UNIVERSIDAD NACIONAL DE INGENIERÍA

FACULTAD DE INGENIERÍA INDUSTRIAL Y DE SISTEMAS

TESIS

DESARROLLO DE UN MODELO PREDICTIVO BASADO EN APRENDIZAJE DE MÁQUINA PARA MEJORAR EL PROCESO DE SELECCIÓN DE PERSONAL EN UNA EMPRESA DE CONSULTORÍA TECNOLÓGICA

PARA OBTENER EL TÍTULO PROFESIONAL DE:

INGENIERO DE SISTEMAS

ELABORADO POR:

RONALDO FARID NOLASCO CHAVEZ

0009-0005-4398-5991

ASESOR:

DR. HILARIO ARADIEL CASTAÑEDA

0000-0001-6921-6721

LIMA - PERÚ

2023

Nolasco Chavez, Ronaldo Farid
rnolascoc@uni.pe
931855474

DEDICATORIA

A mis padres, Aldo y Miriam, les agradezco infinitamente por todas las enseñanzas, amor y apoyo incondicional, es lo que más aprecio.

A mi hermano, Adriano, por su compañía en los momentos difíciles y por sus sabios consejos para el logro de mis objetivos.

A todos mis familiares, amigos y grupos a los que pertenezco, por haber compartido conmigo experiencias increíbles que nunca hubiera imaginado tener.

Por último, a ese samurái de cabello plateado, por nunca rendirse, siempre levantarse, y darlo todo por proteger a sus seres queridos.

AGRADECIMIENTOS

Agradezco al Dr. Glen Rodriguez Rafael, al Dr. Emilio Un Jan Liau Hing, y al Dr. Hilario Aradiel Castañeda, por su ayuda, sugerencias y conocimiento brindado en el proceso de elaboración del presente trabajo.

Agradezco al gerente Rubén Parodi Guerrero por la retroalimentación y sugerencias brindadas a lo largo del desarrollo del trabajo.

Por último, pero no menos importante, agradezco a todos aquellos que me han ayudado, en menor o mayor medida, en el desarrollo de la presente investigación.

Muchas gracias a todos, en verdad, muchas gracias.

RESUMEN

En los últimos años, se ha observado una gran complejidad en el proceso de selección de personal en empresas relacionadas con la consultoría tecnológica. Esto se debe a la gran demanda de candidatos para los puestos en este sector, la alta competencia en términos de conocimientos requeridos y también a los recursos limitados de las empresas contratantes para cumplir con las fechas de entrega de los proyectos.

Es debido a ello que en la presente tesis se presenta un modelo predictivo basado en aprendizaje de máquina, con el fin de predecir un conjunto de candidatos que sean más aptos para el puesto de trabajo, y poder plasmar estos resultados en un reporte para el área de RRHH, el cual usará para tomar la decisión final de contratación.

Se han evaluado los resultados de la solución, obteniendo métricas de exactitud y precisión que superan el 98%, lo que proporciona seguridad en los resultados obtenidos.

De igual manera, se ha evaluado la capacidad para lograr estos resultados en un tiempo mínimo, logrando tiempos de filtrado de candidatos menores a 5 segundos, y también tiempos de generación de reporte menor a 0.1 segundos, ambos tiempos óptimos para el proceso.

Por un lado, esta solución logra resolver el problema para la empresa a la que está aplicado, pero, por otro lado, el aporte al campo de conocimiento hace que sea posible replicar este desarrollo a otras empresas del mismo sector dentro del territorio peruano.

Palabras clave: modelo predictivo, aprendizaje de máquina, proceso de selección de personal, consultoría tecnológica.

ABSTRACT

In recent years, a great complexity has been observed in the personnel selection process in companies related to technology consulting. This is due to the high demand of candidates for positions in this sector, the high competition in terms of required knowledge and also to the limited resources of the hiring companies to meet project deadlines.

It is because of this that in this thesis a predictive model based on machine learning is presented, in order to predict a set of candidates that are more suitable for the job, and to be able to translate these results into a report for the HR area, which they will use to make the final hiring decision.

The results of the solution have been evaluated, obtaining accuracy and precision metrics that exceed 98%, which provides confidence in the results obtained.

Likewise, the ability to achieve these results in minimum time has been evaluated, achieving candidate filtering times of less than 5 seconds, and also report generation times of less than 0.1 seconds, both optimal times for the process.

On the one hand, this solution manages to solve the problem for the company to which it is applied, but, on the other hand, the contribution to the field of knowledge makes it possible to replicate this development to other companies in the same sector within the Peruvian territory.

Keywords: predictive model, machine learning, personnel selection process, technology consulting.

PRÓLOGO

En el presente trabajo se propone la solución a los problemas dentro del proceso de selección de personal, mediante la creación, entrenamiento y uso de un modelo predictivo basado en aprendizaje de máquina.

En el primer capítulo se verá el análisis del problema en cuestión, junto con los objetivos a cumplir, justificación a sustentar, hipótesis a validar y alcances propios de la solución.

En el segundo capítulo se realiza una revisión de la literatura, la cual comprende analizar antecedentes de investigación de diferentes autores relacionados al tópico de estudio, junto con una evaluación comparativa entre estos antecedentes mencionados. Esto seguido del análisis a profundidad de las bases teóricas de la variable dependiente e independiente.

En el tercer capítulo se revisará el tipo, nivel y diseño de la presente investigación, además de analizar la población, muestra, técnicas de recolección de datos y métodos de análisis de datos.

En el cuarto capítulo se detalla la metodología para el desarrollo de la solución, para posteriormente aplicar cada uno de los pasos en secuencia, con el fin de desarrollar la solución al problema de investigación.

En el quinto capítulo se revisarán los resultados de las mediciones de los indicadores antes y después, analizando medidas descriptivas, realizando un análisis de normalidad, completando con pruebas estadísticas para validar las hipótesis planteadas.

En el sexto capítulo se discutirán los resultados obtenidos en el capítulo anterior, además de compararlo con resultados similares obtenidos en la revisión de la literatura.

Se finaliza el trabajo con conclusiones, recomendaciones y referencias bibliográficas.

ÍNDICE

LISTA DE TABLAS

LISTA DE FIGURAS

LISTA DE SÍMBOLOS Y SIGLAS

SIGLAS

G&S	:	Gestión y Sistemas S.A.C.
RRHH	:	Recursos Humanos
BPMN	:	Business Process Model Notation
HR	:	Human Resources
TI	:	Tecnologías de información
CV	:	Curriculum Vitae
PDF	:	Portable Document Format
CSV	:	Comma Separated Values
MTPE	:	Ministerio de Trabajo y Promoción de Empleo
ETC	:	Etcétera
BCG	:	Boston Consulting Group
CRISP-DM	:	Cross Industry Standard Process for Data Mining
ATS	:	Applicant Tracking System

CAPÍTULO I
INTRODUCCIÓN

1.1 GENERALIDADES

Ford (2016) menciona que el rápido crecimiento del sector tecnológico está transformando la economía y generando una demanda creciente de habilidades técnicas y digitales. Las empresas deben adaptar sus procesos de selección de personal para encontrar candidatos con las habilidades necesarias para el trabajo en un entorno cada vez más automatizado.

Por otro lado, Bock (2015) afirma que en el sector TI, las habilidades técnicas son importantes, pero también lo son las habilidades blandas, como el pensamiento crítico, la resolución de problemas y el trabajo en equipo. Los candidatos que demuestren un equilibrio entre habilidades técnicas y habilidades interpersonales destacarán en el proceso de selección.

Este último punto de vista contrasta muy bien con el anterior, ya que ambos, empresa y candidato, en este entorno altamente competitivo, deben poder adaptarse de buena manera a las tendencias y últimas tecnologías para mantenerse vigente en el mercado laboral.

La investigación se desarrolló en la empresa Gestión y Sistemas S.A.C. (de ahora en adelante G&S), es una empresa que se encuentra en el rubro de consultoría tecnológica, ubicada en el distrito de San Juan de Miraflores, Lima, Perú. Es una empresa con más de 13 años de experiencia enfocada en el desarrollo de soluciones informáticas a la medida de sus clientes, utilizando tecnologías y tendencias que demanda el mercado actual.

En la Figura 1.1, se muestran los principales servicios que G&S ofrece a sus clientes:

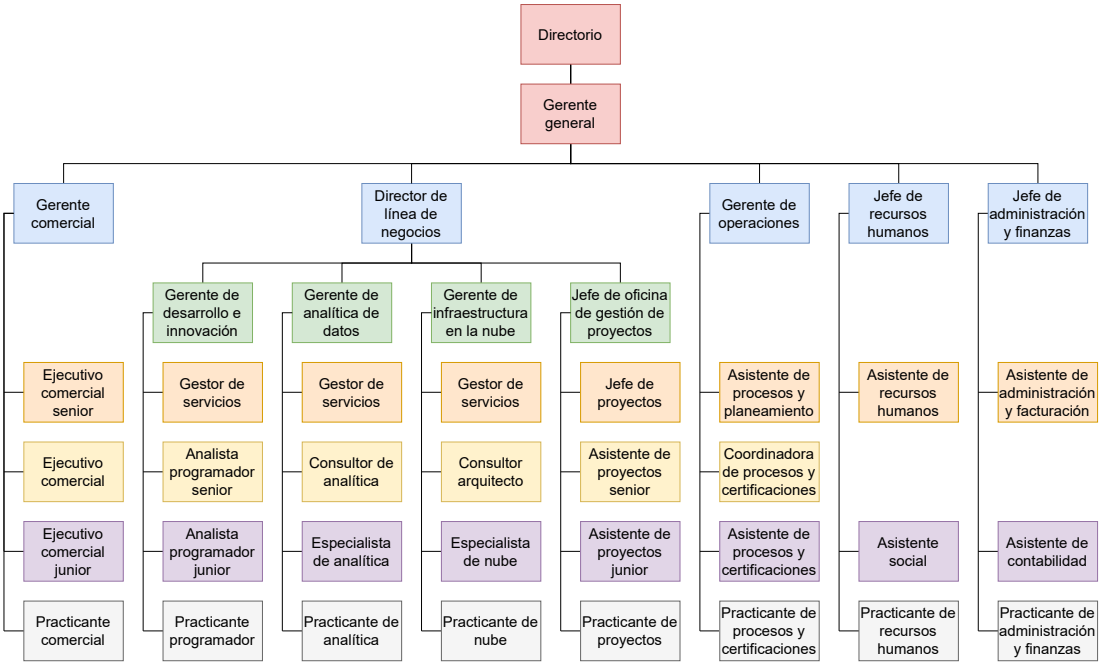
Figura 1.1: Servicios ofrecidos por la empresa

Fuente: La empresa

Existen múltiples áreas o líneas de negocio dentro de la empresa (Desarrollo e innovación, Analítica de datos, Infraestructura en la nube), pero el principal foco del estudio es el área de RRHH, la cual se encuentra en el primer nivel dentro del organigrama de la empresa.

