ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

FACULTAD DE CIENCIAS

CARRERA DE INGENIERÍA EN ESTADÍSTICA INFORMÁTICA

"COMPARATIVA ENTRE REGRESIÓN LOGÍSTICA ORDINAL, REDES NEURONALES ARTIFICIALES Y GRADIENT BOOSTING; EN LA PREDICCIÓN DE LA SATISFACCIÓN LABORAL EN ECUADOR".

Trabajo de titulación:

Tipo: Proyecto de investigación

Presentado para optar al grado académico de:

INGENIERO EN ESTADÍSTICA INFORMÁTICA

AUTOR: VINICIO ALEXANDER ANDRADE SALTOS

DIRECTOR: ING. PABLO JAVIER FLORES MUÑOZ

Riobamba – Ecuador

2020

Yo, Vinicio Alexander Andrade Saltos, declaro que el presente trabajo de titulación es de mi autoría y los resultados del mismo son auténticos. Los textos en el documento que provienen de otras fuentes están debidamente citados y referenciados.

Como autor asumo la responsabilidad legal y académica de los contenidos de este trabajo de titulación; El patrimonio intelectual le pertenece a la Escuela Superior Politécnica de Chimborazo.

Riobamba, 10 de enero de 2020

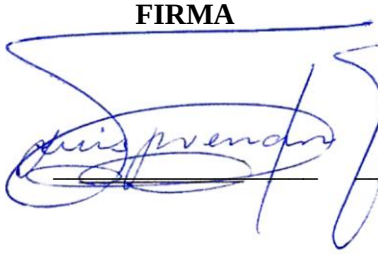
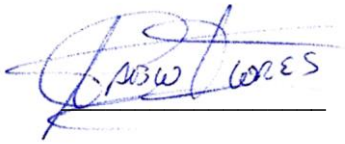
Vinicio Alexander Andrade Saltos
020202511-0

ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

FACULTAD DE CIENCIAS

CARRERA DE INGENIERÍA EN ESTADÍSTICA INFORMÁTICA

El Tribunal del trabajo de titulación certifica que: El trabajo de investigación: Tipo: Proyecto de Investigación, **COMPARATIVA ENTRE REGRESIÓN LOGÍSTICA ORDINAL, REDES NEURONALES ARTIFICIALES Y GRADIENT BOOSTING; EN LA PREDICCIÓN DE LA SATISFACCIÓN LABORAL EN ECUADOR**, realizado por el señor: **VINICIO ALEXANDER ANDRADE SALTOS**, ha sido minuciosamente revisado por los Miembros del Tribunal del trabajo de titulación, El mismo que cumple con los requisitos científicos, técnicos, legales, en tal virtud el Tribunal Autoriza su presentación.

	FIRMA	**FECHA**
Dr. Luis Antonio Vera Rojas **PRESIDENTE DEL TRIBUNAL**		2020 – 01 – 10
Ing. Pablo Javier Flores Muñoz **DIRECTOR DEL TRABAJO DE TITULACIÓN**		2020 – 01 – 10
Dr. Rubén Antonio Pazmiño Maji **MIEMBRO DEL TRIBUNAL**		2020 – 01 – 10

DEDICATORIA

A mi hermano Damián, no importa el sendero que decidas seguir en tu vida, importa que la caminata sea pacífica, con buena compañía y llena de felicidad.

Alexander

AGRADECIMIENTO

A Pilar, quien con un extraordinario esfuerzo ha permitido que finalice mis estudios y jamás ha dejado que me falte nada.

A Tati, quien ha sido mi compañera incondicional estos dos últimos años, los cuales puedo decir, han sido hasta ahora, los mejores años de mi vida.

A mis tutores, Ing. Pablo Flores y Dr. Rubén Pazmiño, de quienes he aprendido mucho y quienes han aportado con su valioso conocimiento y experiencia en la presente investigación.

Finalmente, pero no menos importante, un especial agradecimiento al Dr. Luis Vera, quien ha sido un extraordinario docente, un dirigente ejemplar y un invaluable amigo.

Alexander

TABLA DE CONTENIDO

CAPÍTULO II

CAPÍTULO III

ÍNDICE DE TABLAS

ÍNDICE DE GRÁFICOS

ÍNDICE DE ANEXOS

RESUMEN

La presente investigación tiene como objetivo comparar la calidad predictiva y la demanda de procesamiento de la técnica clásica: regresión logística ordinal y las técnicas de machine learning: redes neuronales artificiales y gradient boosting. El estudio se plantea en un contexto donde el avance tecnológico ha permitido un crecimiento exponencial en la producción de información, la cual requiere ser analizada de forma eficiente, por lo tanto, resulta indispensable identificar las mejores técnicas para el análisis. La comparación se realizó en el marco de la construcción de un modelo que prediga el nivel de satisfacción laboral en jefes de hogar ecuatorianos con un único trabajo. Así, se estudiaron las principales características de ambas metodologías y se identificaron sus equivalencias en terminología. Posteriormente se realizó una comparación cuantitativa de la calidad predictiva, tiempos de procesamiento y pico de memoria RAM asociados a cada uno de los modelos construidos con las tres técnicas, se realizó un proceso de remuestreo mediante ten-fold cross validation y se corrieron 200 modelos por cada técnica para controlar la variabilidad propia del fenómeno bajo estudio. Finalmente se contrastó el nivel de procesamiento generado, tomando en cuenta dos factores: 1) tamaño de muestra (real y aumentada con 37 336 y 373 360 observaciones, respectivamente), y 2) número de núcleos efectivos del procesador (uno y siete). Los resultados mostraron que el error total de predicción para gradient boosting fue del 29.5%, concluyendo así que esta técnica es la más confiable en su tarea predictiva, presentando una alta demanda de procesamiento, la cual disminuye considerablemente al trabajar en paralelo, es decir, al utilizar todos los núcleos del procesador. Se recomienda utilizar gradient boosting en estudios socio – económicos similares al estudio aquí planteado.

Palabras clave: ESTADÍSTICA, MODELO PREDICTIVO, ESTADÍSTICA CLÁSICA, MACHINE LEARNING (METODOLOGÍA), SATISFACCIÓN LABORAL, R Y RSTUDIO (SOFTWARE)

SUMMARY

This research aims to compare the predictive quality and processing demand of the classical technique: ordinal logistic regression and machine learning techniques: artificial neural networks and gradient boosting. The study is set in a context where technological progress has allowed exponential growth in the production of information, which needs to be analyzed efficiently, therefore, it is essential to identify the best techniques for analysis. The comparison was made within the framework of the construction of a model that predicts the level of job satisfaction in Ecuadorian householders with a single job. Therefore, the main characteristics of both methodologies were studied and their equivalences in terminology were identified. Subsequently, a quantitative comparison of the predictive quality was made, processing times and peak RAM associated with each of the models built with the three techniques, a resampling process was performed using ten-fold cross validation and 200 models were run per each technique to control the variability of the phenomenon under study. Finally, the level of processing generated was contrasted, taking into account two factors: 1) sample size (real and increased with 37 336 and 373 360 observations, respectively), and 2) number of effective processor cores (one and seven). The results showed that the total prediction error for gradient boosting was 29.5%, concluding that this technique is the most reliable in its predictive task, presenting a high demand for processing, which decreases considerably when working in parallel, that is, when using all processor cores. It is recommended to use gradient boosting in socio-economic studies like the study proposed here.

Keywords: STATISTICS, PREDICTIVE MODEL, CLASSICAL STATISTICS, MACHINE LEARNING (METHODOLOGY), LABOR SATISFACTION, R Y RSTUDIO (SOFTWARE)

INTRODUCCIÓN

Antecedentes del problema

Antecedentes metodológicos

Se han realizado varios estudios comparativos entre técnicas de predicción mediante clasificación; ya sean técnicas clásicas de estadística multivariante, como análisis de regresión logística; o, metodologías relativamente modernas basadas en sistemas computacionales, como redes neuronales artificiales (ANN por sus siglas en inglés) o gradient boosting. Varios ejemplos sobre investigaciones de tipo comparativo se presentan a continuación: Brown y Mues (2012: p.3447) en su trabajo titulado "Una comparación experimental de algoritmos de clasificación para conjuntos de datos de calificación crediticia desequilibrada", realizan una comparación entre los algoritmos: regresión logística, random forest, redes neuronales, análisis discriminante y máquinas de vectores de soporte; concluyendo que random forest y gradient boosting, presentan el mejor desempeño en clasificación. Así también, Roe et al. (2005: p.583) en su investigación "Boosted decision trees como alternativa para redes neuronales artificiales en la identificación de partículas", comparan ANN y gradient boosting en un estudio físico de identificación de partículas mediante clasificación, en donde, se concluye que los resultados que ofrecen las técnicas de gradient boosting son más eficientes que los resultados obtenidos mediante ANN. Tu (1996, p.1231) en el área de salud, realiza el estudio "Ventajas y desventajas del uso de redes neuronales artificiales versus regresión logística para predecir resultados médicos", entre sus principales conclusiones se observa que ANN son particularmente útiles cuando el objetivo principal es la predicción del resultado, mientras que, regresión logística es una elección razonable cuando el objetivo del estudio es buscar posibles relaciones causales entre variables dependientes. Otro estudio de gran relevancia es el trabajo realizado por Li et al. (2008: p.7) que tiene por nombre "Aprendiendo a clasificar usando clasificación múltiple y gradient boosting", en el cual, se discute el problema del ranking en motores de búsqueda, es decir, cuál es el orden en que deben aparecer los resultados de búsqueda cuando se realiza una nueva consulta, para ello, se compara un modelo de regresión con fines de clasificación mediante rangos, y una variante de este modelo en la que se implementa la técnica de gradient boosting; el estudio concluye que el modelo basado en gradient boosting supera a su par basado en regresión por rangos.

Trabajos de investigación más recientes se han realizado en la búsqueda de un método más confiable en la predicción de diferentes problemas específicos, por ejemplo, Churpek et al. (2016: p.7), en su investigación "Comparación multicéntrica de métodos de machine learning y métodos de regresión convencional para predecir el deterioro clínico en las salas", comparan diferentes métodos predictivos, con el fin de detectar el deterioro clínico de pacientes en una base de datos compuesta por diferentes centros de salud, concluyendo que los métodos de machine learning son

mejores que las técnicas de regresión convencionales. En la estimación de los costos de manufactura de las componentes de un motor de jet, Loyer et al. (2016: p.15) publican el trabajo "Comparación de los métodos de machine learning aplicados a la estimación del costo de fabricación de componentes de motores a reacción" donde se comparan los algoritmos de: modelos aditivos generalizados, support vector regression, gradient boosting, con respecto a técnicas más usuales como: regresión lineal múltiple y redes neuronales artificiales; entre sus conclusiones, se tiene que, en el aspecto metodológico, las técnicas de regresión lineal múltiple y redes neuronales artificiales son superadas, en términos de rendimiento, por las técnicas más recientes desarrolladas en el campo del data mining. Nuevamente en el área de salud, Hung et al. (2017: p.26530) publican "Comparación de deep neural network y otros algoritmos de machine learning para la predicción de accidentes cerebrovasculares en una base de datos de reclamos médicos electrónicos basada en una población a gran escala", en este trabajo se compara el algoritmo de deep neural network (DNN) con otras técnicas predictivas; concluyendo que DNN y gradient boosting tienen una calidad de predicción similar y son mejores que la regresión logística y que el algoritmo de support vector machines. En el estudio de los procesos de manufactura, donde la complejidad de analizar una gran cantidad de pasos previos a la obtención del producto final, es muy alta, Anghel et al. (2018: p.36) plantean el estudio "Predicción de errores en los procesos de fabricación: gradient boosted trees versus deep neural network", en el estudio se plantea una comparación entre dos enfoques: 1) utilizar los algoritmos de Multivariate Adaptive Regression Splines (MARS) y support vector machines (SVM) en tareas de clasificación, 2) utilizar técnicas de redes neuronales artificiales y gradient boosting, en la misma tarea; los resultados del análisis muestran que las técnicas de redes neuronales artificiales y gradient boosting, permiten obtener mejores resultados. Con el objetivo de predecir la satisfacción laboral, Andrade Saltos y Flores M (2018: p.47) publican el estudio "Comparativa entre classification trees, random forest y gradient boosting; en la predicción de la satisfacción laboral en Ecuador", los resultados permiten concluir que las tres técnicas presentan tasas de error similares (aproximadamente 30%), resaltando el algoritmo de Gradient Boosting, con un menor tiempo de procesamiento y una mínima demanda de memoria RAM.

Varias de las técnicas aquí estudiadas han sufrido cambios para mejorar su rendimiento en tareas específicas. Qu et al. (2016: p.15) proponen el algoritmo Product-based Neural Networks (PNN), el cual, demuestra tener un mejor rendimiento que los modelos de deep learning, en la predicción de la respuesta de los usuarios a ciertos estímulos en páginas web, tal respuesta se ve reflejada en los clics del usuario o en las veces que cambia entre las pestañas disponibles. Las redes neuronales artificiales han presentado problemas de sobreajuste y gradientes de alta variación, cuando la información se caracteriza mediante una masiva cantidad de variables y un tamaño muestral pequeño, uno de los casos en los que se presenta esta realidad es en la predicción de fenotipos

usando datos genéticos en el ámbito de la bioinformática; para subsanar este problema, Liu et al. (2017: p.2292) presentan un algoritmo conocido como Deep Neural Pursuit (DNP), el cual, demuestra ser robusto cuando se presentan muchas variables y tiene un buen rendimiento en su etapa de aprendizaje cuando las muestras son pequeñas. En el problema de generar imágenes en alta resolución, teniendo como datos de entrada imágenes de baja resolución, Kim et al. (2016: p.85) proponen un algoritmo de redes neuronales convolucionales con aprendizaje muy profundo, el cual, subsana el problema de una etapa de aprendizaje demasiado lenta y demuestra una calidad superior con respecto a otros métodos existentes.

Con la intención de potenciar la tarea de clasificación, se han realizado investigaciones que combinan las técnicas aquí analizadas, por ejemplo, Zhang et al. (2016: p.1794) analizan el reconocimiento remoto de imágenes en alta resolución vía satélite, la técnica estándar y científicamente aceptada para realizar este tipo de análisis es Redes Convolucionales (un caso particular de Redes Neuronales Artificiales), los autores plantean la posibilidad de enriquecer esta técnica al combinarla con Gradient Boosting, creando así, la técnica conocida como gradient boosting random convolutional network (GBRCN); entre sus hallazgos, se destaca el hecho de que GBRCN genera resultados de mayor calidad en comparación a los métodos tradicionales. Otro ejemplo, es el estudio planteado por Fonseca et al. (2017: p.5), quienes analizan la clasificación de escenas acústicas, en este caso, el objetivo es implementar dos técnicas de machine learning (gradient boosting y convolutional neural network), en diferentes etapas del proceso de gestión de información; su conclusión es que la combinación de estas técnicas ofrece una mejora substancial en el desempeño del análisis. También, Wang et al. (2016: p.1902), en un estudio de sistemas de recomendación de celulares, es decir, cómo encontrar un modelo de celular específico para un usuario con ciertas características de comportamiento; plantean fusionar un modelo basado en un algoritmo de regresión lineal y un algoritmo de gradient boosting, el resultado de este análisis muestra que el modelo fusionado mejora en un 2% su calidad de recomendación con respecto al modelo de gradient boosting y en un 36% con respecto al modelo de regresión lineal. Otro trabajo relevante es el realizado por Xiao et al. (2018: p.7), quienes analizan el diagnóstico de cáncer mediante deep neural networks como herramienta para ensamblar cinco modelos de clasificación: k-nearest neighbour (kNN), support vector machines (SVM), decision trees (DT), random forest (RF) y gradient boosting (GB); tal ensamble tiene el objetivo de predecir el cáncer en condiciones normales y en presencia de tumores, tomando en cuenta tres tipos de tejidos: pulmón, estómago y pecho; se concluye que este método reduce el error de predicción y obtiene mejores resultados.

Antecedentes aplicativos

Varios analistas han visto la necesidad de profundizar en el estudio de la satisfacción laboral, como mencionan Madrigal et al. (2015: p.435), uno de los factores determinantes para conocer la adaptación del profesional al entorno organizacional es la satisfacción laboral. Este hecho es de suma importancia en cualquier organización, desde temas de salud, en donde, si los profesionales sienten y valoran que trabajan en un excelente clima laboral y su trabajo les proporciona un alto grado de satisfacción, entonces los médicos presentan mejores resultados con los pacientes (García et al., 2010: p.208); hasta el ámbito industrial, donde Uríen y Osca (2001: p.326) señalan que, son las tareas con herramientas y, en algunos aspectos las tareas de mejora, las que aumentan la satisfacción y el interés de los empleados a la hora de implementar un sistema de trabajo en equipo o en grupos.

En la búsqueda de los factores que influyen en la satisfacción laboral, Andersen et al. (2017: p.481) pretenden encontrar la oferta de promoción de la salud en el trabajo, que tenga mayor incidencia en la satisfacción laboral, para lo cual, se toma en cuenta: no fumar, dieta sana, ejercicio físico, contacto con profesionales de la salud, chequeos periódicos y ambiente laboral; concluyendo que el apoyo social de los superiores y colegas influye fuertemente en la satisfacción laboral.

Los profesionales de la salud que trabajan en academia, se enfrentan a toda una variedad de actividades, incluyendo: enseñanza a pregrado y posgrado, atención clínica, consultoría (facultad, colegas y estudiantes), participación en investigación y roles administrativos, toda esta presión influye en su satisfacción laboral, con ello en mente, Krueger et al. (2017: p.184) estudian las variables que pudieren influir en la satisfacción, entre sus resultados, se observa que la satisfacción laboral es una construcción multidimensional, que puede estar conformada por: el estado de salud, variables demográficas, experiencias que ha tenido en su vida laboral, nivel de su educación académica y el entorno de trabajo. En otro ámbito, con similar demanda de esfuerzo y dedicación, Villar-Rubio et al. (2015: p.219) estudian la satisfacción laboral en empleados de la administración del sector tributario en España, la importancia de tal investigación radica en la gran cantidad de trabajo que se presenta por persona en este ámbito, ya que, España es uno de los países de la Unión Europea con una de las más bajas proporciones de empleados públicos; los resultados del análisis establecen como variables influyentes en la satisfacción laboral al: género, sensación de ser valorado, autonomía en las decisiones, participación en decisiones relacionadas con las tareas a realizarse, conciencia de los objetivos demandados, cumplimiento de las expectativas de trabajo y rotación adecuada.

En Estados Unidos, se predice una escasez en profesionales del área neurológica, esto limita el acceso de los pacientes a tratamientos de calidad en los trastornos neurológicos, Teixeira-Poit et al. (2017: p.7) estudian la satisfacción laboral en este ámbito profesional, identificando factores

influyentes como las características demográficas y las prácticas laborales, se concluye que para mejorar la satisfacción laboral y aumentar el número de neurólogos a futuro, se deben implementar estrategias como: incrementar las oportunidades de investigación y colaboración, proporcionar recursos necesarios para profesionales que trabajan en áreas rurales pequeñas, y mejorar la remuneración de las neurólogas para lograr equidad con respecto al salario de los neurólogos.

Los adultos con esclerosis múltiple se ven expuestos a una serie de dificultades en el sector laboral, este hecho se ve reflejado en la alta tasa de desempleo en personas con este padecimiento (Julian et al., 2008: p.1358). Un estudio realizado por Li et al. (2017: p.39) pretende determinar la relación entre la satisfacción laboral de los adultos con esclerosis múltiple y cuatro factores: características personales y demográficas, factores externos o de contexto, estado de incapacidad, y variables de percepción vocacional; la investigación señala que los factores en estudio representan un 35% de la variabilidad presente en la satisfacción laboral, lo cual, plantea un punto de inicio para que los consejeros en rehabilitación vocacional puedan lidiar con los problemas usuales en personas con esclerosis múltiple, tales como discriminación y estereotipos sociales.

La trayectoria de satisfacción laboral en adolescentes que empiezan un nuevo trabajo, suele iniciar con una alta satisfacción y luego presenta una fuerte pendiente de caída. Una prominente explicación para este fenómeno es que las expectativas sobre el nuevo trabajo y sus actividades eran muy altas, o que tal vez, al comparar el nuevo trabajo con el anterior, se observan desventajas e inconformidades, Valero y Hirschi (2019: p.9) al estudiar esta realidad, mencionan que aquellos adolescentes con altas percepciones por parte de sus supervisores y apoyo de sus compañeros de trabajo, pueden mantener una alta satisfacción laboral, independientemente del tiempo que lleven en el mismo trabajo.

La Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU) ha tenido un importante protagonismo en los estudios socio - económicos en Ecuador. La ENEMDU, realizada por el Instituto Nacional de Estadísticas y Censos (INEC), ha sido una fuente significativa de información para análisis enfocados en la satisfacción laboral, por ejemplo, Montecé et al. (2016: p.2) vieron la necesidad de estudiar las posibles diferencias existentes en el salario y la satisfacción laboral en Ecuador, tomando en cuenta el género y la etnia de los trabajadores. Otro ejemplo, es el análisis realizado por Hidalgo (2018, p.11) quien se propuso determinar el nivel de satisfacción laboral del personal en la unidad de talento humano de la Universidad Central del Ecuador. También, es importante mencionar el estudio realizado por Cabrera y Huerta (2017: p.2) quienes analizaron la posible incidencia del Clima Laboral en la Satisfacción de los empleados en la empresa Arcgold del Ecuador S.A, con el fin de mejorar su ventaja competitiva y productividad, mediante el buen estado anímico de sus empleados. Finalmente, en un estudio de la satisfacción

laboral presente en jefes de hogar ecuatorianos con un único empleo, Andrade Saltos y Flores M (2018: p.45) plantean que, con base en la información de la ENEMDU de 2017, aproximadamente el 70% de los individuos están contentos con su trabajo.

