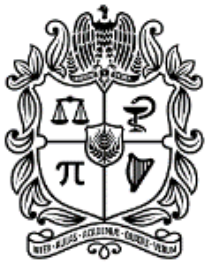

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Modelo de Medición de la Rotación de Personal como Variable de Decisión Estratégica

Carolina Henao Ríos, Esp.

Universidad Nacional de Colombia
Facultad de Minas
Área Curricular de Sistemas e Informática
Medellín, Colombia
2021

Modelo de Medición de la Rotación de Personal como Variable de Decisión Estratégica

Trabajo final presentado como requisito parcial para optar al título de:

Magister en Ingeniería-Analítica

Carolina Henao Ríos, Esp.

Director:

Juan David Velásquez Henao, MSc, PhD

Línea de Investigación:

Analítica

Universidad Nacional de Colombia

Facultad de Minas

Área Curricular de Sistemas e Informática

Medellín, Colombia

2021

Agradecimientos

Este trabajo final de maestría y el proceso de formación que se ha requerido, han sido acompañados por diferentes personas que me parece importante mencionar a continuación:

- Agradezco al profesor Juan David Velásquez por su acompañamiento técnico en todo el proceso de creación de este trabajo y sus enseñanzas en mi proceso de formación.
- Agradezco a Bibiana Cuervo por apoyarme, entenderme y ayudarme en cada etapa de mi proceso de formación.
- Agradezco a Alejando Hincapié compartirme su conocimiento y compañía en este proceso de formación.
- Agradezco a mi familia y amigos por su acompañamiento para traerme hasta donde estoy hoy.

Resumen

Modelo de Medición de la Rotación de Personal como Variable de Decisión Estratégica

Este trabajo final de maestría presenta diferentes modelos de machine learning que buscan predecir la propensión a la renuncia voluntaria de los colaboradores de la Caja de Compensación Comfenalco Antioquia. Para su desarrollo se recolectaron los datos históricos de los colaboradores que han renunciado y de los que permanecen activos en la Caja de Compensación, los cuales fueron verificados, limpiados, transformados y modelados con diferentes técnicas estadísticas en Python. Los resultados obtenidos indican que el mejor modelo de machine learning para pronosticar la propensión a la renuncia con los datos obtenidos es el XGBoost, con una métrica de precisión del 87,3%. Además, se identificó que la estructura organizacional de la Caja de Compensación representa un riesgo para los ejercicios de pronóstico que se deseen realizar, debido a que sus múltiples mutaciones a lo largo de la historia pueden generar desviaciones mayores en la predicción de los modelos y es una variable que debe ser homologada de una manera que pueda tener permanencia en el tiempo.

Palabras clave: machine learning, modelos de pronóstico, renuncia voluntaria

Abstract

Model for Measuring Staff Turnover as a Strategic Decision Variable

This final master's thesis presents different machine learning models that seek to predict the propensity to voluntary resignation of employees of the Comfenalco Antioquia Compensation Fund. For its development, the historical data of the collaborators who have resigned and those who remain active in the Compensation Fund were collected, which were verified, cleaned, transformed, and modeled with different statistical techniques in Python. The results obtained indicate that the best machine learning model to predict the propensity to resign with the data obtained is XGBoost, with a precision metric of 87.3%. In addition, it was identified that the organizational structure of the Compensation Fund represents a risk for the forecasting exercises that are to be carried out, because its multiple mutations throughout history can generate greater deviations in the prediction of the models and is a variable that must be homologated in a way that can have permanence over time.

Keywords: machine learning, forecasting models, voluntary resignation

Tabla de contenido

Agradecimientos 5

Resumen 7

Abstract 8

1. Introducción 12

1.1. Definición del problema de negocio 12

1.2. Definición del problema de analítica 13

1.3. Revisión de literatura 13

1.4. Definición del problema de tesis y pregunta de investigación 14

1.5. Hipótesis 15

1.6. Objetivos 15

1.6.1. Objetivo General 15

1.6.2. Objetivos Específicos 15

1.7. Metodología 15

2. Comprensión del Negocio 17

2.1. Resultados del proyecto 17

2.2. Valoración de la situación actual 18

2.3. Objetivos de la minería de datos 18

2.4. Plan del proyecto 18

3. Comprensión y preparación de los Datos 20

3.1. Recopilación de datos iniciales 20

3.2. Descripción de los datos 21

3.3. Exploración y verificación de la calidad de los datos 25

3.4. Ingeniería de características 31

4. Modelos 36

4.1. Modelos seleccionados 36

4.2. Partición de los datos y métricas de precisión 36

4.3. Entrenamiento de los modelos y búsqueda de parámetros óptimos 37

4.4. Resultados obtenidos 39

4.5. Conclusiones 41

5. Evaluación u obtención de resultados 42
5.1. Evaluación de resultados 42
5.2. Recalibración del modelo 42
5.3. Pasos siguientes 43
6. Despliegue o puesta en producción 44
7. Conclusiones 45
7.1. Respuesta a la pregunta de investigación 45
7.2. Cumplimiento de objetivos 45
7.2.1. Objetivo específico 1: Realizar la recolección, limpieza y depuración de la información disponible para el análisis de la rotación de personal. 45
7.2.2. Objetivo específico 2: Desarrollar un modelo de clasificación que permita identificar qué empleados tienen un alto riesgo de renunciar. 45
7.2.3. Objetivo específico 3: Desarrollar un prototipo de una herramienta de datos que implemente el modelo y permita determinar el riesgo de renuncia de un grupo de empleados en una organización. 46
8. Referencias 47

1. Introducción

En esta propuesta de trabajo se aborda el problema de la rotación de personal, el cual tiene impactos principalmente en la economía y productividad de las empresas. Las organizaciones asumen los costos derivados de los procesos de selección y formación de nuevos empleados, y a mayor rotación de personal habrá mayores costos asociados a este rubro. Así mismo, la rotación de personal afecta la continuidad de los diferentes procesos de las empresas, lo que hace que su productividad disminuya.

Para abordar el problema de rotación de personal desde la analítica, diferentes autores han utilizado técnicas de aprendizaje de máquinas y aprendizaje estadístico como redes neuronales artificiales híbridas, mapas auto-organizativos (SOM), árboles de decisión, máquinas potenciadas por gradiente, regresión logística y k vecinos más próximos, entre otros. Estas técnicas se han utilizado para resolver problemas como la identificación de: las características individuales de los grupos con tendencia a la rotación; de los empleados valiosos susceptibles de abandonar la compañía; de los empleados con una alta probabilidad de abandonar la empresa.

En resumen, se propone como trabajo final el desarrollo de un modelo de clasificación que permita determinar qué empleados están propensos a abandonar la empresa. Para ello, se explorarán técnicas provenientes tanto de la estadística como de la inteligencia artificial. Finalmente, se propone el desarrollo de un prototipo computacional que implemente la metodología.

1.1. Definición del problema de negocio

Los altos índices de rotación de personal han afectado la productividad y economía de las empresas, dado que estas deben hacer inversiones de dinero, tiempo y recursos generando estrategias que atraigan a las personas idóneas para conformar sus equipos de trabajo; las empresas deben invertir en la formación de sus empleados con el fin de actualizar sus conocimientos y mantenerse a la vanguardia en las buenas prácticas y ser competitivos en el mercado. De igual manera, muchas empresas hacen inversiones orientadas al bienestar de sus empleados buscando su fidelización y de esta manera evitar que decidan abandonar la empresa. Sin embargo, todos estos esfuerzos serán en vano si las personas se retiran, puesto que es necesario asumir los costos de selección del nuevo personal, de su entrenamiento y adaptación al puesto de trabajo, más todos los demás costos asociados a la fidelización del empleado.

Además de lo anterior, la rotación de personal hace que en las empresas se disminuya la productividad de sus procesos en general; la continuidad de los procesos se ve afectada, y se interfiere con la transferencia de conocimientos al interior. Cuando las nuevas personas llegan

a las empresas deben pasar un tiempo considerable desarrollando la curva de aprendizaje que es diferente para cada cargo, con los costos que conlleva.

1.2. Definición del problema de analítica

La implementación de estrategias y acciones dirigidas que buscan la reducción de la rotación tienen dos efectos importantes, ayuda a consolidar la estrategia de cultura organizacional; y genera valor agregado para todas las personas que laboran allí, ya que se desprenden acciones que buscan su bienestar.

El uso de datos para tomar decisiones en los procesos de recursos humanos ha venido cobrando relevancia, debido a que en otro tipo de procesos como los core de negocios, donde ha habido mayor protagonismo de los datos, se ha demostrado que los análisis basados en estos han tenido efectos positivos importantes para las empresas.

Las áreas de recursos humanos almacenan un gran volumen de datos con variedad de información, lo cual es el principal insumo para emprender la implementación de herramientas analíticas que permitan su procesamiento y encontrar relaciones que a simple vista no es posible, y de esta manera optimizar recursos económicos y modificar o reforzar las estrategias y acciones que vienen tomando para abordar la problemática.