El organigrama se presenta en la Figura 1.2:

Figura 1.2: Organigrama de la empresa

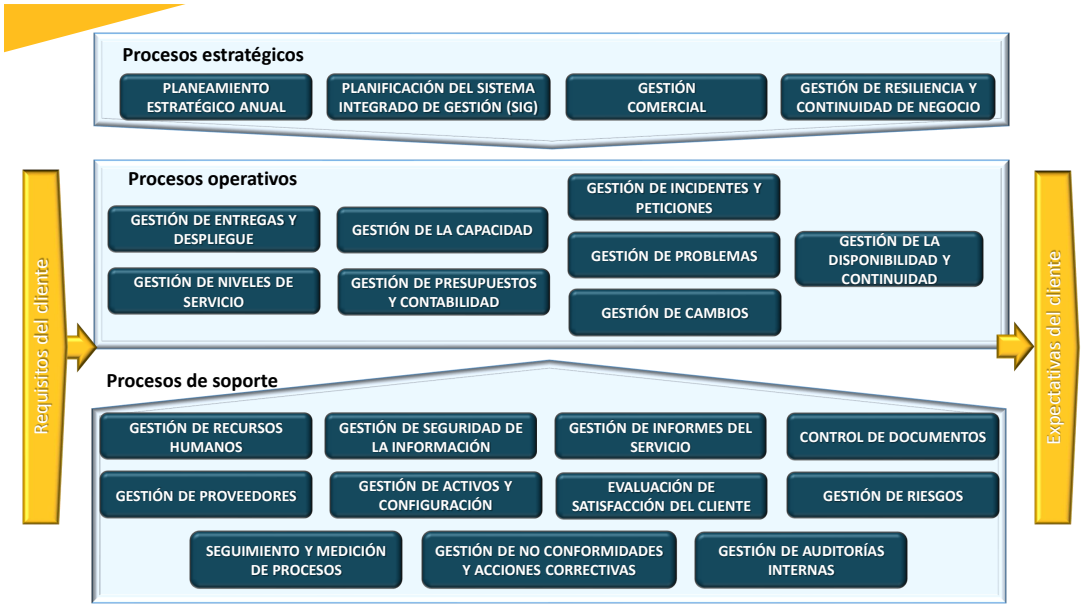


Fuente: La empresa

El macroproceso correspondiente a esta área se encuentra en el mapa de procesos general de la empresa, bajo el nombre de *Gestión de recursos humanos*:

El mapa de procesos se presenta en la Figura 1.3:

Figura 1.3: Mapa de procesos de la empresa



Fuente: La empresa

Dentro de este macroproceso, existe un listado de procesos clave del cual el área se encarga de realizar. De este listado, nos centraremos en el segundo proceso clave, llamado *Selección, reclutamiento e ingreso de personal*, el cual, para la presente investigación, se optó por determinarlo solamente como *Proceso de selección de personal*.

El listado de los procesos del macroproceso de Gestión de los RRHH se presenta en la Tabla 1.1:

Tabla 1.1: Listado de procesos del macroproceso de Gestión de los RRHH

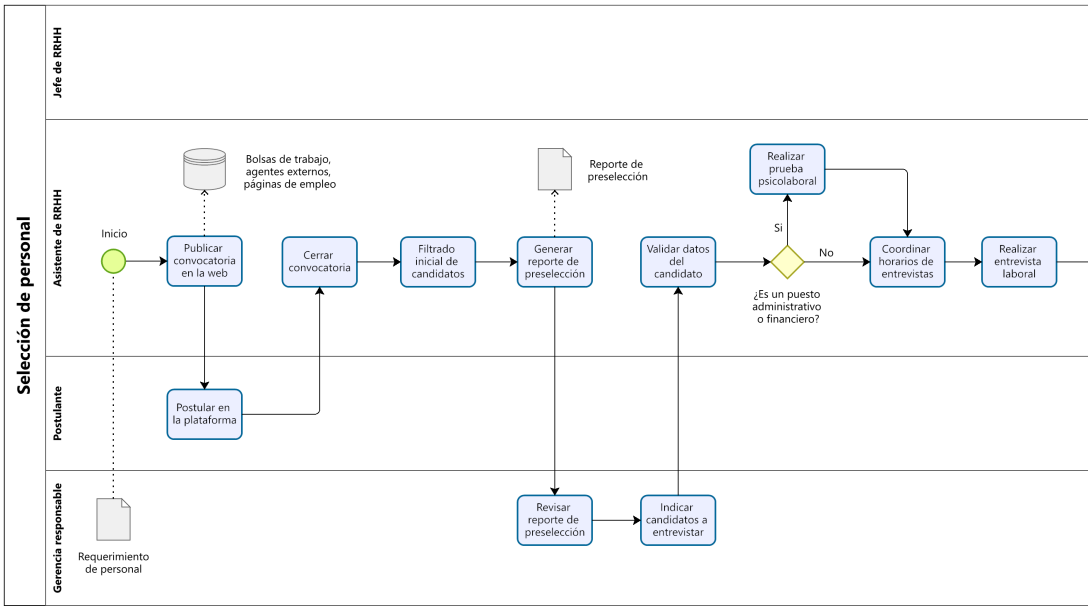
N.º	Proceso	Responsable
1	Requerimiento del personal	Gerente/Área responsable Jefe de RRHH Asistente de RRHH
2	Selección, reclutamiento e ingreso de personal	Jefe de RRHH Postulante Gerente/Área responsable Administrador de soporte Asistente de RRHH
3	Inducción al personal	Asistente de RRHH
4	Planificación de las capacitaciones	Jefe de RRHH Gerentes de línea Colaboradores
5	Ejecución de la capacitación	Asistente de RRHH Responsable de Calidad Colaboradores
6	Evaluación de eficacia de capacitación	Asistente de RRHH Colaborador Responsable de calidad
7	Cambio/promoción de puesto	Jefe/Gerente de área Colaborador Asistente de RRHH Administrador de soporte
8	Desvinculaciones	Jefe/Gerente de área Colaborador Jefe de RRHH
9	Proceso Disciplinario	Responsable del sistema Jefe de RRHH Miembro de Directorio

Fuente: La empresa

El proceso de selección de personal es encargado de integrar y capacitar la nueva fuerza laboral en la empresa.

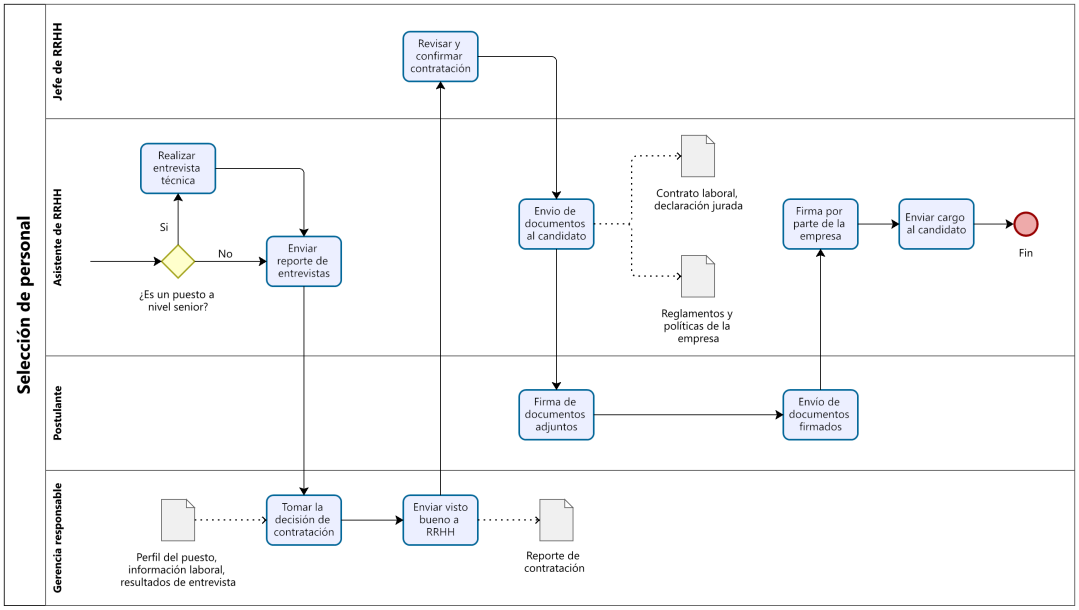
Este proceso se representa mediante los diagramas de BPMN de la Figura 1.4 y Figura 1.5:

Figura 1.4: Diagrama BPMN del proceso de selección de personal (primera parte)



Fuente: La empresa
Elaboración: Propia

Figura 1.5: Diagrama BPMN del proceso de selección de personal (segunda parte)



Fuente: La empresa
Elaboración: Propia

De manera general, las principales actividades del proceso son:

1. Inicia cuando un área de la empresa solicita el requerimiento de personal a RRHH, una vez RRHH aprueba el requerimiento, inicia la búsqueda de candidatos para el puesto.
2. Se ingresa el requerimiento a las bolsas de trabajo de universidades y portales de empleo.
3. Una vez se tengan suficientes postulaciones, RRHH preselecciona a los que cumplan los requisitos en el perfil de puesto, genera un reporte de preselección y lo envía al área solicitante para su evaluación.
4. Posterior a esto, el área indica quienes pasarán a entrevista, y envía su disponibilidad de tiempo para estas entrevistas.
5. El área de RRHH llama al candidato para corroborar sus datos y coordinar un horario para la entrevista, una vez confirmado, se agenda la entrevista. En caso sea necesario, es posible realizar una entrevista técnica adicional.
6. Luego, el entrevistador realiza un informe de la entrevista y lo envía al área solicitante. La jefatura del área solicitante analiza el informe enviado y la información proveída por RRHH, y en base a eso qué candidato es seleccionado o no para el puesto, luego esta decisión se envía a RRHH.
7. RRHH realiza una llamada informando al candidato que fue seleccionado, además de confirmar la fecha de inicio de contrato el contrato, y coordinar el envío del contrato y algunas políticas adicionales.
8. Una vez realizada la firma del contrato por ambas partes, concluye el proceso.

De estas actividades nos centraremos en aquellas que forman parte de la **preselección**, como lo son el **filtrado inicial de candidatos** y la **generación del reporte de preselección**.

1.2 REALIDAD PROBLEMÁTICA

El problema en el proceso de selección es que, desde inicios del año 2021, el número de proyectos activos de G&S con sus clientes se ha incrementado con el tiempo, esto causa que, en un momento determinado, no exista personal suficiente para poder cubrir todos los proyectos. Por lo cual es necesario que se publiquen convocatorias de trabajo continuamente solicitando diferentes perfiles, los cuales se determinan según los requerimientos del cliente en el proyecto.

En la Figura 1.6, se presenta el total de convocatorias activas en el año 2022:

Figura 1.6: Convocatorias activas en el año 2022

Fuente: La empresa
Elaboración: Propia

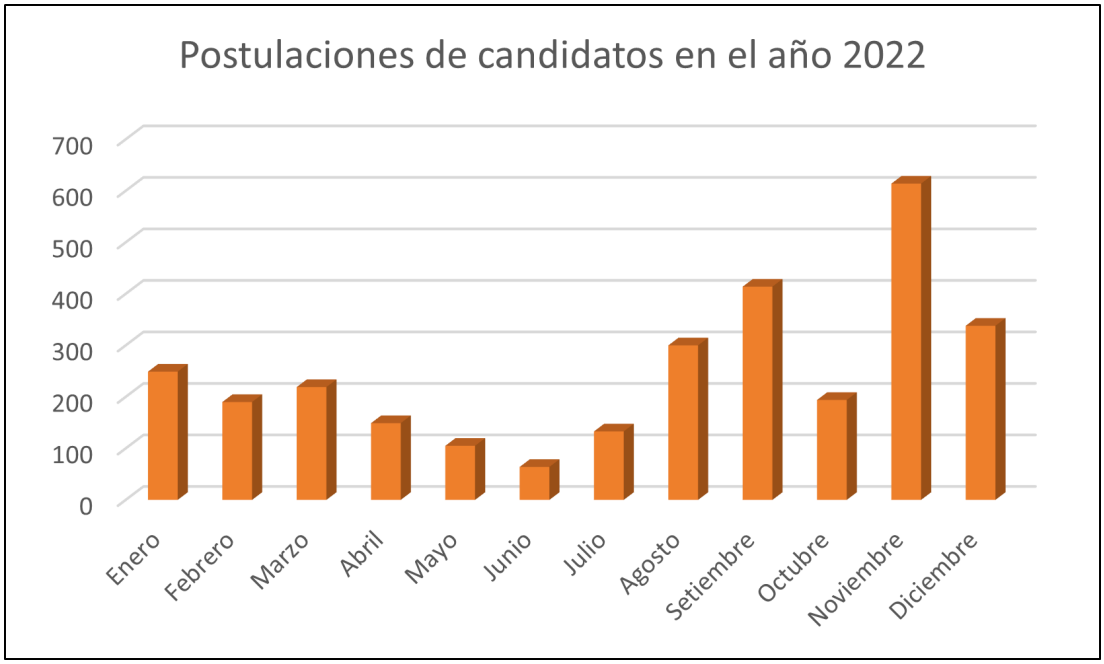
Teniendo así un total de 198 de convocatorias activas en el año 2022.

Adicional al incremento del número de proyectos, existe una limitante, la cual son los recursos de la empresa. G&S no puede incrementar el volumen de convocatorias realizadas indefinidamente, en proporción a los proyectos. En su lugar, en cada proceso de convocatoria, debe ser capaz de elegir al mejor talento disponible, el cual cumpla en mayor medida con los requerimientos del puesto publicado.

