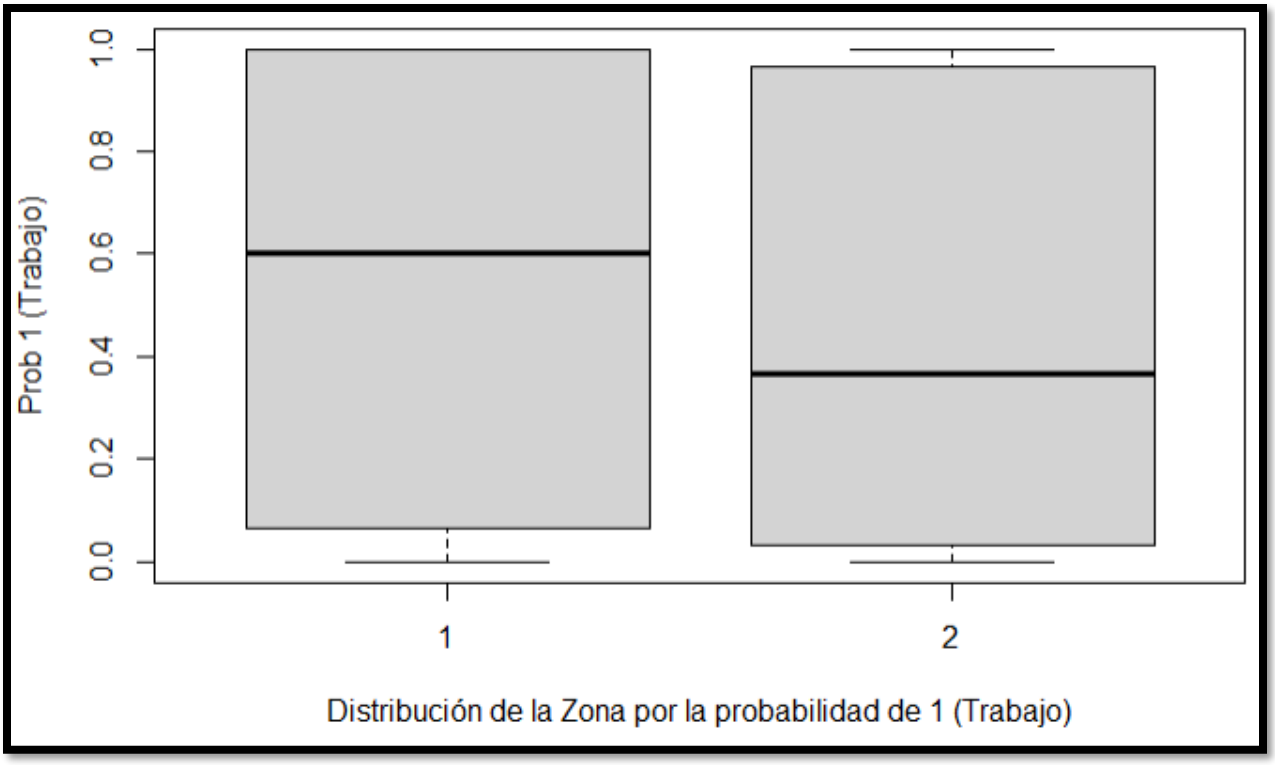


Ilustración 95: Distribución de la *Zona* por la probabilidad de conservar el trabajo
Fuente: Elaboración propia

La segunda variable en examinar es *Zona*, la cual se compone de dos niveles, el nivel urbano (1) y el nivel rural (2). De modo tal que la probabilidad más alta la tienen las observaciones que habitan en la zona urbana, con el 0.60, y muy por debajo de la zona urbana está la zona rural, casi llegando al 0.40.

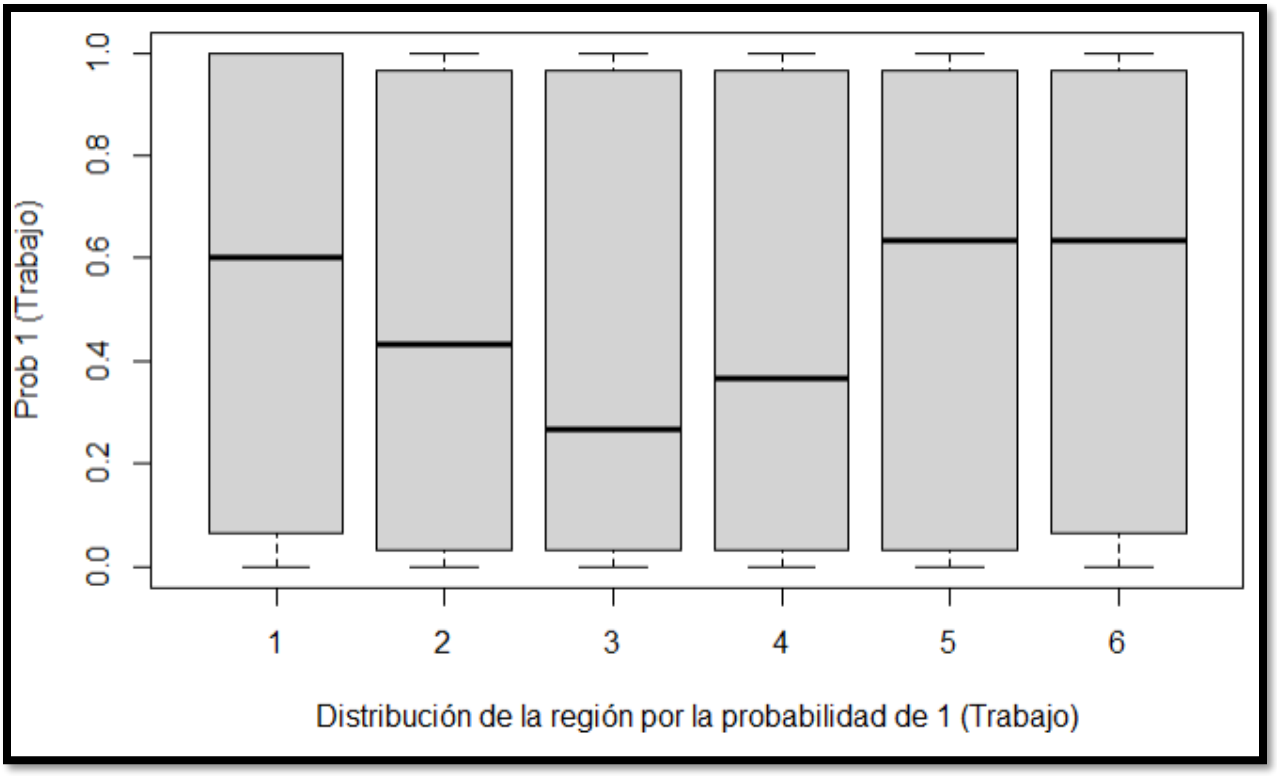


Ilustración 96: Distribución de la *Region* por la probabilidad de conservar el trabajo
Fuente: Elaboración propia

La *Region* es una variable que va de la mano con la *Provincia_nacimiento*, como se aprecia en la ilustración 96, donde los niveles más altos son el 1 (central), el 5 (Huetar Caribe) y el 6 (Huetar Norte); todos muy cercanos al 60 % de probabilidad de encontrar un trabajo que los aleje del desempleo.

4.7. Revaluación de los resultados

Si bien hallar la manera de predecir un comportamiento es el objetivo principal de este proyecto de investigación, también es muy valioso identificar las variables más representativas y eliminar aquellas que no poseen una relación directa con la variable por predecir, como lo son las constantes que identifican los salarios, deducciones o aguinaldos.

Existen muchos modelos de minería de datos, así como formas de parametrizarlos, sin embargo, es necesario probarlos con la finalidad de establecer cuál

de estos es el más idóneo para solucionar el problema expuesto en este trabajo de graduación.

A nivel social, es muy importante indicar cómo todas aquellas personas que cuentan con un nivel educativo completo logran disminuir su probabilidad de quedar sin empleo: entre más alto sea su nivel de educación o preparación académica, mejor es su probabilidad de encontrar trabajo y si a esto se le suma hablar un segundo idioma, se potencia aún más la capacidad de disminuir el porcentaje de desempleo que afecta al pueblo costarricense.

En cuanto a la variable más cuestionada, a saber, *Educacion_titulo*, se demuestra que no tiene el impacto esperado en la presente investigación, este comportamiento se percibe en especial por la gran cantidad de observaciones sin respuesta y el poco aporte al aprendizaje del modelo que esto representa.

Con relación al aspecto demográfico, las provincias de procedencia con la mayor probabilidad de conservar o conseguir trabajo son Heredia y Limón, lo cual es positivo para esta última al ser actualmente afectada por la delincuencia y el desempleo. Por su parte, la zona con la mayor probabilidad es la urbana, superando a la zona rural por más de 20 puntos porcentuales.

La última de las variables demográficas es la región, la cual posee una similitud en sus resultados con los obtenidos por la provincia de nacimiento, donde la provincia de Heredia tiene similitud con la región central de país (60 %) y la provincia de Limón con las regiones Huetar Norte y Huetar Caribe, con una alta probabilidad de ocupar un puesto de trabajo (+60 %).

4.7.1. Revisar el proceso

El proceso de investigación es muy complejo desde el inicio, pero satisfactorio. Cabe señalar que el conjunto de datos no es creado para el propósito de este proyecto de graduación final, pero dada la preocupación que existe en los costarricenses ante el aumento del desempleo, se decide usar dichos datos con el propósito de guiar a las autoridades respectivas para que tomen las mejores decisiones y modifiquen las legislaciones de este país en beneficio de las poblaciones más vulnerables, esto gracias al análisis de variables desarrollado en los puntos anteriores.

El paso más delicado es la selección de las variables independientes en cada conjunto de datos. A pesar de no contener muchas observaciones, alrededor de 100 000, sí existen muchas variables, más de 300, segmentadas en diez grupos. Así, el proceso de selección exige la lectura y análisis de la documentación técnica preparada por los especialistas del INEC sobre el programa de la ECE.

Se emplean diversos criterios de descarte, entre los que se encuentran la cantidad de observaciones sin respuesta y la relación de dependencia con la variable por predecir, como por ejemplo las diversas variables salariales, las cuales se vinculan estrictamente con la variable *Trabajo.* Otro criterio de descarte es la cantidad de niveles contenidos en las variables, pues en algunos casos hay hasta 100 niveles, provocando el aprendizaje del modelo algo engorroso y complejo.

Al momento de comenzar con el aprendizaje de los modelos de minería de datos, se consideran ocho algoritmos estadísticos debido a las lecturas realizadas en el “Estado de la cuestión” de este proyecto, donde investigaciones significativas se perfilan mejor con ciertos modelos de minería de datos, entre estos las redes neuronales y las máquinas de soporte vectorial. Al final, el estudio emplea solo seis algoritmos de aprendizaje y se exponen los resultados en este documento.

Los tiempos de aprendizaje para cada modelo son relativamente cortos, no se superan los 30 minutos en los casos más extensos, dejando el camino abierto para futuros trabajos de investigación y disminuyendo en gran medida las necesidades técnicas de los equipos de infraestructura para el procesamiento de los modelos.

4.7.2. Determinar los próximos pasos

Entre los próximos pasos, se plantea esperar los datos de las encuestas realizadas en el año 2020 y en el primer trimestre del año 2021; luego, analizar si existen cambios en las variables que componen dichas encuestas, o bien, si la relación de las variables independientes con respecto a la variable por predecir cambia en cuanto a su correlación. Además, se debe contabilizar el número de observaciones sin respuesta.

Posteriormente, y sin importar cuál sea el próximo paso, es vital entender que entra en juego un parámetro nunca considerado llamado COVID-19, variable que afecta de gran manera a Costa Rica en el año 2020, influyendo en todas las actividades

económicas desempeñadas por la población costarricense, en todas las regiones, zonas o provincias, sin diferenciar las edades, parentescos, géneros, niveles educativos o títulos académicos. No obstante, es posible ajustar los modelos y prepararlos para nuevos datos y lograr mejores resultados, siempre en beneficio de la población que habita este país.

4.8. Implementación

En la siguiente sección se detalla el mecanismo de implementación, monitoreo y mantenimiento propuesto por el autor de este proyecto de graduación final.