Tal como ya se discutió, las empresas pierden valor debido a la rotación de empleados y se hace importante que los esfuerzos que realizan sean dirigidos a la fidelización de los empleados con la empresa; es por esto, que surge la necesidad de utilizar herramientas estadísticas, de machine learning y de analítica en general, que ayuden a tomar decisiones en las áreas de recursos humanos fundamentadas en los datos para reducir la cantidad de personas que abandonan sus cargos antes de lo esperado.

1.3. Revisión de literatura

En esta sección se presentan los trabajos que aparecen en la literatura más relevante que abordan el problema de rotación de personal en las empresas.

Harris, Craig y Light (2011) plantearon el análisis de datos sobre el talento humano como factor determinante para mejorar el retorno sobre la inversión (ROI) que se hace en el recurso humano de las organizaciones; sin embargo, en la práctica estos análisis han quedado rezagados si son comparados con los esfuerzos que se realizan corrientemente en las áreas que son el core del negocio como marketing, TI y CRM. Estos mismos autores plantean que las empresas líderes están utilizando seis estrategias para mejorar la conexión entre las inversiones en recursos humanos y los retornos comerciales; estas herramientas son: el desarrollo de bases de datos detalladas de los empleados, la segmentación del talento para valorar el recurso humano, las inversiones focalizadas en grupos humanos de interés, la personalización de la propuesta de

valor para el empleado, la planificación de la fuerza laboral a largo plazo para hacerla más eficiente, y el desarrollo de cadenas de suministro de talento humano.

Fan, Chan y Chang (2012) aplicaron minería de datos y machine learning para predecir las tendencias en las tasas de rotación de personal profesional en tecnología para empresas de Taiwan; en ese trabajo, se utilizó una red neuronal artificial híbrida y mapas auto-organizativos (SOM) para estudiar las características individuales de los grupos con tendencia a la rotación. Este estudio reporta como principales factores de rotación la falta de identificación de los empleados con la empresa, las falencias de la cultura de liderazgo y estrategias deficientes de fidelización interna.

Saradhi y Palshikar (2010) plantean que la rotación de clientes es un problema notorio para la mayoría de las industrias, ya que la pérdida de un cliente afecta los ingresos y la imagen de marca, y es difícil adquirir nuevos clientes. Estos autores indican que la rotación de empleados tiene impactos similares a la rotación de clientes, ya que la falta del recurso humano lleva a interrupciones en los procesos, insatisfacción de las personas, y tiempo y esfuerzos perdidos en la búsqueda y capacitación de nuevos empleados. Los autores utilizaron las técnicas de machine learning utilizadas en la predicción del abandono de clientes para desarrollar un modelo que les permitió identificar los empleados valiosos susceptibles de abandonar la compañía. Estos resultados fueron utilizados para diseñar mejores planes de retención de empleados y mejorar la satisfacción de estos.

King (2016) desarrolló un proyecto para la identificación de empleados con una alta probabilidad de abandonar la empresa. El autor utilizó técnicas de modelado como árboles de decisión, máquinas potenciadas con gradiente, regresión logística y k vecinos más próximos, donde el modelo con mejor desempeño fue la máquina potenciada con gradiente. El estudio destaca como principal razón de deserción de los empleados el tiempo en el cargo, es decir, cuanto más tiempo permanezca un empleado en un cargo sin recibir un ascenso, más probabilidades tendrá de abandonar la empresa. Este resultado es consistente con la investigación de Fitz-enz y Mattox (2014), donde los autores identificaron que permanecer demasiado tiempo en un trabajo tiene una alta correlación con la rotación de empleados y otros problemas de personal.

1.4. Definición del problema de tesis y pregunta de investigación

En este trabajo final de maestría se desea construir un modelo de predicción para identificar los colaboradores que están propensos a renunciar voluntariamente en la Caja de Compensación.

La pregunta de investigación que se pretende responde en este trabajo es:

> ¿Es posible construir un modelo de pronóstico que permita determinar cuál es la propensión a la renuncia voluntaria?

1.5. Hipótesis

Es posible desarrollar una herramienta que implemente técnicas de estadística o de machine learning para determinar la propensión a la renuncia de un empleado, a partir de la información histórica del personal y su rotación que tiene una empresa en sus bases de datos.

1.6. Objetivos

1.6.1. Objetivo General

Desarrollar una herramienta que implemente técnicas de estadística o de machine learning para determinar la propensión a la renuncia de un empleado, a partir de la información histórica del personal y su rotación que tiene una empresa en sus bases de datos.

1.6.2. Objetivos Específicos

1. Realizar la recolección, limpieza y depuración de la información disponible para el análisis de la rotación de personal.
2. Desarrollar un modelo de clasificación que permita identificar qué empleados tienen un alto riesgo de renunciar.
3. Desarrollar un prototipo de una herramienta de datos que implemente el modelo y permita determinar el riesgo de renuncia de un grupo de empleados en una organización.

1.7. Metodología

Esta propuesta de trabajo se llevará a cabo de acuerdo con la metodología CRISP-DM, que son las siglas de Cross Industry Standard Process for Data Mining.

Fase 1: Identificación de las necesidades del negocio, la cual consiste en establecer los resultados esperados del proyecto y entender la situación actual básicamente.

Fase 2: Estudio y comprensión de los datos: en esta fase de entendimiento de datos se realizará la colección de datos inicial y continuará con las actividades que permiten familiarizarse con los datos, identificar los problemas de calidad, y descubrir subconjuntos interesantes para formar hipótesis.

Fase 3: Análisis de los datos y selección de características: aquí se realizará la preparación de datos para construir el conjunto final de datos. Las tareas incluyen la selección de variables, registros y atributos, así como la transformación y la limpieza de datos.

Fase 4: Modelado: en esta fase, se seleccionan y aplican las técnicas de modelado que sean pertinentes al problema y se calibran sus parámetros a valores óptimos.

Fase 5: Evaluación u obtención de resultados: esta etapa consiste en revisar los resultados de los modelos aplicados y compararlos con los objetivos de negocio, con el fin de seleccionar el modelo que mejor se adapte.

Fase 6: Despliegue o puesta en producción: el trabajo finalizará con la generación de un informe sobre el modelo implementado.

2. Comprensión del Negocio

En este capítulo se desarrolla la Fase de compresión del negocio de acuerdo con la metodología CRISP-DM. Los objetivos de este capítulo son:

- Establecer los resultados del proyecto.
- Valorar la situación actual.
- Definir los objetivos de la minería de datos.
- Formalizar el plan del proyecto.

2.1. Resultados del proyecto

Tal como se indicó y se justificó en el Capítulo 1 de este trabajo final de maestría, los objetivos son los siguientes:

- El objetivo general es: desarrollar una herramienta que implemente técnicas de estadística o de machine learning para determinar la propensión a la renuncia de un empleado, a partir de la información histórica del personal y su rotación que tiene una empresa en sus bases de datos.
- Los objetivos específicos son:
    - Realizar la recolección, limpieza y depuración de la información disponible para el análisis de la rotación de personal.
    - Desarrollar un modelo de clasificación que permita identificar qué empleados tienen un alto riesgo de renunciar.
    - Desarrollar un prototipo de una herramienta de datos que implemente el modelo y permita determinar el riesgo de renuncia de un grupo de empleados en una organización.

El plan preliminar de desarrollo del proyecto fue discutido en la Sección 1.6, en la cual se establece que se utilizará la metodología CRISP-DM. Los resultados obtenidos al aplicar esta metodología son presentados en este capítulo y los siguientes.

Este trabajo final se considerará exitosa si se obtiene un modelo que permita pronosticar la propensión a la renuncia voluntaria de los colaboradores de la entidad. Es necesario aclarar que, desde el punto de vista de la analítica, el proyecto realmente debería buscar la reducción de la tasa de renuncia de los colaboradores; esto implica que los resultados del modelo deberían ser utilizados para el desarrollo de planes que mejoren las condiciones de los colaboradores, para que estos eviten que tomen la decisión de renunciar. Los resultados también pueden utilizarse para el diseño de planes de retención y permanencia. En otras palabras, el éxito real

del proyecto (mas no de este trabajo final de maestría) está asociado a la reducción de la tasa de renuncia de los colaboradores; sin embargo, esta situación está por fuera del alcance de un trabajo final de maestría debido a los tiempos requeridos para la medición de la efectividad de los planes de retención y permanencia.

2.2. Valoración de la situación actual

Para realizar este proyecto, se encuentran disponibles las bases de datos de personal y de valoración de cargos; parte de la información requerida se encuentra disponible en el sitio web del Ministerio de Educación.

Como requerimientos se espera que el modelo pronostique la propensión a la renuncia voluntaria de los colaboradores. Se espera que la precisión del modelo sea mínima del 70%; este valor fue definido por criterio experto, ya que a la fecha no se cuenta con estudios previos dentro de la empresa.