Un estudio más reciente plantea la necesidad de analizar la responsabilidad social y la gestión de seguridad en el trabajo, para ello, SUASNAVAS et al. (2019: p.6) toman en cuenta la satisfacción laboral, variables socio-demográficas, responsabilidad social empresarial y gestión de la prevención de riesgos laborales; para determinar el desempeño social interno y la gestión de seguridad y salud en las empresas ecuatorianas, utilizando información presente en la ENEMDU de 2018. Entre los resultados obtenidos, resalta el hecho de que la gran mayoría de trabajadores encuestados manifiestan satisfacción laboral, pero es importante contrastar esta realidad con el hecho de que el salario básico actual ($386 USD) no es suficiente en relación a la canasta básica vital $495 USD y la canasta básica familiar $708 USD.

Formulación del problema

¿Cuál técnica de predicción ofrece resultados más confiables en la clasificación de los jefes de hogar con un único trabajo en Ecuador, según su nivel de satisfacción laboral: redes neuronales artificiales, regresión logística ordinal o gradient boosting?

Justificación

Justificación teórica

El avance tecnológico ha permitido un crecimiento exponencial en la producción de información. En la actualidad, casi todo se cuantifica mediante sensores electrónicos, éstos se encuentran en: celulares inteligentes, medios de transporte, electrodomésticos, y demás dispositivos de uso diario. Esta realidad requiere que los conjuntos masivos de datos generados (Big Data) sean estudiados con eficiencia y prontitud. En tal contexto, es imperante para el analista de datos, escoger la metodología idónea para dar solución a los retos planteados en un entorno globalmente competitivo; así, nace la necesidad de conocer si técnicas estadísticas con una concepción clásica, como Regresión Logística Ordinal, son suficientes para afrontar desafíos de tal magnitud; o posiblemente, existen técnicas como Redes Neuronales Artificiales o Gradient Boosting, las cuales aprovechan la enorme capacidad de procesamiento y almacenamiento del hardware actual, y por lo tanto, podrían ofrecer resultados más confiables en su tarea de predicción.

La estadística clásica, ha resultado ser muy útil gracias a su naturaleza inferencial, la cual, nace en un entorno donde el manejo de datos no era precisamente una prioridad en las políticas de estado, por tanto, la recolección y el acceso a información presentaban un alto nivel de dificultad, más aún, cuando no se contaba con el poder tecnológico necesario para la gestión de datos, así, la escasa información existente se recolectaba y analizaba en papel, lo cual representaba un enorme

trabajo para el analista. En tal contexto, el hecho de trabajar con muestras representativas de la población, suponía una ventaja y un ahorro significativo en el trabajo, más aún, cuando el análisis podía obtener resultados que estimen de forma aproximada los parámetros poblacionales; pero, este hecho ocurría siempre y cuando ciertos supuestos matemáticos intrínsecos al modelo propuesto en el estudio, se cumpliesen. Hoy en día la realidad es otra, el manejo de la información se ha convertido en un tema común, tanto así, que grandes conglomerados como Facebook, Amazon o Google; comercian, ya no con bienes o servicios típicos, sino con la información que sus clientes generan, tal información es muy variada y, en general, no presenta patrones o características que se ajusten a técnicas estadísticas clásicas. Los retos que se presentan en el área de análisis de datos, han evolucionado a tal punto, que hoy en día el profesional debe analizar información no estructurada, por ejemplo: vídeos, reconocimiento facial, reconocimiento de voz, imágenes donde cada pixel representa un dato, detección de sonidos específicos en un archivo de audio con mucho ruido, la reacción que presenta un automóvil autónomo cuando transita sin conductor, etc. Es en este punto, donde áreas como Machine Learning, que, en esencia tiene como objetivo que una máquina sea capaz de aprender y reaccionar de forma rápida y eficiente cuando se le proporciona cierta información; presentan una variedad de técnicas de análisis con mayor versatilidad y que están en concordancia con las demandas de un mundo tecnológicamente conectado. La principal técnica utilizada en Machine Learning se conoce como Redes Neuronales Artificiales, la cual, pretende emular las conexiones neuronales de un organismo con vida, en donde, el sistema podrá aprender de la información que percibe, además, será capaz de corregir sus resultados tomando en cuenta los datos históricos que haya procesado. Otra importante herramienta en Machine Learning es el Gradient Boosting, el cual, es una evolución de los árboles de clasificación y ha sido ampliamente utilizado en estudios socio-económicos, donde clasificar a un individuo según ciertas características financieras, representa una ventaja para cualquier empresa.

Un ejemplo de la adaptabilidad de las técnicas antes mencionadas, es el trabajo realizado por Wong et al. (2013: p.235) quienes proponen un modelo basado en redes neuronales cuyo objetivo es reconocer la imagen de la cara de una persona, de entre un conjunto de imágenes de caras de otras personas; lo novedoso de este modelo es que introduce un parámetro que toma en cuenta el color de las imágenes y por ende los diferentes matices de color en la piel de cada persona, de esta forma se consigue un mejor porcentaje de reconocimiento exitoso en las imágenes. Otro ejemplo de la relevancia aplicativa que presentan las técnicas aquí expuestas es la investigación realizada por Atkinson et al. (2012: p.1397), quienes analizan imágenes óseas avanzadas, con el fin de estudiar el riesgo de posibles fracturas y de esta forma poder predecirlas; dicho estudio se realiza mediante la técnica de gradient boosting, la cual demuestra ser superior ante técnicas generalmente utilizadas en procedimientos similares. Finalmente, pero no menos importante, es el trabajo

realizado por Google, en colaboración con la Universidad de Harvard, publicado en la revista Nature a finales de agosto de 2018, en el cual, mediante el uso de redes neuronales, se ha conseguido entrenar un modelo capaz de predecir con precisión los lugares en donde se producirán posibles réplicas después de un sismo (DeVries et al., 2018: p.632).

Tomando en cuenta la utilidad de las técnicas antes mencionadas, es evidente que su aplicación en proyectos de interés nacional, contribuiría de manera sustancial a que las políticas públicas se encaminen en la dirección correcta, decidiendo con base en un análisis estadístico responsable y eficiente, permitiendo así, un desarrollo orgánico de los hogares ecuatorianos y del país en general.

Justificación práctica

La satisfacción laboral es, sin lugar a duda, un factor clave en el desempeño de los trabajadores. Este hecho puede incidir de forma decisiva en la competitividad de una empresa o institución, por ello, es importante que quienes ocupan cargos gerenciales o administrativos, tengan información relevante sobre el estado actual de sus empleados y la visión que presentan con respecto a su medio laboral. De esta forma, es posible planificar proyectos de vinculación entre el sector laboral y la institución o empresa en cuestión. Rowden (2002, p.408) argumenta que la satisfacción laboral es una variable independiente esencial ya que puede guiar el comportamiento de los trabajadores, y de esta forma, afectar el funcionamiento de la organización.

Incrementar la satisfacción en el trabajo debería ser una parte esencial en la misión de las organizaciones (Friday y Friday, 2003: p.428), por lo tanto, las instituciones deben garantizar un entorno de trabajo, en el que sus empleados puedan desenvolverse de forma consistente con sus aspiraciones y necesidades.

El presente estudio pretende construir un modelo predictivo con la técnica de análisis de datos que ofrezca resultados de predicción confiables, dicho modelo servirá para clasificar a un jefe de hogar con un único trabajo en Ecuador, de acuerdo a su nivel de satisfacción laboral (Contento, Poco contento, Descontento pero conforme o Totalmente descontento). La importancia de este trabajo radica en que las empresas o instituciones ecuatorianas podrán conocer qué tan satisfechos están sus empleados, lo cual, será un importante referente para la toma de decisiones en el área de recursos humanos, priorizando que el ambiente laboral esté en concordancia con el artículo 33 de la Constitución del Ecuador, donde se establece que: “El trabajo es un derecho y un deber social, y un derecho económico, fuente de realización personal y base de la economía. El Estado garantizará a las personas trabajadoras el pleno respeto a su dignidad, una vida decorosa, remuneraciones y retribuciones justas y el desempeño de un trabajo saludable y libremente escogido o aceptado”.

Objetivos

Objetivo General

Comparar las técnicas de machine learning: redes neuronales artificiales y gradient boosting, con la técnica clásica: regresión logística ordinal; mediante indicadores de calidad predictiva; con el fin de establecer el modelo que clasifique de forma más confiable a los jefes de hogar con un único trabajo en Ecuador, según su satisfacción laboral.

Objetivos Específicos

- Seleccionar entre las variables de la ENEMDU, aquellas que hayan demostrado tener mayor influencia en la satisfacción laboral, tomando en cuenta trabajos de investigación similares.
- Realizar un análisis exploratorio de datos con el fin de describir las variables en estudio y corregir cualquier irregularidad que pueda presentar la información.
- Construir modelos de predicción mediante las técnicas de clasificación: regresión logística ordinal, redes neuronales artificiales y gradient boosting.
- Comparar la demanda de procesamiento computacional y la calidad predictiva de los tres modelos construidos.

CAPÍTULO I

1. MARCO TEÓRICO REFERENCIAL

1.1. Aspectos socio – económicos

1.1.1. Satisfacción laboral

> La satisfacción laboral se define como el grado en que a las personas les gusta su trabajo. Algunas personas disfrutan el trabajo y lo encuentran como una parte central de la vida. Otros odian trabajar y lo hacen solo porque deben hacerlo. (Spector, 1997, p.7)

1.1.2. Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU)

1.1.2.1. Estructura ENEMDU

El Instituto Nacional de Estadística y Censos (INEC), en su función de proveedor oficial de las estadísticas laborales para Ecuador, realiza la Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU). Esta encuesta se realiza bajo un esquema de panel de viviendas seleccionadas en una submuestra, la cual se mantiene en la muestra durante dos trimestres consecutivos, seguido de un descanso de dos trimestres, y finalmente entran en la muestra por dos últimos trimestres. En este sentido, es posible analizar a las mismas unidades de observación en distintos cohortes temporales. (INEC, 2018, p.4)

Particularmente, la ENEMDU de diciembre de 2018, cuenta con 4 módulos, los cuales tienen el objetivo de agrupar la información según su relación con ciertos fenómenos socio-económicos específicos. Los módulos se enumeran a continuación:

- 15 años
- Financiero

- Vivienda – Hogar
- Confianza del consumidor

1.1.2.2. Historia ENEMDU

En 1987, nace la Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU), en el Instituto Nacional de Empleo, perteneciente al Ministerio de Trabajo. A partir de 1990, se realizan encuestas con periodicidad anual, y con representatividad urbana. (INEC, 2018, p.5)

En 1993 llega al Instituto Nacional de Estadística y Censos (INEC) con la misma metodología, periodicidad y representatividad. A partir de diciembre 2003, la encuesta se realiza con periodicidad trimestral. En diciembre 2007, se realiza una revisión de la metodología vigente y se incorporan mejoras. En diciembre 2013, se realiza la migración al marco de muestreo 2010, mientras que en marzo 2014 se incrementa el tamaño de la muestra. (INEC, 2018, p.5)

1.1.2.3. Objetivo ENEMDU

La ENEMDU tiene por objetivo, reconocer la actividad económica y las fuentes de ingresos de la población, recaba información acorde a las principales categorías poblacionales en relación con el mercado laboral, es decir, recoge características sobre la población económicamente activa (empleados, y desempleados) e inactiva (rentistas, jubilados o pensionados, amas de casa, estudiantes, incapacitados). (INEC, 2018, p.5)

1.1.2.4. Justificación ENEMDU

La Décimo Tercera Conferencia de Estadísticos del Trabajo (13°CIET) llevada a cabo en 1982, señala la importancia de medir la población económicamente activa con tres fines: i) la medición de la magnitud del tiempo de trabajo y de la disponibilidad de las personas de trabajar, a efectos de una evaluación macroeconómica y planificación del desarrollo profesional; ii) la medición de las relaciones entre el empleo, los ingresos y otras características sociales y económicas, que permitan la formulación y evaluación de las políticas de empleo, y iii) la promoción de estudios y ejercicios analíticos al respecto. (INEC, 2018, p.5)

1.2. Teoría estadística y ciencia de datos

1.2.1. Análisis multivariante (AM)

1.2.1.1. Definición

El análisis multivariante (AM) es la parte de la estadística y del análisis de datos que estudia, analiza, representa e interpreta los datos que resultan de observar más de una variable estadística sobre una muestra de individuos. (Cuadras, 2007, p.13)

1.2.1.2. Clasificación estadística

Clasificar los elementos de un conjunto finito consiste en realizar una partición del conjunto en subconjuntos homogéneos, siguiendo un determinado criterio de clasificación. Cada elemento pertenece a un único subconjunto, que a menudo tiene un nombre que lo caracteriza. (Cuadras, 2007, p.187)

1.2.1.3. Regresión logística

Supongamos que hay para el análisis una o más series de ensayos, y la observación en cualquier ensayo toma una de dos formas, como "éxito" o "fracaso", "defectuoso" o "no defectuoso", y así sucesivamente. Denote las posibles observaciones por 0 y 1, cada serie de pruebas da una secuencia de 0's y 1's. Supongamos además que, para cada ensayo, hay una o más variables independientes, y que sospechamos que la probabilidad de que un ensayo en particular arroje el resultado 1, depende de los valores correspondientes de las variables independientes. Los métodos de regresión logística buscan estimar y probar tales dependencias. (Cox, 1958, p.215)

1.2.1.4. Regresión logística multinomial

La regresión logística multinomial es un método de clasificación que generaliza la regresión logística a problemas multiclase, es decir, cuando la variable dependiente presenta más de dos posibles resultados discretos. (Greene, 2003, p.91)

1.2.1.5. Regresión logística ordinal

Los modelos de regresión logística ordinal utilizan la naturaleza ordinal de la variable dependiente mediante la descripción de varios modos de ordenación estocástica, este hecho elimina la necesidad de asignar puntajes o asumir de otra manera la cardinalidad en lugar de la ordinalidad. (McCullagh, 1980, p.109)

1.2.2. Big data

Big data se refiere a volúmenes de datos en el rango de exabytes (10^{18}) y más. Dichos volúmenes exceden la capacidad de la tecnología para almacenar, administrar y procesarlos de manera eficiente. Estas limitaciones solo se descubren mediante un sólido análisis de los datos en sí, las necesidades de procesamiento explícito y la capacidad de las herramientas (hardware, software y métodos) utilizadas para analizarlo. (Kaisler et al., 2013: p.995)

1.2.3. Machine learning

1.2.3.1. Definición

El aprendizaje es un fenómeno multifacético. Los procesos de aprendizaje incluyen la adquisición de nuevos conocimientos declarativos, el desarrollo de habilidades motoras y cognitivas a través de la instrucción o la práctica, la organización de nuevos conocimientos en general, representaciones efectivas y el descubrimiento de nuevos hechos y teorías a través de la observación y la experimentación. Desde el inicio de la era de la informática, los investigadores se han esforzado por implantar tales capacidades en las computadoras. Resolver este problema ha

sido, y sigue siendo, el objetivo de largo alcance más desafiante y fascinante en inteligencia artificial. (AI). El estudio y la modelización informática de los procesos de aprendizaje en sus múltiples manifestaciones constituyen lo que hoy se conoce como Machine Learning. (Michalski et al., 2013: p.3)

1.2.3.2. Redes neuronales artificiales

Redes Neuronales Artificiales son un conjunto de modelos matemáticos compuestos por un gran número de elementos procesales organizados en distintos niveles, los cuales, buscan emular la sinapsis de un ente biológico. De forma más detallada se puede decir que las redes neuronales artificiales son redes interconectadas masivamente en paralelo de elementos simples (usualmente adaptativos) y con organización jerárquica, las cuales intentan interactuar con los objetos del mundo real del mismo modo que lo hace el sistema nervioso biológico. (Matich, 2001, p.8)

1.2.3.3. Aprendizaje profundo

El aprendizaje profundo o deep learning es una forma de inteligencia artificial que permite a las computadoras aprender de la experiencia y comprender el mundo en términos de una jerarquía de conceptos. Debido a que la computadora reúne los conocimientos de la experiencia, no es necesario que un operador humano especifique formalmente todo el conocimiento que necesita ser procesado. La jerarquía de conceptos permite que la computadora aprenda conceptos complicados al construirlos a partir de conceptos más simples, se puede decir, que es una generalización de las redes neuronales artificiales, pero con muchas capas de profundidad. (Goodfellow et al., 2016: p.5)

1.2.3.4. Arboles de clasificación

Un árbol de clasificación tiene como objetivo predecir que cada observación pertenece a la clase más común de observaciones de entrenamiento en la región a la que pertenece. Al interpretar los resultados de un árbol de clasificación, a menudo resulta interesante no solo la predicción de clase correspondiente a una región de nodo terminal particular, sino también las proporciones de clase entre las observaciones de entrenamiento que caen dentro de esa región. (Gareth et al., 2006: p.311)

1.2.3.5. Bagging

Bagging es una técnica que combina las predicciones de múltiples algoritmos de aprendizaje automático para hacer predicciones más precisas que cualquier modelo individual, la cual consiste en construir un número de árboles de clasificación en varias muestras de entrenamiento para finalmente promediar los resultados de cada árbol. Este procedimiento se puede usar para reducir la varianza de aquellos algoritmos que tienen alta variabilidad. (Kuhn y Johnson, 2013: p.192)

1.2.3.6. Random forest

Random forest proporciona una mejora sobre la técnica de bagging mediante un pequeño cambio aleatorio que transforma los árboles. Al igual que en el bagging, construimos un número de árboles de clasificación en las muestras de entrenamiento. Pero al construir estos árboles de clasificación, cada vez que se considera una división en un árbol, se elige una muestra aleatoria de m predictores como candidatos divididos del conjunto completo de predictores p. La división solo permite usar uno de esos m predictores, el número de predictores considerados en cada división es aproximadamente igual a la raíz cuadrada del número total de predictores. (Gareth et al., 2006: p.320)

1.2.3.7. Gradient boosting

Gradient boosting presenta similitudes con la técnica de random forest, En random forest, todos los árboles se crean de forma independiente, cada árbol se crea para tener la profundidad máxima y de esta forma contribuir por igual al modelo final. Sin embargo, en gradient boosting cada árbol depende de los árboles del pasado, se tiene una profundidad mínima y cada modelo contribuye de manera desigual al resultado final. (Kuhn y Johnson, 2013: p.206)

1.2.4. Stratified tenfold cross – validation

El procedimiento estándar para predecir la tasa de error en una técnica de aprendizaje dado un conjunto de datos es el método conocido como Stratified Tenfold Cross – Validation, el cual consiste en dividir aleatoriamente a los datos en diez partes, donde cada clase es representada

aproximadamente en igual proporción a su proporción en el total de datos. Una de las diez partes servirá como conjunto de prueba, mientras que las nueve restantes servirán para entrenar el modelo; después, se generará un nuevo modelo tomando en cuenta otra de las diez partes como conjunto de prueba; la idea es que se creen diez modelos, alternando las diez partes en calidad de conjunto de prueba y las nueve restantes como conjunto de entrenamiento en cada caso; finalmente se promediará la tasa de error de los diez modelos creados y de esta forma se definirá la estimación para el error total de predicción para cierta técnica. (Witten et al., 2016: p.153)

Extensas pruebas en numerosos conjuntos de datos de distinta naturaleza, con diferentes técnicas de aprendizaje, han demostrado que 10 es la cantidad correcta de pliegues para obtener la mejor estimación del error, y también hay cierta evidencia teórica que respalda esto. (Witten et al., 2016: p.153)

CAPÍTULO II

2. MARCO METODOLÓGICO

2.1. Tipo de investigación

La presente investigación es de tipo:

- *Aplicada*, pues la investigación se basa en la aplicación de técnicas pre establecidas, y no pretende construir nuevas.
- *Cuantitativa*, ya que busca comparar diferentes metodologías de predicción mediante un criterio empírico, el cual, se construye con información cuantificable, siguiendo un orden metodológico riguroso y secuencial, tomando en cuenta información correspondiente a investigaciones previas como punto de partida referencial, y generando un modelo que describa el comportamiento de la población en estudio.
- *Explicativa*, ya que existe una variable (explicada, respuesta, dependiente), la cual, será explicada por un conjunto de otras variables (explicativas, predictoras, independientes)
- *Transversal*, ya que el objetivo principal de la investigación se enfoca en un único periodo temporal (diciembre - 2018), y el tiempo no representa un factor a considerarse en la creación del modelo de clasificación.
- *Predictiva*, pues su objetivo es el de construir un modelo que prediga el nivel de satisfacción laboral en nuevos individuos.