4.8.1. Planeamiento de la implementación

Para la implementación del proyecto, en primer lugar es necesario contactar y trabajar con el personal encargado del proyecto de la ECE y presentar los resultados obtenidos; seguidamente, obtener acceso a los datos más recientes, lo cual no representa un reto, dado que la información generada en las encuestas es de carácter público y puede ser descargada del sitio del INEC.

El tercer paso es la entrega del código fuente, el cual está programado en el lenguaje de programación estadístico R y almacenado en un proyecto generado en RStudio, que utiliza un archivo de tipo R-Markdown de fácil ejecución, porque las líneas de código se ejecutan de forma secuencial. Además, el INEC ya cuenta con *software* especializado para el manejo de este tipo de fuentes, como lo es IBM SSPS y RStudio.

4.8.2. Planeamiento del monitoreo y mantenimiento

El mantenimiento de esta investigación va de la mano con la generación de nueva información y gracias al trabajo de campo realizado por el personal del INEC, es posible obtener nuevos comportamientos y originar nuevos aprendizajes cada tres meses.

De ahí que se pretende recomendar puntos de mejora para la ECE, de manera que se vigile más de cerca el desempeño de ciertas variables y se intente hasta donde sea posible de ingresar todas las respuestas de las variables más significativas encontradas en esta investigación, en especial la *Educacion_titulo*, porque entre más

información se ingrese al modelo, mejor va a ser el aprendizaje del algoritmo y, así, se efectúan predicciones más exactas.

5. Propuesta de solución

La propuesta de solución gira en torno a los objetivos propuestos en este documento; en cuanto a estos, el más importante indica que se busca predecir la probabilidad de que un individuo obtenga empleo según sus características sociodemográficas y, en especial, considerando la variable llamada *Educacion_titulo*. Asimismo, de ser posible, guiar a las autoridades respectivas para efectuar un cambio en las políticas educativas y laborales, con el propósito de que estas disminuyan el desempleo en el país.

De ahí que la solución también compara la eficiencia en la sensibilidad de los modelos al utilizar la variable *Educacion_titulo* y al no emplearla, permitiendo de esta manera establecer si el impacto en la probabilidad de conseguir empleo es superior con dicha variable o, por el contrario, si los títulos en la educación formal no son un factor determinante.

A continuación, se exponen dos soluciones al problema, la primera hace énfasis en la variable *Educacion_titulo*, donde el modelo con la exactitud más alta utiliza el algoritmo máquinas de soporte vectorial, con una sensibilidad bastante positiva para este estudio, llegando al 85.87 %, en relación con las demás variables sociodemográficas.

La segunda solución se propone sin la variable *Educacion_titulo*, lo que permite comprobar si esta variable es determinante ante la probabilidad de salir del desempleo. Dicho esto, el modelo de minería de datos con mejor exactitud es el bosque aleatorio, logrando de igual modo la sensibilidad más alta, un 87.15 %, para predecir si el individuo tiene una alta probabilidad de conservar su trabajo.

6. Conclusiones y recomendaciones

En esta sección se describen las conclusiones asociadas a cada uno de los objetivos del proyecto, así como las posibles recomendaciones que se pueden llevar a cabo en la investigación y en la ECE para mejorar los procesos y sus resultados.

6.1. Conclusiones

El desempleo es un problema social que afecta a miles de personas en Costa Rica, por consiguiente, este trabajo brinda una herramienta que posibilita entender comportamientos, analizar patrones, priorizar características y predecir probabilidades en beneficio de una sociedad necesitada de oportunidades laborales.

En relación con el objetivo principal, a saber: “Implementar un modelo de predicción estadística de aprendizaje automático que permita predecir la probabilidad que tiene un sujeto que habita una vivienda individual ocupada de conservar su empleo según las variables sociodemográficas encontradas en las encuestas continuas trimestrales y la influencia positiva que puede ejercer la variable *Educacion_titulo* en dicha probabilidad”, se concluye que sí es posible aplicar modelos de minería de datos para la predicción de la probabilidad de salir del desempleo con la información obtenida de una encuesta continua.

Al respecto, las máquinas de soporte vectorial y el bosque aleatorio son modelos con una exactitud y una sensibilidad bastante buenas y cada una de las variables ayuda muchísimo con el resultado logrado. En cuanto a la calidad de un título educativo en relación con el resultado obtenido, su aporte al modelo no es relevante, no influye realmente en una mejoría de los resultados; no obstante, si la variable en un futuro no tiene un porcentaje tan alto de variables sin respuesta, se puede efectuar el mismo ejercicio en espera de resultados más significativos.

Por su parte, referente al objetivo: “Definir los posibles modelos de predicción y clasificación estadística aplicables a las fuentes de datos existentes”, se consiguen resultados muy positivos. En cuanto a esto, se analizan ocho distintos algoritmos, dos de los mismos son descartados antes del ensayo; luego se trabaja con los restantes seis modelos de minería de datos, cada uno con características muy diferentes y con diversas capacidades de aprendizaje, además son modelos de minería de datos que son vistos y

explicados en las clases de maestría por parte de los profesores a cargo de impartir los cursos. El objetivo se cumple al encontrar los dos modelos que más se ajustan al conjunto de datos, siendo las máquinas de soporte vectorial y los bosques aleatorios.

Para el objetivo: “Definir cuál debe ser la variable por predecir por parte del modelo”, la variable correcta es la llamada *Trabajo*, siendo esta variable binomial con valores de 1 (uno) para las observaciones que tienen trabajo y 0 (cero) para indicar que el individuo se encuentra desempleado al momento de aplicar la ECE.

Acerca del objetivo específico: “Distinguir entre las diferentes opciones de modelaje, según sus características y variables, y las que mejor se ajustan a la predicción de la probabilidad de ser un individuo sin empleo”, se concluye que existen muchos modelos, pero no todos se ajustan a las necesidades de este proyecto de investigación final; sin embargo, después del trabajo de análisis realizado, se demuestra que sí existen dos modelos con la exactitud y la sensibilidad adecuadas para cumplir con este objetivo.

En cuanto al objetivo: “Analizar cada una de las variables cualitativas y cuantitativas que brindan las encuestas del INEC, escogiendo solo las variables con mayor relevancia y correlación”, se cumple a cabalidad. Ahora bien, para culminar este objetivo, se envuelve la documentación técnica y las lecturas oficiales de las preguntas de la ECE que generan las más de 300 variables que conforman el conjunto de datos. Asimismo, se aplican múltiples mecanismos de limpieza y descarte, los cuales demuestran ser muy efectivos, permitiendo optimizar la selección de las variables a un total de 26, cada una con un potencial de aprendizaje muy alto, lo cual enriquece los resultados obtenidos y prepara el camino para futuras investigaciones.

Respecto al objetivo: “Definir cuáles son los valores de cada variable que más importancia tienen en la generación de la probabilidad de encontrar empleo”, se define, por ejemplo, que los individuos nieto, hermano, cuñado u otro no familiar tienen la probabilidad alta de conseguir empleo; como conclusión, sí se cumple el objetivo al demostrar cuáles valores aportan más la predicción del modelo.

Por último, para el objetivo: “Analizar los resultados obtenidos por el modelo y aplicar las correcciones que sean necesarias”, se cumple con lo establecido, eso sí, gracias a las diversas correcciones aplicadas a lo largo de la investigación y que en su mayoría son indicadas por el profesor tutor.

6.2. Recomendaciones

Se le recomienda al INEC velar por la correcta recolección de los datos, en especial respecto a las preguntas que conforman las más de 20 variables independientes utilizadas en esta investigación, y en específico para las educativas, las cuales representan un valor agregado muy alto, no obstante tienen un porcentaje significativo de observaciones sin respuesta, lo que no beneficia el aprendizaje.

Asimismo, se le sugiere al INEC garantizar que los datos obtenidos en censos futuros continúen al alcance de la población general. Existen muchos profesionales en bases de datos, ciencia de datos, entre otros, que pueden contribuir de una forma desinteresada con el crecimiento y bienestar del país.

A todos aquellos estudiantes que desean o tienen la inquietud de crecer en el área de minería de datos, se les aconseja fortalecer sus bases y conocimientos en el área de la estadística, dado que el campo del análisis de datos no es solo aplicar funciones y revisar los resultados, sino ajustar los modelos, entender cómo funcionan y proporcionarle al algoritmo lo que mejor se ajusta para los requerimientos que se desean solventar.

Para evitar costos en temas de licenciamiento, se recomienda usar el lenguaje de programación estadística R y la herramienta de compilación RStudio, por su licencia *open source* o código abierto; igualmente, por la posibilidad que existe de descargar una infinidad de librerías útiles como *dplyr* para el manejo del conjunto de datos, *ggplot2* para la generación de gráficas y *maps* para el uso de coordenadas geográficas y utilizarlas en la creación de gráficas locales o globales.

También se sugiere no perder de vista el sesgo por no respuesta, a pesar de ser un problema complicado de resolver y que perjudica a todos los estudios que obtienen sus datos de las encuestas. El sesgo por no respuesta se presenta en las encuestas por una falta de conciencia sobre la importancia de la información brindada, por un desconocimiento de los beneficios a mediano y largo plazo que puede ofrecer el análisis de los datos, sumado al miedo que existe en relación con posibles situaciones delictivas como estafas, robos o suplantación de identidad. Sin embargo, es posible disminuir el sesgo por no respuesta mediante un rediseño de las preguntas aplicadas o un cambio

en la duración de la encuesta, lo cual reduce el cansancio y la pérdida de la concentración.

En conclusión, se recomienda para todo proyecto de minería de datos implementar la metodología CRISP-DM respetando cada una de sus etapas. Una de las cualidades más relevantes de esta metodología es su capacidad de regresar a etapas pasadas cuando el proyecto se encuentra avanzado, lo cual posibilita ajustar o modificar lo que presente problemas y continuar justo donde es necesario retroceder. Así mismo, cada etapa está bien definida y es de fácil comprensión para los investigadores.

7. Reflexiones finales

El activo más valioso de toda organización o compañía son sus datos, los cuales la mayoría de las veces son recopilados durante años y después almacenados por largos periodos de tiempo, esto en el mejor de los escenarios, pero en otros casos no tan optimistas, luego de un cierto tiempo, la información es desechada sin nunca ser aprovechada.

Este proyecto busca aprovechar la información generada por una organización, exponer la importancia de sus datos y demostrar que si se analiza el pasado de esta, es posible llegar a predecir de una manera muy confiable los futuros pasos de la compañía. La minería de datos, guiada por una excelente metodología de proyectos de análisis de datos como lo es CRISP-DM, puede aportar mucho valor agregado al crecimiento de la organización.

Sin embargo, los esfuerzos se enfocan no solo en ayudar a una organización, sino a toda una población, por lo cual el valor agregado es más significativo. Al respecto, Costa Rica afronta muchos problemas sociales que afectan a toda la población, ya sean de drogas, inseguridad, trata de personas o desempleo, siendo este último el problema que impulsa el desarrollo del proyecto, pues la población se queda sin oportunidades reales de trabajo, lo cual se refleja en el indicador nacional de desempleo que aumenta hasta superar los doce puntos porcentuales, valor que tiene al momento de iniciar el proyecto y sin presentar la afectación actual por la pandemia de la COVID-19.

Se pretende que esta investigación ayude a mejorar la situación del empleo en Costa Rica al brindar una herramienta para que las autoridades tomen decisiones en cuanto a combatir este problema social y puedan replantear las políticas laborales que ataquen las variables más influyentes del desempleo.

8. Trabajos a futuro

Concluida la investigación, se amplían las posibilidades y los alcances, por ejemplo, es perfectamente viable centrarse a estudiar el comportamiento de la población ante otros problemas sociales, así como intentar atacarlos de manera progresiva y guiada por una metodología que favorezca el análisis de la información.

Para esto, es preciso buscar nuevas formas de conseguir la información, no solo limitarse a los datos obtenidos de modo trimestral. Una buena fuente de información en la actualidad son las redes sociales, tales como Facebook y Twitter, donde los individuos comparten estados y situaciones personales con mayor facilidad, en especial porque no sienten la presión o la pena de tener que conversar de sus situaciones o problemas privados con una persona extraña que llega a su casa y realiza una serie de preguntas.