La presente complejidad es adicionada a que, también desde el año 2021, ha presentado una demanda elevada de postulantes en las convocatorias de trabajo, en los diferentes canales de selección que maneja la empresa (bolsas de trabajo, portales de empleo, bandeja de entrada de RRHH, referidos por empleados de la empresa, etc.).

En la Figura 1.7, se presenta el total de postulaciones de candidatos en el año 2022:

Figura 1.7: Postulaciones de candidatos en el año 2022



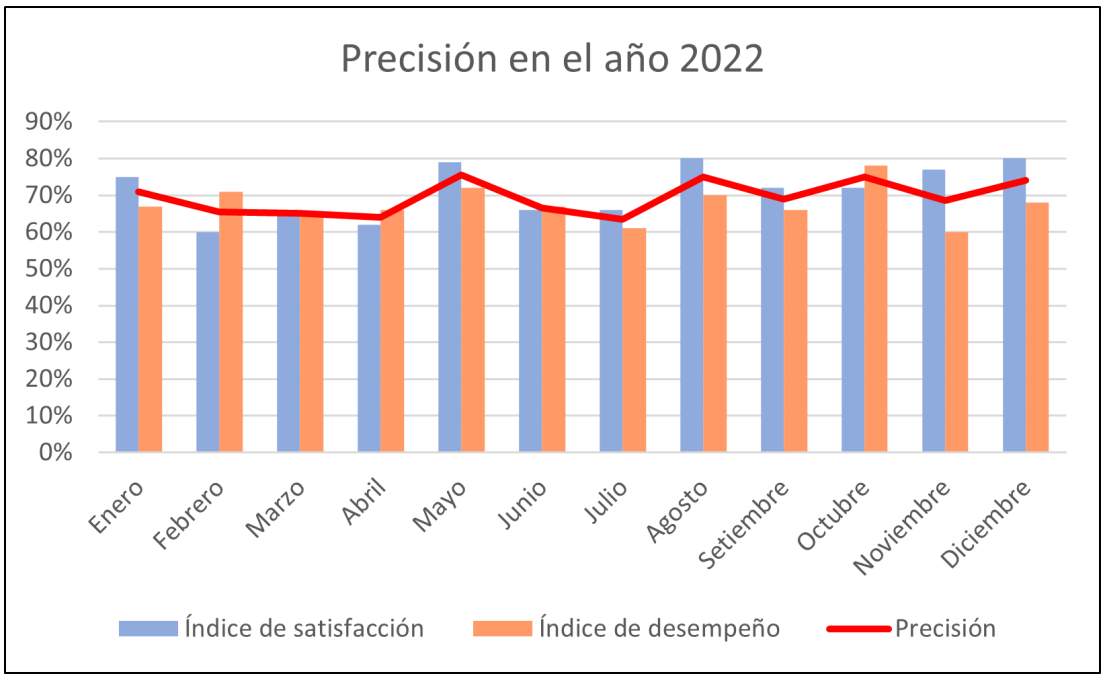
Fuente: La empresa
Elaboración: Propia

Teniendo así un total de 2969 postulaciones en el año 2022.

El área de RRHH, con fines de la investigación, definió la precisión en la selección de personal en la empresa como un promedio de la satisfacción y desempeño de los empleados en la empresa, esto ya que la satisfacción muestra que tan conforme está el empleado dentro de la empresa, y el desempeño es que tan bien está el empleado rindiendo en sus actividades.

En la Figura 1.8, se presenta la precisión de la selección de personal en el año 2022:

Figura 1.8: Precisión de la selección de personal en el año 2022



Fuente: La empresa
Elaboración: Propia

Finalmente, se obtiene en promedio:

- Índice de satisfacción promedio del 71.17%.
- Índice de desempeño promedio del 67.58%.
- Precisión promedio del 69.38%.

Es por ello que se identifica que existe un gran punto de mejora en la precisión en la selección de personal, ya que, aproximadamente, de 10 personas contratadas, 3 de ellas no deberían haber sido realmente contratadas en la empresa.

De igual manera, se ha analizado el tiempo total de selección, el cual cubre todas las fases del proceso de selección antes descrito, además del tiempo de preselección, a nivel de convocatoria y a nivel de candidato.

En la Tabla 1.2, se presentan los tiempos de selección de personal en el año 2022:

Tabla 1.2: Tiempos de selección en el año 2022

Mes	Tiempo de selección por convocatoria (días)	Tiempo de filtrado de candidatos (horas)	Tiempo de generación de reporte (minutos)
Enero	30.00	4.21	9.14
Febrero	31.47	4.26	10.10
Marzo	34.37	4.36	11.34
Abril	19.50	3.78	10.66
Mayo	26.40	4.08	14.69
Junio	28.71	4.17	12.89
Julio	23.00	3.94	9.78
Agosto	28.06	4.14	7.87
Setiembre	24.80	4.02	11.64
Octubre	25.33	4.04	12.49
Noviembre	38.05	4.47	6.99
Diciembre	30.00	4.21	8.97

Fuente: La empresa
Elaboración: Propia

Finalmente, se obtiene en promedio:

- Tiempos de selección (todo el proceso) promedio de 28.31 días.
- Tiempos de filtrado de candidatos promedio de 4.15 horas.
- Tiempos de generación de reporte promedio de 8.31 minutos.

El tiempo de filtrado de candidatos también representa un punto de mejora, debido a la demora de las etapas por cada candidato (revisar el CV, verificar si se ajusta al puesto, corroborarlo con otros candidatos y determinar si continúa en el proceso de selección), esto además del tiempo de generación del reporte (filtrar los candidatos por convocatoria, corroborar los datos ingresados, darle formato al reporte y exportarlo en un archivo PDF).

Todo esto, sumado a que el proceso es realizado de forma manual por el personal del área de RRHH, determina que no solo existan demoras por el proceso en sí, sino que también está sujeta a subjetividad, sesgos y errores propios del personal de RRHH.

1.3 FORMULACIÓN DEL PROBLEMA

1.3.1 Problema principal

- ¿Cómo influye el modelo predictivo en el proceso de selección de personal?

1.3.2 Subproblemas

- ¿Cómo influye el modelo predictivo en la exactitud del proceso de selección de personal?
- ¿Cómo influye el modelo predictivo en la precisión del proceso de selección de personal?
- ¿Cómo influye el modelo predictivo en la sensibilidad del proceso de selección de personal?
- ¿Cómo influye el modelo predictivo en la robustez del proceso de selección de personal?
- ¿Cómo influye el modelo predictivo en el tiempo de filtrado de candidatos del proceso de selección de personal?
- ¿Cómo influye el modelo predictivo en el tiempo de generación de reporte del proceso de selección de personal?

1.4 JUSTIFICACIÓN DEL ESTUDIO

1.4.1 Justificación práctica

Según datos del INEI (2021), en solo el Perú existen más de 2 millones 838 mil empresas activas registradas en el Directorio Central de Empresas y Establecimientos (DCEE), dónde cada una de estas empresas necesita un área o sector que se encargue de la gestión del capital humano, reclutamiento, selección, contratación, abordaje, inducción, capacitación, etc. Es debido a ello que es prioritario conocer a la nueva fuerza laboral ingresante a la empresa y poder mantenerla el mayor tiempo posible, de acuerdo con las metas y objetivos de la organización.

Contar con una solución que pueda pronosticar quien sería el mejor empleado a contratar, el cual esté basado en un sistema regido por reglas y políticas previamente definidas, sería de interés para el directorio de la organización, el cual podría contar con él como un sistema de soporte a las decisiones en el área de RRHH de la empresa, debido a la ausencia de subjetividad, sesgo y error que presentaría, a diferencia del factor humano.

1.4.2 Justificación académica

El uso del aprendizaje de máquina en los últimos años ha sido bastante difundido, analizado e implementado en diversas áreas del conocimiento, como transporte, medicina, genética, marketing, comercio, seguridad, etc. Por lo que es prudente pensar que es posible aplicarlo en un área como lo son los RRHH.

Gartner Inc. (2021) define el concepto de *Analítica de Recursos Humanos* como la colección y aplicación de datos de los talentos para mejorar los resultados del talento crítico y del negocio. Este es un concepto que, principalmente en los últimos 5 años, está generando soluciones efectivas para la toma de decisiones en el área de RRHH, sin embargo, no ha sido explorado a fondo aún, ni en Perú ni en el mundo.

La presente investigación tiene como fin aportar sustancialmente a esa área en conjunto con las tecnologías de inteligencia artificial y aprendizaje de máquina.

1.5 HIPÓTESIS

1.5.1 Hipótesis general

El modelo predictivo mejora el proceso de selección de personal.

1.5.2 Hipótesis específicas

- El modelo predictivo aumenta la exactitud del proceso de selección de personal.
- El modelo predictivo aumenta la precisión del proceso de selección de personal.
- El modelo predictivo aumenta la sensibilidad del proceso de selección de personal.
- El modelo predictivo aumenta la robustez del proceso de selección de personal.
- El modelo predictivo reduce el tiempo de filtrado de candidatos del proceso de selección de personal.
- El modelo predictivo reduce el tiempo de generación de reporte del proceso de selección de personal.

1.6 OBJETIVOS

1.6.1 Objetivo general

Determinar cómo influye el modelo predictivo en el proceso de selección de personal.

1.6.2 Objetivos específicos

- Determinar cómo influye el modelo predictivo en la exactitud del proceso de selección de personal.
- Determinar cómo influye el modelo predictivo en la precisión del proceso de selección de personal.
- Determinar cómo influye el modelo predictivo en la sensibilidad del proceso de selección de personal.
- Determinar cómo influye el modelo predictivo en la robustez del proceso de selección de personal.
- Determinar cómo influye el modelo predictivo en el tiempo de filtrado de candidatos del proceso de selección de personal.
- Determinar cómo influye el modelo predictivo en el tiempo de generación de reporte del proceso de selección de personal.

1.7 LIMITANTES DE LA INVESTIGACIÓN

1.7.1 Limitantes teóricos

- La solución pronosticará una variable categórica dicotómica (0 o 1), con lo cual solo se determinará si se contrata o no al candidato.
- La solución solo considerará dos tipos de orígenes de datos principales:
  - Datos del puesto de trabajo (ID, nombre, tipo, nivel).
  - Datos del candidato (datos personales, experiencia laboral previa, educación previa, habilidades técnicas, habilidades blandas, idiomas).
- Coronel (2021) señala que el proceso de selección centrado en el factor humano es influenciado por el sesgo cognitivo. Es debido a ello que la solución actuará como un soporte a las decisiones del área de RRHH, a fin de evitar subjetividad propia del reclutador en la etapa de selección. De igual forma, la solución predecirá el resultado y dará sugerencias, pero la decisión final la tomará el reclutador.
- Se priorizará maximizar las métricas porcentuales (exactitud, precisión, sensibilidad y robustez) precisión y también reducir las métricas de tiempo (filtrado de candidatos, generación de reporte), despreciando algunas otras métricas no tan relacionadas a éstas.

1.7.2 Limitantes temporales

- Todos los datos que se usarán serán datos de la etapa de preselección, es decir, se usarán los datos que se tengan hasta el momento del cierre de la convocatoria, sin dejar espacio a otros datos posteriores a esta etapa (pruebas psicométricas, pruebas técnicas, entrevistas), esto debido a que solamente se cuentan con los datos de esta etapa de preselección.
- Se usarán datos históricos para el entrenamiento y prueba de la solución. Se tienen datos de los postulantes desde el 21 de junio del 2019 hasta el 22 de mayo del 2023 (4 años). Esto equivalen a aproximadamente 303 convocatorias y a 10563 postulaciones.

1.7.3 Limitantes espaciales

- Pese a que la gran mayoría de candidatos están dentro del territorio peruano, existe un muy reducido porcentaje de candidatos residentes en países extranjeros (menor al 1 % del total). Dependería del tipo de vacante (presencial o remota) para determinar si se puede contratar o no al candidato.