Como entradas al modelo se esperan incluir variables que reflejen la estructura organizacional. El uso de estas variables puede ser un riesgo muy importante para el éxito del trabajo debido a que un cambio profundo en dicha estructura puede invalidar las fuentes de datos con que se entrena el modelo.

No se consideran costos y beneficios, ya que el proyecto es desarrollado como un trabajo final de maestría.

2.3. Objetivos de la minería de datos

El objetivo de la minería de datos en este trabajo final es desarrollar un modelo de clasificación binaria que permita predecir la propensión a la renuncia voluntaria de los colaboradores; este modelo utilizará como variables explicativas relacionadas con salarios, estructura organizacional e información personal del colaborador.

Como criterio de éxito del negocio se define la reducción de la tasa de renuncia del valor actual que es de 12,2% al valor 8%. Esto se logra a partir del desarrollo de planes de retención y permanencia, lo cual está por fuera del alcance de este trabajo final de maestría.

Como criterio de éxito de la minería de datos se establece lograr una precisión mínima del 70%.

2.4. Plan del proyecto

El proyecto será desarrollado dentro de un semestre académico, de acuerdo con lo establecido en la reglamentación de la Universidad.

El proyecto será desarrollado utilizando la librería scikit-learn del lenguaje Python. Específicamente de esta librería se probarán diferentes modelos de clasificación binaria.

3. Comprensión y preparación de los Datos

En este capítulo se desarrollarán las Fases 2 y 3 de la metodología CRISP-DM, la cuales tienen los siguientes objetivos:

- Recopilar los datos iniciales.
- Describir los datos obtenidos.
- Realizar el análisis exploratorio y la verificación de calidad de los datos.
- Describir los datos disponibles.
- Transformar y eliminar características.

3.1. Recopilación de datos iniciales

La información utilizada para el desarrollo de este trabajo fue obtenida de las bases datos de la empresa, la cual contiene la información de los colaboradores que han renunciado voluntariamente y a los que permanecen activos en la Caja de Compensación Comfenalco Antioquia S.A, para el periodo comprendido entre enero de 2010 y abril de 2021. La descripción detallada de los datos aparece en la Sección 3.2.

Los datos utilizados están sujetos a un acuerdo de confidencialidad, bajo el cual la Caja de Compensación los pone a disposición, por lo cual esta información no aparecerá en este trabajo final. La recolección de esta información no significa costos adicionales para la empresa ni para la universidad.

La información inicial fue recolectada en tablas que provienen de las siguientes fuentes:

- Información sobre la estructura organizacional y datos referentes a los colaboradores obtenida de SAP.
- Salarios Mínimos Mensuales Legales Vigentes para cada año desde 2010 hasta 2021 de la página del Ministerio de Educación.
- La valoración realizada por el equipo de Compensación para cada cargo que se ha creado y los salarios referentes a la mediana en el mercado de acuerdo con la valoración del cargo, la cual proviene de un archivo de Excel.

El archivo histórico obtenido de SAP contiene la totalidad de la información disponible con que cuenta la Caja de Compensación.

No se considerarán variables referentes a habilidades blandas como las referentes a reconocimientos a los colaboradores, sistema de medición del desempeño, engagement, riesgo psicosocial y plan carrera. Estos sistemas de medición cuentan con un historial entre 1 y 3 años, por lo que no serían datos suficientes ni significativos para su uso en un modelo.

3.2. Descripción de los datos

La información recolectada para este trabajo contiene un total de 3.414 registros, de los cuales 1.660 corresponden a colaboradores que han renunciado voluntariamente, y 1.754 a colaboradores que permanecen activos en la fecha de corte.

A continuación, se describen cada uno de los datos recopilados:

- Personnel No.: es la identificación de cada colaborador dentro de la organización, toma tantos valores como colaboradores existen.

- Cargo: cargo que ocupa el colaborador.

- Clasificación cargo: determina si el cargo es administrativo o de servicio.

- H mes: son la cantidad de horas de la jornada que tiene un colaborador, puede tomar los siguientes valores en horas:
    - 90.
    - 120.
    - 150.
    - 180.
    - 210.
    - 240.

- Jornada: se refiere al tipo de jornada en que está clasificado el colaborador, toma los siguientes valores:
    - TC cuando el colaborador labora 240 horas al mes.
    - JP cuando el colaborador labora 90, 150, 180 o 210 horas al mes.
    - MT cuando el colaborador labora medio tiempo (120 horas).

- Sueldo: salario que devenga el colaborador en su momento de retiro o en la actualidad estando activo en la empresa. El valor está dado en pesos colombianos.

- SMMLV: Salario Mínimo Mensual Legal Vigente en Colombia dependiendo del año del retiro o de la actualidad en caso de que el colaborador se encuentre activo.

- Cantidad SMMLV: cantidad de SMMLV que recibe o recibió el colaborador.

- Nivel: es el nivel del cargo que ocupa el colaborador, pueden ser:
    - OPERATIVO.

    - ADMIN ASISTENCIAL.
    - PROFESIONAL.
    - COORDINADORES.
    - DIRECTIVO MEDIO.
    - DIRECTIVO.

- Género: es el género del colaborador, puede tomar los valores de:
    - FEMENINO.
    - MASCULINO.

- Edad del colaborador: es la edad del colaborador, está dada en años. En el caso de los colaboradores que renunciaron es la edad que tenían al momento de renunciar.

- Antigüedad: se refiere a la antigüedad del colaborador, está dada en años. En el caso de los colaboradores que renunciaron es la antigüedad que tenían al momento de renunciar.

- Tipo salario: se refiere a si el salario del colaborador es Ordinario o Integral.

- Factor Prestacional: se refiere a los tipos de beneficios extralegales que tiene el colaborador así:
    - R1 para los colaboradores antiguos que tienen mayores beneficios.
    - R2 para los colaboradores vinculados a la Caja que tienen unos beneficios moderados.
    - R3 para aquellos colaboradores que no tienen beneficios extralegales.

- Contrato: se refiere al tipo de vinculación contractual que tiene el colaborador con la Organización, donde se pueden encontrar los siguientes valores:
    - Indefinido.
    - Fijo a 1 año
    - Fijo por obra- labor.
    - Fijo o inf. a 1 año
    - Acdo Coop Trab. Asoc.

- Gerencia, Departamento actual, Departamento actual 2, Coordinación actual, Texto Nivel 7 Un. Organizativa: estas variables se refieren a la dependencia organizacional del colaborador, el área a la que pertenece. Estas variables pueden tomar más de 460 valores diferentes.

- Clasificación área: determina a qué tipo de área pertenece toda la dependencia organizacional del colaborador. Los valores que toma esta variable son:

- INFRAESTRUCTURA.
- SOCIAL.
- ADMINISTRATIVA.
- FINANCIERA.
- MERCADEO.
- VENTAS.
- LOGISTICA.
- CULTURA Y BIBLIOTECAS.
- EDUCACION.
- PARQUES Y HOTELES.
- GESTION HUMANA.
- PROYECTOS.
- CENTRO SERVICIOS.
- TECNOLOGIA.
- VIVIENDA.
- CONVENIO.
- COMUNICACIONES.
- DIRECCION.
- AUDITORIA.
- CONTRATO.
- EMPLEO.
- TRANSFORMACION.
- CULTURA Y BIBLIOTECA.
- FONDO EMPLEADOS.

- Clasificación tipo área: determina si el área es de la prestación de servicios o es un área transversal en la organización. Puede tomar los siguientes valores.
  - Área servicio.
  - Área transversal.

- Región: se refiere a la ubicación geográfica regional del colaborador, puede tomar los siguientes valores:
  - Ant. Aburrá Sur.
  - Ant. Aburrá Medellín.
  - Ant. Oriente.
  - Ant. Aburrá Norte.
  - Ant. Norte.
  - Ant. Occidente.
  - Bogotá.
  - Ant. Urabá.
  - Córdoba.

    - Ant. Suroeste.
    - OBS.Ant. Nordeste.
    - Quindio.
    - OBS.Ant. Bajo Cauca.
    - Ant. Norte-Magd.Medio.
    - Santander.

- Grupo personal: es la forma en la que la Organización formaliza el contrato con el colaborador, puede ser Directa cuando existe vinculación del colaborador a la Organización, o Externa cuando la contratación se hace por medio de terceros o también llamados empresas temporales.

- Ascenso: hace referencia a si el colaborador ha tenido o tuvo cambios de cargos que implican aumento en el nivel de estos. Toma los valores de Si o No.

- Formación: hace referencia al nivel de formación del colaborador, ya sea porque lo haya reportado en algún momento al área de Gestión Humana o porque es el mínimo nivel para ocupar el cargo que desempeña. Puede tomar los valores de:
    - Primaria.
    - Bachiller.
    - Técnico.
    - Tecnólogo.
    - Profesional.
    - Especialista.
    - Máster.