Un tema que preocupa y se vincula con la generación de nueva información es el impacto del caso de la Unidad Presidencial de Análisis de Datos (UPAD), el cual aún se encuentra en investigación pues su creación y administración no se da en el margen de las mejores prácticas, negando así un enorme potencial de análisis de información en beneficio de la población costarricense.

Otro tema que origina preocupación sobre el futuro de esta investigación y de la obtención de datos es la incertidumbre que existe con respecto a la pandemia de la COVID-19. Aunque los datos utilizados en esta investigación no están relacionados con dicha pandemia, como se explica en puntos anteriores, sí se debe considerar que la afectación inicia en el primer trimestre del año 2020 y el personal encargado de recopilar la información ve muy limitado su trabajo de campo y, por ende, la aplicación de las encuestas.

Para terminar, este proyecto de investigación final tiene un alcance público, por lo tanto se facilitan los datos y los resultados conseguidos, con lo cual se espera que muchos otros profesionales se animen a contribuir en la investigación y solución de otros problemas sociales, velando siempre por el bienestar de los habitantes del país mediante la toma de mejores decisiones y la generación de políticas gubernamentales tanto a nivel de Gobierno local como de Gobierno de la República.

Referencias

Aiken, M. (1996). A Neural Network to Predict Civilian Unemployment Rates. *Journal of International Information Management,* *5*(1). Recuperado de: https://scholarworks.lib.csusb.edu/jiim/vol5/iss1/3

Biolchini, J., Gomes, P., Cruz, A. y Horta, G. (2005). *Systematic Review in Software Engineering*. Río de Janeiro, Brasil: Systems Engineering and Computer Science Department.

Curbelo, C. y Hurst, W. (2020). A Machine Learning Approach for Detecting Unemployment Using the Smart Metering Infrastructure. *IEEE Access*. Recuperado de: https://www.researchgate.net/publication/338897258_A_Machine_Learning_Approach_for_Detecting_Unemployment_Using_the_Smart_Metering_Infrastructure

Ford, M. (2013). Could Artificial Intelligence Create an Unemployment Crisis? *Communications of the ACM,* *56*(7), 37-39. Recuperado de: https://cacm.acm.org/magazines/2013/7/165475-could-artificial-intelligence-create-an-unemployment-crisis/fulltext

Hamilton, H. (2018). *ROC Graph.* Recuperado de http://www2.cs.uregina.ca/~dbd/cs831/notes/ROC/ROC.html

Hernández, R. y Mendoza, C. (2018). *Metodología de la investigación. Las rutas cuantitativa, cualitativa y mixta.* Ciudad de México: Editorial McGraw Hill Education.

Hernández, R., Fernández, C. y Baptista, P. (2010). *Metodología de la investigación*, (5° ed.). México: McGraw Hill.

Instituto Nacional de Estadística y Censos (INEC). (2012). *Encuesta Continua de Empleo: Métodos y Procedimientos.* Recuperado de https://www.inec.cr/sites/default/files/documetos-biblioteca-virtual/meecemetodos_01.pdf

Instituto Nacional de Estadística y Censos (INEC). (2016). *Manual de Clasificación Geográfica con Fines Estadísticos de Costa Rica.* Recuperado de http://sistemas.inec.cr/sitiosen/sitiosen/Archivos/Codificador_pa%C3%ADs_2015.pdf

Instituto Nacional de Estadística y Censos (INEC). (2017). *Reseña histórica*. Recuperado de: https://www.inec.cr/resena-historica

Instituto Nacional de Estadística y Censos (INEC). (2018). *Marco orientador y metodología para la implementación del sistema de control interno en el INEC*. Recuperado de: https://www.inec.cr/sites/default/files/documetos-biblioteca-virtual/ciplanidmarcoorientadorsci2018-01.pdf

Instituto Nacional de Estadística y Censos (INEC). (2019). *Tasa de desempleo abierto se situó en 12,4 %*. Recuperado de: https://www.inec.cr/noticia/tasa-de-desempleo-abierto-se-situo-en-124

Instituto Nacional de Estadística y Censos (INEC). (2020). *Encuesta Continua de Empleo, I, II, III y IV trimestres 2019*. Recuperado de http://sistemas.inec.cr/pad5/index.php/catalog/246/get-microdata

Instituto Nacional de Estadística y Censos (INEC). (2020). *Quiénes somos.* Recuperado de: https://www.inec.cr/quienes-somos

Katris, C. (2020). Prediction of Unemployment Rates with Time Series and Machine Learning Techniques. *Computational Economics*. Recuperado de: https://www.researchgate.net/publication/334492803_Prediction_of_Unemployment_Rates_with_Time_Series_and_Machine_Learning_Techniques

Lozada, J. (2014). *Investigación aplicada: definición, propiedad intelectual e industrial.* Quito, Ecuador: Universidad Tecnológica Indoamérica.

Nirmala, C., Roopa, G. y Naveen, K. (2015). *Twitter Data Analysis for Unemployment Crisis.* International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT). IEEE, India. Recuperado de: https://ieeexplore.ieee.org/document/7456920/authors#authors

Olmedo, E. (2014). Forecasting Spanish Unemployment Using Near Neighbour and Neural Net Techniques. *Computational Economics, 43*(2). Recuperado de: https://www.researchgate.net/publication/257552955_Forecasting_Spanish_Unemployment_Using_Near_Neighbour_and_Neural_Net_Techniques

Programa Estado de la Nación. (2019). *Seguimiento al desarrollo humano sostenible*. Costa Rica: Programa Estado de la Nación.

República de Costa Rica. Ley n.° 7839 de 1998. Sistema de Estadística Nacional. 15 del 10 de 1998. La Gaceta n.° 214. Recuperado de: http://www.pgrweb.go.cr/scij/Busqueda/Normativa/Normas/nrm_norma.aspx?param1=NRM&nValor1=1&nValor2=42064&nValor3=0&strTipM=FN

República de Costa Rica. Ley n.° 9694 de 2019. Ley del Sistema de Estadística Nacional. 04 de junio de 2019. La Gaceta n.° 110. Recuperado de: http://www.pgrweb.go.cr/scij/Busqueda/Normativa/Normas/nrm_norma.aspx?param1=NRM&nValor1=1&nValor2=88964&nValor3=116572&strTipM=FN

Santra, A. y Christy, J. (2012). Genetic Algorithm and Confusion Matrix for Document Clustering. *IJCSI International Journal of Computer Science Issues, 9*(2), 322-328. Recuperado de http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.403.2710&rep=rep1&type=pdf

Smart Vision Europa. (2015). *Cross Industry Standard Process for Data Mining model*. Recuperado de: http://crisp-dm.eu/reference-model/

Smart Vision Europe. (2015). *CRISP-DM methodology*. Recuperado de: http://crisp-dm.eu/home/crisp-dm-methodology

Xu, W., Li, Z., Cheng, C. y Zheng, T. (2013). Data Mining for Unemployment Rate Prediction Using Search Engine Query Data. *SOCA 7*, 33–42. Recuperado de: https://doi.org/10.1007/s11761-012-0122-2

Apéndice A

Código fuente programado en R

Código fuente para el análisis, limpieza, exploración y preparación de los datos

```
```{r}
#Lectura de los datos según las fuentes descargadas

library(haven)
library(dplyr)
library(foreign)

#Lectura de las 4 encuestas
Trimestre_1 <- read.spss("I_Trimestre_2019.sav", to.data.frame = TRUE, reencode = NA, use.missings = FALSE)
Trimestre_2 <- read.spss("II_Trimestre_2019.sav", to.data.frame = TRUE, reencode = NA, use.missings = FALSE)
Trimestre_3 <- read.spss("III_Trimestre_2019.sav", to.data.frame = TRUE, reencode = NA, use.missings = FALSE)
Trimestre_4 <- read.spss("IV_Trimestre_2019.sav", to.data.frame = TRUE, reencode = NA, use.missings = FALSE)