2.2. Hipótesis de investigación

Las técnicas de machine learning: redes neuronales artificiales y gradient boosting, ofrecen resultados más confiables que la técnica estadística: regresión logística ordinal, en la predicción de la satisfacción laboral en los jefes de hogar ecuatorianos con un único trabajo.

2.3. Variables en estudio

En la investigación realizada por Andrade Saltos y Flores M (2018: p.44), se seleccionan las variables que tengan una posible relación teórica (desde un punto de vista socio - económico) en la predicción de la Satisfacción Laboral en Ecuador. Tomando en cuenta esta información, para el presente estudio se establecieron: como variable explicada a "p59", la cual caracteriza la satisfacción laboral; y como variables explicativas a: "p02", "p03", "p06", "p14", "p15", "p27", "p42", "p45", "p46", "p10a", "p47a", "p51a", "p71a", "grupo1", "ingrl", "rama1", "secemp". Las variables en estudio se describen en la **Tabla 1-2**.

2.3.1. Operacionalización de variables

Tabla 1-2: Operacionalización de variables en estudio

Código	Variable	Tipo/Escala
p59	**¿Cómo se siente en su trabajo? (Contento, Poco contento, …)**	**Cualitativa/Ordinal**
p02	Sexo (Hombre, Mujer)	Cualitativa/Nominal
p03	Edad [años]	Cuantitativa/De razón
p06	Estado civil (Casado(a), Separado(a), Divorciado(a), …)	Cualitativa/Nominal
p14	Idioma que habla (Sólo lengua indígena, Lengua indígena y español, Sólo español, …)	Cualitativa/Nominal
p15	Cómo se considera (Indígena, Afroecuatoriano, Negro, …)	Cualitativa/Nominal
p27	Desea trabajar más horas (Trabajar más horas en su trabajo actual, Trabajar más horas en otro trabajo, …)	Cualitativa/Nominal
p42	Categoría de ocupación (Empleado de gobierno, Empleado privado, Patrono, …)	Cualitativa/Nominal
p45	¿Cuántos años trabaja? [años]	Cuantitativa/De razón
p46	Sitio de trabajo (Local patrono, Obra en construcción, Se desplaza, En la calle, …)	Cualitativa/Nominal
p10a	Nivel de instrucción (Ninguno, Centro de alfabetización, Primaria, Secundaria, ...)	Cualitativa/Ordinal
p47a	Tamaño de establecimiento (Menos de 100 personas, 100 personas y más)	Cualitativa/Ordinal
p51a	Horas de trabajo principal [horas/semana]	Cuantitativa/De razón
p71a	Recibió ingresos derivados del capital (Sí, No)	Cualitativa/Nominal
grupo1	Grupo de Ocupación (Empleados de Oficina, Fuerzas Armadas, Profesionales Científicos e Intelectuales, …)	Cualitativa/Nominal
ingrl	Ingreso del trabajo [dólares]	Cuantitativa/De razón
rama1	Rama de actividad (Industrias manufactureras, Actividades financieras y de seguros, ...)	Cualitativa/Nominal
secemp	Sectores de los Empleados (Sector Formal, Sector Informal, Sector Doméstico, ...)	Cualitativa/Nominal

Fuente: INEC, 2018 (Encuesta Nacional de Empleo, Desempleo y Subempleo. Diciembre 2018)
Realizado por: Andrade Saltos, Vinicio, 2020

2.4. Población y muestra

La ***unidad de análisis*** se definió como el jefe de hogar que tenga un único trabajo en Ecuador; por lo tanto, la ***población*** en estudio fue definida como el conjunto de todos los jefes de hogar con un único trabajo en Ecuador, en diciembre de 2018. La información para la ***muestra*** se obtuvo mediante datos secundarios, es decir, información que fue recolectada y tabulada previamente, en este caso, por el Instituto Nacional de Estadísticas y Censos (INEC) mediante la Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU), realizada en diciembre del 2018 (http://www.ecuadorencifras.gob.ec/enemdu-2018/), tal encuesta utiliza un esquema probabilístico de panel de viviendas seleccionadas en una submuestra, con un total de individuos muestreados de 59 350.

2.5. Análisis Estadístico

2.5.1. Instrumentos de procesamiento y análisis de información

En la presente investigación se utilizó el software estadístico R (versión 3.5.3), mediante su Entorno de Desarrollo Integrado conocido como RStudio (versión 1.2.1335), durante todo el procesamiento, análisis y obtención de resultados. La interfaz de trabajo fue codificada en UTF-8, garantizando así que se conserven los caracteres propios del idioma español (tildes, letra ñ, etc.).

2.5.2. Obtención y análisis inicial de datos

Los datos correspondientes a la ENEMDU de diciembre - 2018 fueron descargados del sitio oficial del INEC, en formato SPSS (extensión .sav):

http://www.ecuadorencifras.gob.ec/documentos/web-inec/EMPLEO/2018/Diciembre-2018/BDD_ENEMDU_2018_12_SPSS.zip

Se cargó la información del archivo "201812_enemdubdd_15anios.sav", correspondiente al módulo "15 años". Se filtraron aquellos individuos que sean jefes de hogar y tengan un único trabajo, obteniendo una muestra de 12 464 unidades de análisis. Posteriormente, se seleccionaron las 18 variables en estudio (**Tabla 1-2**).

Los 17 individuos que respondieron “No sabe, no responde” en la variable respuesta (p59), fueron eliminados, dejando una muestra total de 12 447 unidades de análisis.

Finalmente, se transformaron las variables cuantitativas: “p03”, “p45”, “p51a”, “ingrl”; a tipo *numérico*, ya que, al importar la base de datos, todas las variables fueron reconocidas como variables de tipo *factor*.

2.5.3. Análisis de datos faltantes (NA´s)

Se analizó la posible existencia de datos faltantes, así, se encontró que las variables “p03”, “p27” y “ingrl” presentaron 1, 187 y 446 NA´s, lo cual, equivale a 0,008%, 1,5% y 3,58% de los datos, respectivamente.

Tomando en cuenta la baja proporción de NA’s, se decidió imputarlos mediante la *mediana* para las variables cuantitativas (“ingrl” y “p03”) y la *moda* para la variable cualitativa (“p27”).

2.5.4. Análisis exploratorio de datos (AED)

Se describió brevemente la evolución histórica que ha presentado la satisfacción laboral en el Ecuador, mediante un gráfico de barras agrupado. La información temporal corresponde a las ENEMDU de diciembre de: 2014, 2015, 2016, 2017 y 2018.

Se realizó un resumen estadístico descriptivo de todas las variables en estudio. El cual, para variables cuantitativas, consta de: media, mediana, valores mínimo y máximo, desviación estándar, rango intercuartílico, coeficiente de variación, número de valores distintos en cada variable, e histograma. Y para variables cualitativas consta de: frecuencia absoluta con su correspondiente porcentaje y diagrama de barras.

Finalmente, se estudió la dispersión de las variables cuantitativas, mediante diagramas de caja y bigotes (boxplot).

2.5.5. Preprocesamiento de datos

Las técnicas utilizadas, en su tarea de construir modelos de predicción, tienen en común que las variables explicativas cuantitativas, se ven ponderadas gracias a la magnitud de sus intensidades,

sobreestimando o subestimando su influencia en el modelo; es decir, aquellas variables que tengan valores en conjunto mayores (por ejemplo, la variable "ingresos") automáticamente tendrán una mayor influencia en el modelo, con respecto a aquellas variables con valores en conjunto menores (por ejemplo, la variable "edad"). Este hecho afecta el desempeño del modelo ya que no permite estudiar la naturaleza intrínseca de las variables y su influencia en el fenómeno estudiado. Por tal motivo, se decidió realizar una estandarización de todas las variables cuantitativas (técnica muy utilizada en el preprocesamiento de datos en modelos de machine learning); en este caso, se utilizó un tipo de estandarización que realiza un cambio de escala basándose en el valor mínimo y máximo de cada variable. La cual se define a continuación:

Sea *X* una variable estadística cuantitativa continua, la variable estandarizada *X'* se define como:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Dicha técnica garantizó que todas las variables cuantitativas continuas tengan un rango entre 0 y 1, sin perder su naturaleza intrínseca, dejándolas en iguales condiciones para interactuar en la creación del modelo predictivo.

Otro problema muy usual en técnicas de este tipo, es que las variables explicativas cualitativas suelen presentar demasiadas categorías, las cuales, en su mayoría, tienen frecuencias absolutas asociadas muy bajas. Por ejemplo, la variable "Sitio de trabajo" presenta 12 categorías, pero se espera que la mayoría de individuos se ubiquen solo en un pequeño subconjunto de éstas. Con esto en mente, se creó una técnica de visualización basada en un gráfico de Pareto, pero enfocada en el número de categorías y en la frecuencia porcentual acumulada asociada a cada una de ellas. Tal técnica tiene como argumentos de entrada: una variable de tipo factor y opcionalmente, el nombre de dicha variable; mientras que, como elemento de salida, se obtiene una gráfica de barras con las frecuencias porcentuales de las categorías de dicha variable, ordenadas de forma descendente, además de una ojiva que representa las frecuencias porcentuales acumuladas, con el número de cada categoría en los nodos de la ojiva y una recta horizontal que corta al eje *y* en el valor 90%. Se programó esta técnica como una función en RStudio, con el nombre *pareto_plot(x, title)*, su objetivo es el de visualizar, en las variables explicativas, aquellas categorías que en conjunto acumulan al menos el 90% (porcentaje referencial) del total de individuos, para posteriormente, colapsar las categorías que se crea conveniente en una única categoría llamada "Otra". De esta forma, es posible reducir dramáticamente el número de categorías presentes en las variables de tipo factor, sin dejar de lado aquellas categorías que sean representativas en la muestra de individuos, así, el desempeño de las técnicas en estudio mejora y, el tiempo y procesamiento que toma construir cada uno de los modelos se ve disminuido. Para complementar el análisis, se programó una función con el nombre *collapse_factor(x, freq)*, la cual, dada una

variable de tipo factor (*x*) y una proporción (*freq*), se encarga de conservar las categorías que, en conjunto, tienen una frecuencia porcentual menor a *freq*%, colapsando las demás categorías en una única categoría llamada "Otra". El análisis de Pareto para categorías y el posterior colapso de aquellas que sean poco representativas se realizó únicamente en las variables explicativas que tengan más de tres categorías.

Finalmente, mediante programación en R, se estableció la característica intrínseca de ordinalidad en la variable respuesta: p59 (¿Cómo se siente en su trabajo?), transformándola a tipo *Ord.factor*.

2.6. Metodologías de estadística clásica y de machine learning (Análisis cualitativo)

La presente investigación tiene como principal objetivo comparar las técnicas: regresión logística ordinal (estadística clásica) versus redes neuronales articiales y gradient boosting (ambas técnicas de machine learning), con esto en mente, se decidió caracterizar de forma general tanto a las técnicas de estadística clásica como a las técnicas de machine learning, con el fin de contextualizar el posterior análisis y aportar información relevante al estudio. Dicha caracterización se llevó a cabo en dos fases:

1) Se realizó un estudio de la terminología utilizada en ambas metodologías, y sus correspondientes equivalencias. Para ello se recopiló información de dos revisiones sistemáticas publicadas en 1994 y 2014. La intención además fue comparar la evolución de la terminología en el transcurso del tiempo.

2) Se compararon cualitativamente las características generalizables de mayor relevancia para las dos metodologías en estudio.

2.7. Preparación de información y construcción de modelos de clasificación

El código de R utilizado para la preparación de información y posterior construcción de los tres modelos predictivos se muestra en el **Anexo A**.

2.7.1. Muestras de entrenamiento y de prueba

Mediante la técnica conocida como *Stratified Tenfold Cross–Validation*, se dividió aleatoriamente a la muestra de individuos en 10 partes de aproximadamente el mismo tamaño, donde cada parte, se mantuvo proporcional a las categorías de la variable respuesta p59 (¿Cómo se siente en su trabajo?). Así, fue posible entrenar 9 partes y evaluar 1 parte, alternadamente, para un total de 10 modelos con cada una de las tres técnicas en estudio (Gradient boosting, Redes neuronales artificiales y Regresión logística ordinal). Así, se garantizó una medida más robusta de estimación para el error total de predicción, siendo éste, el promedio de los errores de predicción de los 10 modelos construidos con cada técnica.

2.7.2. Construcción de modelos de clasificación

2.7.2.1. Gradient boosting

La técnica de Gradient boosting es una extensión de la técnica de árboles de clasificación, por lo tanto, al ajustar un modelo con dicha técnica, el algoritmo selecciona aleatoriamente un subconjunto de variables, para de entre ellas, escoger la que divida de forma contundente a la muestra en cada uno de los nodos de clasificación que se van formando. Esta aleatoriedad propia del algoritmo se controló al construir 200 modelos, lo cual, se traduce en 200 bosques aleatorios con estructuras diferentes. De entre los 200 modelos, se seleccionó aquel que presenta el menor error total de predicción.

2.7.2.2. Redes neuronales artificiales

Al igual que en el apartado anterior, la técnica de Redes neuronales artificiales construye un modelo mediante un desarrollo aleatorio; en este caso, la aleatorización se encuentra en el proceso conocido como descenso del gradiente (gradient descent), el cual, busca minimizar los errores existentes en la etapa de aprendizaje; para ello, tal proceso parte de un conjunto de parámetros definidos en un inicio de forma aleatoria, con la intención de que dichos parámetros varíen de tal manera que el error del modelo llegue a ser mínimo. La aleatoriedad en los parámetros iniciales del modelo, se controló al construir, análogamente al apartado anterior, 200 modelos, para finalmente seleccionar aquel que presenta el menor error total de predicción.

2.7.2.3. Regresión logística ordinal

Mediante regresión logística ordinal se ajustó un modelo con odds proporcionales, utilizando la función "logit". A diferencia de los dos procedimientos expuestos en los apartados anteriores, esta técnica no presenta aleatoriedad intrínseca en su aplicación, este hecho se debe a que, al no ser un proceso iterativo, las ecuaciones que describen el modelo, están preestablecidas mediante fórmulas matemáticas que son generales para cualquier conjunto de datos que cumpla con los supuestos asociados al modelo, con esto en mente, el paso previo a la aplicación de esta técnica fue la comprobación de supuestos, los cuales se enumeran a continuación:

1. La variable dependiente ha sido medida en una escala ordinal.
2. Una o más de las variables independientes pueden ser de tipo: categórica, ordinal o continua.
3. No existe multicolinealidad.
4. Los odds son proporcionales (paralelismo).

2.8. Desempeño de las técnicas utilizadas en la construcción de los modelos de clasificación

El análisis del desempeño de las técnicas bajo estudio se realizó desde dos diferentes enfoques:

1. Se construyó la matriz de confusión asociada a cada modelo generado, la cual, se calculó con base en porcentajes, de esta forma se aseguró que la interpretación sea más simple e intuitiva. Subsecuentemente, se analizó el error total de predicción, para así, tener un criterio de evaluación que, en conjunto, describa claramente la calidad predictiva de cada modelo.
2. Se analizó de forma descriptiva la demanda de procesamiento computacional de cada una de las técnicas bajo estudio. Dicho análisis se llevó a cabo tomando en cuenta dos características que influyen en la demanda de procesamiento: 1) Tiempo que tarda cada modelo en ser compilado (segundos), 2) Pico de memoria RAM que se genera al procesar cada técnica (mebibytes).

   Para el cálculo del tiempo de compilación y pico de memoria RAM, se consideraron tres factores que, se presume, pueden generar variaciones en la demanda de procesamiento:

   a) *Tipo de muestra*. _ Se tomaron en cuenta dos muestras: 1) la muestra final con la cual se crearon los modelos bajo análisis, a la que se denominó "Muestra Real", 2) una

muestra con una numerosidad mucho mayor, la cual se etiquetó como "Muestra Aumentada", específicamente el tamaño de esta última es 10 veces el tamaño de la muestra real; la muestra aumentada se creó mediante un remuestreo aleatorio de la muestra real, con reposición. La intención de este procedimiento es analizar el comportamiento computacional de las técnicas aquí estudiadas, es decir, observar cómo se desempeña cada algoritmo al enfrentarse a una muestra masiva de datos. Es importante aclarar que la muestra aumentada tiene como único propósito comparar la demanda de procesamiento de cada técnica, mas no, mejorar la calidad predictiva de los modelos.

b) *Número de núcleos del procesador*. _ En el software estadístico R, por defecto las sentencias se ejecutan de forma secuencial, es decir, cuando se realizan varias tareas, éstas se desarrollan una tras otra, una nueva tarea empieza únicamente cuando la anterior es completada, por ello R utiliza un único núcleo del procesador de forma pre establecida; este hecho representa un particular problema cuando se construyen modelos mediante técnicas de remuestreo como el *Stratified Tenfold Cross–Validation* (técnica que se utiliza en el presente estudio), ya que los modelos finales construidos mediante técnicas de machine learning, son el resultado de correr decenas e incluso cientos de modelos, hasta minimizar el error de predicción. Lo ideal en este tipo de escenarios es que los algoritmos se ejecuten en paralelo, haciendo uso de todos los núcleos con los que cuenta el procesador, de esta forma, es posible ejecutar varias tareas simultáneamente, lo cual, a su vez, representa un ahorro importante en el coste de procesamiento asociado a cada técnica. Es así que, se decidió analizar la demanda de procesamiento al trabajar con un único núcleo (configuración por defecto), en comparación con la demanda que se obtiene al trabajar con todos los núcleos del procesador (configuración personalizada).

c) *Técnica utilizada*. _ La técnica a utilizarse (gradient boosting, redes neuronales artificiales, regresión logística ordinal) es el factor principal bajo estudio, ya que el objetivo de este apartado es conocer cómo se comporta cada técnica, en cuanto a su respectiva demanda de procesamiento.

Se corrieron diez veces las sentencias correspondientes a cada combinación de: tipo de muestra, número de núcleos y técnica utilizada, se registró el tiempo y pico de memoria RAM, y con los diez datos obtenidos para cada combinación, se calculó su correspondiente promedio. Posteriormente se realizó un gráfico de medias, tanto para el tiempo de procesamiento como para el pico de memoria RAM, dicho gráfico se estructuró para que sea posible visualizar los cambios en el rendimiento computacional, al cambiar los niveles de los factores antes mencionados.

Es importante resaltar el hecho de que el análisis de la demanda de procesamiento computacional depende directamente del hardware de cada computador, por lo tanto, los resultados aquí obtenidos son específicos para la computadora con la que se realizaron los cálculos, en todo caso, tales resultados pueden entenderse como un marco referencial a la hora de extrapolar su usabilidad a otros dispositivos informáticos. Con esto en mente, se presentan a continuación las especificaciones del computador (laptop) con el que se realizó este estudio:

- Procesador: Intel® Core™ i7-9750H 2.6GHz (siete núcleos)
- Memoria RAM: Samsung DDR4-2666 16GB
- Almacenamiento: 512GB M.2 SSD
- Tarjeta gráfica: NVIDIA® GeForce® GTX 1660Ti GDDR6 6GB
- Sistema operativo: Windows 10 Pro

Además, se controlaron ciertos factores en el experimento, con el fin de no incluir variabilidad externa en los resultados y que éstos puedan ser extrapolables. Dichos factores se presentan a continuación:

- *Conexión a energía eléctrica.* _ El computador se mantuvo conectado a energía eléctrica mientras se realizaban las pruebas, ya que las computadoras portátiles disminuyen su poder de procesamiento al alimentarse únicamente de la batería.
- *Acceso a internet.* _ No se permitió el acceso a internet mientras se realizaban las pruebas, ya que, al estar conectado a internet, el sistema operativo Windows en ocasiones descarga automáticamente actualizaciones del sistema, lo cual representa una carga extra en el procesamiento.
- *Estrangulamiento térmico*. _ Las pruebas se efectuaron con períodos de pausa entre cada corrida de las sentencias que se encargan de construir los modelos, el objetivo de esta medida es evitar el estrangulamiento térmico, es decir, evitar que el procesador alcance altas temperaturas por el uso constante y por ende disminuya su desempeño.
- *Programas en segundo plano*. _ En todo momento se controló la posible ejecución de programas en segundo plano, permitiendo únicamente que se ejecuten los programas que necesita el sistema operativo para funcionar y el software necesario para realizar las pruebas (R y RStudio).

2.9. Post-procesamiento de datos

Una vez identificada la técnica que genera un modelo predictivo con el mínimo error total de predicción, se realizó un análisis de reducción de la dimensionalidad, el objetivo de este análisis es potenciar el modelo para que obtenga mayor confiabilidad predictiva con la mínima demanda de información. La reducción de la dimensionalidad se realizó mediante el análisis de curvas ROC para cada una de las variables explicativas. Usualmente cuando la variable respuesta tiene dos clases, se aplica una serie de cortes a los datos de las variables predictoras para predecir cada clase, así, se calculan la sensibilidad y la especificidad para cada corte y en base a estas medidas se construye la curva ROC. Para poder comparar la importancia individual de cada variable predictora, se calcula el área bajo la curva ROC generada por cada predictor, este cálculo se realiza mediante la regla trapezoidal. En los problemas donde la variable respuesta presenta más de dos clases, como el caso que se está tratando en el presente estudio, la variable explicada se descompone en variables dicotómicas (binarias) y se comparan de dos en dos los resultados. Posteriormente, se realiza un cambio de escala en la importancia que presenta cada variable explicativa, con el fin de que se mantenga en un rango de 0 a 100, donde una puntuación de 100 se interpreta como la importancia más alta.