CAPÍTULO II
FUNDAMENTO TEÓRICO

2.1 ANTECEDENTES DE INVESTIGACIÓN

2.1.1 Revisión de métodos

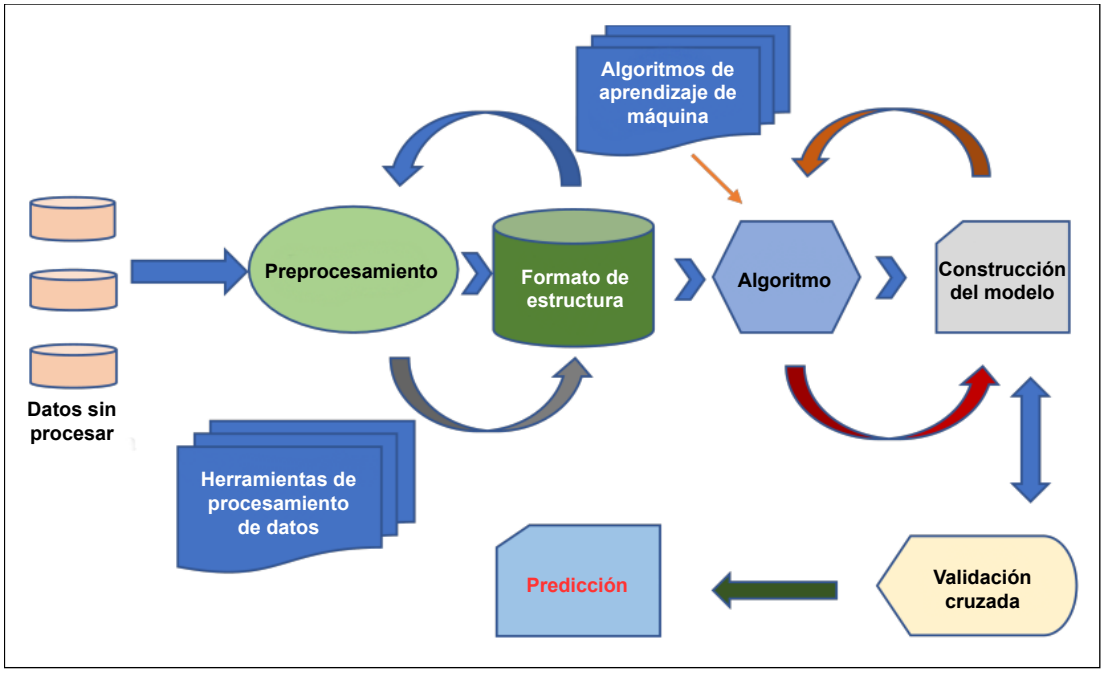
2.1.1.1 Antecedente 1

En el artículo de Jagan Mohan Reddy et al. (2020), se comenta que el reclutamiento es el aspecto principal de todas las organizaciones comerciales, esto debido a que depende principalmente del desempeño de los empleados.

De igual forma, remarca que no solo la decisión queda en la empresa, sino que paralelamente el candidato puede estar buscando otras oportunidades, tomando en cuenta el salario, compensaciones, puestos de trabajo e incentivos, es por ello que es muy importante poder formar esta relación contractual con el mejor candidato lo más pronto posible.

Los autores proponen un nuevo diseño de sistema que podría aportar a la solución, el cual se encuentra en la Figura 2.1:

Figura 2.1: Diseño del sistema del artículo 1



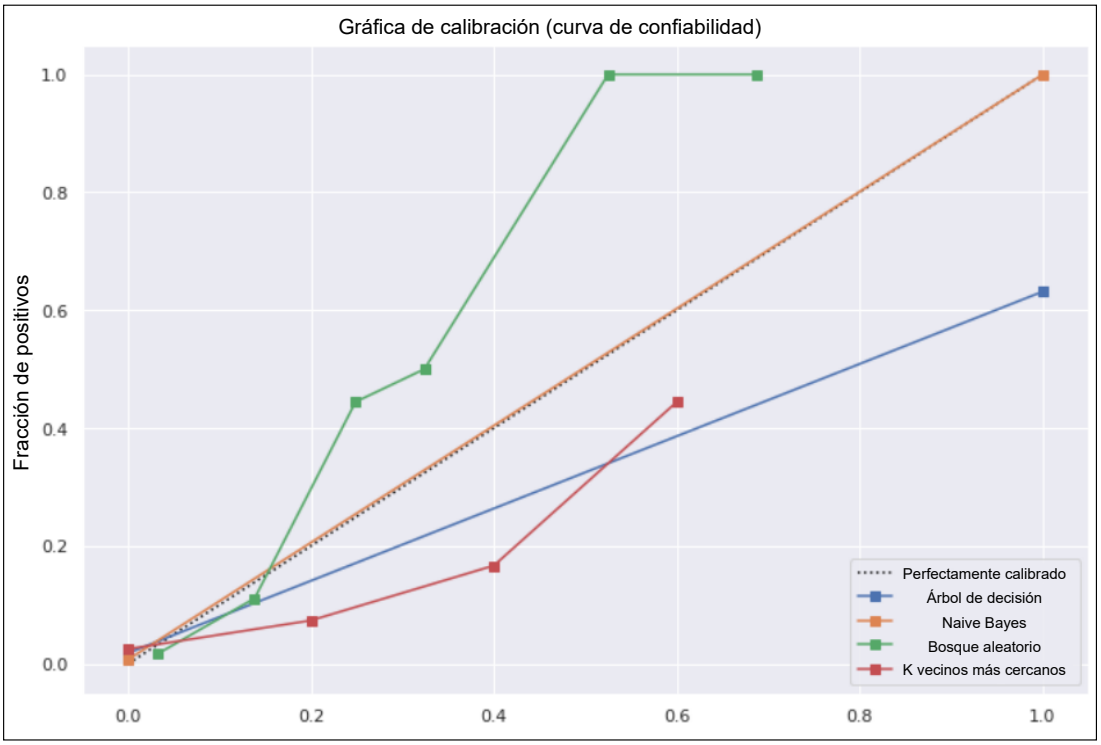
Fuente: Jagan Mohan Reddy et al. (2020)

Los autores usaron 4 algoritmos para el entrenamiento del modelo:

- Árbol de decisión.
- Bosque aleatorio.
- Naive Bayes Gaussiano.
- K vecinos más cercanos.

En las métricas, se tiene una gráfica de calibración del modelo, en la cual se compara la fracción de positivos con respecto al valor promedio pronosticado, en esta se evidencia que el algoritmo Naive Bayes Gaussiano (GNB) es el que está más cerca de ser perfectamente calibrado. Esta gráfica se encuentra en la Figura 2.2:

Figura 2.2: Gráfica de calibración del artículo 1



Fuente: Jagan Mohan Reddy et al. (2020)

Finalmente, se tiene la evaluación comparativa final de los modelos, en los cuales el que tiene mayor exactitud, precisión y recuperación es el GNB. Según el autor, esto es producido porque no hay mucha interdependencia entre características, ya que este algoritmo, por naturaleza, asume que todas las características son independientes. Esta comparación se encuentra en la Tabla 2.1:

Tabla 2.1: Resultados del artículo 1

Modelo	Exactitud	Precisión	Sensibilidad
Árbol de decisión	0.96	0.75	0.7
Bosque aleatorio	0.95	1	0.17
Naive Bayes Gaussiano	0.99	1	0.88
K vecinos más cercanos	0.93	0.44	0.23

Fuente: Jagan Mohan Reddy et al. (2020)

Se tienen las siguientes observaciones sobre el artículo:

- Es bastante completo en el sentido de plantear y fundamentar la necesidad y la solución del problema.
- Hay ciertas variables del modelo que no se consideraron y que podrían ser claves al momento de la selección, como el máximo grado académico conseguido, número de proyectos en los que participó, número de investigaciones, su rol en la última empresa en la que trabajó y número de certificaciones.
- Existe falta de visualización de datos para poder entender completamente las variables del modelo, como lo pueden ser mapas de calor, diagramas de caja o matrices de correlación.

2.1.1.2 Antecedente 2

En el artículo de Pessach et al. (2020), se analiza la dificultad del proceso de selección, y se apoyan en un estudio del BCG señalando que el proceso de contratación es el que tiene mayor impacto en el crecimiento de ingresos y utilidades, a comparación de otros procesos de RRHH.

Ellos plantean el uso del modelo Red Bayesiana de Orden Variable (VOBN), los cuales son una extensión de las redes bayesianas, se utilizan en el aprendizaje de máquina y amplían los modelos de matriz de pesos de posición, los modelos de Markov y los modelos de redes bayesianas.

Ellos abarcan el problema desde 2 perspectivas, en las cuales para cada una tiene una solución:

- Local: Se calcula el nivel de adecuación o afinidad de un empleado a un puesto específico para el que es contratado (Se utiliza el modelo VOBN). Esta se describe en la Figura 2.3:

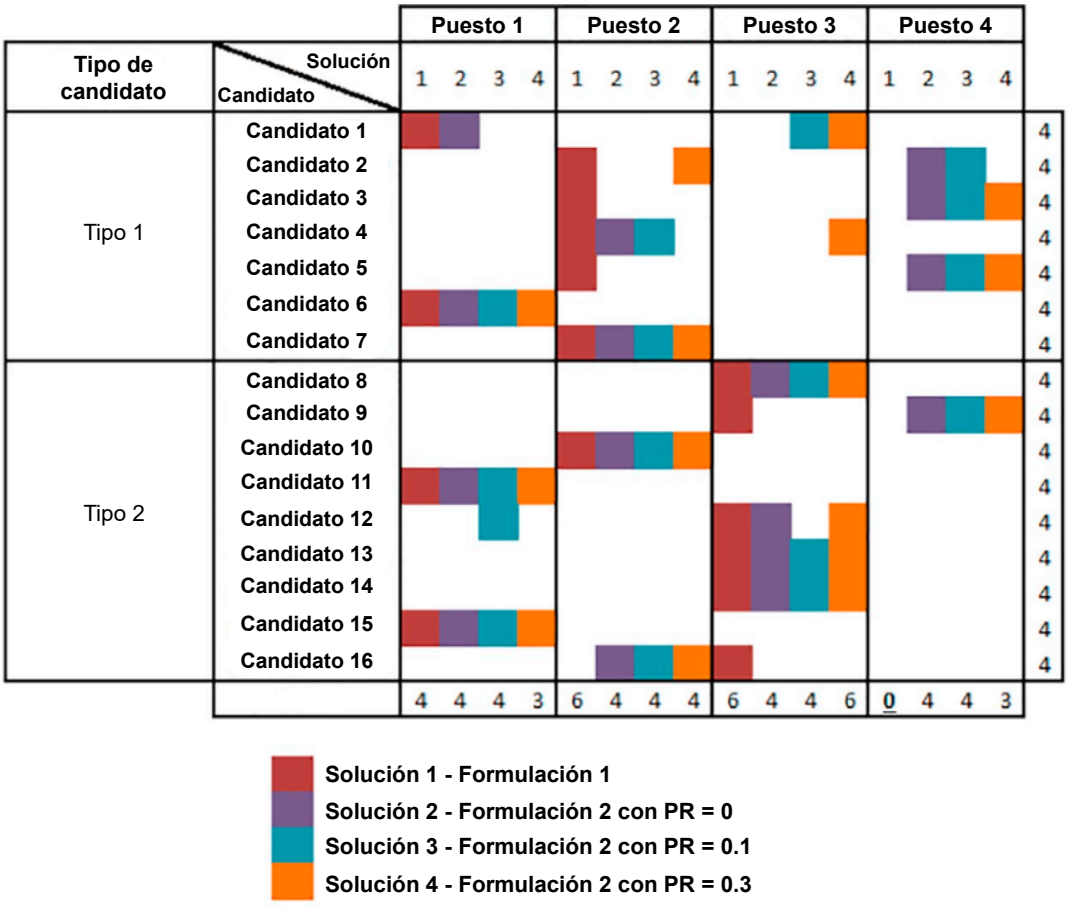
Figura 2.3: Resultados del enfoque local del artículo 2

Tipo de candidato	ID de candidato	Puesto de trabajo 1	Puesto de trabajo 2	Puesto de trabajo 3	Puesto de trabajo 4
Tipo 1	Candidato 1	0.67	0.68	0.68	0.48
	Candidato 2	0.47	0.68	0.48	0.56
	Candidato 3	0.53	0.68	0.39	0.56
	Candidato 4	0.61	0.68	0.51	0.55
	Candidato 5	0.63	0.68	0.48	0.57
	Candidato 6	0.67	0.68	0.51	0.51
	Candidato 7	0.55	0.68	0.45	0.55
Tipo 2	Candidato 8	0.66	0.68	0.85	0.58
	Candidato 9	0.48	0.68	0.85	0.60
	Candidato 10	0.56	0.73	0.85	0.60
	Candidato 11	0.69	0.68	0.85	0.55
	Candidato 12	0.69	0.58	0.86	0.59
	Candidato 13	0.64	0.68	0.86	0.59
	Candidato 14	0.59	0.62	0.86	0.48
	Candidato 15	0.84	0.68	0.85	0.54
	Candidato 16	0.62	0.69	0.83	0.54

Fuente: Pessach et al. (2020)

- Global: Se utilizan los datos procesados de manera local, junto con la consideración de cumplimiento de objetivos organizacionales propuestos por la alta dirección. (Se utiliza el modelo VOBN y un modelo de programación lineal). Esta se describe en la Figura 2.4:

Figura 2.4: Resultados del enfoque global del artículo 2



Fuente: Pessach et al. (2020)

Finalmente, se compara el modelo VOBN con otros algoritmos:

- Máquina de aumento de gradiente.
- Bosque aleatorio.
- Regresión logística.
- Máquina de vectores de soporte.
- C45.
- CHAID.
- Naive Bayes.
- CART.