#Unión de 4 datasets

Encuesta_2019 <- bind_rows(Trimestre_1, Trimestre_2, Trimestre_3, Trimestre_4)
Enc2019_SeccionAnalisis <- Encuesta_2019[, c(7:28, 181, 182, 189, 190, 192, 193, 196)]
```
```

```
###### Se aplica un filtro para las observaciones que no son númericas y que pueden presentar error al momento de convertir.
Edad_Filtrada <- filter(Enc2019_SeccionAnalisis, Edad != "Menor de un año")
Edad_Filtrada <- filter(Edad_Filtrada, Edad != "97 años y más Menor de 15 años con edad ignorada")
Edad_Filtrada <- filter(Edad_Filtrada, Edad != "Mayor de 15 años con edad ignorada")

###### Se crea un subset con la función select de dplyr
Enc2019_SeccionAnalisis <- select(Edad_Filtrada, Relacion_parentesco, Sexo, Edad, Estado_conyugal, Lugar_nacimiento, Permanencia_pais, Permanencia_intension, Permanencia_motivo, Seguro, Tipo_seguro, Regimen_pension, Plan_voluntario, Educacion_asiste, Educacion_nivel_grado, Educacion_titulo, Educacion_codigotitulo, EducacionNoregular_asiste, EducacionNoregular_codigo, EducacionNoregular_institucion, Idioma, Idioma_cual, Trabajo, Region, Zona, Poblacion_joven, Poblacion_adulto, Provincia_nacimiento, Pais_nacimiento, Nivel_educativo) #De este forma se extrae el subset con la funcion select de Dplyr

#Enc2019_Poblacion <- select(Encuesta_2019, Poblacion_adulto, Poblacion_joven)

##### Se realiza una conversión del factor Edad a uno numérico

Enc2019_SeccionAnalisis$Edad <- as.numeric(Enc2019_SeccionAnalisis$Edad)

##### Se filtran las edades que son menores a 18 años.

Enc2019_SeccionAnalisis <- filter(Enc2019_SeccionAnalisis, Edad >= 18)

#### Generación de agrupación de la edad en factores.

Enc2019_SeccionAnalisis$EdadAgrupada <- cut(Enc2019_SeccionAnalisis$Edad, breaks = seq(17,98,by=10), right = TRUE)
```

```
Enc2019_SeccionAnalisis %>% group_by(EdadAgrupada)
```

```{r}
#### Sección para la presentación de las gráficas para la exploración de los datos.

library(ggplot2)
library(maps)

### generación de un gráfico que nuestra la cantidada de observaciones que trabaja y la cantidad de observaciones que no trabajan.

ggplot(Encuesta_2019, aes(x = Trabajo, fill=Trabajo)) + geom_bar() + geom_text(aes(y=..count.., label=..count..,colour = factor(Trabajo)),
              stat="count", color="white",
              hjust=1.0, size=3) + theme(legend.position="none") + coord_flip() + xlab("Trabaja ?") +
         ylab("Cantidad de Observaciones")

### Visualización de la cantidad de observaciones por Genero

ggplot(Encuesta_2019, aes(x = Sexo)) + geom_bar(aes(fill = Sexo)) + scale_fill_manual(values = c("skyblue", "royalblue"), limits = c("Hombre", "Mujer"), breaks =c("Hombre", "Mujer"), name = "Genero", labels = c("Hombre", "Mujer")) + ylab("Cantidad de Observaciones") + ggtitle("Cantidad de individuos por género.")##scale_fill_brewer(palette = "Greens")

### Visualizaciones del conjunto de datos por edad.

hist(Enc2019_SeccionA$Edad, xlab = 'Edad', ylab = 'Cantidad', main = 'Histograma por Edad', col = "skyblue", lwd = 5,)
```

```
### Se pretende crear un mapa con la nacionalidad de los individuos.

Encuesta_2019.region <- dplyr::count(Encuesta_2019, Pais_nacimiento)
Encuesta_2019.region <- dplyr::rename(Encuesta_2019.region, region = Pais_nacimiento)
Encuesta_2019.region <- dplyr::rename(Encuesta_2019.region, value = n)

levels(Encuesta_2019.region$region)[levels(Encuesta_2019.region$region)=="Estados Unidos"] <- "USA"
levels(Encuesta_2019.region$region)[levels(Encuesta_2019.region$region)=="México"] <- "Mexico"

world_map <- map_data

Encuesta_2019.region.map <- dplyr::left_join(Encuesta_2019.region, world_map, by = "region")

ggplot(Encuesta_2019.region.map, aes(long, lat, group = group))+
  geom_polygon(aes(fill = value), color = "white")+
  scale_fill_viridis_c(option = "C") + ylab("Latitud") + xlab("Longitud") + ggtitle("Cantidad de individuos por país de nacimiento")

ggplot(Encuesta_2019, aes(x = Pais_nacimiento, fill=Pais_nacimiento)) + geom_bar() + geom_text(aes(y=..count.., label=..count..,colour = factor(Pais_nacimiento)),
            stat="count", color="black",
            hjust=1.0, size=3) + theme(legend.position="none") + coord_flip() + xlab("País Nacimiento") +
        ylab("Cantidad de Observaciones") + ggtitle("Cantidad de individuos por país de nacimiento")
```

```
### Generción de la información por titulo academico.

ggplot(Encuesta_2019, aes(x = Educacion_titulo, fill=Educacion_titulo)) + geom_bar() + geom_text(aes(y=..count.., label=..count..,colour = factor(Educacion_titulo)),
            stat="count", color="black",
            hjust=1.0, size=3) + theme(legend.position="none") + coord_flip() + xlab("Educación Título") +
      ylab("Cantidad de Observaciones") + ggtitle("Individuos según el Título de Educación")

### Generación de la visualización para la variable Educacion Asiste.

levels(Enc2019_SeccionAnalisis$Educacion_asiste)[levels(Enc2019_SeccionAnalisis$Educacion_asiste)=="...educación abierta en instituciones para presentar examenes ante el MEP)?"] <- "...Sin examenes del MEP?"
levels(Enc2019_SeccionAnalisis$Educacion_asiste)[levels(Enc2019_SeccionAnalisis$Educacion_asiste)=="...otro tipo de formación no formal (especifique)"] <- "...otro tipo de formación no formal"

ggplot(Enc2019_SeccionAnalisis, aes(x = Educacion_asiste, fill=Educacion_asiste)) + geom_bar() + geom_text(aes(y=..count.., label=..count..,colour = factor(Educacion_asiste)),
            stat="count", color="black",
            hjust=1.0, size=3) + theme(legend.position="none") + coord_flip() + xlab("Educación Asiste") +
      ylab("Cantidad de Observaciones") + ggtitle("Individuos según la asistencia de la Educación")

#########################################################
```

```
ggplot(Encuesta_2019, aes(x = Idioma, fill=Idioma)) + geom_bar() + geom_text(aes(y=..count.., label=..count..),
            stat="count", color="black",
            hjust=1.0, size=3) + theme(legend.position="none") + xlab("Segundo Idioma") +
      ylab("Cantidad de Observaciones") + ggtitle("Individuos según la capacidad de hablar un segundo idioma") #+ scale_fill_manual(values = c("royalblue", "deepskyblue4", "skyblue"), limits = c("Sí", "No", "Ignorado"), breaks =c("Sí", "No", "Ignorado"), name = "Genero", labels = c("Sí", "No", "Ignorado"))

#########################################################

### Generación de la visualización de la Zona Geografica.

ggplot(Encuesta_2019, aes(x = Zona, fill=Zona)) + geom_bar() + geom_text(aes(y=..count.., label=..count..),
            stat="count", color="black",
            hjust=0.5, size=4) + theme(legend.position="none") + xlab("Zona") +
      ylab("Cantidad de Observaciones") + ggtitle("Individuos según la Zona Geográfica de Costa Rica") + scale_fill_manual(values = c("bisque1", "bisque4"), limits = c("Urbana", "Rural"), breaks =c("Urbana", "Rural"), name = "Zona", labels = c("Urbana", "Rural"))

#########################################################

### Generación de graficos por Región Geográfica

labels.regiones <- c("Central", "Chorotega", "Pacífico central", "Brunca", "Huetar caribe", "Huetar norte")
```

```
ggplot(Encuesta_2019, aes(x = Region, fill=Region)) + geom_bar() + geom_text(aes(y=..count.., label=..count..,colour = factor(Region)),
                stat="count", color="black",
                hjust=0.5, size=3) + theme(legend.position="none") + xlab("Región Geográfica") +
          ylab("Cantidad de Observaciones") + ggtitle("Individuos según la Región Geográfica") + scale_fill_manual(values = c("cadetblue", "cadetblue1", "cadetblue2", "cadetblue3", "cadetblue4", "aquamarine4"), limits = labels.regiones, breaks = labels.regiones, name = "Regiones", labels = labels.regiones)

##########################################################

### Generación de la provincia - Provincia_nacimiento

Encuesta_2019.provincia <- dplyr::count(Encuesta_2019, Provincia_nacimiento)

Encuesta_2019.provincia <- dplyr::rename(Encuesta_2019.provincia, value = n)

ggplot(Encuesta_2019.provincia, aes(y = Provincia_nacimiento, x = value, fill=Provincia_nacimiento)) + geom_dotplot(binaxis = "y", stackdir = "center") + xlab("Cantidad de Observaciones") + ylab("Provincia") + ggtitle("Individuos según la Provincia de Nacimiento") + geom_text(aes(x= value, label= value),
                color="black", hjust=1.5, size=3)

##########################################################

### Obserciones graficas del Nivel_educativo

Encuesta_2019.Nivel_Educativo <- dplyr::count(Enc2019_SeccionAnalisis, Nivel_educativo)
```

```
Encuesta_2019.Nivel_Educativo <- dplyr::rename(Encuesta_2019.Nivel_Educativo, value = n)

ggplot(Encuesta_2019.Nivel_Educativo, aes(y = Nivel_educativo, x = value, fill=Nivel_educativo)) + geom_dotplot(binaxis = "y", stackdir = "center") + xlab("Cantidad de Observaciones") + ylab("Nivel Educativo") + ggtitle("Individuos según el Grado Educativo Alcanzado") + geom_text(aes(x= value, label= value),
                color="black", hjust=1.5, size=3)

########################################################## Comparativo de Trabajo con Nivel Educativo

mosaicplot(Encuesta_2019$Trabajo ~ Encuesta_2019$Nivel_educativo,
        main = 'Proporción de Trabajo y Nivel Educativo',
        ylab = 'Nivel Educativo',
        xlab = 'Trabaja?',
        col = c('navy', 'royalblue'),
        las = 1)
        ###set_labels=list(Nivel_educativo = c("1", "2", "3", "4M", "5F", "6", "7M", "8F", "9")))
########################################################## Comparativo de Trabajo con el Idioma

mosaicplot(Encuesta_2019$Trabajo ~ Encuesta_2019$Idioma,
        main = 'Proporción de Trabajo y Nivel Educativo',
        ylab = 'Nivel Educativo',
        xlab = 'Trabaja?',
        col = c('navy', 'royalblue'),
        las = 1)
        ###set_labels=list(Nivel_educativo = c("1", "2", "3", "4M", "5F", "6", "7M", "8F", "9")))
```

```
```

```{r}
# Limpieza de los datos
Enc2019_SeccionAnalisis <- Enc2019_SeccionAnalisis[!is.na(Enc2019_SeccionAnalisis$Educacion_asiste), ]
#View(data_is.na)

Enc2019_SeccionAnalisis <- Enc2019_SeccionAnalisis[!is.na(Enc2019_SeccionAnalisis$Nivel_educativo), ]