Los resultados de la importancia de cada variable explicativa (dividida en sus correspondientes categorías) se visualizaron mediante dos gráficos de barras, en el primero se presentaron las 35 variables más importantes, mientras que, en el segundo, se observaron las 25 variables menos importantes.

Finalmente, se eliminaron aquellas variables explicativas que presentaron, en conjunto, categorías poco importantes. El proceso de eliminación se realizó para cada variable, siempre y cuando el error total de predicción del modelo no se vea fuertemente afectado.

El estudio de importancia de variables sirvió también como un punto de partida para conocer aquellas variables que tienen mayor relevancia en la creación de un modelo que describa el comportamiento de la satisfacción laboral en Ecuador.

2.10. Construcción de una aplicación web interactiva para medir la satisfacción laboral

El modelo predictivo con el menor error total de predicción y los resultados obtenidos fueron materializados en una aplicación web interactiva, la cual, tiene como objetivo medir el nivel de satisfacción laboral en los jefes de hogar ecuatorianos con un único trabajo. Para su construcción

se utilizó el paquete *shiny* de R, el cual, combina el potencial estadístico del software, con la versatilidad de internet.

La interfaz de la aplicación cuenta con una breve descripción, las instrucciones de uso y un formulario definido según las variables identificadas como importantes en el modelo predictivo más confiable. Al llenar el formulario, el modelo se encarga de predecir el nivel de satisfacción laboral del individuo que está haciendo uso de la aplicación.

Es posible ingresar a la aplicación web desde cualquier dispositivo con acceso a internet, por medio del siguiente link: https://vaashub.shinyapps.io/satisfaccion_laboral/. La interfaz de usuario para ordenadores y dispositivos inteligentes se muestra en el **Anexo B**.

Es importante aclarar que dicha aplicación tendrá asociado el error total de predicción del modelo con el cual fue creada, por lo que sus estimaciones no serán perfectas. Además, el modelo se basa en el supuesto de que los datos recopilados no presentan sesgo de muestreo, y ya que los datos no fueron recolectados por el investigador, sino que se utilizó información recopilada por el INEC, este supuesto no puedo ser corroborado, lo cual, pudiere traducirse en un error extra en cada predicción.

CAPÍTULO III

3. MARCO DE RESULTADOS, DISCUSIÓN Y ANÁLISIS DE RESULTADOS

3.1. Evolución histórica de la satisfacción laboral en Ecuador

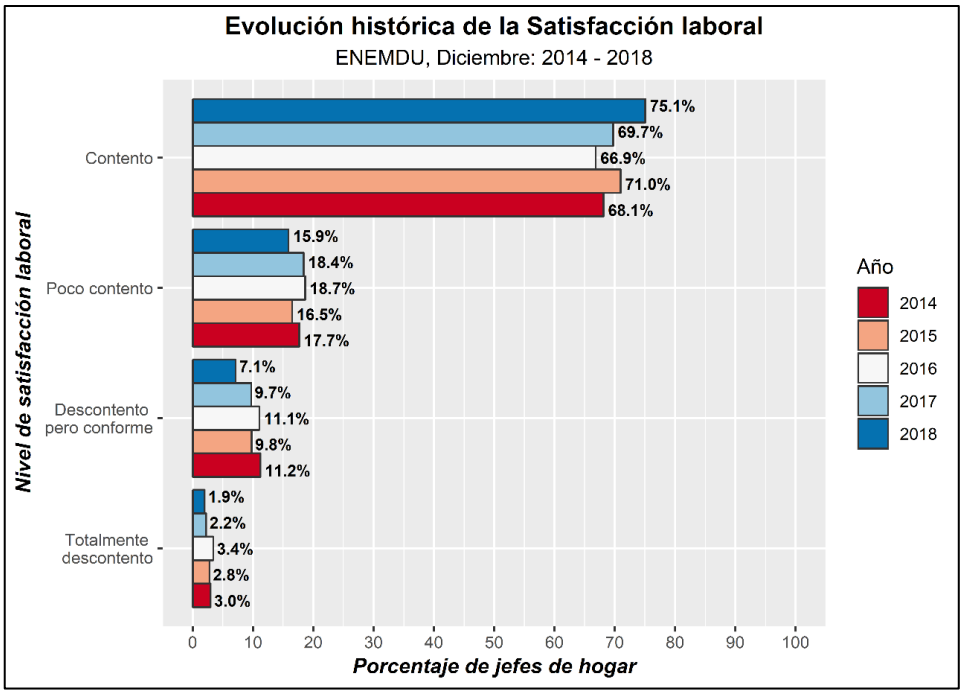


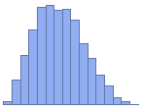
Gráfico 1-3: Revisión histórica de la Satisfacción laboral (gráfico de barras agrupado)

Realizado por: Andrade Saltos, Vinicio, 2020

Como se observa en el **Gráfico 1-3**, la satisfacción laboral en Ecuador (variable respuesta) ha sido un fenómeno estable a lo largo del tiempo, es decir, el patrón de proporciones es muy similar desde diciembre del 2014 hasta diciembre del 2018, siendo el nivel "Contento" el de mayor frecuencia, seguido de los niveles "Poco contento", "Descontento pero conforme" y siempre con una proporción muy baja, el nivel "Totalmente descontento". Se observa además que, en comparación con los demás años, en 2016 se presenta la menor proporción de jefes de hogar que están contentos con su trabajo (66,9%) y la mayor proporción de aquellos que están totalmente descontentos con su trabajo (3,4%), mientras que, en 2018 se presenta la mayor proporción de jefes de hogar que están contentos con su trabajo (75,1%) y la menor proporción de aquellos que están totalmente descontentos con su trabajo (1,9%).

3.2. Análisis descriptivo

Tabla 1-3: Resumen descriptivo

Variable	Estadísticas / Valores	Freq (% de válidos)	Gráfico
p59 [factor]	1. Contento	9344 (75.1%)	
	2. Descontento pero conforme	886 (7.1%)	
	3. Poco contento	1975 (15.9%)	
	4. Totalmente descontento	242 (1.9%)	
p02 [factor]	1. Hombre	9566 (76.8%)	
	2. Mujer	2881 (23.2%)	
p03 [num]	Mean (sd) : 48.9 (14.2) min < med < max: 16 < 48 < 94 IQR (CV) : 21 (0.3)	79 valores distintos	
p06 [factor]	1. Casado(a)	5464 (43.9%)	
	2. Separado(a)	1200 (9.6%)	
	3. Divorciado(a)	582 (4.7%)	
	4. Viudo(a)	755 (6.1%)	
	5. Unión libre	3102 (24.9%)	
	6. Soltero(a)	1344 (10.8%)	
p14 [factor]	1. Sólo lengua indígena	19 (0.1%)	
	2. Lengua indígena y español	784 (6.3%)	
	3. Sólo español	11350 (91.2%)	
	4. Español e idioma extranje	286 (2.3%)	
	5. Lengua indígena e idioma	3 (0.0%)	
	6. Idioma extranjero	5 (0.0%)	
	7. No habla	0 (0.0%)	
p15 [factor]	1. Indígena	908 (7.3%)	
	2. Afroecuatoriano	120 (1.0%)	
	3. Negro	133 (1.1%)	
	4. Mulato	127 (1.0%)	
	5. Montubio	758 (6.1%)	
	6. Mestizo	10213 (82.0%)	
	7. Blanco	182 (1.5%)	
	8. Otro, cual	6 (0.0%)	
p27 [factor]	1. Trabajar más horas en su	1030 (8.3%)	
	2. Trabajar más horas en otr	979 (7.9%)	
	3. Cambiar el trabajo por ot	658 (5.3%)	
	4. No desea trabajar más hor	9780 (78.6%)	
p42 [factor]	1. Empleado de gobierno		
	2. Empleado privado	1036 (8.3%)	
	3. Empleado terciarizado	3158 (25.4%)	
	4. Jornalero o peón	0 (0.0%)	
	5. Patrono	1446 (11.6%)	
	6. Cuenta Propia	631 (5.1%)	
	7. Trabajador del hogar no r	5745 (46.2%)	
	8. Trabajador no del hogar n	159 (1.3%)	
	9. Ayudante no remunerado de	11 (0.1%)	
	10. Empleado(a) Doméstico(a)	3 (0.0%) 258 (2.1%)	
p45 [num]	Mean (sd) : 17.8 (15.7) min < med < max: 0 < 14 < 80 IQR (CV) : 22 (0.9)	80 valores distintos	

p46 [factor]	1. Local patrono	3720	(29.9%)
	2. Obra en construcción	636	(5.1%)
	3. Se desplaza	1139	(9.2%)
	4. En la calle	220	(1.8%)
	5. Kiosco calle	46	(0.4%)
	6. Local propio o arrendado	1430	(11.5%)
	7. Local cooperativa u asoci	0	(0.0%)
	8. Vivienda distinta a la su	720	(5.8%)
	9. Su vivienda	993	(8.0%)
	10. Finca o terreno	2269	(18.2%)
	[2 más…]	1274	(10.2%)
p10a [factor]	1. Ninguno	574	(4.6%)
	2. Centro de alfabetización	66	(0.5%)
	3. Jardín de infantes	0	(0.0%)
	4. Primaria	5232	(42.0%)
	5. Educación Básica	115	(0.9%)
	6. Secundaria	3972	(31.9%)
	7. Educación Media	157	(1.3%)
	8. Superior no universitario	226	(1.8%)
	9. Superior Universitario	1877	(15.1%)
	10. Post-grado	228	(1.8%)
p47a [factor]	1. Menos de 100	10277	(82.6%)
	2. 100 y más	2170	(17.4%)
p51a [num]	Mean (sd) : 39.8 (13.1) min < med < max: 1 < 40 < 108 IQR (CV) : 10 (0.3)	83 valores distintos	
p71a [factor]	1. Si	395	(3.2%)
	2. No	12052	(96.8%)
grupo1 [factor]	1. Personal direct./admin. p	209	(1.7%)
	2. Profesionales científicos	874	(7.0%)
	3. Técnicos y profesionales	560	(4.5%)
	4. Empleados de oficina	324	(2.6%)
	5. Trabajad. de los servicio	2242	(18.0%)
	6. Trabajad. calificados agr	3515	(28.2%)
	7. Oficiales operarios y art	1797	(14.4%)
	8. Operadores de instalac. m	1264	(10.2%)
	9. Trabajadores no calificad	1613	(13.0%)
	10. Fuerzas Armadas	49	(0.4%)
	11. No especificado	0	(0.0%)
ingrl [num]	Mean (sd) : 519.6 (989.4) min < med < max: 0 < 390 < 83000 IQR (CV) : 400 (1.9)	1323 valores distintos	
rama1 [factor]	1. A. Agricultura, ganadería		
	2. B. Explotación de minas y	3853	(31.0%)
	3. C. Industrias manufacture	105	(0.8%)
	4. D. Suministros de electri	1311	(10.5%)
	5. E. Distribución de agua,	44	(0.4%)
	6. F. Construcción	55	(0.4%)
	7. G. Comercio, reparación v	854	(6.9%)
	8. H. Transporte y almacenam	2038	(16.4%)
		949	(7.6%)
	9. I. Actividades de alojami	659	(5.3%)
	10. J. Información y comunica	112	(0.9%)
		2467	(19.8%)
	[12 más…]		
secemp [factor]	1. Sector Formal	6024	(48.4%)
	2. Sector Informal	5637	(45.3%)
	3. Empleo Doméstico	258	(2.1%)
	4. No Clasificados por Secto	528	(4.2%)

Fuente: INEC, 2018 (Encuesta Nacional de Empleo, Desempleo y Subempleo. Diciembre 2018)
Realizado por: Andrade Saltos, Vinicio, 2020

Como se observa en la **Tabla 1-3**, el estudio se realiza con 4 variables cuantitativas y 14 cualitativas.

Las características que destacan entre las variables cuantitativas son:

- Las variables edad (p03) y horas de trabajo por semana (p51a) presentan histogramas que reflejan una posible aproximación a la distribución normal. Mientras que, las variables años de trabajo (p45) e ingresos del trabajo (ingrl) presentan sesgo de cola derecha, siendo esta última la que tiene un sesgo extremadamente pronunciado.
- El promedio de edad entre los jefes de hogar con un único trabajo es de 48 años, con una media de 40 horas semanales de trabajo, aproximadamente. Se tiene que el 50% de los individuos no presentan más de 14 años de trabajo ni un ingreso mensual mayor a 390 dólares.

Con respecto a las variables cualitativas, se tiene que:

- Los jefes de hogar con un único trabajo, en su mayoría: están contentos con su trabajo, son hombres, están casados, hablan únicamente español, se consideran mestizos, no desean trabajar más horas, tienen cuenta propia como categoría de ocupación, trabajan en el local del patrono, tienen educación primaria, trabajan en un lugar con menos de 100 personas, no han recibido ingresos derivados del capital, son trabajadores calificados agropecuarios y pesqueros, trabajan dentro del sector formal y su rama de actividad es la agricultura, ganadería, caza, silvicultura, y pesca.
- Se observa además que las variables que presentan un mayor número de categorías son: “rama1”, “p46”, “grupo1”, “p42” y “p10a” con 22, 12, 11, 10 y 10 categorías, respectivamente.

3.3. Análisis de dispersión

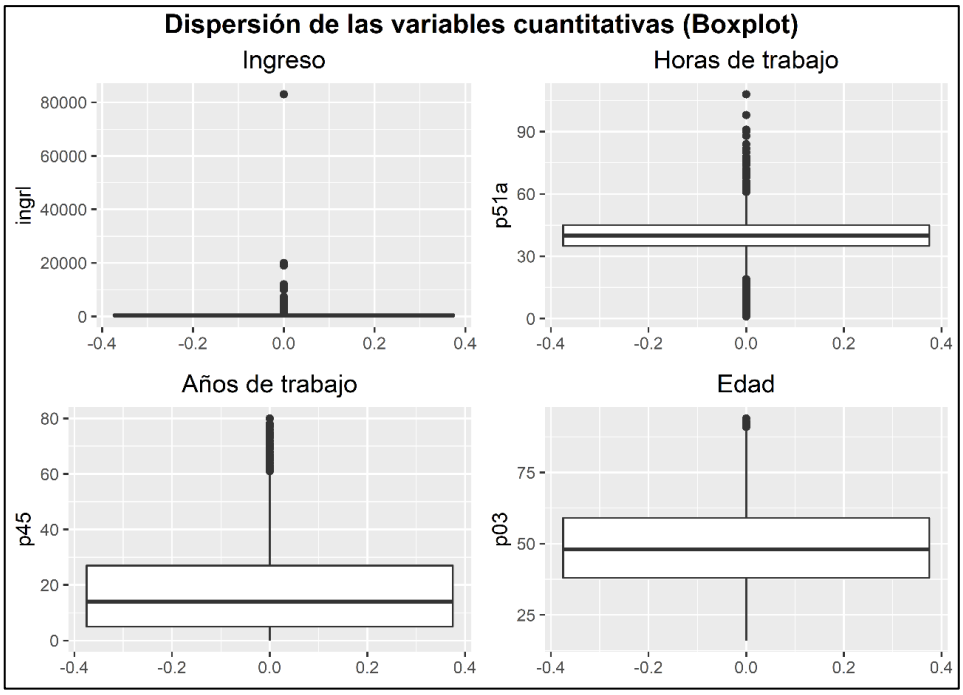


Gráfico 2-3: Dispersión de variables cuantitativas (boxplot)

Realizado por: Andrade Saltos, Vinicio, 2020

Como se observa en el **Gráfico 2-3**, las variables cuantitativas presentan individuos que se sospecha podrían tener valores atípicos (desde un punto de vista univariante). Si bien las tres técnicas aquí estudiadas son técnicas no paramétricas, se consideró importante para el desempeño de los modelos predictivos, eliminar aquellos individuos con valores extremadamente atípicos, en este caso, la variable "Ingreso" es aquella que presenta mayor variabilidad, este hecho está en total concordancia con el fenómeno económico conocido como acumulación del capital, el cual, tiene como consecuencia que muchos individuos tengan muy pocos ingresos y muy pocos individuos tengan ingresos estratosféricos. Por lo tanto, se decidió eliminar a aquellos individuos con ingresos mensuales mayores a 10 000 dólares; curiosamente, pero sin ser una sorpresa, únicamente 11 individuos presentaron ingresos superiores a este techo, dejando así, una muestra final de 12 436 unidades de análisis.

3.4. Análisis de Pareto para categorías

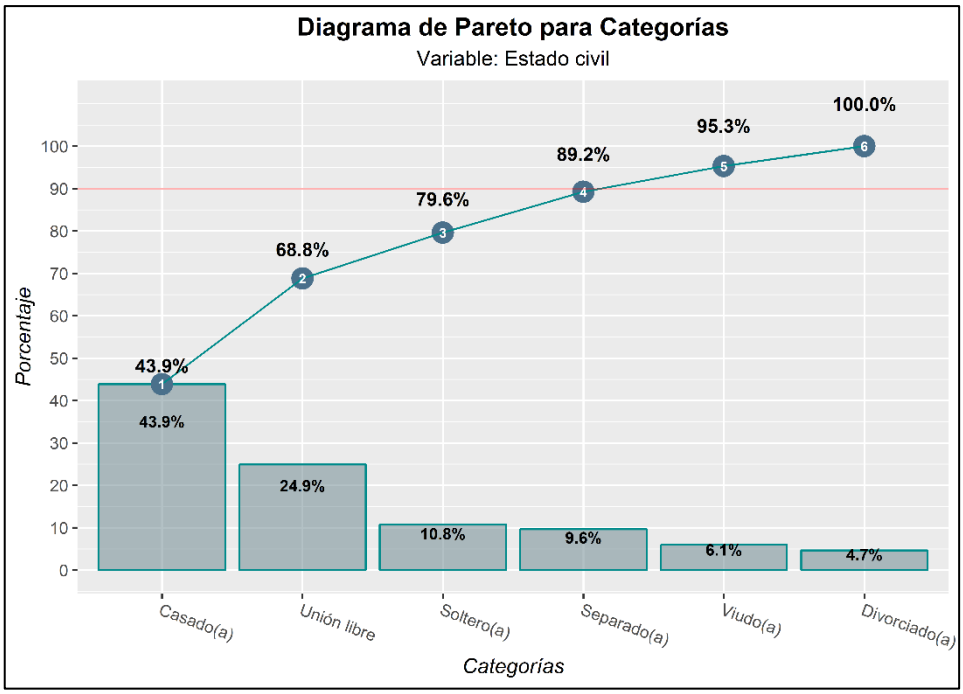


Gráfico 3-3: Pareto para categorías, variable "Estado civil"

Realizado por: Andrade Saltos, Vinicio, 2020

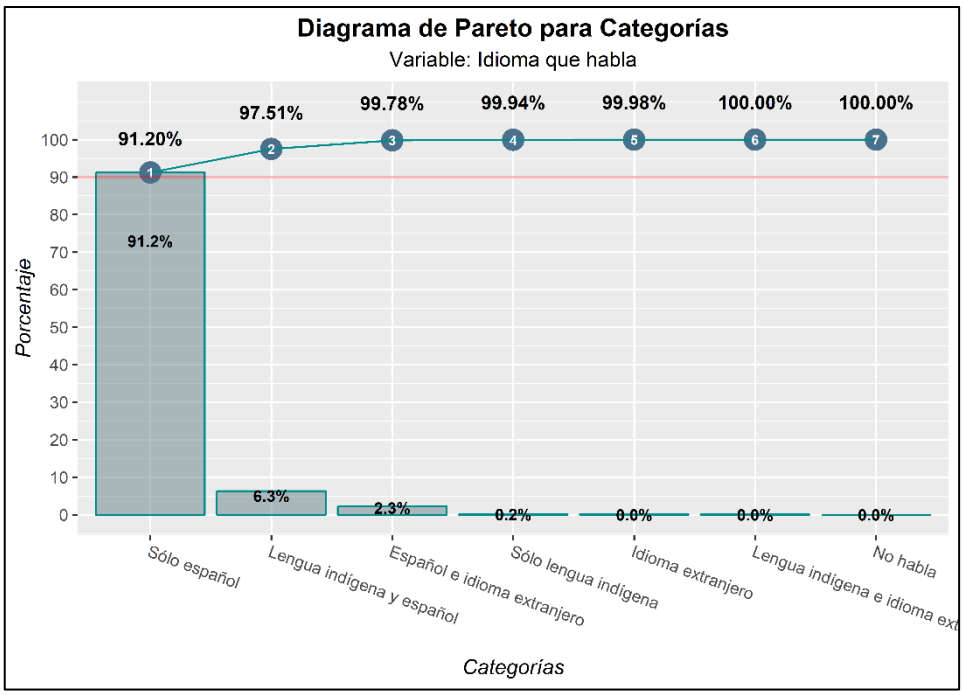


Gráfico 4-3: Pareto para categorías, variable "Idioma que habla"