Finalmente, para fines de interpretabilidad, el modelo VOBN es mejor que todos los demás.

Se tienen las siguientes observaciones sobre el artículo:

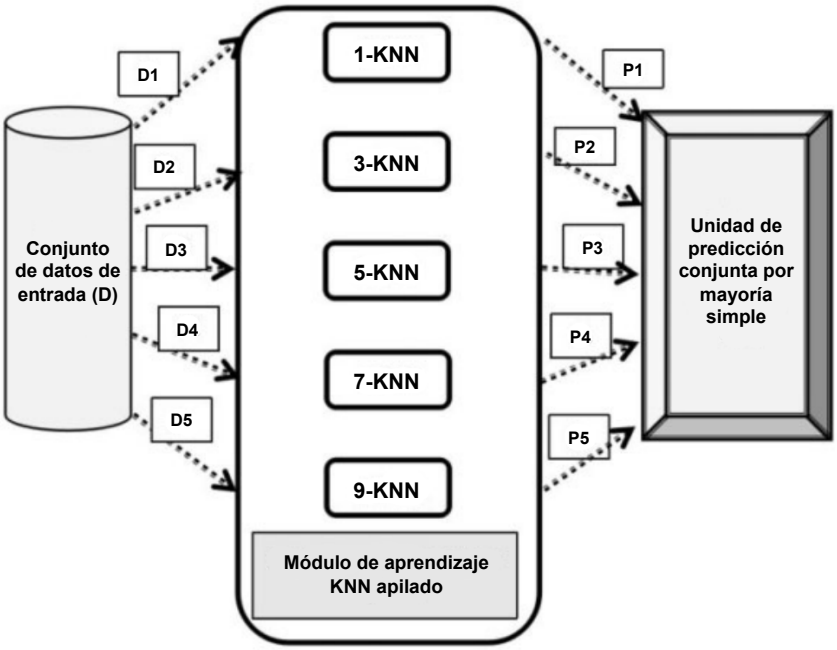
- Los autores amplían mucho el horizonte de esta rama de la analítica mediante el uso de capas o niveles para la selección de personal, la cual es adaptable según las necesidades de la organización.
- Descubrir patrones entre los atributos que podrían generarse a modo de reporte en la ejecución del modelo podrían ser muy importantes y claves para cada tipo de puesto a seleccionar (administrativo, técnico, operativo, de apoyo), ya que el reclutador podría encontrar y entender cierta correlación, la cual podría generar confianza en el sistema.
- Es un artículo bastante completo, pero es necesario leerlo varias veces para poder comprender todo lo que los autores plantean.

2.1.1.3 Antecedente 3

En el artículo de Mishra et al. (2021), se analiza el crecimiento del sector tecnológico y la gran demanda que tiene hoy en día. Debido a la gran diversidad de puestos y capacidades, concluyen que es difícil elegir a los mejores candidatos a puestos.

Es debido a ello que los autores plantean el uso de 5 variantes del modelo KNN, cada uno con diferente valor K, de tal forma que la etiqueta de salida final es determinada por mayoría simple entre estas 5 variantes. Los autores lo denominan modelo KNN apilado, este modelo se encuentra en la Figura 2.5:

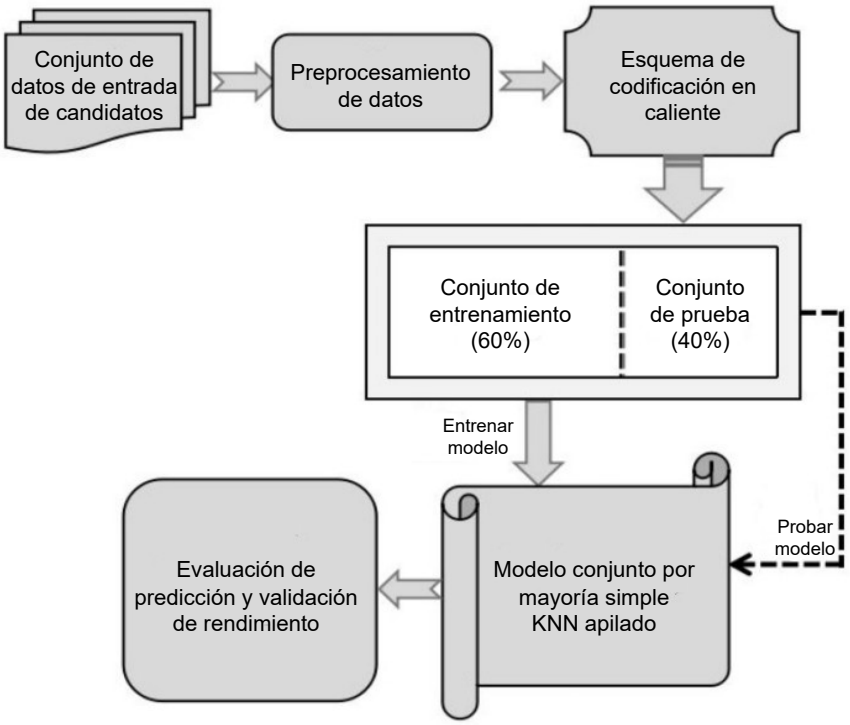
Figura 2.5: Modelo KNN apilado del artículo 3



Fuente: Mishra et al. (2021)

El diseño completo del sistema se encuentra en la Figura 2.6:

Figura 2.6: Diseño de sistema del artículo 3



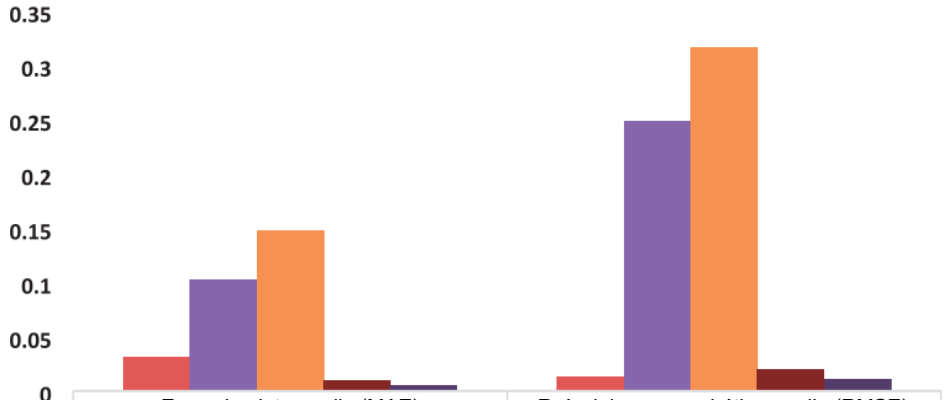
Fuente: Mishra et al. (2021)

Ellos comparan el modelo KNN apilado con otros algoritmos, como son:

- Redes neuronales.
- Naive Bayes.
- Árbol de decisión.
- K vecinos más cercanos.

Finalmente, se evidencia que el modelo KNN apilado tiene una menor tasa de error que los otros 4. Esta comparación de tasa de error se encuentra en la Figura 2.7:

Figura 2.7: Comparación de tasa de error entre modelos del artículo 3



	Error absoluto medio (MAE)	Raíz del error cuadrático medio (RMSE)
Red neuronal	**0.0315**	**0.0129**
Naive Bayes	**0.1026**	**0.2487**
Árbol de decisión	**0.1476**	**0.3159**
KNN	**0.0097**	**0.0197**
KNN apilado	**0.0048**	**0.0102**

Fuente: Mishra et al. (2021)

De igual manera, ellos calculan algunas métricas para este algoritmo KNN apilado, como la exactitud, especificidad, sensibilidad y robustez. Esta comparación de métricas se encuentra en la Tabla 2.2:

Tabla 2.2: Resultados del artículo 3

Tamaño de la muestra	Exactitud (%)	Especificidad (%)	Sensibilidad (%)	Robustez (%)
100	96.2	94.2	94.8	94.5
200	95.4	95.6	95.2	95.4
300	94.8	93.6	93.9	93.7
400	95.8	93.8	95.1	94.6
500	93.2	95.2	92.9	94.3
600	86.2	85.6	86.6	86.2
700	92.8	94.4	94.6	94.5
800	94.4	95.2	94.8	95.1
900	93.8	94.6	95.1	94.8
1000	88.5	87.8	85.6	86.3
1100	94.7	96.2	93.8	94.2
1200	96.1	95.2	93.2	94.5
1300	**96.96**	**96.38**	**96.02**	**96.26**

Fuente: Mishra et al. (2021)

Se tienen las siguientes observaciones sobre el artículo:

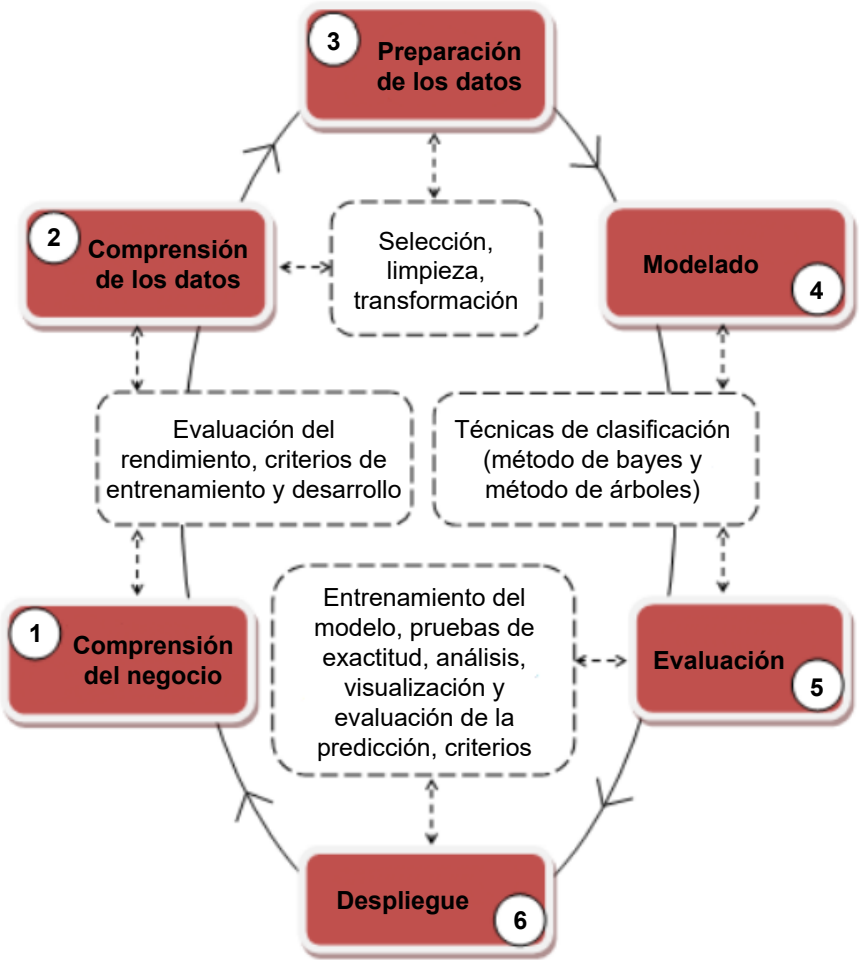
- Los autores explican de manera concisa el problema descrito, un meta modelo basado en un algoritmo existente y demostrar con métricas que es muy conveniente usarlo en organizaciones relacionadas a las TI.
- Fue preciso haber definido brevemente el funcionamiento del algoritmo, a fin de que alguien relacionado y no tan relacionado al rubro de TI pueda entender el artículo con facilidad.
- El artículo debería haber presentado más visualización de datos para poder entender la relación entre los atributos, como algún mapa de calor, diagrama de cajas, diagrama de violín, etc.