###### Se crea un subset con la función select de dplyr
Enc2019_SeccionAnalisis <- select(Enc2019_SeccionAnalisis, Relacion_parentesco, Sexo, Edad, Estado_conyugal, Lugar_nacimiento, Permanencia_pais, Permanencia_intension, Permanencia_motivo, Seguro, Tipo_seguro, Regimen_pension, Plan_voluntario, Educacion_asiste, Educacion_titulo, EducacionNoregular_asiste, Idioma, Idioma_cual, Trabajo, Region, Zona, Poblacion_joven, Poblacion_adulto, Provincia_nacimiento, Pais_nacimiento, Nivel_educativo)

Enc2019_SeccionAnalisis$Tipo_Poblacion <- coalesce(Enc2019_SeccionAnalisis$Poblacion_adulto, Enc2019_SeccionAnalisis$Poblacion_joven)

#summary(Enc2019_SeccionAnalisis$Relacion_parentesco)
#str(Enc2019_SeccionAnalisis$Relacion_parentesco)

Enc2019_SeccionAnalisis <- select(Enc2019_SeccionAnalisis, Relacion_parentesco, Sexo, Edad, Estado_conyugal, Lugar_nacimiento, Permanencia_pais, Permanencia_intension, Permanencia_motivo, Seguro, Tipo_seguro,
```

```
Regimen_pension, Plan_voluntario, Educacion_asiste, Educacion_titulo, EducacionNoregular_asiste, Idioma, Idioma_cual, Trabajo, Region, Zona, Provincia_nacimiento, Pais_nacimiento, Nivel_educativo, Tipo_Poblacion)

##### Quitar los niveles de las variables que se refieren a Ignorado.

Enc2019_SeccionAnalisis$Trabajo <- droplevels(Enc2019_SeccionAnalisis$Trabajo)
Enc2019_SeccionAnalisis$Estado_conyugal <- droplevels(Enc2019_SeccionAnalisis$Estado_conyugal)
Enc2019_SeccionAnalisis$Lugar_nacimiento <- droplevels(Enc2019_SeccionAnalisis$Lugar_nacimiento)
Enc2019_SeccionAnalisis$Permanencia_pais <- droplevels(Enc2019_SeccionAnalisis$Permanencia_pais)
Enc2019_SeccionAnalisis$Permanencia_intension <- droplevels(Enc2019_SeccionAnalisis$Permanencia_intension)
Enc2019_SeccionAnalisis$Permanencia_motivo <- droplevels(Enc2019_SeccionAnalisis$Permanencia_motivo)
Enc2019_SeccionAnalisis$Seguro <- droplevels(Enc2019_SeccionAnalisis$Seguro)
Enc2019_SeccionAnalisis$Tipo_seguro <- droplevels(Enc2019_SeccionAnalisis$Tipo_seguro)
Enc2019_SeccionAnalisis$Plan_voluntario <- droplevels(Enc2019_SeccionAnalisis$Plan_voluntario)
Enc2019_SeccionAnalisis <- Enc2019_SeccionAnalisis %>% filter(Educacion_asiste != "Ignorado")
Enc2019_SeccionAnalisis$Educacion_asiste <- droplevels(Enc2019_SeccionAnalisis$Educacion_asiste)
Enc2019_SeccionAnalisis$EducacionNoregular_asiste <- droplevels(Enc2019_SeccionAnalisis$EducacionNoregular_asiste)
Enc2019_SeccionAnalisis$Idioma <- droplevels(Enc2019_SeccionAnalisis$Idioma)
Enc2019_SeccionAnalisis$Idioma_cual <- droplevels(Enc2019_SeccionAnalisis$Idioma_cual)
```

```
# Conversión de los datos a numeros y evitar la confusión del modelo.

## Relación_Parentesco

levels(Enc2019_SeccionAnalisis$Relacion_parentesco)[levels(Enc2019_SeccionAnalisis$Relacion_parentesco)=="Jefe(a)"] <- "1"
levels(Enc2019_SeccionAnalisis$Relacion_parentesco)[levels(Enc2019_SeccionAnalisis$Relacion_parentesco)=="Esposo(a) o compañero(a)"] <- "2"
levels(Enc2019_SeccionAnalisis$Relacion_parentesco)[levels(Enc2019_SeccionAnalisis$Relacion_parentesco)=="Hijo(a)"] <- "3"
levels(Enc2019_SeccionAnalisis$Relacion_parentesco)[levels(Enc2019_SeccionAnalisis$Relacion_parentesco)=="Yerno o nuera"] <- "4"
levels(Enc2019_SeccionAnalisis$Relacion_parentesco)[levels(Enc2019_SeccionAnalisis$Relacion_parentesco)=="Nieto(a)"] <- "5"
levels(Enc2019_SeccionAnalisis$Relacion_parentesco)[levels(Enc2019_SeccionAnalisis$Relacion_parentesco)=="Padre o Madre"] <- "6"
levels(Enc2019_SeccionAnalisis$Relacion_parentesco)[levels(Enc2019_SeccionAnalisis$Relacion_parentesco)=="Hermano(a)"] <- "7"
levels(Enc2019_SeccionAnalisis$Relacion_parentesco)[levels(Enc2019_SeccionAnalisis$Relacion_parentesco)=="Cuñado(a)"] <- "8"
levels(Enc2019_SeccionAnalisis$Relacion_parentesco)[levels(Enc2019_SeccionAnalisis$Relacion_parentesco)=="Otro familiar"] <- "9"
levels(Enc2019_SeccionAnalisis$Relacion_parentesco)[levels(Enc2019_SeccionAnalisis$Relacion_parentesco)=="Otro no familiar"] <- "10"
levels(Enc2019_SeccionAnalisis$Relacion_parentesco)[levels(Enc2019_SeccionAnalisis$Relacion_parentesco)=="Servicio doméstico sin otro lugar donde vivir o su familiar"] <- "11"
levels(Enc2019_SeccionAnalisis$Relacion_parentesco)[levels(Enc2019_SeccionAnalisis$Relacion_parentesco)=="Pensionista sin otro lugar donde vivir o su familiar"] <- "12"
```

```
levels(Enc2019_SeccionAnalisis$Relacion_parentesco)[levels(Enc2019_SeccionAnal
isis$Relacion_parentesco)=="Nuevo jefe(a)"] <- "13"
levels(Enc2019_SeccionAnalisis$Relacion_parentesco)[levels(Enc2019_SeccionAnal
isis$Relacion_parentesco)=="Hijastro(a)"] <- "14"
levels(Enc2019_SeccionAnalisis$Relacion_parentesco)[levels(Enc2019_SeccionAnal
isis$Relacion_parentesco)=="Suegro(a)"] <- "15"
levels(Enc2019_SeccionAnalisis$Relacion_parentesco)[levels(Enc2019_SeccionAnal
isis$Relacion_parentesco)=="Unión libre con personas del mismo sexo"] <- "16"
levels(Enc2019_SeccionAnalisis$Relacion_parentesco)[levels(Enc2019_SeccionAnal
isis$Relacion_parentesco)=="Nuevo conyugue"] <- "17"
levels(Enc2019_SeccionAnalisis$Relacion_parentesco)[levels(Enc2019_SeccionAnal
isis$Relacion_parentesco)=="Nuevo jefe(a) o conyugue"] <- "18"
levels(Enc2019_SeccionAnalisis$Relacion_parentesco)[levels(Enc2019_SeccionAnal
isis$Relacion_parentesco)=="Nuevo cónyugue"] <- "19"
levels(Enc2019_SeccionAnalisis$Relacion_parentesco)[levels(Enc2019_SeccionAnal
isis$Relacion_parentesco)=="Nuevo jefe(a) o cónyugue"] <- "20"

## Sexo

#summary(Enc2019_SeccionAnalisis$Sexo)
#str(Enc2019_SeccionAnalisis$Sexo)

levels(Enc2019_SeccionAnalisis$Sexo)[levels(Enc2019_SeccionAnalisis$Sexo)=="H
ombre"] <- "1"
levels(Enc2019_SeccionAnalisis$Sexo)[levels(Enc2019_SeccionAnalisis$Sexo)=="M
ujer"] <- "0"

## Estado_conyugal

#summary(Enc2019_SeccionAnalisis$Estado_conyugal)
#str(Enc2019_SeccionAnalisis$Estado_conyugal)
```

```
levels(Enc2019_SeccionAnalisis$Estado_conyugal)[levels(Enc2019_SeccionAnalisis
$Estado_conyugal)=="...en unión libre o juntado?"] <- "1"
levels(Enc2019_SeccionAnalisis$Estado_conyugal)[levels(Enc2019_SeccionAnalisis
$Estado_conyugal)=="...casado(a)?"] <- "2"
levels(Enc2019_SeccionAnalisis$Estado_conyugal)[levels(Enc2019_SeccionAnalisis
$Estado_conyugal)=="...divorciado(a)?"] <- "3"
levels(Enc2019_SeccionAnalisis$Estado_conyugal)[levels(Enc2019_SeccionAnalisis
$Estado_conyugal)=="...separado(a)?"] <- "4"
levels(Enc2019_SeccionAnalisis$Estado_conyugal)[levels(Enc2019_SeccionAnalisis
$Estado_conyugal)=="...viudo(a)?"] <- "5"
levels(Enc2019_SeccionAnalisis$Estado_conyugal)[levels(Enc2019_SeccionAnalisis
$Estado_conyugal)=="...soltero(a)?"] <- "6"
levels(Enc2019_SeccionAnalisis$Estado_conyugal)[levels(Enc2019_SeccionAnalisis
$Estado_conyugal)=="...casado(a) con cónyuge no residente?"] <- "7"
levels(Enc2019_SeccionAnalisis$Estado_conyugal)[levels(Enc2019_SeccionAnalisis
$Estado_conyugal)=="... unión libre con compañero(a) del mismo sexo?"] <- "8"

## Lugar_nacimiento

#summary(Enc2019_SeccionAnalisis$Lugar_nacimiento)
#str(Enc2019_SeccionAnalisis$Lugar_nacimiento)

levels(Enc2019_SeccionAnalisis$Lugar_nacimiento)[levels(Enc2019_SeccionAnalisis
$Lugar_nacimiento)=="En este mismo cantón"] <- "1"
levels(Enc2019_SeccionAnalisis$Lugar_nacimiento)[levels(Enc2019_SeccionAnalisis
$Lugar_nacimiento)=="En otro cantón"] <- "2"
levels(Enc2019_SeccionAnalisis$Lugar_nacimiento)[levels(Enc2019_SeccionAnalisis
$Lugar_nacimiento)=="En otro país"] <- "3"

## Permanencia_pais
```

```
#summary(Enc2019_SeccionAnalisis$Permanencia_pais)
#str(Enc2019_SeccionAnalisis$Permanencia_pais)

levels(Enc2019_SeccionAnalisis$Permanencia_pais)[levels(Enc2019_SeccionAnalisis$Permanencia_pais)=="Menos de un año"] <- "1"
levels(Enc2019_SeccionAnalisis$Permanencia_pais)[levels(Enc2019_SeccionAnalisis$Permanencia_pais)=="Un año o más"] <- "2"

## Permanencia_intension

#summary(Enc2019_SeccionAnalisis$Permanencia_intension)
#str(Enc2019_SeccionAnalisis$Permanencia_intension)

levels(Enc2019_SeccionAnalisis$Permanencia_intension)[levels(Enc2019_SeccionAnalisis$Permanencia_intension)=="...Costa Rica?"] <- "1"
levels(Enc2019_SeccionAnalisis$Permanencia_intension)[levels(Enc2019_SeccionAnalisis$Permanencia_intension)=="...otro país?"] <- "2"

## Permanencia_motivo

#summary(Enc2019_SeccionAnalisis$Permanencia_motivo)
#str(Enc2019_SeccionAnalisis$Permanencia_motivo)

levels(Enc2019_SeccionAnalisis$Permanencia_motivo)[levels(Enc2019_SeccionAnalisis$Permanencia_motivo)=="Trabajo"] <- "1"
levels(Enc2019_SeccionAnalisis$Permanencia_motivo)[levels(Enc2019_SeccionAnalisis$Permanencia_motivo)=="Otro"] <- "2"

## Seguro
```

```
#summary(Enc2019_SeccionAnalisis$Seguro)
#str(Enc2019_SeccionAnalisis$Seguro)

levels(Enc2019_SeccionAnalisis$Seguro)[levels(Enc2019_SeccionAnalisis$Seguro)==
="Sí"] <- "1"
levels(Enc2019_SeccionAnalisis$Seguro)[levels(Enc2019_SeccionAnalisis$Seguro)==
="No"] <- "0"

## Tipo_seguro

#summary(Enc2019_SeccionAnalisis$Tipo_seguro)
#str(Enc2019_SeccionAnalisis$Tipo_seguro)

levels(Enc2019_SeccionAnalisis$Tipo_seguro)[levels(Enc2019_SeccionAnalisis$Tipo
_seguro)=="...asalariado?"] <- "1"
levels(Enc2019_SeccionAnalisis$Tipo_seguro)[levels(Enc2019_SeccionAnalisis$Tipo
_seguro)=="...mediante convenio como asociaciones, sindicatos, cooperativas, etc?"]
<- "2"
levels(Enc2019_SeccionAnalisis$Tipo_seguro)[levels(Enc2019_SeccionAnalisis$Tipo
_seguro)=="...cuenta propia o voluntario?"] <- "3"
levels(Enc2019_SeccionAnalisis$Tipo_seguro)[levels(Enc2019_SeccionAnalisis$Tipo
_seguro)=="...pensionado de la CCSS, Magisterio, u otro?"] <- "4"
levels(Enc2019_SeccionAnalisis$Tipo_seguro)[levels(Enc2019_SeccionAnalisis$Tipo
_seguro)=="...familiar de asegurado directo o pensionado?"] <- "5"
levels(Enc2019_SeccionAnalisis$Tipo_seguro)[levels(Enc2019_SeccionAnalisis$Tipo
_seguro)=="..asegurado por el Estado, incluye familiar de asegurado por el Estado?"]
<- "6"
levels(Enc2019_SeccionAnalisis$Tipo_seguro)[levels(Enc2019_SeccionAnalisis$Tipo
_seguro)=="..pensionado del régimen no contributivo monto básico, gracia o guerra?"]
<- "7"
```

```
levels(Enc2019_SeccionAnalisis$Tipo_seguro)[levels(Enc2019_SeccionAnalisis$Tipo_seguro)==''..seguro privado o del extranjero?''] <- ''8''
levels(Enc2019_SeccionAnalisis$Tipo_seguro)[levels(Enc2019_SeccionAnalisis$Tipo_seguro)==''..otras formas como seguro de estudiante, de refugiado y otros?''] <- ''9''