Realizado por: Andrade Saltos, Vinicio, 2020

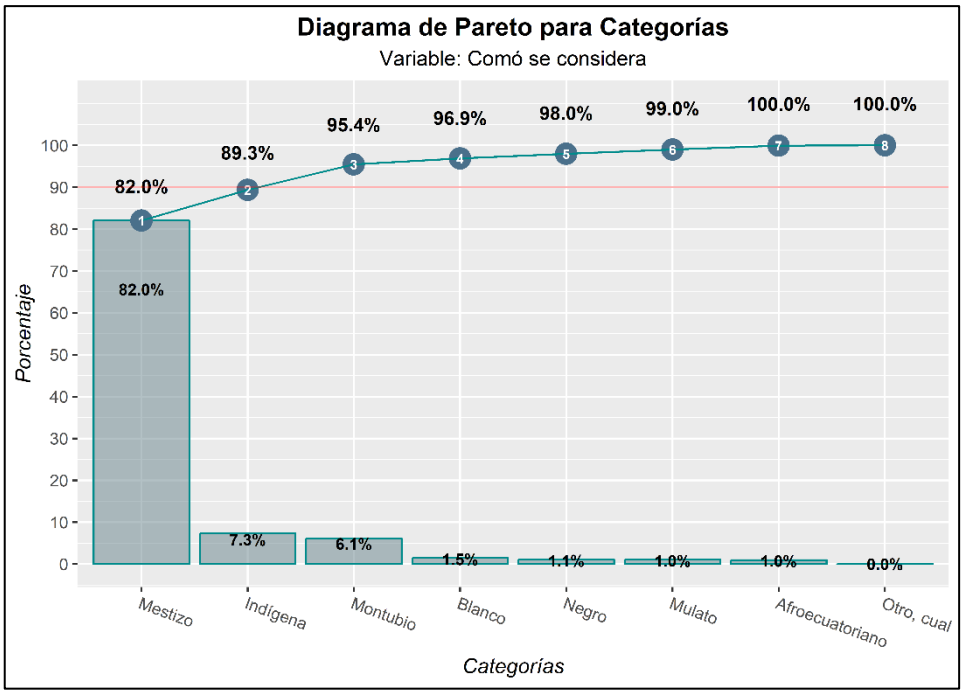


Gráfico 5-3: Pareto para categorías, variable "Cómo se considera"

Realizado por: Andrade Saltos, Vinicio, 2020

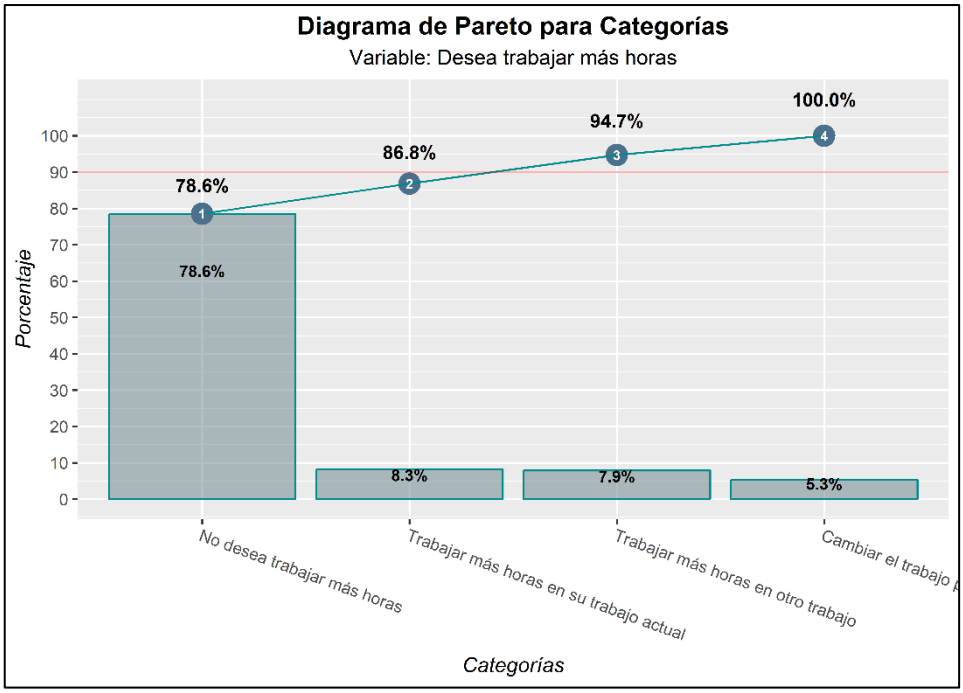


Gráfico 6-3: Pareto para categorías, variable "Desea trabajar más horas"

Realizado por: Andrade Saltos, Vinicio, 2020

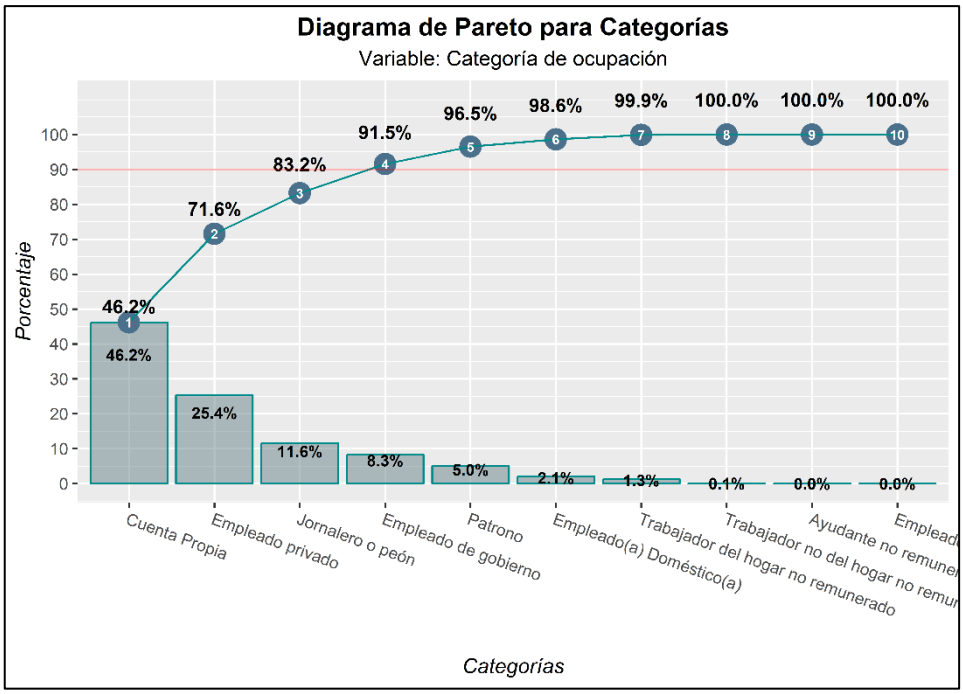


Gráfico 7-3: Pareto para categorías, variable "Categoría de ocupación"

Realizado por: Andrade Saltos, Vinicio, 2020

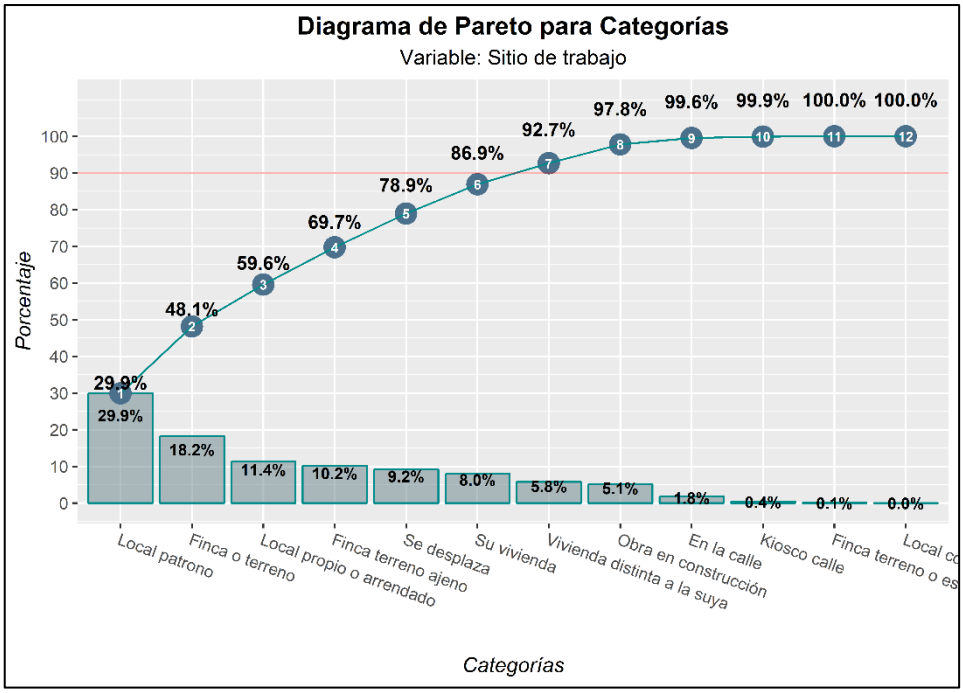


Gráfico 8-3: Pareto para categorías, variable "Sitio de trabajo"

Realizado por: Andrade Saltos, Vinicio, 2020

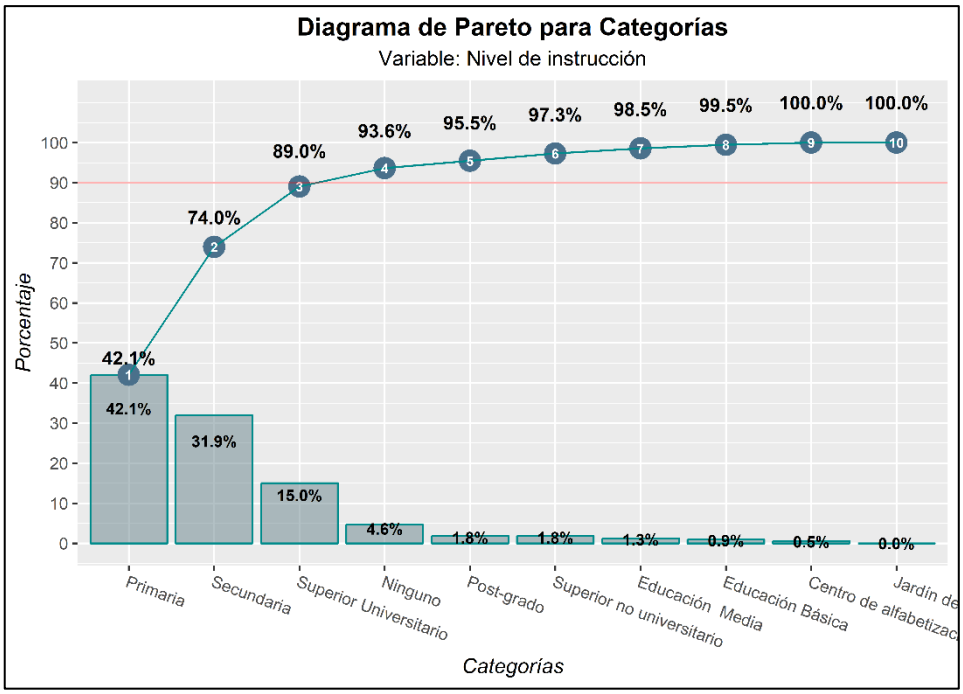


Gráfico 9-3: Pareto para categorías, variable "Nivel de instrucción"

Realizado por: Andrade Saltos, Vinicio, 2020

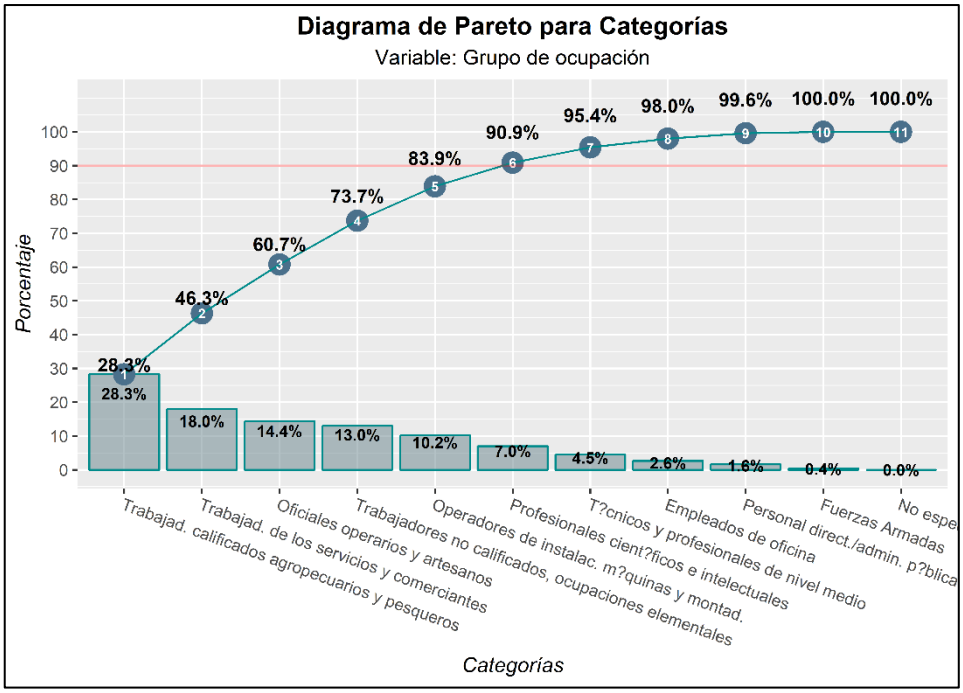


Gráfico 10-3: Pareto para categorías, variable "Grupo de ocupación"

Realizado por: Andrade Saltos, Vinicio, 2020

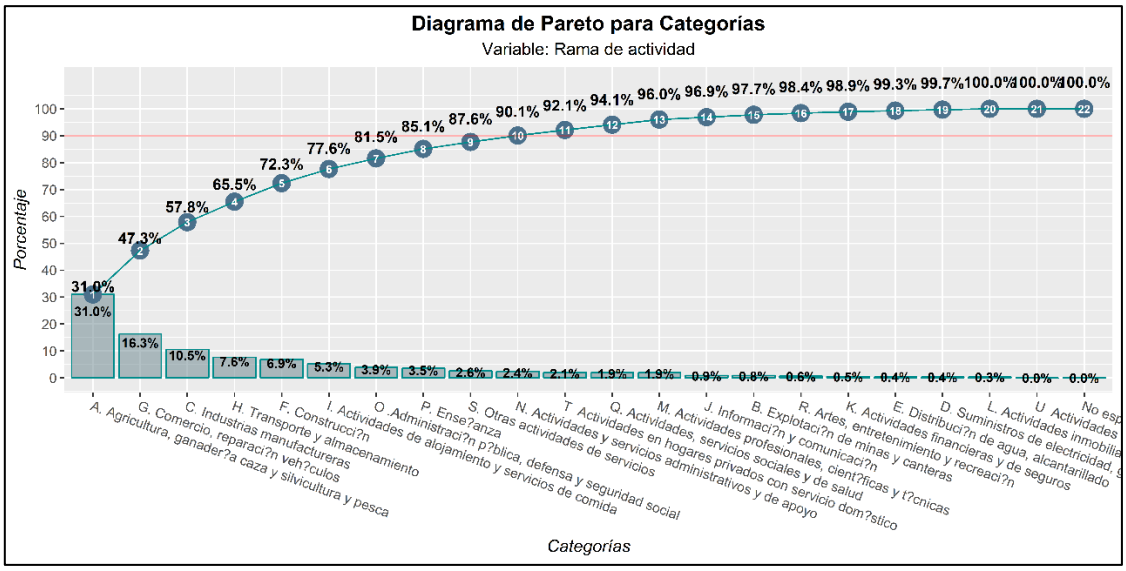


Gráfico 11-3: Pareto para categorías, variable "Rama de actividad"

Realizado por: Andrade Saltos, Vinicio, 2020

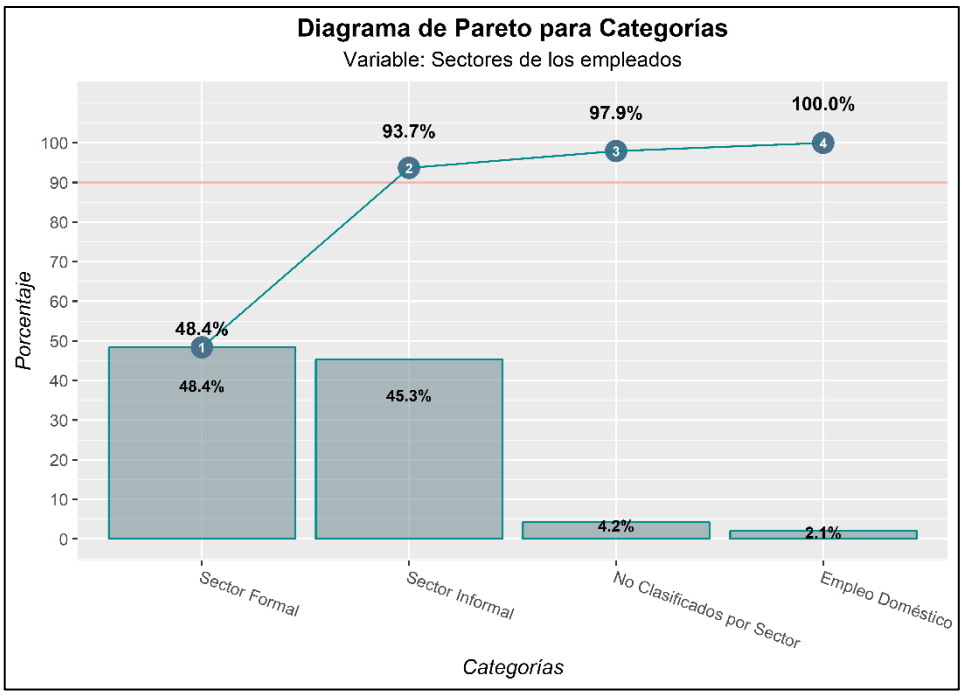


Gráfico 12-3: Pareto para categorías, variable "Sectores de los empleados"

Realizado por: Andrade Saltos, Vinicio, 2020

Como se observa en los gráficos: **Gráfico 3-3** al **Gráfico 12-3**, un gran número de variables presentan categorías poco representativas, por lo tanto, se decidió colapsarlas. La descripción de las variables y sus correspondientes categorías colapsadas se muestran en la **Tabla 2-3**.

Tabla 2-3: Descripción de variables y sus correspondientes categorías colapsadas

Variable	Número total de categorías	Número de categorías colapsadas	Porcentaje acumulado de las categorías colapsadas
Idioma que habla (p14)	7	5	2,49%
Cómo se considera (p15)	8	5	4,60%
Categoría de ocupación (p42)	10	5	3,50%
Sitio de trabajo (p46)	12	5	7,30%
Nivel de instrucción (p10a)	10	6	6,40%
Grupo de ocupación (grupo1)	11	5	9,10%
Rama de actividad (rama1)	22	12	9,90%

Fuente: Análisis de Pareto para categorías
Realizado por: Andrade Saltos, Vinicio, 2020

Es importante aclarar que se omitió el proceso de colapso de categorías en aquellos casos donde las categorías a ser colapsadas eran únicamente dos.

3.5. Caracterización de las metodologías de estadística clásica y de machine learning

3.5.1. Terminología utilizada en técnicas clásicas de estadística y su equivalente en técnicas de machine learning

La metodología estadística y la metodología de machine learning, guardan una estrecha relación en cuanto a los objetivos que se buscan alcanzar mediante su aplicación, y también al procedimiento de gestión de datos. Es así que ambas metodologías presentan equivalencias en su terminología. En la **Tabla 3-3** se muestran las equivalencias más representativas:

Tabla 3-3: Equivalencia en terminología

Metodología Estadística	Metodología de Machine Learning
Variable independiente	Variable de entrada, características
Variable dependiente	Variable de salida, etiquetas
Residuos	Errores
Proceso de estimación	Proceso de entrenamiento, aprendizaje y adaptabilidad
Criterio de estimación	Función de error, función de coste
Observaciones	Pares de entrenamiento
Parámetros por estimar	Pesos
Transformaciones	Enlaces funcionales
Análisis exploratorio de datos	Limpieza, depuración de datos
Modelo	Redes, grafos
Gráficas estadísticas	Visualización de datos
Técnicas de dependencia	Aprendizaje supervisado

Técnicas de interdependencia	Aprendizaje no supervisado

Fuente: Adaptado de Sarle (1994) y James et al. (2014)
Realizado por: Andrade Saltos, Vinicio, 2020

Al tomar en cuenta la evolución de la terminología, resultó que entre 1994 y 2014 no existieron cambios importantes, las expresiones siguen siendo las mismas y su usabilidad continua en vigencia.

3.5.2. Comparativa cualitativa entre metodología de estadística clásica y metodología de machine learning

En la **Tabla 4-3** se presentan las principales características que estructuran la metodología de estadística clásica, en comparación con aquellas características propias de machine learning.