- Fecha retiro: corresponde al año en que el colaborador se retiró, en caso de que continúe activo tomará el valor ACTIVO.

- Categoría: hace referencia a la categoría numérica que arroja la metodología de valoración de cargos que tiene implementada la empresa, es necesaria para poder determinar salarios competitivos en el mercado.

- Posicionamiento: hace referencia a la comparación del salario del colaborador con el salario de referencia que tiene el cargo que desempeña en el mercado. Este valor está dado en porcentaje.

- Rango posicionamiento: es el agrupamiento de los porcentajes del posicionamiento, así:

    - Sub-pagado, donde se agrupan los valores de posicionamiento menores a 80%, lo que significa que el colaborador no está siendo compensado de una manera competitiva con el mercado.
    - Equilibrado, donde se agrupan los valores de posicionamientos entre 80% y 120%, que indican que el salario del colaborador es competitivo en el mercado.
    - Sobre-pagado, que agrupa los valores de posicionamiento superiores a 120%, que indican que el colaborador está siendo mejor pago que la mediana del mercado.

- RENUNCIA: indica si el colaborador permanece ACTIVO en la Organización o si presentó RENUNCIA VOLUNTARIA en algún momento. Es la variable dependiente.

3.3. Exploración y verificación de la calidad de los datos

Una vez descrita la información disponible, se procede a explorar y verificar los datos, lo que permite hacer un primer acercamiento, y determinar la consistencia y completitud de la información. La data recopilada está compuesta inicialmente por 34 variables y 3.414 registros.

La variable dependiente corresponde al campo RENUNCIA e indica si el colaborador ha renunciado voluntariamente a la Caja o permanece activo; esta columna registra 1.660 casos activos y 1.754 casos de renuncia voluntaria. No se tuvieron en cuenta los colaboradores que por otros motivos fueron retirados de la Caja, porque no contribuyen a la clasificación que se desea predecir, ni los que renunciaron antes del 2010, ya que no existe registro de ellos.

En el análisis preliminar se encontró que la variable Coordinación actual tiene 659 valores nulos y la variable Texto Nivel 7 Un.Organizaiva tiene 2.591 valores nulos, para un total de 3.250 valores nulos. Por otra parte, se verificó que la variable Personnel No. contenga valores únicos.

También se encontró que 27 de las 34 variables disponibles son categóricas.

A continuación, se describe la distribución de los colaboradores activos y que han renunciado en las diferentes variables que se tuvieron en cuenta para este modelamiento.

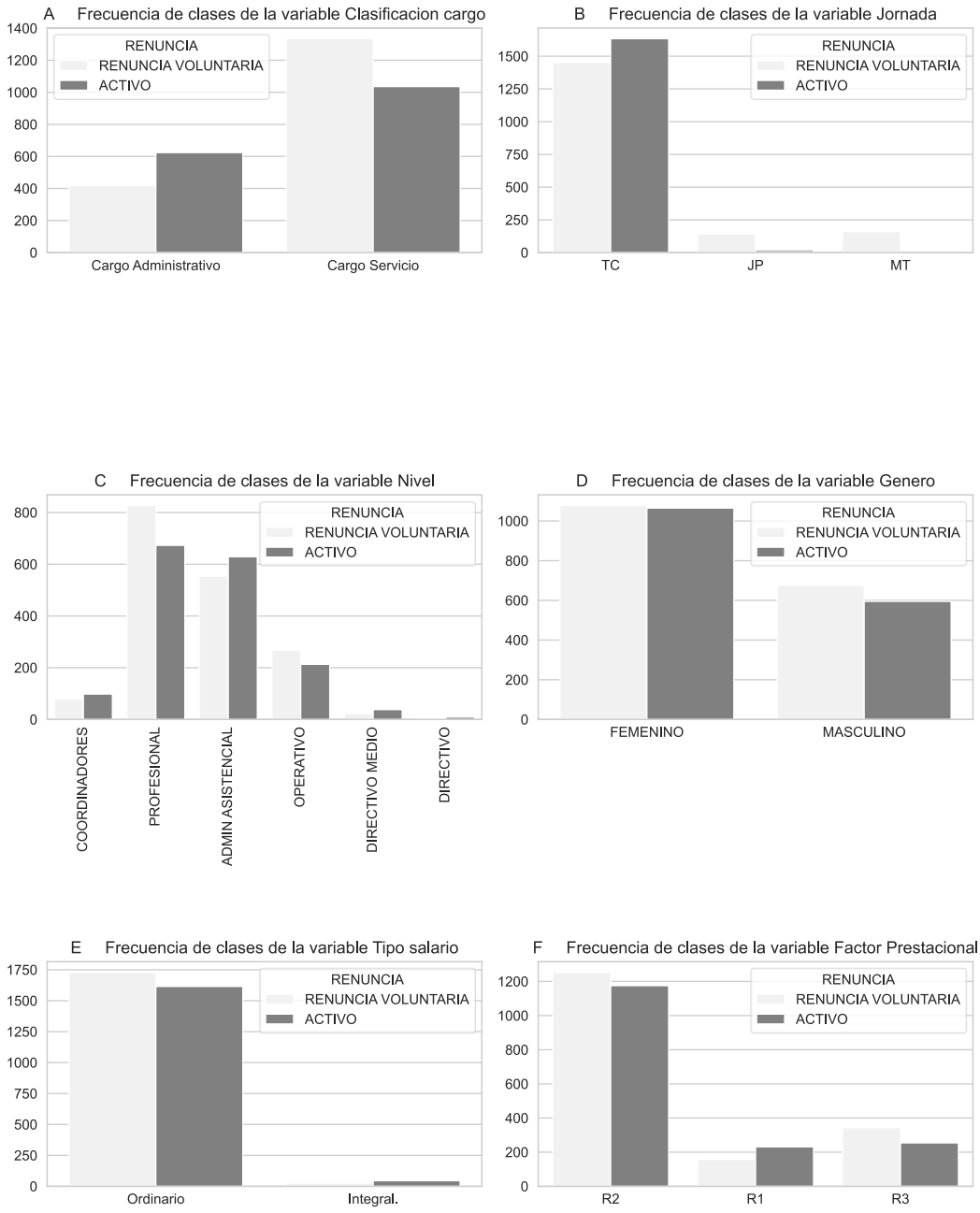

Figura 1. Frecuencias de variables categóricas 1.

Figura 1. Continuación

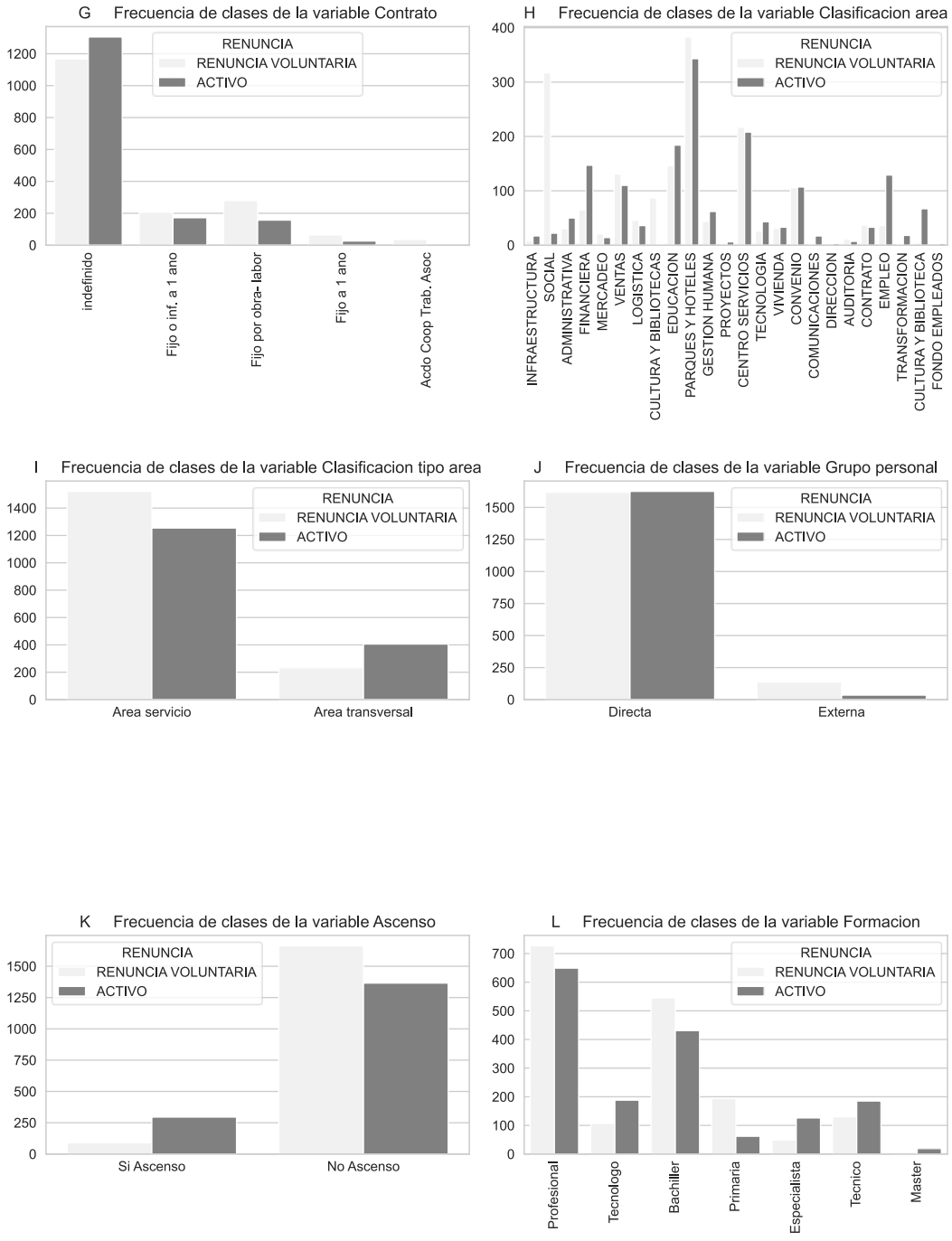

Figura 1. Frecuencias de variables categóricas 2.