2.1.1.4 Antecedente 4

En el artículo de Jantawan y Tsai (2013), se revisa que debido al incremento de la tasa de desempleo a nivel mundial y el incremento de número de estudiantes graduados de institutos o universidades, los autores sostienen que la empleabilidad de estos egresados se encuentra en peligro.

Ellos plantean el uso de la metodología CRISP-DM para elaborar su propuesta de solución, esto debido a que es un conjunto de métodos ya probados y validados previamente. El esquema de esta metodología se encuentra en la Figura 2.8:

Figura 2.8: Metodología a usar del artículo 4



Fuente: Jantawan y Tsai (2013)

Se realizan 10 modelos en 2 grupos de tipos de algoritmos, como lo son:

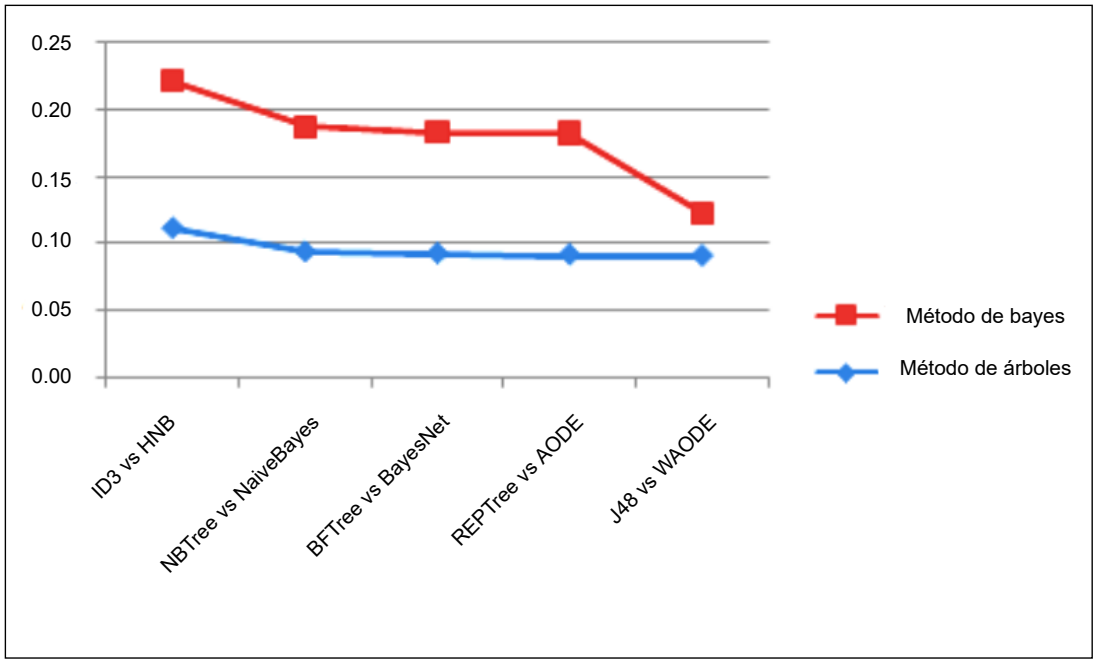
- Modelos de árboles:
  - ID3
  - J48
  - BFTree
  - NBTree

    - REPTree
- Modelos de Bayes:
    - AODE
    - BayesNet
    - HNB
    - Naive Bayes
    - WAODE

Cada uno de éstos fue implementado en el Entorno de Waikato para el análisis del conocimiento (WEKA), la cual es una plataforma de software que contiene implementaciones de algoritmos minería de datos y aprendizaje de máquina, escrita en Java y desarrollada en una universidad de Nueva Zelanda.

Finalmente se concluye que el error es menor en los modelos de árboles a comparación de los bayesianos, además que el modelo de árbol con mayor exactitud es el J48. Esta comparación se encuentra en la Figura 2.9:

Figura 2.9: Comparación de raíz del error cuadrático medio del artículo 4



Fuente: Jantawan y Tsai (2013)

Se tienen las siguientes observaciones sobre el artículo:

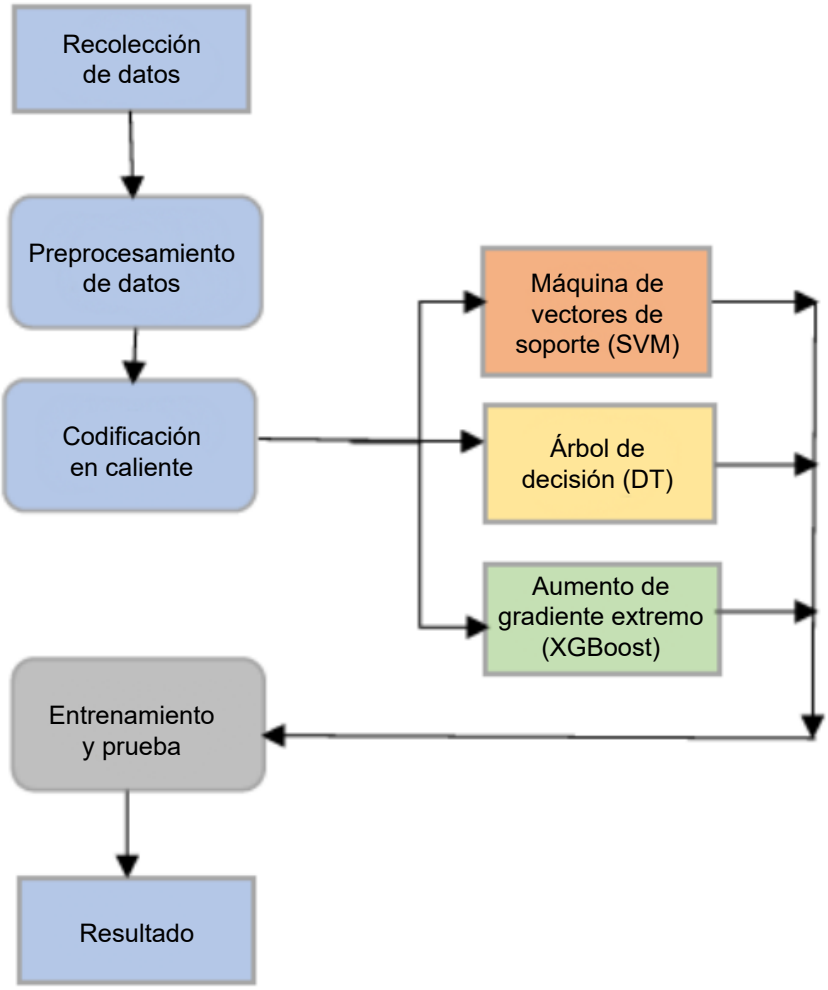
- Al tener una metodología soportada y probada previamente, es más sencillo poder desarrollar una nueva propuesta de solución a un problema en específico.
- Aporta otro punto de vista al problema de la contratación, tomándolo más por el lado de la empleabilidad de los egresados en lugar de la selección de personal.
- El artículo debería haber presentado más visualización de datos para poder entender la relación entre los atributos, como algún mapa de calor, diagrama de cajas, diagrama de violín, etc.
- No señala exactamente la proporción de datos usados para el entrenamiento y prueba.
- Al presentar varios algoritmos diferentes, ayuda a considerar más opciones para el desarrollo de mi tesis.

2.1.1.5 Antecedente 5

En el artículo de Roy et al. (2018), se sostiene que, por un lado, los estudiantes deben prepararse desde etapas tempranas de su formación para adaptarse a la vida laboral y evaluar constantemente su desempeño, y, por otro lado, los reclutadores primero evalúan a los candidatos en diferentes parámetros (habilidades, talentos e intereses) y luego eligen en que puesto de trabajo deben mantener al candidato según su desempeño, y en base al aporte que puede realizar a la organización.

Ellos proponen un modelo de aprendizaje de máquina para predecir la rama o carrera a la cual el estudiante mejor cumple con los criterios de selección, esta carrera se encuentra directamente relacionada con ciencias de la computación (administrador de base de datos, desarrollador, gerente de pruebas, gerente de redes, científico de datos, etc.), para ello consideran de 3 diferentes algoritmos de aprendizaje de máquina. El diagrama de flujo propuesto se encuentra en la Figura 2.10:

Figura 2.10: Diagrama de flujo del proceso del artículo 5



Fuente: Roy et al. (2018)

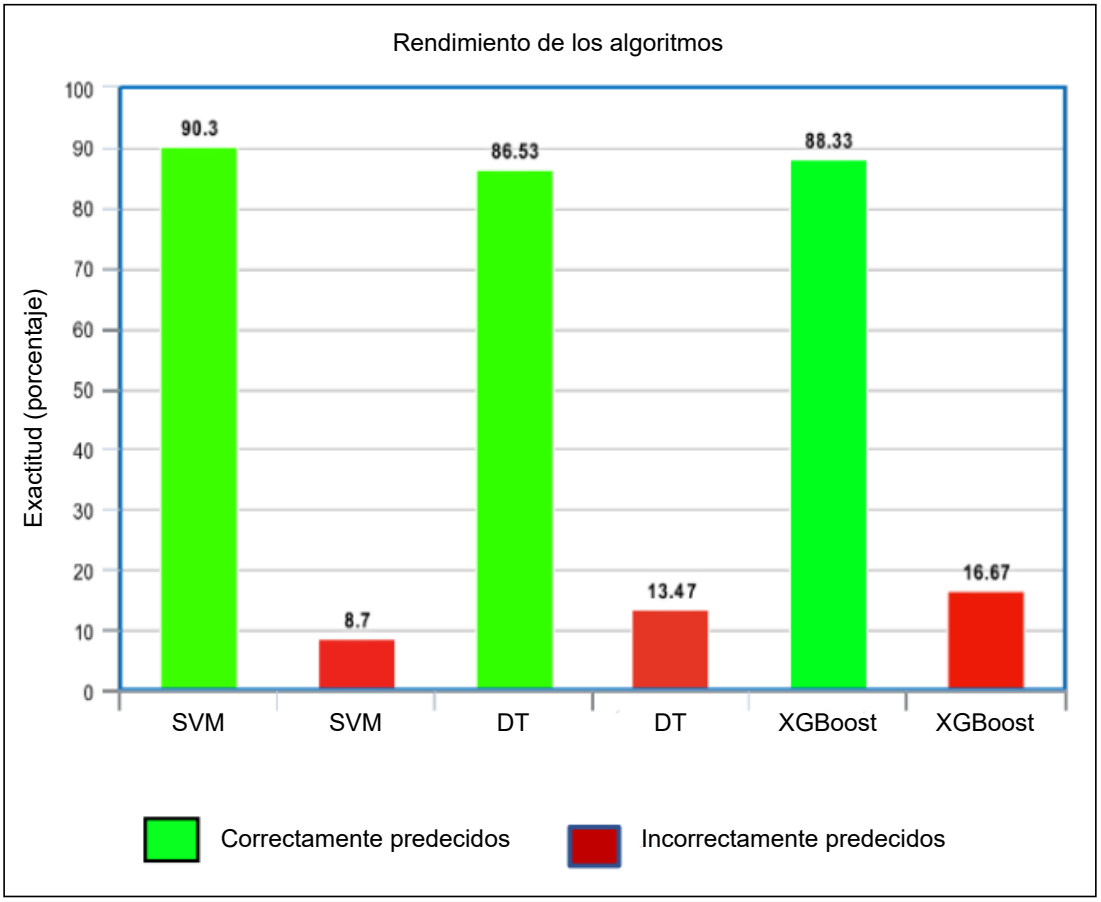
Los 3 algoritmos que se definen en la investigación son:

- Máquina de vectores de soporte (SVM)
- Árbol de decisión (DT)
- Aumento de gradiente extremo (XGBoost)

De igual manera, recalcan que aquel de los 3 algoritmos que alcance una mayor precisión, se usará en una fase posterior de la investigación.