Tabla 4-3: Comparativa cualitativa

	Metodología Estadística Clásica	**Metodología de Machine Learning**
Enfoque	Los métodos estadísticos generalmente están definidos en términos de ecuaciones matemáticas.	La metodología de machine learning, usualmente se define con base en su arquitectura y sus algoritmos de aprendizaje
Datos	En general, se da un tratamiento a los datos con el supuesto de que fueron generados mediante procesos estocásticos.	Se asume que la información fue generada por mecanismos de los cuales se desconoce su dinámica.
Objetivos	En el proceso de generar un modelo predictivo, se pretenden estudiar ciertas características sobre los datos y su estructura, por ejemplo, interpretar efectos marginales y los signos de los coeficientes, estudiar elasticidad y propiedades de los estimadores, etc.	Su principal objetivo a la hora de modelar un determinado fenómeno bajo estudio, es consolidar estimadores eficientes en términos de calidad predictiva,.
Proceso a seguir al modelar un fenómeno	Se pretende construir un modelo final que por sí solo se encargue de modelar la realidad.	Usualmente resulta en un conjunto de modelos interconectados que caracterizan la realidad, estudiando por separado ciertas partes naturalmente definidas en el fenómeno bajo análisis.
Flexibilidad	Varias técnicas inferenciales asumen ciertos supuestos que condicionan la naturaleza de los datos.	La forma funcional del modelado se aproxima mediante la fase de entrenamiento, lo cual es independiente de la naturaleza intrínseca de los datos.

Interpretabilidad	Las ecuaciones que describen el fenómeno son de fácil interpretación.	Los procesos generados al modelar un fenómeno en ocasiones son tan complejos que no es posible interpretarlos, es decir se convierten en black boxes.
Curva de aprendizaje	Generalmente, el aprendizaje de estadística se lleva a cabo de forma explícitamente teórica, con connotaciones matemáticas altamente formales.	El aprendizaje de machine learning principalmente se realiza mediante software especializado, el cual se encarga de simplificar la mayor parte del trabajo.

Fuente: Adaptado de Karlaftis y Vlahogianni (2011)
Realizado por: Andrade Saltos, Vinicio, 2020

3.6. Modelos de clasificación

3.6.1. Gradient boosting

La técnica de Gradient boosting presentó un error total de predicción de 24,27%, la **Tabla 5-3** muestra la matriz de confusión asociada a este modelo.

Tabla 5-3: Matriz de confusión Gradient boosting

	Totalmente descontento	Descontento pero conforme	Poco contento	Contento
Totalmente descontento	0.02	0.01	0.03	0.02
Descontento pero conforme	0.00	0.01	0.01	0.03
Poco contento	0.55	1.09	2.73	2.03
Contento	1.38	6.02	13.11	72.97

Fuente: Modelo Gradient boosting
Realizado por: Andrade Saltos, Vinicio, 2020

3.6.2. Redes neuronales artificiales

La técnica de Redes neuronales artificiales presentó un error total de predicción de 24,47%, la **Tabla 6-3** muestra la matriz de confusión asociada a este modelo.

Tabla 6-3: Matriz de confusión Redes neuronales artificiales

	Totalmente descontento	Descontento pero conforme	Poco contento	Contento
Totalmente descontento	0.00	0.00	0.00	0.00
Descontento pero conforme	0.00	0.00	0.00	0.00
Poco contento	0.44	0.96	2.40	1.92
Contento	1.50	6.17	13.48	73.13

Fuente: Modelo RNA
Realizado por: Andrade Saltos, Vinicio, 2020

3.6.3. Regresión logística ordinal

Comprobación de supuestos

Como se observa en la operacionalización de variables (**Tabla 1-2**), los supuestos 1 y 2 asociados a esta técnica, son evidentes; en contraste, los supuestos 3 y 4 conllevan un análisis más profundo:

Multicolinealidad

El supuesto de multicolinealidad se comprobó mediante el factor de inflación de la varianza (VIF), el cual, cuantifica la intensidad de la multicolinealidad existente en la construcción de un modelo. El **Gráfico 13-3** muestra los resultados del VIF por cada una de las variables explicativas.

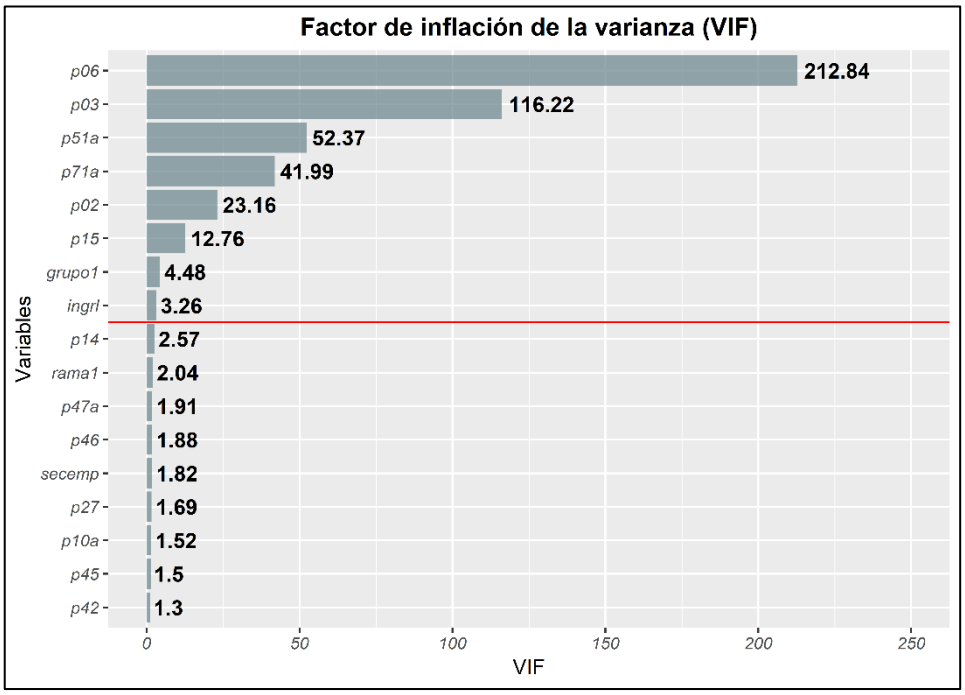


Gráfico 13-3: Factor de inflación de la varianza (VIF)

Realizado por: Andrade Saltos, Vinicio, 2020

Las variables que presentan valores de VIF mayores que 3 se consideran variables con problemas de multicolinealidad. Como se observa en el **Gráfico 13-3**, existen 8 variables con este problema, por lo que se decidió eliminar aquellas variables que no sean significativas en el modelo y que además tengan un VIF mayor a 3, de esta forma se eliminaron las variables: "p02", "p03", "p06", "p15" y "p71a".

Odds proporcionales (paralelismo)

El supuesto de paralelismo se comprobó mediante un test desarrollado en el programa "Métodos cuantitativos en ciencias sociales" (QMSS) de la universidad de Columbia, utilizando la función *propOddsTest(fit, relaxed.fit)* disponible en el paquete QMSS (github). Tal función compara un modelo que asume el supuesto de paralelismo, con un modelo similar que no está sujeto a la asunción de paralelismo. Dicho test establece como hipótesis nula que el modelo bajo el supuesto de paralelismo se ajusta de mejor manera a los datos en estudio. El resultado del test mostró un valor p contundente, de aproximadamente 1, por lo que se concluyó que el supuesto de paralelismo sí se cumple.

Resultados (Regresión logística ordinal)

Luego de comprobar y subsanar las violaciones de los supuestos asociados a la técnica de Regresión logística ordinal (Odds proporcionales), se procedió a construir y evaluar el modelo. Así, se obtuvo un error total de predicción de 24,69%, la **Tabla 7-3** muestra la matriz de confusión asociada a este modelo.

Tabla 7-3: Matriz de confusión Regresión logística ordinal

	Totalmente descontento	Descontento pero conforme	Poco contento	Contento
Totalmente descontento	0.00	0.00	0.00	0.00
Descontento pero conforme	0.05	0.08	0.09	0.07
Poco contento	0.12	0.18	0.53	0.28
Contento	1.78	6.86	15.25	74.70

Fuente: Modelo Regresión logística ordinal
Realizado por: Andrade Saltos, Vinicio, 2020

3.7. El problema de un conjunto desbalanceado de datos y el sobreajuste

En la construcción de modelos de clasificación, se dice que un conjunto de datos está desbalanceado, cuando la variable respuesta presenta una frecuencia absoluta muy alta en un pequeño número de sus categorías. Como se observa en el **Gráfico 14-3**, la variable respuesta p59 (satisfacción laboral), se compone de un 75,1% de individuos que se identifican con la

categoría “Contento”, mientras que las otras tres categorías suman en conjunto únicamente el 24,9% del total de individuos. Este hecho se traduce en un problema de datos desbalanceados (imbalanced data).

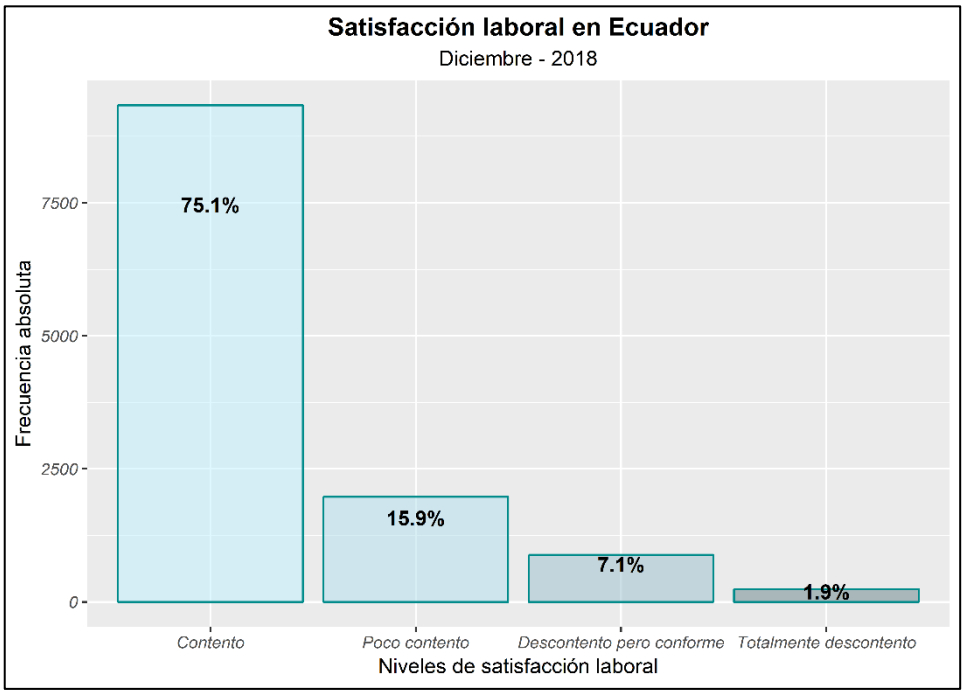


Gráfico 14-3: Satisfacción laboral en Ecuador (Diciembre – 2018)

Realizado por: Andrade Saltos, Vinicio, 2020

No es casualidad que, en las tres técnicas analizadas, el error total de predicción sea muy similar, aproximadamente 24,5%, es decir la calidad predictiva es de aproximadamente el 75,5%. Este fenómeno es una consecuencia del problema de datos desbalanceados, ya que, como se observa en las tablas: **Tabla 5-3**, **Tabla 6-3**, **Tabla 7-3**; los modelos generados fallan en gran medida a la hora de predecir las categorías “Poco contento”, “Descontento pero conforme” y “Totalmente descontento”, y únicamente aciertan cuando se trata de predecir la categoría “Contento”. Por lo tanto, los modelos no están clasificando con base en las variables explicativas, sino que, se enfocan de forma altamente ponderada en las frecuencias absolutas de las categorías de la variable respuesta. Así, los modelos no son generalizables, pues prácticamente clasifican a todos los nuevos individuos dentro de la categoría “Contento”, sin tomar en cuenta sus características medidas por las variables explicativas. En machine learning, a este fenómeno se le conoce como sobreajuste (overfitting), es decir, los modelos se “sobre ajustan” a los datos de entrenamiento y logran un error total de predicción relativamente bajo; pero, cuando se generaliza la predicción para nuevos individuos, el modelo presenta una tasa correcta de clasificación muy baja.

Existen varios procedimientos que buscan subsanar el problema de datos desbalanceados. En la presente investigación se optó por utilizar la técnica conocida como “sobre muestreo” (oversampling). Dicha técnica consiste en obtener una muestra aleatoria con reemplazo, de cada

una de las categorías que tienen una frecuencia absoluta menor a la frecuencia absoluta de la categoría más representativa, en este caso, la categoría "Contento"; de tal manera que la frecuencia absoluta de todas las categorías llegue a ser la misma (igual a la frecuencia absoluta de la categoría "Contento"). Así, se refuerza la naturaleza de las categorías menos representativas, se garantiza que la influencia de las variables explicativas sobre la variable respuesta no se subestime en el ajuste de los modelos, y finalmente, no se pierde información con respecto a la muestra original.

Al aplicar la técnica de oversampling, la numerosidad de la matriz de datos ascendió a 37 336 filas, de tal forma que cada categoría de la variable respuesta obtuvo una frecuencia absoluta de 9 334 observaciones. Posteriormente se volvieron a construir los modelos con las tres técnicas en estudio, siguiendo el procedimiento expuesto en los apartados correspondientes a cada técnica. Los resultados de los nuevos modelos se presentan a continuación:

3.7.1. Gradient boosting (datos balanceados)

El nuevo modelo construido mediante la técnica de gradient boosting presentó un error total de predicción de 29,29%, la matriz de confusión asociada a este modelo se presenta en la **Tabla 8-3**.

Tabla 8-3: Matriz de confusión Gradient boosting (datos balanceados)

	Totalmente descontento	Descontento pero conforme	Poco contento	Contento
Totalmente descontento	23.61	1.11	1.81	1.29
Descontento pero conforme	0.38	17.38	3.44	3.70
Poco contento	0.51	2.76	14.09	4.38
Contento	0.50	3.75	5.66	15.63

Fuente: Modelo Gradient boosting (datos balanceados)
Realizado por: Andrade Saltos, Vinicio, 2020

3.7.2. Redes neuronales artificiales (datos balanceados)

El nuevo modelo construido mediante la técnica de redes neuronales artificiales presentó un error total de predicción de 54,11%, la matriz de confusión asociada a este modelo se presenta en la **Tabla 9-3**.

Tabla 9-3: Matriz de confusión Redes neuronales artificiales (datos balanceados)

	Totalmente descontento	Descontento pero conforme	Poco contento	Contento
Totalmente descontento	19.66	4.71	5.81	3.93
Descontento pero conforme	2.21	7.23	5.86	4.26
Poco contento	1.22	4.04	4.56	2.38
Contento	1.92	9.02	8.78	14.44

Fuente: Modelo RNA (datos balanceados)
Realizado por: Andrade Saltos, Vinicio, 2020

3.7.3. Regresión logística ordinal (datos balanceados)

El nuevo modelo construido mediante la técnica de regresión logística ordinal presentó un error total de predicción de 64,12%, la matriz de confusión asociada a este modelo se presenta en la **Tabla 10-3**.

Tabla 10-3: Matriz de confusión Regresión logística ordinal (datos balanceados)

	Totalmente descontento	Descontento pero conforme	Poco contento	Contento
Totalmente descontento	13.34	8.22	7.95	2.58
Descontento pero conforme	4.30	3.78	4.64	3.20
Poco contento	5.05	6.41	6.08	6.54
Contento	2.31	6.59	6.33	12.68

Fuente: Modelo Regresión logística ordinal (datos balanceados)
Realizado por: Andrade Saltos, Vinicio, 2020

3.8. Demanda de procesamiento computacional en la construcción de los modelos de clasificación

3.8.1. Tiempo

Una vez consolidado un conjunto de datos balanceado, se analizó la demanda de procesamiento que cada una de las técnicas bajo estudio presenta. En la **Tabla 11-3**, se muestran los promedios de los tiempos (segundos) que tardó cada técnica en procesar su respectivo modelo.

Los resultados de los promedios se dividen según las combinaciones de:

a) Tipo de muestra: Real con 37 336 observaciones (2.8 mb) y Aumentada con 373 360 observaciones (28.4 mb).
b) Número de núcleos del procesador: Un núcleo y Siete núcleos

c) Técnica bajo análisis: Gradien boosting (Xgboost), Redes neuronales artificiales (Nnet) y Regresión logística ordinal (Vglm).

Tabla 11-3: Promedio del tiempo de procesamiento

Muestra	Núcleos	Técnica	Media
Aumentada	Siete	Nnet	1865.02
Aumentada	Siete	Vglm	2018.60
Aumentada	Siete	Xgboost	6594.53
Aumentada	Uno	Nnet	8610.93
Aumentada	Uno	Vglm	278.45
Aumentada	Uno	Xgboost	9582.38
Real	Siete	Nnet	149.59
Real	Siete	Vglm	787.45
Real	Siete	Xgboost	692.13
Real	Uno	Nnet	717.45
Real	Uno	Vglm	23.70
Real	Uno	Xgboost	1256.58

Fuente: Tiempo de procesamiento (segundos)
Realizado por: Andrade Saltos, Vinicio, 2020

Los resultados presentes en la **Tabla 11-3**, se resumieron en el **Gráfico 15-3**, con la intención de que su interpretación sea más intuitiva.

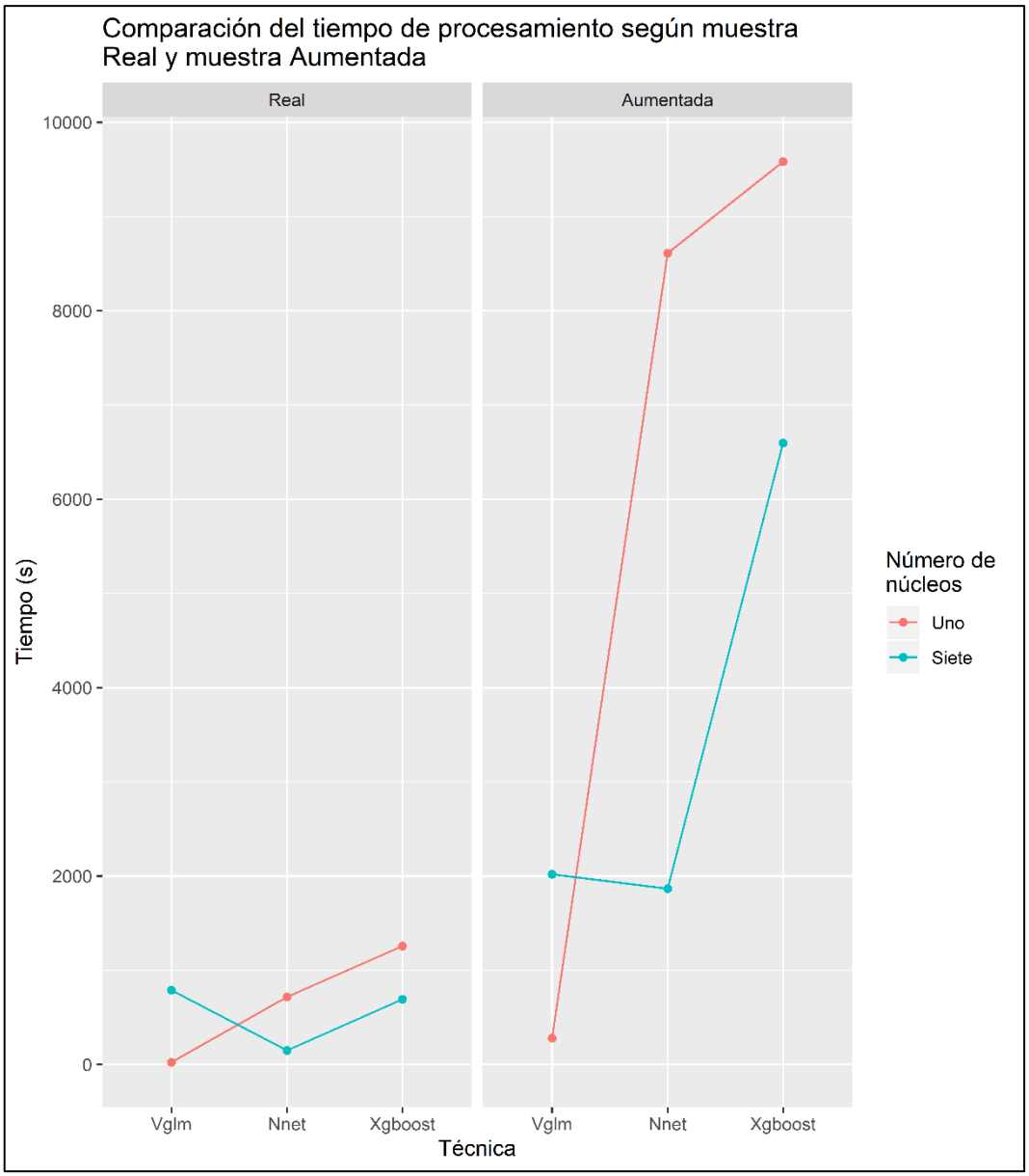


Gráfico 15-3: Promedio del tiempo de procesamiento

Realizado por: Andrade Saltos, Vinicio, 2020

3.8.2. Pico de memoria RAM

Un proceso análogo al anterior se llevó a cabo para el análisis del promedio de pico de memoria RAM (mebibytes) que produjo cada técnica a la hora de procesar su respectivo modelo. Los resultados de los promedios se muestran en la **Tabla 12-3**.