Figura 1. Continuación

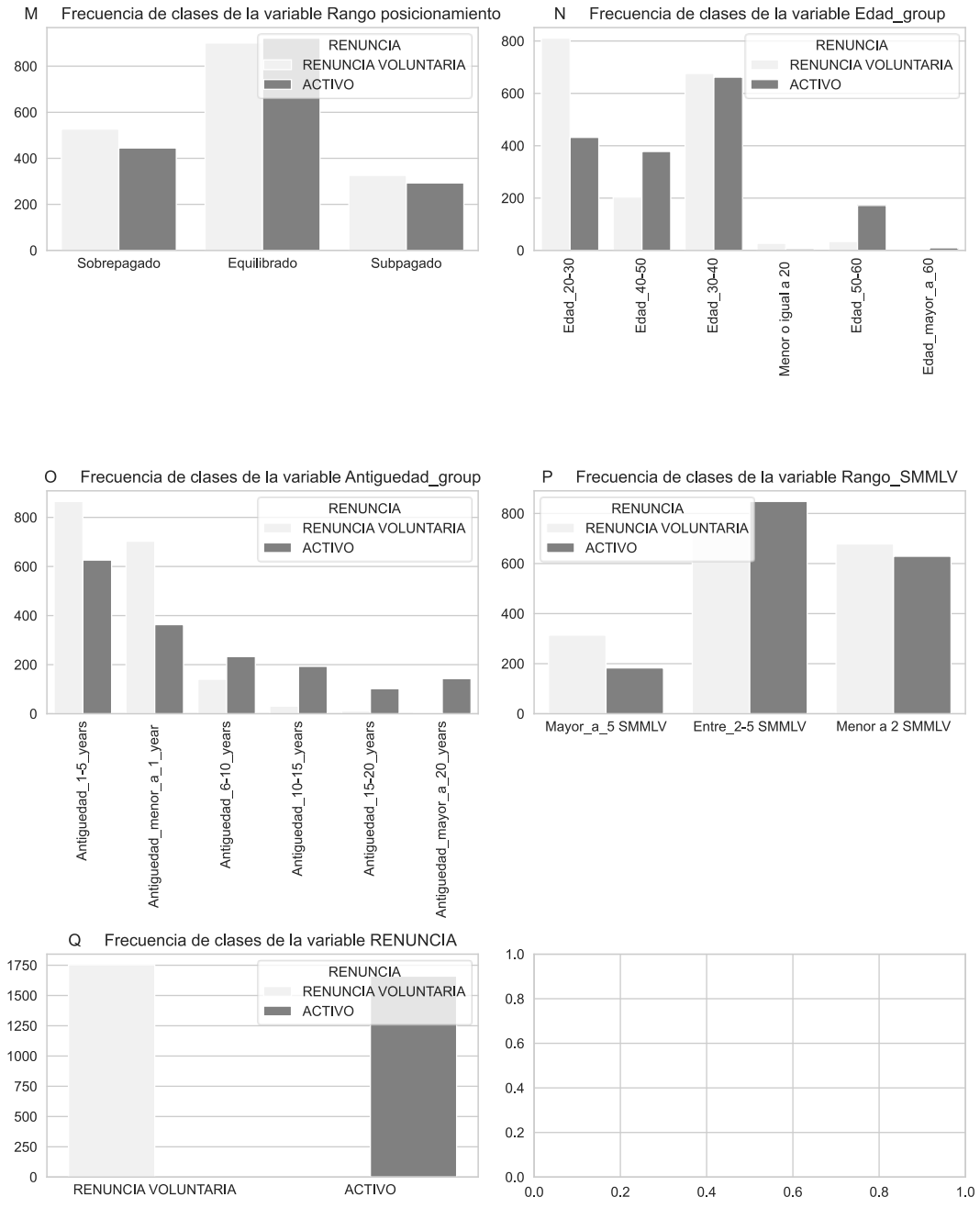

Figura 1. Frecuencias de variables categóricas 3.

En la Figura 1 se observa lo siguiente:

- (A) La mayoría de los colaboradores tienen cargos orientados al servicio, y es en este tipo de cargo donde hay una mayor proporción de renuncias voluntarias respecto a los colaboradores activos.
- (C) La mayor cantidad de colaboradores tienen cargos de nivel profesional y administrativo asistencial, y en el nivel profesional se observa una mayor rotación de personal.
- (D) Hay más colaboradores de género femenino que masculino, sin embargo, la renuncia voluntaria es equiparable con los activos en cada género.
- (F) La mayor cantidad de colaboradores a lo largo de la historia de la Caja han tenido un factor prestacional R2, y que la menor cantidad de renuncias voluntarias respecto a los colaboradores activos se ha dado en los colaboradores con factor prestacional R1, lo que indica que tener una mayor cantidad de beneficios puede evitar que los colaboradores quieran renunciar voluntariamente.
- (G) La mayoría de colaborados activos y que han renunciado han tenido un contrato a término indefinido, sin embargo, la proporción de renuncias voluntarias es mayor respecto a los colaboradores que continúan activos en el resto de los tipos de contrato. Lo anterior indica que esta es una variable importante para que los colaboradores tomen la decisión de renunciar o no.
- (I) Hay una menor cantidad de renuncias voluntarias en las áreas transversales, mientras que en las áreas de servicio la cantidad de colaboradores que han renunciado es mayor a los que permanecen activos.
- (J) Existe una mayor cantidad de contratación directa de los colaboradores, y también una proporcionalidad entre los que han renunciado y los que permanecen activos, es en la contratación externa, por temporal, donde se observa que hay una mayor cantidad de renuncias que de colaboradores activos.
- (K) Hay una menor cantidad de colaboradores que han tenido ascensos dentro de la Caja, y que hay una mayor proporción de renuncias voluntarias entre los colaboradores que no han tenido ascensos.
- (L) La mayor cantidad de colaboradores son profesionales, seguidos por los bachilleres, y en ambos grados de formación se observa una mayor proporción de renuncias voluntarias.
- (M) La mayor parte de colaboradores se encuentran en un posicionamiento equilibrado, y que en este hay valores semejantes entre renuncias voluntarias y activos, es en los colaboradores sobre pagados donde se identifica una mayor proporción de colaboradores que han renunciado voluntariamente sobre los que permanecen activos.