Finalmente, una vez concluido los pasos del diagrama, se presentan los resultados en la Figura 2.11:

Figura 2.11: Exactitud de los modelos del artículo 5



Fuente: Roy et al. (2018)

Se evidencia que el algoritmo SVM es el que tiene mayor exactitud, contando con un valor de 90.3 %, seguido por el XGBoost (88.33 %) y el DT (86.53 %). Es por ello que las predicciones posteriores se realizarían usando el modelo entrenado en el algoritmo SVM.

Se tienen las siguientes observaciones sobre el artículo:

- Al final de los resultados se menciona el desarrollo de una aplicación web para el ingreso de parámetros de entrada y el resultado final, sin embargo, no se presenta ninguna figura o tabla relacionada a ello.
- Solamente presenta la exactitud del modelo, omite otras métricas como precisión, sensibilidad, significancia y Valor-F.
- No describe en términos sencillos todos los algoritmos a implementar en la solución ni las fórmulas usadas en éstos.
- No describe exactamente los atributos a usar en el entrenamiento del modelo.

- En la gráfica inicial se presenta la aplicación de redes neuronales, sin embargo, no se detalla ningún uso.

2.1.2 Evaluación comparativa

Se realizará una comparación de los métodos revisados en el estado del arte. Se definirán criterios, luego niveles para los criterios, se analizará la consistencia de los criterios, se definirán sus pesos y luego se evaluarán los diferentes artículos en base a los criterios, según el Proceso Jerárquico Analítico (AHP).

2.1.2.1 Criterios

- **Exactitud del modelo (C1):** Exactitud del modelo seleccionado por el autor.
- **Tiempo de ejecución (C2):** Tiempo total de predicción de un nuevo dato ingresado en el modelo.
- **Fuente primaria de datos (C3):** Origen principal de los datos usados.
- **Rubro primario de datos (C4):** Sector de empleo principal de los datos.
- **Semejanza cultural de datos (C5):** Semejanza de los datos usados del autor con respecto a la realidad del Perú.

2.1.2.2 Niveles

Se definen los niveles por cada criterio definido, en la Tabla 2.3, Tabla 2.4, Tabla 2.5, Tabla 2.6 y Tabla 2.7:

- **Exactitud del modelo (C1):**

Tabla 2.3: Exactitud del modelo

Exactitud del modelo	Nivel
[0%, 70%>	1
[70%, 80%>	2
[80%, 90%>	3
[90%, 95%>	4
[95%, 100%>	5

Fuente: La empresa
Elaboración: Propia

- **Tiempo de ejecución (C2):**

Tabla 2.4: Tiempo de ejecución

Tiempo de ejecución	Nivel
Desconocido	1
<60s, +∞>	2
<10s, 60s]	3
<3s, 10s]	4
<0s, 3s]	5

Fuente: La empresa
Elaboración: Propia

- **Fuente primaria de datos (C3):**

Tabla 2.5: Fuente primaria de datos

Fuente primaria de datos	Nivel
Desconocido	1
Datos de prueba	2
Información en la red	3
Universidades	4
Empresas	5

Fuente: La empresa
Elaboración: Propia

- **Rubro primario de datos (C4):**

Tabla 2.6: Rubro primario de datos

Rubro primario de datos	Nivel
Cualquiera/desconocido	1
Cargos administrativos u operativos	3
Relacionados a TI	5

Fuente: La empresa
Elaboración: Propia

- **Semejanza cultural de datos (C5):**

Tabla 2.7: Semejanza cultural de datos

Semejanza cultural de datos	Nivel
Desconocido	1
Países primer mundo	2
Países tercer mundo no Latinoamérica	3
Países tercer mundo Centroamérica	4
Países tercer mundo Sudamérica	5

Fuente: La empresa
Elaboración: Propia

2.1.2.3 Consistencia de los criterios

Se toma como base la escala de Saaty para poder determinar la importancia relativa entre criterios, la cual se encuentra en la Tabla 2.8:

Tabla 2.8: Escala de Saaty

Intensidad	Definición
1	Igual importancia
3	Moderada importancia
5	Importancia fuerte
7	Importancia muy fuerte
9	Importancia extrema
2,4,6,8	Valores intermedios

Fuente: Saaty (1980)

Se analizan las comparaciones pareadas de los criterios, la cual se encuentra en la Tabla 2.9:

Tabla 2.9: Matriz de comparaciones pareadas

	C1	C2	C3	C4	C5	Suma	Peso
C1	1.000	2.000	2.000	3.000	1.000	9.000	0.388
C2	0.500	1.000	3.000	1.000	2.000	7.500	0.324
C3	0.500	0.333	1.000	0.500	1.000	3.333	0.144
C4	0.333	1.000	2.000	1.000	2.000	6.333	0.273
C5	1.000	0.500	1.000	0.500	1.000	4.000	0.173
Suma	3.333	4.833	9.000	6.000	7.000	23.167	1.000

Fuente: La empresa
Elaboración: Propia

De esta tabla, obtenemos que el criterio 1 (Exactitud) es el más importante, seguido de tiempo de ejecución, rubro primario de datos, semejanza cultural de datos y fuente primaria de datos.

Esto es debido a que la parte interesada valora más una solución que siempre devuelva la mejor respuesta posible con un tiempo adecuado, a diferencia de una solución muy rápida, pero que no tiene una precisión de la selección del candidato elevada.

Para validar la consistencia de las comparaciones entre criterios, se calculan los valores propios de la matriz de comparación, y se obtiene que el mayor valor propio que tiene es: $\lambda_{max} = 5.357$

Para el cálculo del índice de consistencia, se utiliza la Ecuación 2.1:

$$CI = \frac{\lambda_{max} - n}{n - 1} \tag{2.1}$$

Dónde:

- CI: Índice de consistencia de la matriz
- λ_{max}: Máximo valor propio de la matriz
- n: Orden de la matriz

De esta ecuación, se obtiene un índice de consistencia $CI = 0.089$

De igual manera, para el cálculo de la relación de consistencia, se utiliza la Ecuación 2.2:

$$RC = \frac{CI}{RI} \tag{2.2}$$

Dónde:

- RC: Relación de consistencia de la matriz
- CI: Índice de consistencia de la matriz
- RI: Índice aleatorio (para n = 5, su valor es 1.12)

Finalmente, se obtiene como relación de consistencia $RC = 0.080 = 8.0\%$

Al ser la relación de consistencia $RC < 10\%$, se determina que **la matriz es consistente**.

2.1.2.4 Pesos de los criterios

De la matriz de comparaciones pareadas, se obtiene el peso de los criterios, los cuales se encuentran en la Tabla 2.10:

Tabla 2.10: Pesos de los criterios

Sigla	Criterio	Peso
C1	Exactitud del modelo	0.388
C2	Tiempo de ejecución	0.324
C3	Fuente primaria de datos	0.144
C4	Rubro primario de datos	0.273
C5	Semejanza cultural de datos	0.173

Fuente: La empresa
Elaboración: Propia

2.1.2.5 Evaluación

Se comparan los 5 artículos revisados en el estado del arte.

1. Artículo 1 (A1).
2. Artículo 2 (A2).
3. Artículo 3 (A3).
4. Artículo 4 (A4).
5. Artículo 5 (A5).

La comparación se realiza asignando un puntaje del 1 al 5 a cada uno de los 5 artículos, bajo los 5 criterios mencionados, de tal forma que se calcula un ponderado total por cada artículo y, el que tenga mayor puntaje, será el que sea elegido como referencia principal para la investigación.

Esta comparación se describe en la Tabla 2.11:

Tabla 2.11: Comparación de artículos

	Pesos	**A1**	**A2**	**A3**	**A4**	**A5**
C1	0.388	5.000	2.000	5.000	5.000	4.000
C2	0.324	1.000	1.000	1.000	1.000	1.000
C3	0.144	2.000	5.000	4.000	4.000	1.000
C4	0.273	1.000	3.000	5.000	1.000	5.000
C5	0.173	3.000	2.000	3.000	3.000	1.000
Total	1.129	3.345	2.986	**4.727**	3.633	3.561

Fuente: La empresa
Elaboración: Propia

2.1.2.6 Decisión

De acuerdo con los puntajes obtenidos, se elige el Artículo 3.

2.1.3 Usos alternativos o aplicaciones varias

2.1.3.1 Dentro de selección de personal

El uso de enfoques clásicos de aprendizaje de máquina para la selección de personal es una herramienta que podría aplicarse no solo para G&S, sino también para cualquier otra empresa

de consultoría tecnológica o desarrollo de software presente en el mercado peruano. En ese caso tenemos algunas que destacan como:

- NTT Data
- Conastec
- Accenture Perú
- Xentic
- Innovation Hub Consulting

Este sistema podría integrarse fácilmente con estas entidades, debido a que la gran mayoría de las empresas debe tener un área de RRHH, que incluye el proceso de reclutamiento y selección de personal. Además de que los atributos de los datos de entrada son características generales manejados por muchas empresas del sector. Mishra et al. (2021) ejemplifica un caso para una organización de TI real que presenta ese problema.

2.1.3.2 Dentro del área de RRHH

Alejándonos un poco más, también es posible utilizar el aprendizaje de máquina para la solución de otros problemas dentro del área de RRHH, como lo son:

- **Deserción laboral**: Consiste en la renuncia voluntaria de un empleado de su puesto de trabajo empleado al no contar con las condiciones óptimas para ejercer su labor. Zhao et al. (2019) analiza esta casuística utilizando diferentes algoritmos.
- **Rendimiento del personal**: Producto del trabajo de un empleado o de un grupo de empleados. Lather et al. (2019) ataca el problema analizando diferentes evaluaciones que se pueden tomar, además de presentar un flujo de trabajo y resultados.
- **Ausentismo laboral**: Consiste en el abandono del lugar de empleo y de los deberes inherentes al mismo. También puede ser definido como la ausencia de una persona en su puesto de trabajo durante las horas que debería estar presente. Wahid et al. (2019) trabaja esta problemática abarcándola a nivel de horas, días y semanas, usando algoritmos basados en árboles.

2.1.4 Software o sistemas existentes

Rosales y Parrales (2022) señala que hay empresas que generalmente deciden renunciar al proceso de selección, por lo cual optan por contratar empresas externas especialistas en esto. Debido a esa necesidad es que actualmente existen software, programas o aplicaciones que permiten gestionar todo el proceso de reclutamiento y selección de las empresas, también denominados Sistemas de Seguimiento de Candidatos (ATS), como son:

- SmartRecruiters
    - Su principal pilar es la búsqueda de candidatos para un trabajo, el sistema guía a los posibles aplicantes a la empresa a conocerla y verificar el estado de las convocatorias. Además, centraliza documentos de los candidatos, permite la gestión de convocatorias por diversas personas al mismo tiempo, además de trazar el flujo de contratación.
    - Tiene varias sedes en América del Norte y Europa: USA, Polonia, Alemania, Reino Unido, etc.
    - Usa un modelo por suscripción junto con un periodo de prueba, el precio depende del número de empleados de la empresa, pero el mínimo precio es de $100.
    - Funciona en plataforma web y móvil.
- Zoho Recruit
    - Permite buscar, rastrear y contratar a los mejores candidatos para los puestos. También incluye publicaciones en bolsas de trabajo, creación de evaluaciones previas a la selección, análisis de hojas de vida, generación de reportes y seguimiento de las entrevistas.
    - Su sede principal se encuentra en Chennai, India.
    - Presenta un periodo de prueba de 15 días, una vez concluya se tiene una suscripción mínima de $25 por reclutador por mes.
    - Funciona en plataforma web y móvil.
- Jobvite
    - Usa un modelo de Compromiso Continuo de Candidatos (CCC), el cual está centrado en el candidato y ayuda a las empresas a atraer candidatos con experiencias significativas en el momento adecuado, de la manera correcta.
    - Su sede principal se encuentra en Indiana, USA.
    - Presenta un periodo de prueba de un mes, para luego cobrar $500 mensuales para empresas con más de 100 empleados.
    - Funciona en plataforma web y móvil.
- JazzHR
    - Presenta un sistema intuitivo de seguimiento de candidatos, permite a los reclutadores y gerentes construir un proceso escalable y efectivo que genere contrataciones de calidad.
    - Su sede principal se encuentra en Pennsylvania, USA.