Tabla 12-3: Promedio de pico de memoria RAM

Muestra	Núcleos	Técnica	Media
Aumentada	Siete	Nnet	1478.17
Aumentada	Siete	Vglm	4901.03

Aumentada	Siete	Xgboost	647.36
Aumentada	Uno	Nnet	2402.26
Aumentada	Uno	Vglm	6145.88
Aumentada	Uno	Xgboost	719.14
Real	Siete	Nnet	177.30
Real	Siete	Vglm	493.94
Real	Siete	Xgboost	83.47
Real	Uno	Nnet	256.87
Real	Uno	Vglm	605.48
Real	Uno	Xgboost	115.37

Fuente: Pico de memoria RAM (mebibytes)

Realizado por: Andrade Saltos, Vinicio, 2020

Y su respectiva visualización en el **Gráfico 16-3**.

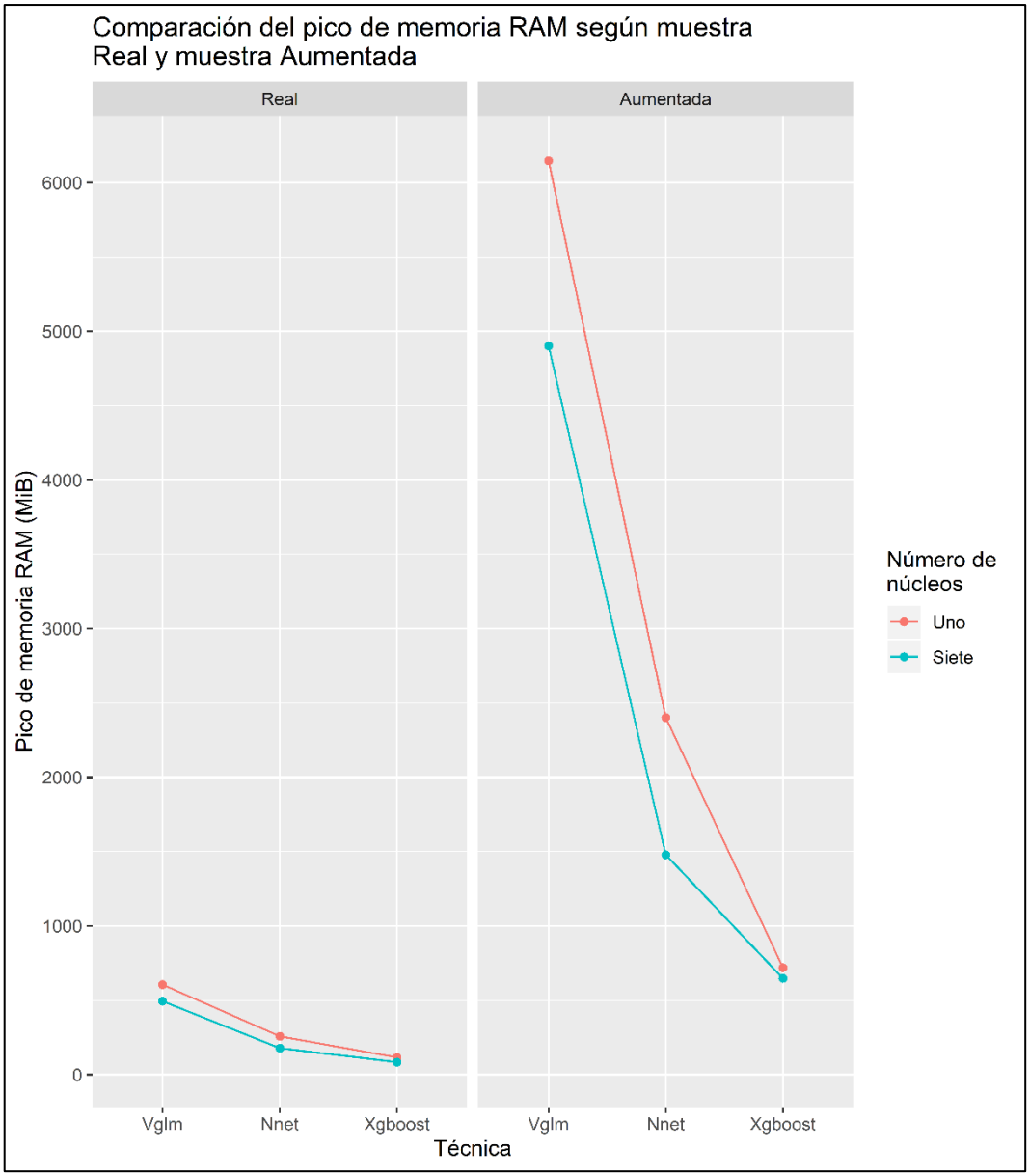


Gráfico 16-3: Promedio de pico de memoria RAM

Realizado por: Andrade Saltos, Vinicio, 2020

Los datos correspondientes a las 10 corridas de cada modelo, tanto para tiempo de procesamiento como para pico de memoria RAM, se muestran en el **Anexo C**.

3.9. El modelo de clasificación con resultados predictivos más confiables

Tomando en cuenta la matriz de confusión de cada modelo construido con un conjunto de datos balanceados (**Tabla 8-3**, **Tabla 9-3** y **Tabla 10-3**), es posible estudiar la calidad predictiva real de cada uno de los modelos generados, ya que, al aplicar la técnica de oversampling, los modelos logran clasificar con base en la interacción de las variables explicativas y la variable respuesta, es así, que los nuevos errores totales de predicción reflejan verdaderamente cómo se comportaría cada modelo al predecir el nivel de satisfacción laboral en nuevos individuos. Al analizar los errores totales de predicción, es posible observar que la técnica de gradient boosting es, con una diferencia contundente, superior a las demás técnicas en la predicción de la satisfacción laboral en Ecuador. Esta realidad concuerda con lo expuesto por Chollet y Allaire (2018, p.23), quienes destacan que en la popular competición Kaggle (competencia para elegir la mejor técnica de machine learning, para una tarea específica), desde que apareció por primera vez el algoritmo de gradient boosting en 2014, éste se convirtió en uno de los mejores, o probablemente el mejor, en la tarea de analizar información no perceptual.

Reducción de la dimensionalidad

Con la intención de potenciar el modelo de predicción construido mediante la técnica de gradient boosting, se decidió identificar las variables más importantes (**Gráfico 17-3**) y por consecuencia las menos importantes (**Gráfico 18-3**), así, fue posible eliminar aquellas variables que tengan una importancia extremadamente baja, sin alterar significativamente el error total de predicción del modelo.

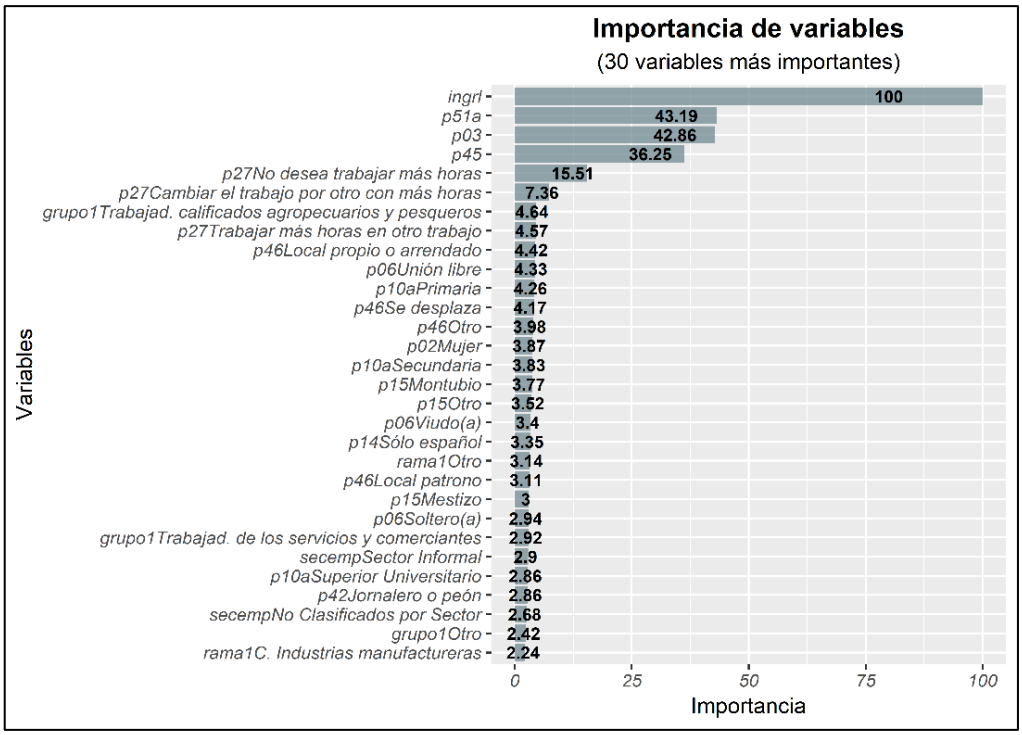


Gráfico 17-3: Variables más importantes (Gradient boosting)

Realizado por: Andrade Saltos, Vinicio, 2020

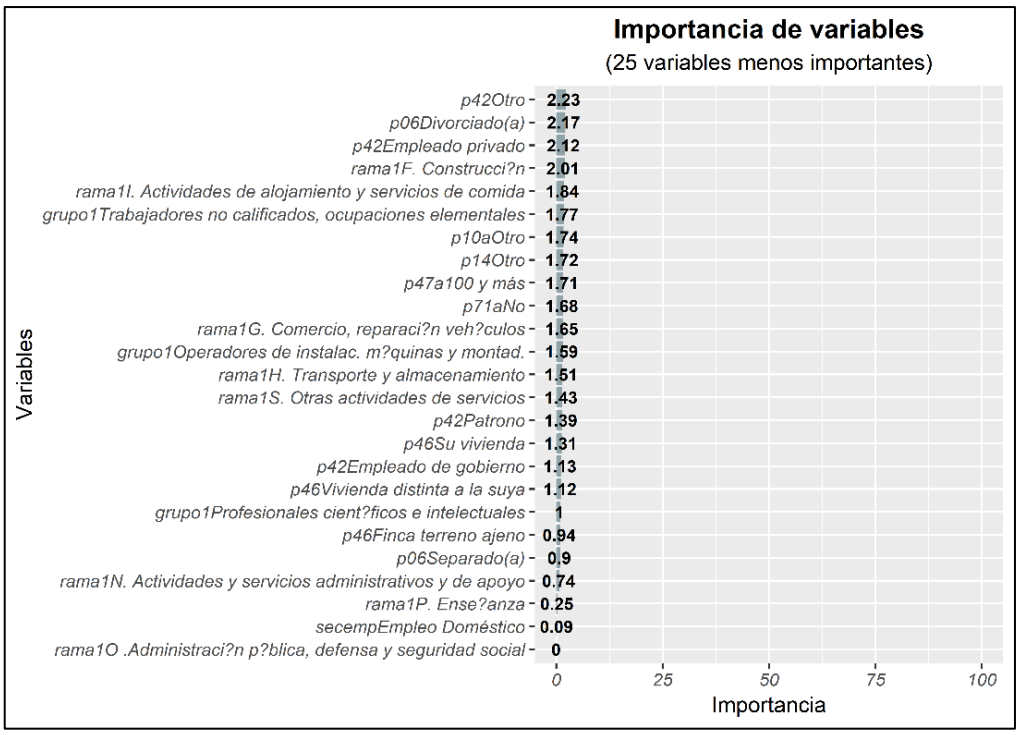


Gráfico 18-3: Variables menos importantes (Gradient boosting)

Realizado por: Andrade Saltos, Vinicio, 2020

Después de suprimir una a una las variables, en orden ascendente de importancia, y medir el error total de predicción después de cada eliminación, se decidió prescindir permanentemente de las variables: “rama1”, “p71a” y “p47a”, manteniendo un error total de predicción de 29,50%, el cual,

es aproximadamente el mismo del modelo construido con todas las variables (29,29%), así se maximiza la calidad predictiva del modelo con el mínimo número posible de variables a ser medidas.

Por lo tanto, el modelo predictivo que presenta resultados más confiables, se construyó mediante la técnica de gradient boosting, dicho modelo está conformado por 15 variables (1 respuesta y 14 explicativas), es generalizable (libre de overfitting) y tiene una calidad predictiva real de 70,50% a la hora de estimar el nivel de satisfacción laboral en nuevos individuos.

Con respecto a las técnicas: regresión logística ordinal y redes neuronales artificiales, se realizó un análisis similar de importancia de variables. Es importante aclarar que dicho análisis tuvo como único objetivo observar cómo se comportan los modelos generados con estas técnicas en el contexto de las variables que los componen. Tomando en cuenta que dichos modelos tienen una calidad predictiva muy mala, los resultados de la importancia de variables deben entenderse como valores que no representan correctamente la realidad del fenómeno bajo estudio (satisfacción laboral). Aún así, sirven para tener una idea de lo ocurrido con respecto a cada modelo. Las 5 variables más importantes para dichas técnicas se presentan a continuación:

- Regresión logística ordinal: “ingrl”, “p51a”, “p45”, “p42Otro”, “p27No desea trabajar más horas”.
- Redes neuronales artificiales: “ingrl”, “p51Otro”, “grupo1Otro”, “p27Cambiar el trabajo por otro con más horas”, “p15mestizo”.

CONCLUSIONES

1. Las técnicas de machine learning y estadística clásica guardan una importante similitud, ya que ambas tienen como objetivo general analizar ciertos fenómenos reales, mediante la construcción de modelos predictivos o de estimación. Pero, al analizar estas metodologías con mayor profundidad, es posible observar que la estadística tiene un trasfondo matemático muy rígido, el cual se basa en supuestos pre establecidos para generar resultados que puedan ser considerados como verdaderos, además, una característica importante en su metodología es la interpretabilidad, caracterizando ciertas particularidades específicas del fenómeno bajo estudio; por su parte, las técnicas de machine learning se fundamentan en el desarrollo y avance computacional, convirtiéndolo en un medio para potenciar el análisis de información, así, es posible brindar mayor libertad al tipo de datos con los que se está trabajando y se pueden generar modelos con una alta calidad predictiva, dejando de lado la interpretación de las características intrínsecas del fenómeno bajo estudio.
2. La técnica de regresión logística ordinal presenta el menor tiempo de procesamiento cuando se trabaja con un único núcleo del procesador, tanto para la muestra real (37 336 observaciones), como para la muestra aumentada (373 360 observaciones). Al trabajar con todos los núcleos del procesador, esta técnica presenta el tiempo de procesamiento más alto en la muestra real (con un aumento de hasta el 3222% con respecto al desempeño obtenido con un único núcleo), y el segundo tiempo más alto en la muestra aumentada (con un aumento de hasta el 624% con respecto al desempeño obtenido con un único núcleo). Este comportamiento se debe al hecho de que, al ser una técnica que basa toda su estructura en una única ecuación matemática, necesita todo el poder de procesamiento enfocado en la construcción de dicha ecuación, por lo que al utilizar varios núcleos del procesador, el poder de procesamiento se dispersa y el tiempo de compilación aumenta; este resultado demuestra que el hecho de utilizar un mayor número de núcleos del procesador, no garantiza una mayor eficiencia en el desempeño computacional, todo dependerá de la naturaleza intrínseca de la técnica utilizada. La técnica de regresión logística ordinal presenta en todo momento la mayor demanda de memoria RAM, por lo tanto, sus tiempos de procesamiento podrían mejorar al contar con una memoria RAM de mayor capacidad.
3. Al utilizar un único núcleo del procesador, la técnica de gradient boosting presenta los tiempos más altos de compilación. El tiempo de procesamiento en la muestra real crece en promedio un 730% al trabajar con la muestra aumentada. Ejecutar esta técnica con todos los núcleos del procesador resulta en un ahorro importante de tiempo (en promedio 32.77%), este fenómeno se repite con la técnica de redes neuronales artificiales; al ser ambas técnicas de machine learning, su construcción se basa en el remuestreo, generando varios modelos

parciales hasta consolidar un modelo final, por lo que, al procesarlas en paralelo, la eficiencia computacional mejora. Tanto la técnica de gradient boosting como la técnica de redes neuronales artificiales presentan una demanda de memoria RAM considerablemente menor a la utilizada por la técnica de regresión logística ordinal.

4. En la tarea de clasificar a los jefes de hogar ecuatorianos con un único trabajo, según su nivel de satisfacción laboral: contento, poco contento, descontento pero conforme, y totalmente descontento; la técnica de gradient boosting presenta un error total de predicción de 29,50%, el cual es mucho menor a los errores obtenidos mediante las técnicas de: redes neuronales artificiales y regresión logística ordinal (54,11% y 64,12%, respectivamente), por lo que esta técnica puede ser considerada como la más confiable en términos predictivos, según el contexto de la presente investigación.
5. El estudio de las variables más importantes en la construcción del modelo predictivo mediante gradient boosting, permite concluir que la satisfacción laboral en Ecuador está fuertemente relacionada con: el sueldo que percibe el trabajador (ingreso), el número de horas que trabaja, la edad del empleado, y el número de años que lleva trabajando.

RECOMENDACIONES

1. Tomando en cuenta que la técnica de gradient boosting resultó ser la más confiable a la hora de predecir la satisfacción laboral en Ecuador; se recomienda su aplicación en bases de datos socio – económicas similares a la utilizada en el presente estudio.
2. Ya que la técnica de gradient boosting presentó un error total de predicción considerablemente bajo, se recomienda compararla con técnicas que resultan de la fusión o evolución de algoritmos ya establecidos, ejemplos de estas técnicas son: support vector regression, convolutional neural networks, deep neural pursuit, product-based neural networks, gradient boosting random convolutional network (GBRCN), entre otras; de tal forma que sea posible conocer si la combinación de técnicas de clasificación, resulta en la construcción de modelos con una mejor calidad predictiva.
3. Una de las características a destacar en las técnicas de machine learning, es la incapacidad de interpretar a profundidad los modelos construidos, esto se debe al alto nivel de complejidad que puede llegar a alcanzar la estructura interna de cada modelo. Por ejemplo, en la presente investigación se utilizó la técnica de gradient boosting, la cual puede entenderse como un conjunto de decenas o incluso cientos de árboles de clasificación; cada uno de estos árboles puede llegar a tener, por sí solo, una alta complejidad en su estructura, este hecho puede visualizarse en el siguiente árbol de clasificación, construido para clasificar a un jefe de hogar ecuatoriano según su satisfacción laboral:

   https://1drv.ms/b/s!AiQ1DuwoZ0ivgngCE3gIbi-Z-Lpl?e=cIAqrF

   Tomando en cuenta esta realidad, se recomienda plantear investigaciones enfocadas en la construcción de técnicas que garanticen la interpretabilidad de los modelos de machine learning, puesto que, además de alcanzar un alto nivel predictivo, es importante también entender el por qué un fenómeno bajo estudio se comporta de cierta manera.

GLOSARIO

[A]

Análisis discriminante. _ Técnica de análisis multivariante que construye un modelo capaz de clasificar a individuos, asignándolos a diferentes categorías de una variable dependiente, según ciertas características que son descritas por variables independientes.

[B]

Big data. _ Conjunto de datos con una numerosidad extremadamente grande (terabytes de información), que requieren ser analizados en el menor tiempo posible.

Black box. _ Modelo de machine learning que por la complejidad de su estructura, no es posible entender su funcionalidad interna.

[C]

Classification trees. _ Árboles de decisión que tienen como tarea clasificar a individuos, mediante una variable objetivo que toma valores categóricos.

Convolutional neural networks. _ Tipo de deep learning, que se enfoca principalmente en el análisis de información visual como imágenes, esta técnica se caracteriza por tener un mayor grado de complejidad con respecto a otros tipos de deep learning.

Curvas ROC. _ Gráfico que describe la calidad predictiva de un modelo de clasificación binaria.

[D]

Data mining. _ Conjunto de técnicas que se encargan de descubrir patrones en el área de big data.

Data science. _ Campo multidisciplinario que utiliza métodos procesos y algoritmos científicos para extraer conocimiento de los datos.

Decision trees. _ Algoritmo de flujo basado en nodos que deciden según frecuencias de individuos, analizando posteriormente las consecuencias de cada decisión.

Deep learning. _ Redes neuronales artificiales con más de una capa oculta, también se conoce como Deep neural network.

Deep neural pursuit. _ Modelo basado en deep neural network, enfocado en analizar conjuntos de datos con un tamaño muestral pequeño y una alta dimensionalidad.

[F]

Factor de inflación de la varianza (VIF). _ Medida de la intensidad de la multicolinealidad en un análisis supervisado (análisis de dependencia).

[G]

Github. _ Empresa que provee almacenamiento mediante servidores, para alojar y compartir proyectos de desarrollo de software en su versión de control.

Gradient descent. _ Algoritmo de optimización iterativa en primera clase que

tiene como objetivo encontrar el mínimo de una función.

Gradient boosting. _ Random forest que procesa la información histórica de su construcción, con la finalidad de reducir el error de total de predicción.

[H]

Hardware. _ Parte física del computador o dispositivos electrónicos.