## Regimen_pension

#summary(Enc2019_SeccionAnalisis$Regimen_pension)
#str(Enc2019_SeccionAnalisis$Regimen_pension)

levels(Enc2019_SeccionAnalisis$Regimen_pension)[levels(Enc2019_SeccionAnalisis$Regimen_pension)==''Ninguno''] <- ''1''
levels(Enc2019_SeccionAnalisis$Regimen_pension)[levels(Enc2019_SeccionAnalisis$Regimen_pension)==''Régimen de IVM de la CCSS?''] <- ''2''
levels(Enc2019_SeccionAnalisis$Regimen_pension)[levels(Enc2019_SeccionAnalisis$Regimen_pension)==''Otro régimen (Magisterio, Poder Judicial, Hacienda, etc.)?''] <- ''3''

## Plan_voluntario

#summary(Enc2019_SeccionAnalisis$Plan_voluntario)
#str(Enc2019_SeccionAnalisis$Plan_voluntario)

levels(Enc2019_SeccionAnalisis$Plan_voluntario)[levels(Enc2019_SeccionAnalisis$Plan_voluntario)==''Sí''] <- ''1''
levels(Enc2019_SeccionAnalisis$Plan_voluntario)[levels(Enc2019_SeccionAnalisis$Plan_voluntario)==''No''] <- ''0''

## Educacion_asiste

#summary(Enc2019_SeccionAnalisis$Educacion_asiste)
```

```
#str(Enc2019_SeccionAnalisis$Educacion_asiste)

levels(Enc2019_SeccionAnalisis$Educacion_asiste)[levels(Enc2019_SeccionAnalisis
$Educacion_asiste)=="...escuela?"] <- "1"
levels(Enc2019_SeccionAnalisis$Educacion_asiste)[levels(Enc2019_SeccionAnalisis
$Educacion_asiste)=="...colegio?"] <- "2"
levels(Enc2019_SeccionAnalisis$Educacion_asiste)[levels(Enc2019_SeccionAnalisis
$Educacion_asiste)=="...parauniversitaria o universitaria?"] <- "3"
levels(Enc2019_SeccionAnalisis$Educacion_asiste)[levels(Enc2019_SeccionAnalisis
$Educacion_asiste)=="...enseñanza especial?"] <- "4"
levels(Enc2019_SeccionAnalisis$Educacion_asiste)[levels(Enc2019_SeccionAnalisis
$Educacion_asiste)=="...educación abierta en instituciones para presentar examenes
ante el MEP)?"] <- "5"
levels(Enc2019_SeccionAnalisis$Educacion_asiste)[levels(Enc2019_SeccionAnalisis
$Educacion_asiste)=="...otro tipo de formación no formal (especifique)"] <- "6"
levels(Enc2019_SeccionAnalisis$Educacion_asiste)[levels(Enc2019_SeccionAnalisis
$Educacion_asiste)=="...No asiste"] <- "7"

## Educacion_titulo

#summary(Enc2019_SeccionAnalisis$Educacion_titulo)
#str(Enc2019_SeccionAnalisis$Educacion_titulo)

levels(Enc2019_SeccionAnalisis$Educacion_titulo)[levels(Enc2019_SeccionAnalisis$
Educacion_titulo)=="No tiene título"] <- "1"
levels(Enc2019_SeccionAnalisis$Educacion_titulo)[levels(Enc2019_SeccionAnalisis$
Educacion_titulo)=="Técnico, perito no universitario"] <- "2"
levels(Enc2019_SeccionAnalisis$Educacion_titulo)[levels(Enc2019_SeccionAnalisis$
Educacion_titulo)=="Profesorado, diplomado o técnico universitario"] <- "3"
levels(Enc2019_SeccionAnalisis$Educacion_titulo)[levels(Enc2019_SeccionAnalisis$
Educacion_titulo)=="Bachillerato"] <- "4"
```

```
levels(Enc2019_SeccionAnalisis$Educacion_titulo)[levels(Enc2019_SeccionAnalisis$
Educacion_titulo)=="Licenciatura"] <- "5"
levels(Enc2019_SeccionAnalisis$Educacion_titulo)[levels(Enc2019_SeccionAnalisis$
Educacion_titulo)=="Especialización"] <- "6"
levels(Enc2019_SeccionAnalisis$Educacion_titulo)[levels(Enc2019_SeccionAnalisis$
Educacion_titulo)=="Maestría y doctorado"] <- "7"
levels(Enc2019_SeccionAnalisis$Educacion_titulo)[levels(Enc2019_SeccionAnalisis$
Educacion_titulo)=="Maestria"] <- "8"
levels(Enc2019_SeccionAnalisis$Educacion_titulo)[levels(Enc2019_SeccionAnalisis$
Educacion_titulo)=="Doctorado"] <- "9"
levels(Enc2019_SeccionAnalisis$Educacion_titulo)[levels(Enc2019_SeccionAnalisis$
Educacion_titulo)=="No especificado"] <- "10"

## EducacionNoregular_asiste

#summary(Enc2019_SeccionAnalisis$EducacionNoregular_asiste)
#str(Enc2019_SeccionAnalisis$EducacionNoregular_asiste)

levels(Enc2019_SeccionAnalisis$EducacionNoregular_asiste)[levels(Enc2019_Secci
onAnalisis$EducacionNoregular_asiste)=="Sí"] <- "1"
levels(Enc2019_SeccionAnalisis$EducacionNoregular_asiste)[levels(Enc2019_Secci
onAnalisis$EducacionNoregular_asiste)=="No"] <- "0"

## Idioma

#summary(Enc2019_SeccionAnalisis$Idioma)
#str(Enc2019_SeccionAnalisis$Idioma)

levels(Enc2019_SeccionAnalisis$Idioma)[levels(Enc2019_SeccionAnalisis$Idioma)==
"Sí"] <- "1"
```

```
levels(Enc2019_SeccionAnalisis$Idioma)[levels(Enc2019_SeccionAnalisis$Idioma)==
"No"] <- "0"

## Idioma_cual

#summary(Enc2019_SeccionAnalisis$Idioma_cual)
#str(Enc2019_SeccionAnalisis$Idioma_cual)

levels(Enc2019_SeccionAnalisis$Idioma_cual)[levels(Enc2019_SeccionAnalisis$Idio
ma_cual)=="...inglés?"] <- "1"
levels(Enc2019_SeccionAnalisis$Idioma_cual)[levels(Enc2019_SeccionAnalisis$Idio
ma_cual)=="...francés?"] <- "2"
levels(Enc2019_SeccionAnalisis$Idioma_cual)[levels(Enc2019_SeccionAnalisis$Idio
ma_cual)=="...alemán?"] <- "3"
levels(Enc2019_SeccionAnalisis$Idioma_cual)[levels(Enc2019_SeccionAnalisis$Idio
ma_cual)=="...otro?"] <- "4"
levels(Enc2019_SeccionAnalisis$Idioma_cual)[levels(Enc2019_SeccionAnalisis$Idio
ma_cual)=="...español?"] <- "5"

## Trabajo

#summary(Enc2019_SeccionAnalisis$Trabajo)
#str(Enc2019_SeccionAnalisis$Trabajo)

levels(Enc2019_SeccionAnalisis$Trabajo)[levels(Enc2019_SeccionAnalisis$Trabajo)
=="Sí"] <- "1"
levels(Enc2019_SeccionAnalisis$Trabajo)[levels(Enc2019_SeccionAnalisis$Trabajo)
=="No"] <- "0"

## Region
```

```
#summary(Enc2019_SeccionAnalisis$Region)
#str(Enc2019_SeccionAnalisis$Region)

levels(Enc2019_SeccionAnalisis$Region)[levels(Enc2019_SeccionAnalisis$Region)=
="Central"] <- "1"
levels(Enc2019_SeccionAnalisis$Region)[levels(Enc2019_SeccionAnalisis$Region)=
="Chorotega"] <- "2"
levels(Enc2019_SeccionAnalisis$Region)[levels(Enc2019_SeccionAnalisis$Region)=
="Pacífico central"] <- "3"
levels(Enc2019_SeccionAnalisis$Region)[levels(Enc2019_SeccionAnalisis$Region)=
="Brunca"] <- "4"
levels(Enc2019_SeccionAnalisis$Region)[levels(Enc2019_SeccionAnalisis$Region)=
="Huetar caribe"] <- "5"
levels(Enc2019_SeccionAnalisis$Region)[levels(Enc2019_SeccionAnalisis$Region)=
="Huetar norte"] <- "6"

## Zona

#summary(Enc2019_SeccionAnalisis$Zona)
#str(Enc2019_SeccionAnalisis$Zona)

levels(Enc2019_SeccionAnalisis$Zona)[levels(Enc2019_SeccionAnalisis$Zona)=="Ur
bana"] <- "1"
levels(Enc2019_SeccionAnalisis$Zona)[levels(Enc2019_SeccionAnalisis$Zona)=="R
ural"] <- "2"

## Provincia Nacimiento

#summary(Enc2019_SeccionAnalisis$Provincia_nacimiento)
#str(Enc2019_SeccionAnalisis$Provincia_nacimiento)
```

```
levels(Enc2019_SeccionAnalisis$Provincia_nacimiento)[levels(Enc2019_SeccionAnal
isis$Provincia_nacimiento)=="San José"] <- "1"
levels(Enc2019_SeccionAnalisis$Provincia_nacimiento)[levels(Enc2019_SeccionAnal
isis$Provincia_nacimiento)=="Alajuela"] <- "2"
levels(Enc2019_SeccionAnalisis$Provincia_nacimiento)[levels(Enc2019_SeccionAnal
isis$Provincia_nacimiento)=="Cartago"] <- "3"
levels(Enc2019_SeccionAnalisis$Provincia_nacimiento)[levels(Enc2019_SeccionAnal
isis$Provincia_nacimiento)=="Heredia"] <- "4"
levels(Enc2019_SeccionAnalisis$Provincia_nacimiento)[levels(Enc2019_SeccionAnal
isis$Provincia_nacimiento)=="Guanacaste"] <- "5"
levels(Enc2019_SeccionAnalisis$Provincia_nacimiento)[levels(Enc2019_SeccionAnal
isis$Provincia_nacimiento)=="Puntarenas"] <- "6"
levels(Enc2019_SeccionAnalisis$Provincia_nacimiento)[levels(Enc2019_SeccionAnal
isis$Provincia_nacimiento)=="Limón"] <- "7"
levels(Enc2019_SeccionAnalisis$Provincia_nacimiento)[levels(Enc2019_SeccionAnal
isis$Provincia_nacimiento)=="No especificado"] <- "8"

## Provincia Nacimiento

#summary(Enc2019_SeccionAnalisis$Pais_nacimiento)
#str(Enc2019_SeccionAnalisis$Pais_nacimiento)

levels(Enc2019_SeccionAnalisis$Pais_nacimiento)[levels(Enc2019_SeccionAnalisis$
Pais_nacimiento)=="Costa Rica"] <- "1"
levels(Enc2019_SeccionAnalisis$Pais_nacimiento)[levels(Enc2019_SeccionAnalisis$
Pais_nacimiento)=="El Salvador"] <- "2"
levels(Enc2019_SeccionAnalisis$Pais_nacimiento)[levels(Enc2019_SeccionAnalisis$
Pais_nacimiento)=="Honduras"] <- "3"
levels(Enc2019_SeccionAnalisis$Pais_nacimiento)[levels(Enc2019_SeccionAnalisis$
Pais_nacimiento)=="Nicaragua"] <- "4"
```

```
levels(Enc2019_SeccionAnalisis$Pais_nacimiento)[levels(Enc2019_SeccionAnalisis$
Pais_nacimiento)=="Panama"] <- "5"
levels(Enc2019_SeccionAnalisis$Pais_nacimiento)[levels(Enc2019_SeccionAnalisis$
Pais_nacimiento)=="Colombia"] <- "6"
levels(Enc2019_SeccionAnalisis$Pais_nacimiento)[levels(Enc2019_SeccionAnalisis$
Pais_nacimiento)=="Estados Unidos"] <- "7"
levels(Enc2019_SeccionAnalisis$Pais_nacimiento)[levels(Enc2019_SeccionAnalisis$
Pais_nacimiento)=="México"] <- "8"
levels(Enc2019_SeccionAnalisis$Pais_nacimiento)[levels(Enc2019_SeccionAnalisis$
Pais_nacimiento)=="Venezuela"] <- "9"
levels(Enc2019_SeccionAnalisis$Pais_nacimiento)[levels(Enc2019_SeccionAnalisis$
Pais_nacimiento)=="Otro país"] <- "10"
levels(Enc2019_SeccionAnalisis$Pais_nacimiento)[levels(Enc2019_SeccionAnalisis$
Pais_nacimiento)=="No especificado"] <- "11"

## Provincia Nacimiento

#summary(Enc2019_SeccionAnalisis$Nivel_educativo)
#str(Enc2019_SeccionAnalisis$Nivel_educativo)

levels(Enc2019_SeccionAnalisis$Nivel_educativo)[levels(Enc2019_SeccionAnalisis$
Nivel_educativo)=="Ninguno"] <- "1"
levels(Enc2019_SeccionAnalisis$Nivel_educativo)[levels(Enc2019_SeccionAnalisis$
Nivel_educativo)=="Primaria incompleta"] <- "2"
levels(Enc2019_SeccionAnalisis$Nivel_educativo)[levels(Enc2019_SeccionAnalisis$
Nivel_educativo)=="Primaria completa"] <- "3"
levels(Enc2019_SeccionAnalisis$Nivel_educativo)[levels(Enc2019_SeccionAnalisis$
Nivel_educativo)=="Secundaria incompleta"] <- "4"
levels(Enc2019_SeccionAnalisis$Nivel_educativo)[levels(Enc2019_SeccionAnalisis$
Nivel_educativo)=="Secundaria completa"] <- "5"
```

```
levels(Enc2019_SeccionAnalisis$Nivel_educativo)[levels(Enc2019_SeccionAnalisis$
Nivel_educativo)=="Universitario sin título"] <- "6"
levels(Enc2019_SeccionAnalisis$Nivel_educativo)[levels(Enc2019_SeccionAnalisis$
Nivel_educativo)=="Universitario con título"] <- "7"
levels(Enc2019_SeccionAnalisis$Nivel_educativo)[levels(Enc2019_SeccionAnalisis$
Nivel_educativo)=="No especificado"] <- "8"

## Tipo_Poblacion

#summary(Enc2019_SeccionAnalisis$Tipo_Poblacion)
#str(Enc2019_SeccionAnalisis$Tipo_Poblacion)

levels(Enc2019_SeccionAnalisis$Tipo_Poblacion)[levels(Enc2019_SeccionAnalisis$T
ipo_Poblacion)=="Población adulta"] <- "1"
levels(Enc2019_SeccionAnalisis$Tipo_Poblacion)[levels(Enc2019_SeccionAnalisis$T
ipo_Poblacion)=="Población joven"] <- "2"
####