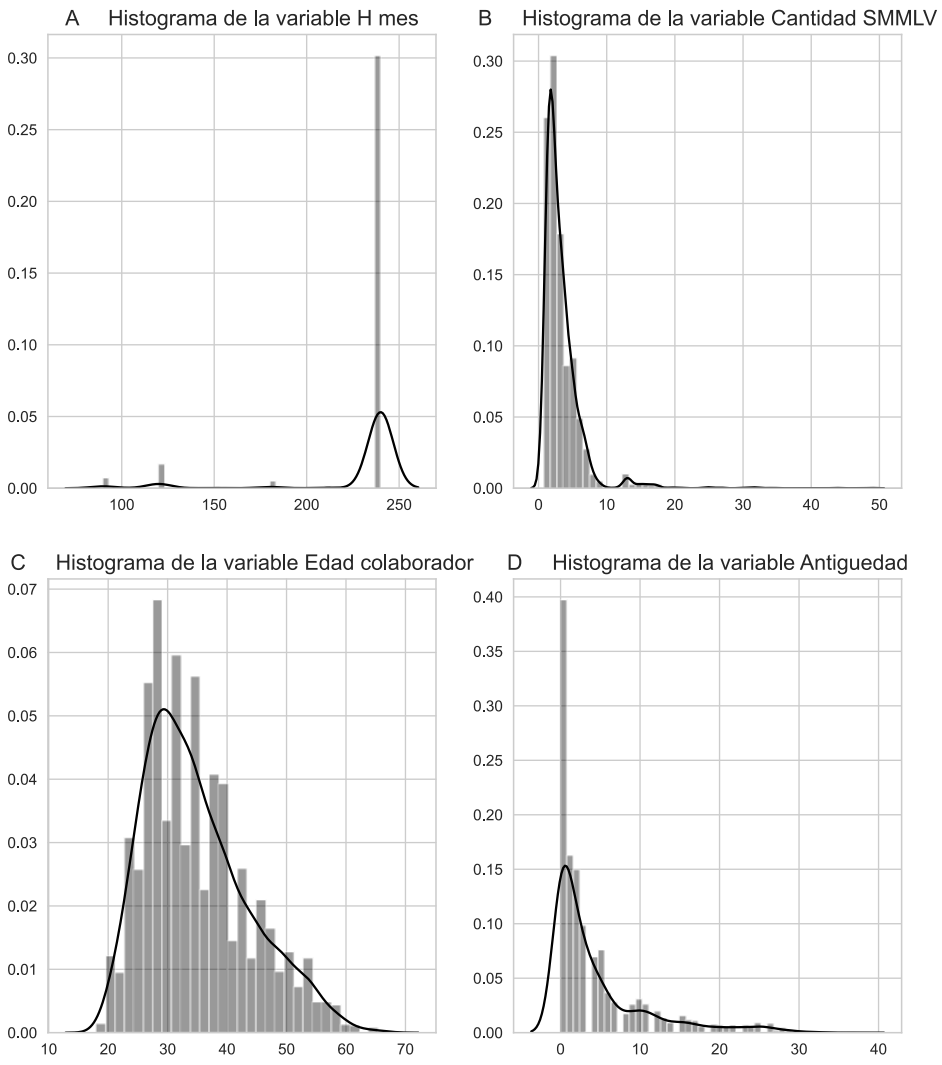

Figura 2. Histogramas variables numéricas.

En la Figura 2 se observa lo siguiente:

- (A) La mayor cantidad de colaboradores tienen asignada una jornada de 240 horas al mes.

- (B) La mayor cantidad de SMMLV que se asignan están por debajo o iguales a 5 SMMLV.
- (B) Como mínimo los colaboradores devengan 1 SMMLV, lo que indica una buena calidad de los datos. También se puede identificar que la mayor cantidad de colaboradores devengan alrededor de los 2 SMMLV.
- (C) La mayor cantidad de colaboradores están entre los mayores a 20 hasta los 40 años.
- (C) La edad mínima de los colaboradores está alrededor de los 18 años y la máxima a de los 62 años, lo que indica que está dentro de los valores esperados de la mayoría de edad y la edad de pensión de los hombres, rango en el cual se espera que se desarrolle la vida productiva laboralmente.
- (D) La mayoría de los colaboradores tienen una antigüedad en la Caja de Compensación menor a 1 año. También se puede observar que a medida que la antigüedad aumenta la cantidad de colaboradores disminuye, lo cual es normal en la dinámica de las organizaciones.

La anterior exploración y verificación de datos evidencia que se deben corregir los valores nulos y replantear el uso de algunas variables que han tenido cambios al pasar el tiempo desde 2010 hasta la actualidad y que para efectos de la predicción podrían generar ruido en el modelo y que en el futuro con valores nuevos en estas variables dicho modelo pierda vigencia.

3.4. Ingeniería de características

Las transformaciones aplicadas y las nuevas variables creadas para ser usadas en los modelos de clasificación son las siguientes:

- Se eliminaron las siguientes variables:
  - Personnel No.
  - Cargo.
  - Sueldo.
  - SMMLV.
  - Gerencia.
  - Departamento actual.
  - Departamento 2 actual.
  - Coordinación actual.
  - Texto Nivel 7 Un.Organizativa.
  - Región.
  - Txt.subd.pers.
  - Fecha retiro.
  - Categoría.

- Salario REF.
- Posicionamiento.

La eliminación de estas variables es validada con expertos en el proceso de Gestión Humana el cual monitorea el problema de rotación de personal por renuncia voluntaria.

- La variable edad del colaborador se recodificó en 6 intervalos:
  - Menor o igual a 20.
  - Edad 20-30.
  - Edad 30-40.
  - Edad 40-50.
  - Edad 60-60.
  - Edad mayor a 60.

- La variable antigüedad del colaborador se recodificó en 6 intervalos por años:
  - Antigüedad menor a 1.
  - Antigüedad 1-5.
  - Antigüedad 5-10.
  - Antigüedad 10-15.
  - Antigüedad 15-20.
  - Antigüedad mayor a 20.

- La variable cantidad de SMMLV se recodificó en 3 intervalos:
  - Menor a 2 SMMLV
  - Entre 2 y 5 SMMLV
  - Mayor a 5 SMMLV

De acuerdo con lo anterior, la cantidad de registros se conserva y las variables se reducen a 21.

- Se aplicó la técnica OneHotEncoder a las 17 variables categóricas del dataset:
  - Clasificación cargo
  - Jornada
  - Nivel
  - Genero
  - Tipo salario
  - Factor Prestacional
  - Contrato
  - Clasificación área
  - Clasificación tipo área
  - Grupo personal
  - Ascenso

- Formación
- Rango posicionamiento
- Rango edad
- Rango antigüedad
- Rango SMMLV
- RENUNCIA

Cuando se aplica la técnica OneHotEncoder al campo RENUNCIA se crean dos campos nuevos, llamados ACTIVO y RENUNCIA VOLUNTARIA. Ya que ambos codifican la misma información se elimina el campo ACTIVO.

Es así, que las 17 variables categóricas se convierten en 79 variables dummies, para un total de 83 variables disponibles para utilizar en la aplicación de los modelos de machine learning.

- Las variables dummies creadas fueron las siguientes:
  - Cargo Administrativo
  - Cargo Servicio
  - JP
  - MT
  - TC
  - ADMIN ASISTENCIAL
  - COORDINADORES
  - DIRECTIVO
  - DIRECTIVO MEDIO
  - OPERATIVO
  - PROFESIONAL
  - FEMENINO
  - MASCULINO
  - Integral.
  - Ordinario
  - R1
  - R2
  - R3
  - Acdo Coop Trab. Asoc
  - Fijo a 1 año
  - Fijo o inf. a 1 año
  - Fijo por obra- labor
  - Indefinido
  - ADMINISTRATIVA
  - AUDITORIA

- CENTRO SERVICIOS
- COMUNICACIONES
- CONTRATO
- CONVENIO
- CULTURA Y BIBLIOTECA
- CULTURA Y BIBLIOTECAS
- DIRECCION
- EDUCACION
- EMPLEO
- FINANCIERA
- FONDO EMPLEADOS
- GESTION HUMANA
- INFRAESTRUCTURA
- LOGISTICA
- MERCADEO
- PARQUES Y HOTELES
- PROYECTOS
- SOCIAL
- TECNOLOGIA
- TRANSFORMACION
- VENTAS
- VIVIENDA
- Área servicio
- Área transversal
- Directa
- Externa
- No Ascenso
- Si Ascenso
- Bachiller
- Especialista
- Máster
- Primaria
- Profesional
- Técnico
- Tecnólogo
- Equilibrado
- Sobre-pagado
- Sub-pagado
- Edad_20-30
- Edad_30-40
- Edad_40-50

- Edad_50-60
- Edad_mayor_a_60
- Menor o igual a 20
- Antiguedad_1-5
- Antiguedad_10-15
- Antiguedad_15-20
- Antiguedad_6-10
- Antiguedad_mayor_a_20
- Antiguedad_menor_a_1
- Entre_2-5 SMMLV
- Mayor_a_5 SMMLV
- Menor a 2 SMMLV
- RENUNCIA VOLUNTARIA

4. Modelos

En este capítulo se desarrollará la Fase 4 de la metodología CRISP-DM, la cual tiene los siguientes objetivos:

- Seleccionar las técnicas de modelado.
- Generar métricas de evaluación del desempeño de los modelos.
- Evaluar los modelos.

4.1. Modelos seleccionados

Como ya se mencionó anteriormente, este es un problema de clasificación binaria, donde se desea pronosticar la propensión de los colaboradores a renunciar voluntariamente o a permanecer activos en la Caja. De acuerdo con lo anterior, se consideran los siguientes modelos:

- Regresión Logística.
- Bosque Aleatorio.
- XGBoost.
- K Vecinos más Cercanos.
- Árbol de Decisión.

4.2. Partición de los datos y métricas de precisión

La base de datos para este trabajo contiene 3.414 registros de colaboradores, de los cuales 1.660 han renunciado voluntariamente 1.754 que permanecen activos; la totalidad de los registros fueron segmentados así: 80% para el entrenamiento de los modelos, y el 20% restante para la evaluación de estos.

En este trabajo se utilizarán las métricas definidas a continuación:

- Matriz de confusión: esta métrica entrega una tabla con la cantidad de registros de la predicción clasificados como Verdaderos Negativos [VN], Verdaderos Positivos [VP], Falsos Positivos [FP] y Falsos Negativos [FN].