[K]

k-nearest neighbor. _ Algoritmo de machine learning que se enfoca en el reconocimiento de patrones.

[L]

Laptop. _ Computadora portátil, la cual se pueden transportar con facilidad para ser utilizada en diferentes ambientes o condiciones.

Logit. _ Función definida como la inversa de la función logística sigmoidal o transformación logística.

[M]

Machine learning. _ Estudio científico de algoritmos y modelos estadísticos con el objetivo de que sistemas computacionales realicen efectivamente una tarea sin necesidad de instrucciones explícitas.

Mebibyte. _ Múltiplo de byte, utilizado para medir información informática, se abrevia MiB, 1 mebibyte es equivalente a 1 048 576 bytes.

Megabyte. _ Múltiplo de byte, utilizado para medir información informática, se abrevia MB, 1 megabyte es equivalente a 1 000 000 bytes.

Modelos aditivos generalizados. _ Extensión de los modelos lineales generalizados, donde el predictor lineal no está restringido a ser lineal en las covariables de X.

Multivariate adaptive regression splines. _ Técnica de regresión no paramétrica, la cual se puede entender como una extensión de los modelos lineales, sin estar sujeta a ciertas restricciones.

[O]

Odds proporcionales. _ Modelo de regresión logística ordinal donde los odds son proporcionales a las categorías de la variable respuesta.

Ord.factor. _ Objeto del lenguaje de programación R, que además de ser factor, tiene la característica de ordinalidad asociada a sus valores.

Overfitting. _ Error de modelado que ocurre cuando la función construida es muy eficiente en términos de la muestra de entrenamiento, pero presenta un error muy alto con nuevos datos, es decir, no es generalizable.

Oversampling. _ Tipo de remuestreo que tiene por objetivo solucionar el problema de datos desbalanceados en una variable respuesta categórica.

[P]

Product-based neural networks. _ Extensión de las redes neuronales artificiales que tiene por objetivo modelar eficientemente las variables independientes con un alto número de categorías.

[R]

Random forest. _ Método de clasificación o regresión que se construye mediante varios árboles de decisión.

Regresión lineal. _ Modelo matemático utilizado para modelar la relación entre una variable dependiente y una o varias variables independientes.

Regresión logística ordinal. _ Tipo de regresión logística que se encarga de modelar una variable dependiente cualitativa medida en escala ordinal.

[S]

Stratified tenfold cross – validation. _ Algoritmo que tiene por objetivo predecir la tasa de error en una técnica de aprendizaje dado un conjunto de datos.

Software. _ Conjunto de instrucciones, datos o programas, usados para realizar operaciones en un computador o ejecutar tareas específicas.

Support vector machines. _ Modelo de aprendizaje supervisado basado en un clasificador lineal binario no probabilístico.

BIBLIOGRAFÍA

ANDERSEN, L.L., et.al. Job satisfaction is more than a fruit basket, health checks and free exercise: Cross-sectional study among 10,000 wage earners. *Scandinavian Journal of Public Health*, 2017. vol. 45, no. 5, pp. 476-484. [Consulta: 10 abril 2019]. ISSN 1403-4948. DOI 10.1177/1403494817698891. Disponible en: http://journals.sagepub.com/doi/10.1177/1403494817698891.

ANDRADE SALTOS, V.A. y FLORES M, P. Comparativa entre classification trees, random forest y gradient boosting; en la predicción de la satisfacción laboral en Ecuador. *Ciencia Digital* 2018. vol. 2, no. 4.1. SE-Artículos. DOI 10.33262/cienciadigital.v2i4.1..189. Disponible en: http://www.cienciadigital.org/revistascienciadigital/index.php/CienciaDigital/article/view/189.

ANGHEL, I., et.al. Prediction of Manufacturing Processes Errors: Gradient Boosted Trees Versus Deep Neural Networks. *2018 IEEE 16th International Conference on Embedded and Ubiquitous Computing (EUC)* [en línea]. S.l.: IEEE, 2018. pp. 29-36. [Consulta: 10 abril 2019]. ISBN 978-1-5386-8296-8. DOI 10.1109/EUC.2018.00012. Disponible en: https://ieeexplore.ieee.org/document/8588845/.

ATKINSON, E.J., et.al. Assessing fracture risk using gradient boosting machine (GBM) models. *Journal of Bone and Mineral Research* [en línea], 2012. vol. 27, no. 6, pp. 1397-1404. [Consulta: 19 septiembre 2018]. ISSN 08840431. DOI 10.1002/jbmr.1577. Disponible en: http://doi.wiley.com/10.1002/jbmr.1577.

BROWN, I. y MUES, C. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications* [en línea], 2012. vol. 39, no. 3, pp. 3446-3453. [Consulta: 11 agosto 2018]. ISSN 0957-4174. DOI 10.1016/J.ESWA.2011.09.033. Disponible en: https://www.sciencedirect.com/science/article/pii/S095741741101342X.

CABRERA JARAMILLO, G.T. y HUERTA VILLIGUA, Y.L. Clima laboral y su incidencia en la satisfacción en el trabajo de la empresa Arcgold del Ecuador S.A. Sucursal Orellana. [en línea], 2017. [Consulta: 9 agosto 2018]. Disponible en:

http://repositorio.ug.edu.ec/handle/redug/22764.

CHOLLET, F. y ALLAIRE, J.J. *Deep Learning with R, Ch. 5.4*. 2018. S.l.: Shelter Island: Manning Publications Company. 2018.

CHURPEK, M.M., et.al. Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards. *Critical care medicine* [en línea], 2016. vol. 44, no. 2, pp. 368-74. [Consulta: 9 abril 2019]. ISSN 1530-0293. DOI 10.1097/CCM.0000000000001571. Disponible en: http://www.ncbi.nlm.nih.gov/pubmed/26771782.

COX, D.R. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)* [en línea], 1958. pp. 215-242. [Consulta: 13 agosto 2018]. ISSN 0035-9246. DOI 10.2307/2983890. Disponible en: https://www.jstor.org/stable/2983890.

CUADRAS, C. *Nuevos métodos de análisis multivariante* [en línea]. S.l.: s.n. 2007. [Consulta: 13 agosto 2018]. Disponible en: http://www.est.uc3m.es/esp/nueva_docencia/getafe/estadistica/analisis_multivariante/doc_generica/archivos/metodos.pdf.

DEVRIES, P.M.R., et.al. Deep learning of aftershock patterns following large earthquakes. *Nature* [en línea], 2018. vol. 560, no. 7720, pp. 632-634. [Consulta: 20 septiembre 2018]. ISSN 0028-0836. DOI 10.1038/s41586-018-0438-y. Disponible en: https://www.nature.com/articles/s41586-018-0438-y.

FONSECA, E., et.al 1975- y SERRA, X., Acoustic scene classification by ensembling gradient boosting machine and convolutional neural networks. [en línea], 2017 [Consulta: 9 abril 2019]. Disponible en: https://repositori.upf.edu/handle/10230/33454.

FRIDAY, S.S. y FRIDAY, E. Racioethnic perceptions of job characteristics and job satisfaction. *Journal of Management Development* [en línea], 2003. vol. 22, no. 5-6, pp. 426-442. [Consulta:

15 julio 2018]. ISSN 02621711. DOI 10.1108/02621710310474778. Disponible en: https://www.emeraldinsight.com/doi/10.1108/02621710310474778.

GARCÍA-POZO, A., MORO-TEJEDOR, M.N. y MEDINA-TORRES, M. Evaluación y dimensiones que definen el clima y la satisfacción laboral en el personal de enfermería. *Revista de Calidad Asistencial* [en línea], 2010. vol. 25, no. 4, pp. 207-214. [Consulta: 9 agosto 2018]. ISSN 1134282X. DOI 10.1016/j.cali.2010.02.003. Disponible en: https://www.sciencedirect.com/science/article/pii/S1134282X10000448.

GARETH, J., et.al. *An Introduction to Statistical Learning* [en línea]. 2006. S.l.: s.n. [Consulta: 13 agosto 2018]. ISBN 9780387781884. Disponible en: https://link.springer.com/content/pdf/10.1007/978-1-4614-7138-7.pdf.

GOODFELLOW, I., et.al. *Deep learning* [en línea]. 2016. S.l.: s.n. [Consulta: 13 agosto 2018]. Disponible en: https://www.synapse.koreamed.org/Synapse/Data/PDFData/1088HIR/hir-22-351.pdf.

GREENE, W.H. Econometric analysis, 5th. *Ed.. Upper Saddle River, NJ*, 2003. pp. 89-140.

HIDALGO, E.T. Niveles de satisfacción laboral en los trabajadores de la Universidad Central Del Ecuador,(Unidad de gestión de Personal docente y administrativo). [en línea], 2018. [Consulta: 9 agosto 2018]. Disponible en: http://www.dspace.uce.edu.ec/handle/25000/14516.

HUNG, C.-Y., et.al. Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database. *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* [en línea]. S.l.: IEEE, 2017. pp. 3110-3113. [Consulta: 9 abril 2019]. ISBN 978-1-5090-2809-2. DOI 10.1109/EMBC.2017.8037515. Disponible en: https://ieeexplore.ieee.org/document/8037515/.

INEC. *Metodología para el cálculo de la matriz de transición laboral Septiembre 2018 –*

diciembre 2018 [en línea]. 2018. Quito: INEC. Disponible en: http://www.ecuadorencifras.gob.ec/documentos/web-inec/EMPLEO/2018/Matrices_de_Transicion/Septiembre-2017_Diciembre-2017/Documento_Metodologico_MTL_sep2017_dic2017.pdf.

JAMES, G., et.al. An Introduction to Statistical Learning with Applications in R. Springer. , 2014.

JULIAN, L.J., et.al. Employment in multiple sclerosis. *Journal of Neurology* [en línea], 2008. vol. 255, no. 9, pp. 1354-1360. [Consulta: 11 abril 2019]. ISSN 0340-5354. DOI 10.1007/s00415-008-0910-y. Disponible en: http://link.springer.com/10.1007/s00415-008-0910-y.

KAISLER, S., et.al. Big Data: Issues and Challenges Moving Forward. *2013 46th Hawaii International Conference on System Sciences* [en línea]. S.l.: IEEE, 2013. pp. 995-1004. [Consulta: 13 agosto 2018]. ISBN 978-1-4673-5933-7. DOI 10.1109/HICSS.2013.645. Disponible en: http://ieeexplore.ieee.org/document/6479953/.

KARLAFTIS, M.G. y VLAHOGIANNI, E.I. Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. [en línea], 2011. [Consulta: 26 noviembre 2019]. DOI 10.1016/j.trc.2010.10.004. Disponible en: http://www.elsevier.com/copyright.

KIM, J., KWON LEE, J. y MU LEE, K. *Accurate Image Super-Resolution Using Very Deep Convolutional Networks* [en línea]. 2016. S.l.: s.n. [Consulta: 9 abril 2019]. Disponible en: https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Kim_Accurate_Image_Super-Resolution_CVPR_2016_paper.html.

KRUEGER, P., et.al. Predictors of job satisfaction among academic family medicine faculty. *Canadian Family Physician*, 2017. vol. 63, no. 3.

KUHN, M. y JOHNSON, K.. *Applied predictive modeling* [en línea]. 2013. S.l.: s.n. [Consulta:

13 agosto 2018]. Disponible en: https://link.springer.com/content/pdf/10.1007/978-1-4614-6849-3.pdf.

LI, J., et.al. Factors Influencing Job Satisfaction for Employed Adults With Multiple Sclerosis. *Rehabilitation Counseling Bulletin* [en línea], 2017. vol. 61, no. 1, pp. 28-40. [Consulta: 11 abril 2019]. ISSN 0034-3552. DOI 10.1177/0034355216662616. Disponible en: http://journals.sagepub.com/doi/10.1177/0034355216662616.

LI, P., WU, Q. PROCESSING, C.B.-A. in neural information y undefined, [sin fecha]. Mcrank: Learning to rank using multiple classification and gradient boosting. *papers.nips.cc* [en línea], 2008. [Consulta: 11 agosto 2018]. Disponible en: http://papers.nips.cc/paper/3270-mcrank-learning-to-rank-using-multiple-classification-and-gradient-boosting.pdf.

LIU, B., et.al. Deep Neural Networks for High Dimension, Low Sample Size Data. [en línea]. 2017. S.l.: [Consulta: 9 abril 2019]. Disponible en: https://pdfs.semanticscholar.org/76f7/55fed7bf1cea8dad35c16bf518eab158c13e.pdf.

LOYER, J.-L., et.al. Comparison of Machine Learning methods applied to the estimation of manufacturing cost of jet engine components. *International Journal of Production Economics* [en línea], 2016. vol. 178, pp. 109-119. [Consulta: 9 abril 2019]. ISSN 0925-5273. DOI 10.1016/J.IJPE.2016.05.006. Disponible en: https://www.sciencedirect.com/science/article/pii/S0925527316300731.

MATICH, D.J. Redes Neuronales: Conceptos Básicos y Aplicaciones. *Historia* [en línea], 2001. pp. 55. [Consulta: 13 agosto 2018]. Disponible en: ftp://decsai.ugr.es/pub/usuarios/castro/Material-Redes-Neuronales/Libros/matich-redesneuronales.pdf.

MCCULLAGH, P. Regression Models for Ordinal Data. *Journal of the Royal Statistical Society. Series B (Methodological)* [en línea], 1980. vol. 42, no. 2, pp. 109-142. [Consulta: 12 agosto 2018]. ISSN 00359246. DOI 10.2307/2984952. Disponible en: https://www.jstor.org/stable/2984952.

MICHALSKI, R., CARBONELL, J. y MITCHELL, T. *Machine learning: An artificial intelligence approach* [en línea]. 2013. S.l.: s.n. [Consulta: 13 agosto 2018]. Disponible en: https://books.google.com.ec/books?hl=en&lr=&id=-eqpCAAAQBAJ&oi=fnd&pg=PA2&dq=machine+learning+concepts&ots=Wl2PLz4Gn7&sig=HjeDeSdLYVhb7mnv_ZsDEdEjvA4.

MONTECÉ, D.S.C., PIGUAVE, E.U.E. y RODRÍGUEZ, Z.M. DIFERENCIAS SALARIALES Y SATISFACCIÓN LABORAL ENTRE GÉNEROS Y ETNIAS EN EL ECUADOR. *Revista Científica ECOCIENCIA*, 2016. vol. 3, no. 4.

QU, Y., et.al. Product-Based Neural Networks for User Response Prediction. *2016 IEEE 16th International Conference on Data Mining (ICDM)* [en línea]. 2016. S.l.: IEEE, pp. 1149-1154. [Consulta: 9 abril 2019]. ISBN 978-1-5090-5473-2. DOI 10.1109/ICDM.2016.0151. Disponible en: http://ieeexplore.ieee.org/document/7837964/.

ROE, B.P., et.al. Boosted decision trees as an alternative to artificial neural networks for particle identification. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* [en línea], 2005. vol. 543, no. 2-3, pp. 577-584. [Consulta: 11 agosto 2018]. ISSN 0168-9002. DOI 10.1016/J.NIMA.2004.12.018. Disponible en: https://www.sciencedirect.com/science/article/pii/S0168900205000355.

ROWDEN, R.W. The relationship between workplace learning and job satisfaction in U.S. small to midsize businesses. *Human Resource Development Quarterly* [en línea], 2002. vol. 13, no. 4, pp. 407-425. [Consulta: 15 julio 2018]. ISSN 1044-8004. DOI 10.1002/hrdq.1041. Disponible en: http://doi.wiley.com/10.1002/hrdq.1041.

RUZAFA MARTÍNEZ, M., et.al. Satisfacción laboral de los profesionales de enfermería españoles que trabajan en hospitales ingleses. *Gaceta sanitaria: Organo oficial de la Sociedad Española de Salud Pública y Administración Sanitaria* [en línea], 2008. vol. 22, no. 5, pp. 434-442. [Consulta: 9 agosto 2018]. ISSN 0213-9111. Disponible en: http://dialnet.unirioja.es/servlet/articulo?codigo=2732516&orden=260729&info=link%5Cnhttp://dialnet.unirioja.es/servlet/extart?codigo=2732516.

SARLE, W. Neural networks and statistical models. [en línea], 1994. [Consulta: 26 noviembre 2019]. Disponible en: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.27.699.

SPECTOR, P.E. *Job Satisfaction: Aplication, Assesment, Causes, and Consequences* [en línea]. 1997. S.l.: s.n. [Consulta: 13 agosto 2018]. Disponible en: https://books.google.com.ec/books?hl=en&lr=&id=_-AXCgAAQBAJ&oi=fnd&pg=PR7&dq=ob+satisfaction:+Application,+assessment,+causes+and+consequences&ots=epFnGrWe9g&sig=1LKAadOxiHjs4ZrYPwPf550TBqc.

SUASNAVAS, P.R., et.al. Responsabilidad social y gestión de la seguridad y salud en el trabajo: panorama actual de las empresas ecuatorianas. *Revista ESPACIOS* [en línea], 2019. vol. 40, no. 04. [Consulta: 11 abril 2019]. Disponible en: http://es.revistaespacios.com/a19v40n04/19400418.html.

TEIXEIRA-POIT, S.M., et.al. Factors influencing professional life satisfaction among neurologists. *BMC Health Services Research* [en línea], 2017. vol. 17, no. 1, pp. 409. [Consulta: 11 abril 2019]. ISSN 1472-6963. DOI 10.1186/s12913-017-2343-8. Disponible en: http://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-017-2343-8.

TU, J. V. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology* [en línea], 1996. vol. 49, no. 11, pp. 1225-1231. [Consulta: 11 agosto 2018]. ISSN 0895-4356. DOI 10.1016/S0895-4356(96)00002-9. Disponible en: https://www.sciencedirect.com/science/article/pii/S0895435696000029.

URÍEN, B. y OSCA, A. Cambios en las tareas y su repercusión en la satisfacción laboral: un estudio en empresas de automoción Changes in task and their effect in job satisfaction: A study in automotive companies. *Revista de Psicología Social* [en línea], 2001. vol. 16, no. 3, pp. 315-329. [Consulta: 9 agosto 2018]. ISSN 02134748. DOI 10.1174/021347401317351062. Disponible en: https://www.tandfonline.com/doi/abs/10.1174/021347401317351062.

VALERO, D. y HIRSCHI, A. To hangover or not: trajectories of job satisfaction in adolescent

workforce newcomers. *European Journal of Work and Organizational Psychology* [en línea], 2019. vol. 28, no. 2, pp. 150-163. [Consulta: 11 abril 2019]. ISSN 1359-432X. DOI 10.1080/1359432X.2018.1564278. Disponible en: https://www.tandfonline.com/doi/full/10.1080/1359432X.2018.1564278.

VILLAR-RUBIO, E., DELGADO-ALAMINOS, J. y BARRILAO-GONZÁLEZ, P. Job Satisfaction Among Spanish Tax Administration Employees: A Logistic Regression Analysis. *Journal of Labor Research* [en línea], 2015. vol. 36, no. 2, pp. 210-223. [Consulta: 11 abril 2019]. ISSN 0195-3613. DOI 10.1007/s12122-015-9202-3. Disponible en: http://link.springer.com/10.1007/s12122-015-9202-3.

WANG, Y., et.al. A mobile recommendation system based on logistic regression and Gradient Boosting Decision Trees. *2016 International Joint Conference on Neural Networks (IJCNN)* [en línea]. 2016. S.l.: IEEE, pp. 1896-1902. [Consulta: 9 abril 2019]. ISBN 978-1-5090-0620-5. DOI 10.1109/IJCNN.2016.7727431. Disponible en: http://ieeexplore.ieee.org/document/7727431/.

WITTEN, I., et.al. *Data Mining: Practical machine learning tools and techniques* [en línea]. 2016. S.l.: s.n. [Consulta: 20 septiembre 2018]. Disponible en: https://books.google.com.ec/books?hl=en&lr=&id=1SylCgAAQBAJ&oi=fnd&pg=PP1&dq=data+mining+practical+machine+learning+tools&ots=8ICPufkAvb&sig=qj3-va9qMH_5EgPUZbRJ9xpVk0o.

WONG, W.K., et.al. Quaternionic Fuzzy Neural Network for View-Invariant Color Face Image Recognition. *Complex-Valued Neural Networks* [en línea]. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2013. pp. 235-278. [Consulta: 19 septiembre 2018]. Disponible en: http://doi.wiley.com/10.1002/9781118590072.ch10.

XIAO, Y., et.al. A deep learning-based multi-model ensemble method for cancer prediction. *Computer Methods and Programs in Biomedicine* [en línea], 2018. vol. 153, pp. 1-9. [Consulta: 10 abril 2019]. ISSN 0169-2607. DOI 10.1016/J.CMPB.2017.09.005. Disponible en: https://www.sciencedirect.com/science/article/pii/S0169260717304947.

ZHANG, F., DU, B. y ZHANG, L. Scene Classification via a Gradient Boosting Random

Convolutional Network Framework. *IEEE Transactions on Geoscience and Remote Sensing* [en línea], 2016. vol. 54, no. 3, pp. 1793-1802. [Consulta: 9 abril 2019]. ISSN 0196-2892. DOI 10.1109/TGRS.2015.2488681. Disponible en: http://ieeexplore.ieee.org/document/7310864/.