```
```

Código fuente para el modelado

```
##### Modelado de los datos #####
#librerias

library(traineR)
library(rpart)
library(rpart.plot)
library(ROCR)
library(plotROC)
library(dplyr)
library(dummies)
```
```

```
library(readr)
library(FactoMineR)
library(factoextra)
library(rattle)
library(randomForest)

### Conversión de las variables Factor a Dummy, con excepción de la variable
respuesta.
set.seed(3427)
#datos <- Enc2019_SeccionAnalisis[, c(1:2, 4:5, 9:10, 12:13, 15:16, 19:23)]
datos <- Enc2019_SeccionAnalisis[, c(1:2, 4:5, 9:10, 12:16, 19:23)] ## Se agrega la
variable Educacion_Titulo
datos2 <- dummy.data.frame(datos, sep = ".")
datos2$Trabajo <- Enc2019_SeccionAnalisis[, "Trabajo"]#datos[, "Trabajo"]

#str(datos2)

#Enc2019_SeccionAnalisis_Dummy <-
dummy.data.frame(Enc2019_SeccionAnalisis[, c(1:17, 19:25)], sep = ".")

muestra <- sample(1:nrow(datos2),floor(nrow(datos2)*0.30))
ttesting <- datos2[muestra,]
taprendizaje <- datos2[-muestra,]

#str(Enc2019_SeccionAnalisis_Dummy)

#Creando testing y training, se elige al 30%

#Enc2019_SeccionAnalisis_Muestra <-
sample(1:nrow(Enc2019_SeccionAnalisis),floor(nrow(Enc2019_SeccionAnalisis) *
0.30))

#Enc2019_SeccionAnalisis_Testing <-
Enc2019_SeccionAnalisis[Enc2019_SeccionAnalisis_Muestra,]

#Enc2019_SeccionAnalisis_Aprendizaje <- Enc2019_SeccionAnalisis[-
Enc2019_SeccionAnalisis_Muestra,]


#grafico para ver como se comparta la prediccion

plotROC <- function(prediccion, real, adicionar = FALSE, color = "red") {
  pred <- ROCR::prediction(prediccion, real)
  perf <- ROCR::performance(pred, "tpr", "fpr")
  plot(perf, col = color, add = adicionar, main = "Curva ROC")
  segments(0, 0, 1, 1, col='black')
```

```
  grid()
}

#### Modelos ####

################################################ k-nearest neighbors
algorithm
# train.knn escoje el k usando leave-one-out crossvalidation

modelo<-train.knn(Trabajo~., data=taprendizaje, kmax=9)

summary(modelo)

prediccion <- predict(modelo, ttesting, type = "prob") # Para que me retorne la
probabilidad

summary(prediccion)

Clase1 <- ttesting[, "Trabajo"]
head(Clase1)
Score1 <- prediccion$prediction[,2]
head(Score1)
Corte <- 0.5
Prediccion <- ifelse(Score1 > Corte, "1", "0")
MC1 <- table(Clase1, Pred = factor(Prediccion, levels = c("0", "1")))
general.indexes(mc=MC1)
# Genera el gráfico
plotROC(Score1,Clase1)

################################## Arboles de decision

modelo.rpart <- train.rpart(formula = Trabajo~.,
                    data = taprendizaje,
                    minsplit = 2)

summary(modelo.rpart)

modelo.rpart

# Visualizar gráficamente el modelo #
prp(modelo.rpart)

fancyRpartPlot(modelo.rpart)
#######################################
```

```
prediccion.rpart <- predict(modelo.rpart, ttesting, type = "prob") # Para que me retorne la probabilidad

#Ver las predicciones
head(prediccion.rpart)
summary(prediccion.rpart)

Clase2 <- ttesting[, "Trabajo"]
head(Clase2)
Score2 <- prediccion.rpart$prediction[,2]
head(Score2)
Corte <- 0.5
Prediccion <- ifelse(Score2 > Corte, "0", "1")
MC2 <- table(Clase2, Pred = factor(Prediccion, levels = c("1", "0")))
general.indexes(mc=MC2)
plotROC(Score2,Clase2)

########################################################## Bosques aleatorios

set.seed(3427)
datos <- Enc2019_SeccionAnalisis[, c(1:2, 4:5, 9:10, 12:13, 15:16, 18:23)]# Sin incluir la variable de Educación Titulo
#datos <- Enc2019_SeccionAnalisis[, c(1:2, 4:5, 9:10, 12:16, 18:23)]# Incluyendo la variable Educación Titulo

muestra <- sample(1:nrow(datos),floor(nrow(datos)*0.30))
ttesting <- datos[muestra,]
taprendizaje <- datos[-muestra,]

modelo.RF <- train.randomForest(formula = Trabajo~., data = taprendizaje, importance = T, na.action = na.omit, ntree = 30)

print(modelo.RF)
summary(modelo.RF)
modelo.RF

#getTree(modelo.RF, 1, labelVar=TRUE)

plot(modelo.RF, type = "l", main = "Bosque Aleatorio")

plot(modelo.RF)

varImpPlot(modelo.RF, sort = TRUE, n.var = min(30, nrow(modelo.RF$importance)), main = "Importancia de las Variables")

table(ttesting$Trabajo, prediccion.RF$prediction[, 2] >= 0.5)
```

```
prediccion.RF <- predict(modelo.RF, ttesting, type = "prob", na.action = na.omit) # Para que me retorne la probabilidad.

odds.RF <- predict(modelo.RF, na.omit(ttesting))

plot(odds.RF$prediction, ttesting_NaOmit$Trabajo, xlab = "Prediciones", ylab = "Trabajo?")

length(odds.RF$prediction)
length(ttesting_NaOmit$Trabajo)

summary(prediccion.RF$prediction)
summary(odds.RF$prediction)

Clase3 <- ttesting[, "Trabajo"]
head(Clase3)
Score3 <- prediccion.RF$prediction[,2]
head(Score3)
Corte <- 0.5
Prediccion <- ifelse(Score3 > Corte, "0", "1")
MC3 <- table(Clase3, Pred = factor(Prediccion, levels = c("1", "0")))
general.indexes(mc=MC3)
plotROC(Score3,Clase3)
ttesting$probabilidad <- prediccion$prediction

########################################################## maquina de soporte vectorial

#datos <- Enc2019_SeccionAnalisis[, c(1:2, 4:5, 9:10, 12:13, 15:16, 18:23)]# Sin incluir la variable de Educación Titulo
set.seed(3427)
datos <- Enc2019_SeccionAnalisis[, c(1:2, 4:5, 9:10, 12:16, 18:23)]# Incluyendo la variable Educación Titulo

muestra <- sample(1:nrow(datos),floor(nrow(datos)*0.30))
ttesting <- datos[muestra,]
taprendizaje <- datos[-muestra,]

modelo.msv <- train.svm(Trabajo~.,data = taprendizaje)

prediccion.msv <- predict(modelo.msv, ttesting, type = "prob") # Para que me retorne la probabilidad.

odds.msv <- predict(modelo.msv, ttesting) # Para que me retorne la posibilidad.
```

```
length(Clase4)
length(Score4)
length(Prediccion)
length(na.omit(ttesting))

summary(modelo.msv)

ttesting_NaOmit <- na.omit(ttesting)
Clase4 <- ttesting_NaOmit[, "Trabajo"]

head(Clase4)
Score4 <- prediccion.msv$prediction[,2]
head(Score4)
Corte <- 0.5
Prediccion <- ifelse(Score4 > Corte, "0", "1")
MC4 <- table(Clase4, Pred = factor(Prediccion, levels = c("1", "0")))
general.indexes(mc=MC4)
plotROC(Score4,Clase4)



##################################################################

########################################################## #Impulso
Adaptativo

datos <- Enc2019_SeccionAnalisis[, c(1:2, 4:5, 9:10, 12:13, 15:16, 18:23)]# Sin incluir
la variable de Educación Titulo

#datos <- Enc2019_SeccionAnalisis[, c(1:2, 4:5, 9:10, 12:16, 18:23)]# Incluyendo la
variable Educación Titulo

muestra <- sample(1:nrow(datos),floor(nrow(datos)*0.30))
ttesting <- datos[muestra,]
taprendizaje <- datos[-muestra,]

modelo.ada <- train.ada(formula = Trabajo~., data = taprendizaje)
prediccion.ada <- predict(modelo.ada, ttesting, type = "prob") # Para que me retorne
la probabilidad

summary(modelo.ada)
print(modelo.ada)

varplot(modelo.ada, plot.it = TRUE, type = c("none","scores"),max.var.show=30)
```

```
table(ttesting$Trabajo, prediccion.ada$prediction[, 2] >= 0.5)

Clase5 <- ttesting[, "Trabajo"]
head(Clase5)
Score5 <- prediccion.ada$prediction[,2]
head(Score5)
Corte <- 0.5
Prediccion <- ifelse(Score5 > Corte, "1", "0")
MC5 <- table(Clase5, Pred = factor(Prediccion, levels = c("1", "0")))
general.indexes(mc=MC5)
plotROC(Score5,Clase5)


################################################## bayes
set.seed(3427)
datos <- Enc2019_SeccionAnalisis[, c(1:2, 4:5, 9:10, 12:13, 15:16, 18:23)]# Sin incluir la variable de Educación Titulo

#datos <- Enc2019_SeccionAnalisis[, c(1:2, 4:5, 9:10, 12:16, 18:23)]# Incluyendo la variable Educación Titulo

muestra <- sample(1:nrow(datos),floor(nrow(datos)*0.30))
ttesting <- datos[muestra,]
taprendizaje <- datos[-muestra,]

modelo.bayes <- train.bayes(Trabajo~.,data=taprendizaje)

prediccion.bayes <- predict(modelo.bayes, ttesting, type = "prob") # Para que me retorne la probabilidad

table(ttesting$Trabajo, prediccion.bayes$prediction[, 2] >= 0.5)

summary(modelo.bayes)
print(modelo.bayes)
modelo.bayes

Clase7 <- ttesting[, "Trabajo"]
head(Clase7)
Score7 <- prediccion.bayes$prediction[,2]
head(Score7)
Corte <- 0.5
Prediccion <- ifelse(Score7 > Corte, "1", "0")
MC7 <- table(Clase7, Pred = factor(Prediccion, levels = c("1", "0")))
general.indexes(mc=MC7)
plotROC(Score7,Clase7)
```

```
################################################## redes neuronales

set.seed(3427)
#datos <- Enc2019_SeccionAnalisis[, c(1:2, 4:5, 9:10, 12:13, 15:16, 18:23)]# Sin
incluir la variable de Educación Titulo

datos <- Enc2019_SeccionAnalisis[, c(1:2, 4:5, 9:10, 12:16, 18:23)]# Incluyendo la
variable Educación Titulo

muestra <- sample(1:nrow(datos),floor(nrow(datos)*0.30))
ttesting <- datos[muestra,]
taprendizaje <- datos[-muestra,]

modelo.nnet <- train.nnet(formula = Trabajo~.,data = taprendizaje, size = 1)

prediccion.nnet <- predict(modelo, ttesting, type = "prob", na.action = na.omit) # Para
que me retorne la probabilidad

summary(modelo.nnet)
print(modelo.nnet)

plot(modelo.nnet)

Clase8 <- ttesting[, "Trabajo"]
head(Clase8)
Score8 <- prediccion.nnet$prediction[,2]
head(Score8)
Corte <- 0.5
Prediccion <- ifelse(Score8 > Corte, "1", "0")
MC8 <- table(Clase8, Pred = factor(Prediccion, levels = c("1", "0")))
general.indexes(mc=MC8)
plotROC(Score8,Clase8)

########################################################## redes
neuronales 2
library(neuralnet)
set.seed(3427)
datos <- Enc2019_SeccionAnalisis[, c(1:2, 4:5, 9:10, 12:13, 15:16, 18:23)]# Sin incluir
la variable de Educación Titulo
#datos <- Enc2019_SeccionAnalisis[, c(1:2, 4:5, 9:10, 12:16, 18:23)]# Incluyendo la
variable Educación Titulo

muestra <- sample(1:nrow(datos),floor(nrow(datos)*0.30))
ttesting <- datos[muestra,]
taprendizaje <- datos[-muestra,]
```

```
datos.aprendizaje.red <- model.matrix(~.,
                                      data = taprendizaje)

datos.test.red <- model.matrix(~.,
                               data = ttesting)

colnames(datos.aprendizaje.red) <- make.names(colnames(datos.aprendizaje.red))
colnames(datos.test.red) <- make.names(colnames(datos.test.red))

#Construir MOdelo

modelo.nnet1 <- neuralnet(Trabajo0~.,
                          data = datos.aprendizaje.red, hidden = 1)

predicciones.nnet1 <- compute(modelo.nnet1, datos.test.red)

table(datos.test.red[, 'Trabajo0'], predicciones.nnet1$net.result >= 0.5)

modelo.nnet2 <- neuralnet(Trabajo0~.,
                          data = datos.aprendizaje.red, hidden = 2)

predicciones.nnet2 <- compute(modelo.nnet2, datos.test.red)

table(datos.test.red[, 'Trabajo0'], predicciones.nnet2$net.result >= 0.5)

summary(modelo.nnet1)
summary(modelo.nnet2)

# ver detalles del modelo
print(modelo.nnet2)

neuralnet::gwplot(modelo.nnet2)

plot(modelo.nnet1)

plot(modelo.nnet2)