- Exactitud:

$$Exactitud = \frac{VP + VN}{VP + VN + FP + FN}$$

- Precisión:

$$Precisión = \frac{VP}{VP + FP}$$

- Sensibilidad:

$$Sensibilidad = \frac{VP}{VP + FN}$$

- Valor F:

$$Valor\ F = \frac{Precisión * Sensibilidad}{Precisión + Sensibilidad}$$

- Área bajo la curva:

$$Área\ bajo\ la\ curva = \frac{VP}{VP + FN} + \frac{VN}{FP + VN} - 1$$

4.3. Entrenamiento de los modelos y búsqueda de parámetros óptimos

Para el entrenamiento de cada modelo y la obtención de los parámetros óptimos se utilizaron las técnicas de particionamiento de grupos estratificados (implementada en la función StratifiedKFold de la librería sklearn) y de búsqueda en malla con validación cruzada (implementada en la función GridSearchCV).

En la Figura 3 se presenta el fragmento de código donde se especifica el proceso de búsqueda del modelo óptimo utilizando las técnicas mencionadas anteriormente.

```
#REGRESIÓN LOGÍSTICA
params_lr = {"solver": ["newton-cg","lbfgs","liblinear","sag","saga"],
             "C":np.logspace(1,10,2), "penalty":["l1", "l2", "elasticnet", "none"]}

kfold = StratifiedKFold(n_splits=2, shuffle=True, random_state=42)

lr_model = GridSearchCV(LogisticRegression(random_state=24),  #Busca los parámetros óptimos
                        param_grid=params_lr,
                        scoring=metricas, refit="f1",
                        return_train_score=True,
                        cv=kfold,n_jobs=-1,verbose=3)

results_lr = lr_model.fit(X_train, Y_train)

#BOSQUE ALEATORIO

n_estimators = [10,100,256] # Number of trees in random forest
max_features = ['auto','sqrt','log2'] # Number of features to consider at every split
max_depth = [16,32,64,128] # Maximum number of levels in tree
min_samples_split = [4,8,12] # Minimum number of samples required to split a node
min_samples_leaf = [1,3,5] # Minimum number of samples required at each leaf node
bootstrap = [False,True] # Method of selecting samples for training each tree

grid_dict = {'n_estimators': n_estimators,
               'max_features': max_features,
               'max_depth': max_depth,
               'min_samples_split': min_samples_split,
               'min_samples_leaf': min_samples_leaf,
               'bootstrap': bootstrap}

kfold = StratifiedKFold(n_splits=2, shuffle=True, random_state=42)

rf_model = GridSearchCV(estimator = RandomForestClassifier(random_state=24),  #Busca los parámetros óptimos
                           param_grid = grid_dict,
                           scoring=metricas,
                           refit="f1",
                           return_train_score=True,
                           cv=kfold, n_jobs=-1,verbose=1)
results_rf = rf_model.fit(X_train, Y_train)
```

Figura 3. Código entrenamiento de modelos 1.

Figura 3. continuación:

```
#AUMENTO DE GRADIENTE
n_estimators = [128,256,512]
learning_rate = [0.5,0.1,0.01]
tuple_eval = [(X_test, Y_test)]
params_xg = {"learning_rate":learning_rate,
             "n_estimators": n_estimators}

kfold = StratifiedKFold(n_splits=2, shuffle=True, random_state=42)

#Busca los parámetros óptimos
xg_model = GridSearchCV(xgb.XGBClassifier(random_state=42), params_xg, n_jobs=-1, cv=kfold, scoring=metricas, refit="f1")

results_xg = xg_model.fit(X_train, Y_train, early_stopping_rounds=10, eval_set=tuple_eval)

#K VECINOS MÁS CERCANOS

n_neighbors = [1,2,3,4,5,6,7,8,9,10]
algorithm = ['auto','ball_tree','kd_tree','brute']

params_knn = {'n_neighbors': n_neighbors,
            'algorithm': algorithm}

kfold = StratifiedKFold(n_splits=2, shuffle=True, random_state=42)

knn_model = GridSearchCV(KNeighborsClassifier(), param_grid=params_knn, #Busca los parámetros óptimos
                                            scoring=metricas, refit="f1",
                                            cv=kfold)

results_knn = knn_model.fit(X_train, Y_train)

#ÁRBOL DE DECISIÓN

criterion = ['gini','entropy']
splitter = ['best','random']
max_features = ['auto', 'sqrt', 'log2', 'None']

params_dt = {'criterion': criterion,
            'splitter': splitter,
            'max_features': max_features}

kfold = StratifiedKFold(n_splits=2, shuffle=True, random_state=42)

dt_model = GridSearchCV(DecisionTreeClassifier(random_state=42), param_grid=params_dt, #Busca los parámetros óptimos
                                            scoring=metricas, refit="f1",
                                            cv=kfold)

results_dt = dt_model.fit(X_train, Y_train)
```

Figura 3. Código entrenamiento de modelos 2.

Los modelos mencionados fueron evaluados con las siguientes transformaciones para los datos:

- Opción 1: OneHotEncoder
- Opción 2: OneHotEncoder + StandardScaler
- Opción 3: OneHotEncoder + StandardScaler + PCA
- Opción 4: OneHotEncoder + PCA

4.4. Resultados obtenidos

De acuerdo con la información anterior se obtuvieron los siguientes resultados después de correr los modelos en diferentes oportunidades.

En la Tabla 1 se presentan los resultados de las métricas obtenidos por los modelos en las diferentes transformaciones realizadas.

Tabla 1. Métricas modelos

Opción	Métrica	Regresión Logística	Bosque Aleatorio	XGBoost	K Vecinos más Cercanos	Árbol de Decisión
OneHotEncoder (Opción 1)	Exactitud	78,0%	83,6%	86,7%	76,4%	79,2%
	Precisión	77,6%	84,0%	87,3%	77,7%	80,0%
	Sensibilidad	79,4%	83,5%	86,1%	74,8%	78,5%
	Valor F1	78,5%	83,7%	86,7%	76,2%	79,2%
	Área bajo la curva ROC	78,0%	83,6%	86,7%	76,4%	79,2%
OneHotEncoder + StandardScaler (Opción 2)	Exactitud	78,5%	83,4%	86,7%	74,1%	79,2%
	Precisión	78,8%	83,9%	87,3%	73,7%	80,0%
	Sensibilidad	80,0%	83,1%	86,1%	75,6%	78,5%
	Valor F1	79,0%	83,5%	86,7%	74,7%	79,2%
	Área bajo la curva ROC	78,5%	83,5%	86,7%	74,1%	79,2%
OneHotEncoder + StandardScaler + PCA (Opción 3)	Exactitud	77,0%	80,0%	81,2%	75,8%	72,8%
	Precisión	76,5%	80,5%	82,6%	76,8%	73,7%
	Sensibilidad	78,5%	80,0%	79,7%	74,8%	71,6%
	Valor F1	77,5%	80,2%	81,1%	75,8%	72,6%
	Área bajo la curva ROC	77,0%	80,0%	81,3%	75,8%	72,8%
OneHotEncoder + PCA (Opción 4)	Exactitud	70,1%	69,4%	70,9%	68,8%	65,7%
	Precisión	67,4%	69,3%	69,9%	68,4%	66,3%
	Sensibilidad	79,1%	70,7%	74,2%	71,0%	65,5%
	Valor F1	72,8%	70,0%	72,0%	69,7%	65,9%
	Área bajo la curva ROC	70,0%	69,3%	70,8%	68,8%	65,7%

En la Tabla 1 se observa que:

- Las transformaciones presentan métricas similares y no hay mucha variabilidad entre las métricas para un mismo clasificador.
- Las métricas obtenidas con la Opción 2 de transformación presentan una disminución en el desempeño de los modelos respecto a la Opción 1, siendo la Opción 4 la que presenta mayor disminución respecto a las demás transformaciones.
- El modelo K Vecinos más Cercanos presenta la mayoría de las métricas con el desempeño más bajo en todas las transformaciones.
- En todas las transformaciones el modelo XGBoost es el que presenta mejores desempeños en la mayoría de las métricas, obteniendo el mejor para todas las métricas en la Opción 1, aplicando solo la transformación OneHotEncoder.

Los modelos fueron ejecutados en un computador con las siguientes especificaciones:

- Memoria RAM: 20,0 GB
- Procesador: Intel(R) Core(TM) i5-8250U CPU @ 1.60GHz 1.80 GHz
- Tipo de sistema: Sistema operativo de 64 bits, procesador x64
- Tipo de disco: unidad de disco estándar de 1 TB

Los tiempos para la ejecución de los modelos bajo las diferentes transformaciones se reportan con timeit y fueron los siguientes:

- Regresión Logística: 17 segundos.
- Bosque Aleatorio: 225 segundos.
- XGBoost: 17 segundos.
- K Vecinos más Cercanos: 22 segundos.
- Árbol de Decisión: 0,6 segundos.