```{r}
################## Gráficos sobre Predicción de la probabilida de la máquina de
soporte vectorial #################

length(prediccion.msv$prediction)
head(prediccion.msv)
```
```

```
boxplot(prediccion.msv$prediction[, "1"] ~ ttesting_NaOmit$Tipo_seguro, ylab="Prob 1 (Trabajo)", xlab="Distribución del Tipo de Seguro por la Probabilidad de 1 (Trabajo)")

boxplot(prediccion.msv$prediction[, "1"] ~ ttesting_NaOmit$Educacion_titulo, ylab="Prob 1 (Trabajo)", xlab="Distribución de la Educación Titulo por la probabilidad de 1 (Trabajo)")

boxplot(prediccion.msv$prediction[, "1"] ~ ttesting_NaOmit$Estado_conyugal, ylab="Prob 1 (Trabajo)", xlab="Distribución del Estado Conyugal por la probabilidad de 1 (Trabajo)")

boxplot(prediccion.msv$prediction[, "1"] ~ ttesting_NaOmit$Relacion_parentesco, ylab="Prob 1 (Trabajo)", xlab="Distribución de la Relacion_parentesco por la probabilidad de 1 (Trabajo)")

boxplot(prediccion.msv$prediction[, "1"] ~ ttesting_NaOmit$Educacion_asiste, ylab="Prob 1 (Trabajo)", xlab="Distribución de la Educación Asiste por la probabilidad de 1 (Trabajo)")

boxplot(prediccion.msv$prediction[, "1"] ~ ttesting_NaOmit$Nivel_educativo, ylab="Prob 1 (Trabajo)", xlab="Distribución del Nivel Educativo por la probabilidad de 1 (Trabajo)")

boxplot(prediccion.msv$prediction[, "1"] ~ ttesting_NaOmit$Region, ylab="Prob 1 (Trabajo)", xlab="Distribución de la región por la probabilidad de 1 (Trabajo)")

boxplot(prediccion.msv$prediction[, "1"] ~ ttesting_NaOmit$Zona, ylab="Prob 1 (Trabajo)", xlab="Distribución de la Zona por la probabilidad de 1 (Trabajo)")

boxplot(prediccion.msv$prediction[, "1"] ~ ttesting_NaOmit$Sexo, ylab="Prob 1 (Trabajo)", xlab="Distribución de la Zona por la probabilidad de 1 (Trabajo)")

caret::featurePlot(prediccion.msv$prediction[, "1"],
ttesting_NaOmit$Estado_conyugal, plot = "box")

caret::featurePlot(prediccion.msv$prediction[, 2], ttesting_NaOmit$Sexo, plot = "box")

################## Gráficos sobre Predicción de probabilidad de Bosques Aleatorios #################

length(prediccion.RF$prediction)
head(prediccion.RF)

boxplot(prediccion.RF$prediction[, "1"] ~ ttesting$Tipo_seguro, ylab="Prob 1 (Trabajo)", xlab="Distribución del Tipo de Seguro por la Probabilidad de 1 (Trabajo)")
```

```
boxplot(prediccion.RF$prediction[, "1"] ~ ttesting$Estado_conyugal, ylab="Prob 1 (Trabajo)", xlab="Distribución del Estado Conyugal por la probabilidad de 1 (Trabajo)")

boxplot(prediccion.RF$prediction[, "1"] ~ ttesting$Relacion_parentesco, ylab="Prob 1 (Trabajo)", xlab="Distribución de la Relación Parentesco por la probabilidad de 1 (Trabajo)")

boxplot(prediccion.RF$prediction[, "1"] ~ ttesting$Educacion_asiste, ylab="Prob 1 (Trabajo)", xlab="Distribución de la Educación Asiste por la probabilidad de 1 (Trabajo)")

boxplot(prediccion.RF$prediction[, "1"] ~ ttesting$Nivel_educativo, ylab="Prob 1 (Trabajo)", xlab="Distribución del Nivel Educativo por la probabilidad de 1 (Trabajo)")

boxplot(prediccion.RF$prediction[, "1"] ~ ttesting$Region, ylab="Prob 1 (Trabajo)", xlab="Distribución de la región por la probabilidad de 1 (Trabajo)")

boxplot(prediccion.RF$prediction[, "1"] ~ ttesting$Zona, ylab="Prob 1 (Trabajo)", xlab="Distribución de la Zona por la probabilidad de 1 (Trabajo)")

boxplot(prediccion.RF$prediction[, "1"] ~ ttesting$Sexo, ylab="Prob 1 (Trabajo)", xlab="Distribución del Género por la probabilidad de 1 (Trabajo)")

boxplot(prediccion.RF$prediction[, "1"] ~ ttesting$Idioma, ylab="Prob 1 (Trabajo)", xlab="Distribución del Idioma por la probabilidad de 1 (Trabajo)")

boxplot(prediccion.RF$prediction[, "1"] ~ ttesting$Provincia_nacimiento, ylab="Prob 1 (Trabajo)", xlab="Distribución de la Provincia Nacimiento por la probabilidad de 1 (Trabajo)")

boxplot(prediccion.RF$prediction[, "1"] ~ ttesting$Nivel_educativo, ylab="Prob 1 (Trabajo)", xlab="Distribución del Nivel Educativo por la probabilidad de 1 (Trabajo)")

```
```
```

Apéndice B

Diccionario de datos

El diccionario de datos nace con el propósito de evitar confusión en los modelos, en especial porque en algunos niveles existen descripciones muy extensas y eso perjudica la comprensión de los factores. En esta sección se explica la transformación de los datos.

Variable	Valor Original	Valor Asignado
Relación_Parentesco	Jefe(a)	1
	Esposo(a) o compañero(a)	2
	Hijo(a)	3
	Yerno o nuera	4
	Nieto(a)	5
	Padre o Madre	6
	Hermano(a)	7
	Cuñado(a)	8
	Otro familiar	9
	Otro no familiar	10
	Servicio doméstico sin otro lugar donde vivir o su familiar	11
	Pensionista sin otro lugar donde vivir o su familiar	12
	Nuevo jefe(a)	13
	Hijastro(a)	14
	Suegro(a)	15
	Unión libre con personas del mismo sexo	16
	Nuevo conyugue	17
	Nuevo jefe(a) o conyugue	18
	Nuevo cónyugue	19
	Nuevo jefe(a) o cónyugue	20

Variable	Valor Original	Valor Asignado
Sexo	Nombre	1
	Mujer	0

Variable	Valor Original	Valor Asignado
Estado_Conyugal	...en unión libre o juntado?	1
	...casado(a)?	2
	...divorciado(a)?	3
	...separado(a)?	4
	...viudo(a)?	5
	...soltero(a)?	6
	...casado(a) con cónyuge no residente?	7
	... unión libre con compañero(a) del mismo sexo?	8

Variable	Valor Original	Valor Asignado
Lugar_Nacimiento	En este mismo cantón	1
	En otro cantón	2
	En otro país	3

Variable	Valor Original	Valor Asignado
Permanencia_pais	Menos de un año	1
	Un año o más	2

Variable	Valor Original	Valor Asignado
Permanencia_motivo	Trabajo	1
	Otro	2

Variable	Valor Original	Valor Asignado
Seguro	Sí	1
	No	0

Variable	Valor Original	Valor Asignado
Tipo_Seguro	...asalariado?	1
	...mediante convenio como asociaciones, sindicatos, cooperativas, etc?	2
	...cuenta propia o voluntario?	3
	...pensionado de la CCSS, Magisterio, u otro?	4
	...familiar de asegurado directo o pensionado?	5
	..asegurado por el Estado, incluye familiar de asegurado por el Estado?	6
	..pensionado del régimen no contributivo monto básico, gracia o guerra?	7
	..seguro privado o del extranjero?	8
	..otras formas como seguro de estudiante, de refugiado y otros?	9

Variable	Valor Original	Valor Asignado
Regimen_pension	Ninguno	1
	Régimen de IVM de la CCSS?	2
	Otro régimen (Magisterio, Poder Judicial, Hacienda, etc.)?	3

Variable	Valor Original	Valor Asignado
Plan_Voluntario	Sí	1
	No	0

Variable	Valor Original	Valor Asignado
Educacion_asiste	...escuela?	1
	...colegio?	2
	...parauniversitaria o universitaria?	3
	...enseñanza especial?	4
	...educación abierta en instituciones para presentar examenes ante el MEP)?	5
	...otro tipo de formación no formal (especifique)	6
	...No asiste	7

Variable	Valor Original	Valor Asignado
Educación_titulo	No tiene título	1
	Técnico, perito no universitario	2
	Profesorado, diplomado o técnico universitario	3
	Bachillerato	4
	Licenciatura	5
	Especialización	6
	Maestría y doctorado	7
	Maestria	8
	Doctorado	9
	No especificado	10

Variable	Valor Original	Valor Asignado
EducaciónNoRegular asiste	Sí	1
	No	0

Variable	Valor Original	Valor Asignado
Idioma	Sí	1
	No	0

Variable	Valor Original	Valor Asignado
Idioma_cual	...inglés?	1
	...francés?	2
	...alemán?	3
	...otro?	4
	...español?	5

Variable	Valor Original	Valor Asignado
Trabajo	Sí	1
	No	0

Variable	Valor Original	Valor Asignado
Región	Central	1
	Chorotega	2
	Pacífico central	3
	Brunca	4
	Huetar caribe	5
	Huetar norte	6

Variable	Valor Original	Valor Asignado
Zona	Urbana	1
	Rural	2

Variable	Valor Original	Valor Asignado
Provincia nacimiento	San José	1
	Alajuela	2
	Cartago	3
	Heredia	4
	Guanacaste	5
	Puntarenas	6
	Limón	7
	No especificado	8

Variable	Valor Original	Valor Asignado
País Nacimiento	Costa Rica	1
	El Salvador	2
	Honduras	3
	Nicaragua	4
	Panama	5
	Colombia	6
	Estados Unidos	7
	México	8
	Venezuela	9
	Otro país	10
	Servicio doméstico sin otro lugar donde vivir o su familiar	11

Variable	Valor Original	Valor Asignado
Nivel Educativo	Ninguno	1
	Primaria incompleta	2
	Primaria completa	3
	Secundaria incompleta	4
	Secundaria completa	5
	Universitario sin título	6
	Universitario con título	7
	No especificado	8

Variable	Valor Original	Valor Asignado
Tipo Población	Población Adulta	1
	Población joven	2