4.5. Conclusiones

Respecto a los resultados presentados anteriormente se puede llegar a la conclusión que para este dataset es suficiente con realizar la transformación de variables con OneHotEncoder, debido a que no hay una mejora en el desempeño de los modelos evidenciada con las otras transformaciones realizadas. Además, se encontró que los tiempos de ejecución de todos los modelos fueron razonables y por lo tanto pueden ser reestimados fácilmente. Se concluye también, de acuerdo con los resultados presentados al inicio de este capítulo, que el modelo con mejor desempeño es XGBoost utilizando las variables obtenidas luego de la transformación con OneHotEncoder.

5. Evaluación u obtención de resultados

En este capítulo se desarrollará la Fase 5 de la metodología CRISP-DM, la cual tiene los siguientes objetivos:

- Evaluación de los resultados.
- Recalibración del modelo.
- Determinación de los pasos siguientes.

5.1. Evaluación de resultados

Como se evidenció en el Capítulo anterior el modelo con mejor desempeño es el XGBoost utilizando solo la transformación de los datos con la técnica OneHotEncoder, el cual tiene los siguientes resultados:

- Exactitud: 86,7%.
- Precisión: 87,3%.
- Sensibilidad: 86,7%.
- Valor F1: 86,7%.
- Área bajo la curva ROC: 86,7%.

Estos resultados cumplen con el criterio propuesto para este trabajo final de maestría, donde se espera que el modelo elegido pronostique la propensión a la renuncia voluntaria de los colaboradores con una precisión mínima del modelo del 70%, lo que indica que el modelo puede ser utilizado en producción.

5.2. Recalibración del modelo

El modelo puede ser recalibrado cuando se tengan al menos 170 datos nuevos (5% de la muestra original), de manera que haya suficientes datos para que sea posible un cambio significativo en los resultados de los modelos.

Se requiere una recalibración completa del modelo cuando se presenten estructuras organizativas muy diferentes a las actuales.

En un futuro se puede estudiar una forma diferente o más precisa de clasificar las áreas, con el fin de buscar que la estructura organizacional no pierda vigencia debido a las diferentes mutaciones que surgen a partir de los cambios de estrategias organizacionales, servicios prestados y/o procesos adoptados dentro de la Caja de Compensación, ya que estos cambios

representan un riesgo para el modelo porque al crear áreas nuevas dentro de la estructura organizacional, estas pueden generar desviaciones mayores en la predicción de los modelos.

Es importante que se estudie la posibilidad de crear modelos con históricos más pequeños que el actual, de manera que se puedan incluir variables referentes a habilidades blandas como reconocimientos a los colaboradores, sistema de medición del desempeño, engagement, riesgo psicosocial y plan carrera.

5.3. Pasos siguientes

Con la precisión y relevancia de los resultados obtenidos con el modelo XGBoost, la herramienta puede ser desplegada o puesta en producción para predecir la propensión a la renuncia voluntaria de los colaboradores de la Caja de Compensación.

6. Despliegue o puesta en producción

Este capítulo se trata sobre la Fase 6 de la metodología CRISP-DM. Los objetivos de esta fase son:

- Conceptos básicos sobre despliegue
- Planificación de despliegue
- Planificación del control y del mantenimiento
- Creación de un informe final
- Revisión final del proyecto

Debido a que el despliegue está más allá del alcance de este trabajo final de maestría, no se va a desarrollar esta parte de la metodología. La información con la que se cuenta es sensible y están sujetos a un acuerdo de confidencialidad donde será la Caja de Compensación quien decida el despliegue o la puesta en producción de la herramienta.

7. Conclusiones

7.1. Respuesta a la pregunta de investigación

La pregunta de investigación formulada en este trabajo fue:

> ¿Es posible construir un modelo de pronóstico que permita determinar cuál es la propensión a la renuncia voluntaria?

Como se puede observar en este trabajo final de maestría, se encontró que el modelo XGBoost permite pronosticar la propensión a la renuncia voluntaria de los colaboradores de la Caja de Compensación con una precisión de 87,3%, en el cual para determinar los parámetros óptimos se utilizaron las técnicas StratifiedKFold y GridSearchCV

7.2. Cumplimiento de objetivos

7.2.1. Objetivo específico 1: Realizar la recolección, limpieza y depuración de la información disponible para el análisis de la rotación de personal.

Como se pudo observar en el Capítulo 3 de este trabajo final, se presentó la recopilación de los datos iniciales, la descripción de estos, la exploración y verificación de su calidad, y finalmente su transformación para poder ser utilizados en los modelos de pronóstico definidos. En el desarrollo del capítulo mencionado se puede observar que la fuente de datos fue el ERP de la Caja de Compensación principalmente, con un histórico desde enero de 2010 a abril de 2021; también se puede observar que se eliminaron 15 variables en su mayoría relacionadas con la estructura organizacional por problemas con su continuidad en el tiempo y el riesgo que puede generar esto en los modelos de pronóstico, también se recodificaron 3 variables llevándolas a rangos de valores. El resultado de lo anterior fue obtener un dataset sin valores nulos ni atípicos, por lo que finalmente las variables se transformaron con la técnica OneHotEncoder para ser procesadas con los modelos de pronóstico.

7.2.2. Objetivo específico 2: Desarrollar un modelo de clasificación que permita identificar qué empleados tienen un alto riesgo de renunciar.

Como se pudo observar en el Capítulo 4 de este trabajo final se presentaron los cinco modelos probados para predecir la propensión a la renuncia voluntaria de los colaboradores de la Caja

de Compensación, se explicó la partición de los datos en 80% para entrenamiento y el 20% restante para la validación, se enunciaron las 6 métricas utilizadas para la evaluación del desempeño de los modelos y las 4 opciones de transformación que se realizaron a las variables con el fin de hallar los mejores desempeños. Finalmente, en este capítulo se concluye que el modelo XGBoost con la trasformación realizada en la Opción 1, fue el que obtuvo el mejor desempeño de todas las métricas utilizadas.

7.2.3. Objetivo específico 3: Desarrollar un prototipo de una herramienta de datos que implemente el modelo y permita determinar el riesgo de renuncia de un grupo de empleados en una organización.

Como se pudo observar en el Capítulo 5 de este trabajo final se realizó la evaluación de los resultados obtenidos de acuerdo con los modelos probados y sus métricas, allí se ratificó que el modelo con el mejor desempeño para predecir la propensión a la renuncia voluntaria de los colaboradores de la Caja de Compensación fue el XGBoost; además, se planteó que el modelo podía ser recalibrado en el momento en que se puedan obtener al menos un 5% más de registros de los 3.414 utilizados para el desarrollo de este trabajo (170 registros nuevos). Finalmente, se especificó que de acuerdo con los resultados obtenidos la herramienta diseñada puede ser puesta en producción.

8. Referencias

Moreno. N., Aplicaciones de la Analítica y la Minería de Datos en la Gestión de Recursos Humanos, Medellín, Colombia 2018

Ministerio de Educación, Histórico del Salario mínimo en Colombia (1894-2021). Obtenido de: https://ole.mineducacion.gov.co/portal/Contenidos/Documento/388408:Historico-del-Salario-minimo-en-Colombia-1894-2019

King, K.G, "Data Analytics in Human Resources: A Case Study and Critical Review", HR. Develop. Rev., vol. 15, no. 4, pp. 487-495, December 2016

Jac Fitz-enz and John Mattox, "Predictive Analytics for Human Resources", Hoboken, NJ: John Wiley, 2014.

J.V. Román, J.C.G. Cristóbal and J.A.G. Vázquez, "TALENT + Tecnologías avanzadas para la Gestión del Talento", Procesamiento de Lenguaje Natural, vol. 57, pp. 159-162, September 2016

J.G. Harris, E. Craig and D.A. Light, "Talent and analytics: new approaches, higher ROI", J. of Bus. Strat., vol. 32, no. 6, pp. 4-13, October 2011

L.F. Chen and C.F. Chien "Manufacturing intelligence for class prediction and rule generation to support human capital decisions for high-tech industries", Flex. Serv. Manuf. J., vol. 23, no. 3, pp. 263-289, September 2011

H. Min and A. Emam, "Developing the profiles of truck drivers for their successful recruitment and retention: A data mining approach", Intl. J. of Phys. Distrib. Log. Mgmt., vol. 33, no. 2, pp. 149-162, March 2003

C.Y. Fan, P.S. Fan, T.Y. Chan and S.H. Chang, "Using hybrid data mining and machine learning clustering analysis to predict the turnover rate for technology professionals", Expert Syst. with Appl., vol. 39, no. 10, pp. 8844-8851, August 2012

V.V. Saradhi and G.K. Palshikar, "Employee churn prediction", Expert Syst. with Appl., vol. 38, no. 3, pp. 1999-2006, March 2011