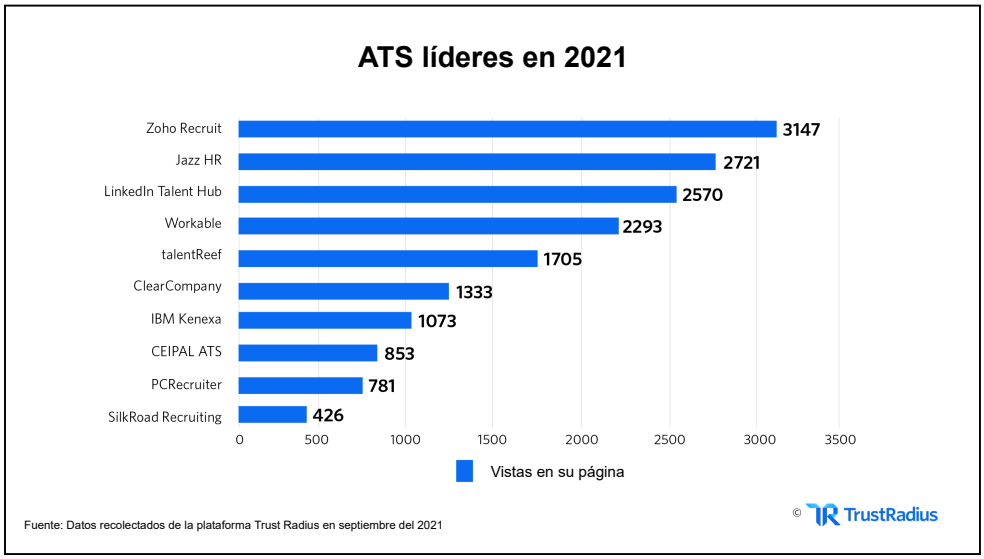
  - Presenta 3 planes de suscripción diferentes, de $39, de $239 y de $359 mensuales, en los cuales las funcionalidades se van incrementando a medida que sube el plan.
  - Solo funciona en plataforma web.

- Workable

  - Provee herramientas de contratación, facilita el manejo de etapas de la convocatoria, cuenta con herramientas automatizadas y basadas en inteligencia artificial que buscan y sugieren candidatos, simplifican la toma de decisiones y agilizan el proceso de contratación.
  - Su sede principal se encuentra en Massachusetts, USA.
  - Presenta 3 planes de suscripción diferentes, de $129, de $279 y de $559 mensuales, en los cuales las funcionalidades se van incrementando a medida que sube el plan.
  - Funciona en plataforma web y móvil.

Una comparación de éstos ATS mediante el número de visualizaciones en su página se describe en la Figura 2.12:

Figura 2.12: Comparación de ATS 2021



Fuente: Sadler (2021)

2.2 BASES TEÓRICAS

2.2.1 Variable dependiente: Proceso de selección de personal

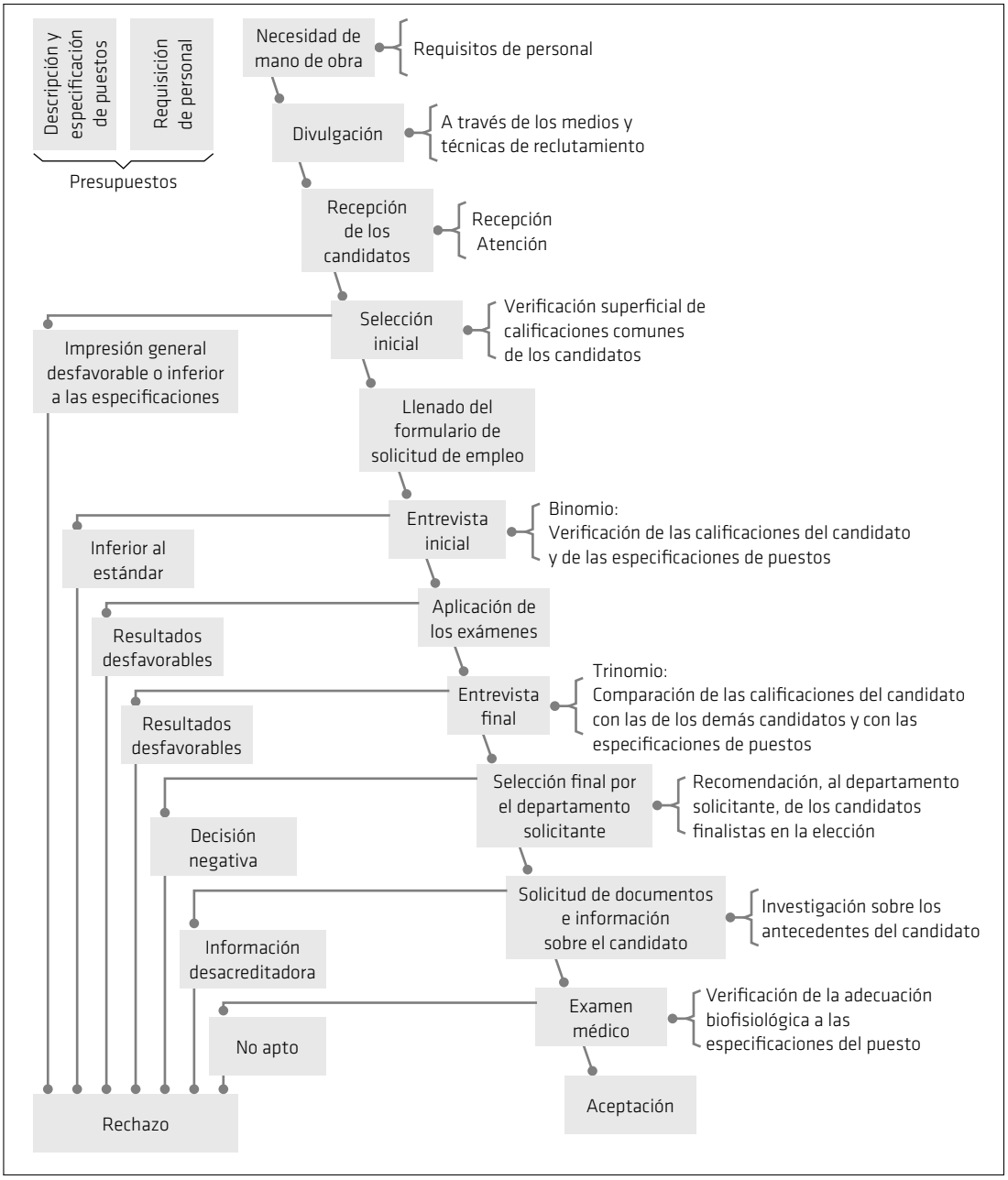
2.2.1.1 Selección de personal

Chiavenato (2011) define la selección de personal como el proceso que utiliza una organización para escoger, entre una lista de posibles candidatos, a la persona que mejor cumple con los criterios de selección para el puesto de trabajo disponible, dadas las condiciones actuales en el mercado laboral.

2.2.1.2 Fases de la selección de personal

Chiavenato (2011) define que el proceso de selección de personal consta de 12 pasos, los cuales se presentan en la Figura 2.13:

Figura 2.13: Proceso de selección de personal



Fuente: Chiavenato (2011)

2.2.1.2.1 Necesidad de mano de obra

Un departamento de la empresa determina que requiere contratar personal para poder cubrir una vacante, o ampliar su fuerza laboral. Se definen los requisitos y características del puesto.

2.2.1.2.2 Divulgación

La empresa publica ofertas de trabajo en una variedad de sitios web, que incluyen bolsas de trabajo, sitios de redes sociales, clasificados de periódicos y carteles de la empresa, esto con el fin de atraer a una gran cantidad de candidatos.

2.2.1.2.3 Recepción de los candidatos

La empresa recibe las hojas de vida y las solicitudes de los candidatos interesados en el puesto. Estos documentos se examinan para ver si cumplen con los requisitos establecidos.

2.2.1.2.4 Preselección

Se realiza una primera evaluación de los candidatos, tomando en cuenta su experiencia laboral, formación académica, habilidades, referencias, y otros criterios definidos previamente por el área de RRHH. Aquellos que sean más adecuados para el puesto continúan en el proceso.

2.2.1.2.5 Llenado del formulario de solicitud de empleo

En caso el reclutador requiera datos adicionales de los candidatos, los cuales no pudo recabar en la etapa anterior, puede hacerlo mediante el envío de un formulario y su posterior recepción.

2.2.1.2.6 Entrevista inicial

Se convoca a los candidatos para una entrevista inicial, que puede ser presencial o virtual. Además de analizar las habilidades y competencias de los candidatos, el objetivo es aprender más sobre ellos y determinar si serían adecuados para la empresa y el puesto.

2.2.1.2.7 Aplicación de los exámenes

Se utilizan para evaluar las habilidades técnicas, los conocimientos o las aptitudes de un candidato, según el tipo de puesto y las políticas de la empresa. Pueden ser pruebas escritas, prácticas o psicométricas.

2.2.1.2.8 Entrevista final

Los candidatos que han avanzado en las otras etapas son invitados a una entrevista final, durante la cual se examinan a fondo las facetas más particulares del puesto y se evalúa su idoneidad para el puesto. En esta fase también podrán participar directivos u otros profesionales relacionados con la toma de decisiones.

2.2.1.2.9 Fase final

Una vez concluida la entrevista final, el departamento solicitante revisa a los candidatos finalistas y toma la decisión final de seleccionar al más adecuado para el puesto. Esto implica tener en cuenta las habilidades técnicas y comunicativas para el puesto en cuestión.

2.2.1.2.10 Solicitud de documentos e información sobre el candidato

Una vez elegido el candidato, la empresa solicita la entrega de documentos, incluyendo transcripciones de antecedentes y cartas de recomendación de empleadores anteriores, con el fin de confirmar su validez y cumplimiento de requisitos.

2.2.1.2.11 Examen médico

Dependiendo del puesto, el empleador puede exigir que el solicitante se someta a un examen médico para confirmar su aptitud física para el puesto y su estado general de salud. Estas pueden ser pruebas de ojos, oídos, drogas u otras condiciones médicas pertinentes.

2.2.1.2.12 Aprobación

La decisión de contratación la toma la empresa después de que el candidato haya completado con éxito todos los pasos previos, incluido el examen médico. Se realiza una oferta de trabajo formal después de una evaluación final para garantizar que el candidato cumpla con todos los requisitos.

2.2.1.3 Tipos de selección de personal

Según Jaime (2023), los 3 tipos principales de selección de personal son:

- Interno: Busca el talento dentro de la empresa. Puede ser un empleado al cual se le haga un cambio de puesto o bien una persona relacionada con la empresa pero que actualmente no ejerce dentro de ella.
- Externo: Buscar talento fuera de la empresa, ya sea que esté buscando trabajo de manera activa o pasiva, es decir, no están buscando trabajo, pero están dispuestos a aplicar a una oportunidad laboral.
- Mixto: El reclutamiento de personal puede ser externo e interno simultáneamente al utilizar una combinación de ambos tipos. Se pueden combinar distintos métodos de reclutamiento para poder así buscar al candidato perfecto sin importar si está dentro o fuera de la empresa.

2.2.1.4 Canales de selección de personal

Según (Andrés, 2023), algunos de los canales principales de selección de personal son:

- Agencias de reclutamiento externo
- Bolsas de trabajo (universidades, institutos, escuelas de negocios)
- Ferias laborales
- Medios de comunicación (revistas, periódicos)
- Páginas de empleo (LinkedIn, Bumeran)
- Redes sociales

2.2.1.5 Dimensión: Preselección

Para definir los siguientes indicadores, nos apoyaremos de la matriz de confusión de orden 2, propuesta por (Kohavi y Provost, 1998):

Kohavi y Provost (1998) la define como una matriz que muestra las clasificaciones pronosticadas y reales, de 2 o más clases. La matriz de confusión de orden 2 se presenta en la Tabla 2.12:

Tabla 2.12: Matriz de confusión

		Valor real	
		Positivo	**Negativo**
Valor pronosticado	**Positivo**	Verdaderos positivos (VP)	Falsos positivos (FP)
	Negativo	Falsos negativos (FN)	Verdaderos negativos (VN)

Fuente: Kohavi y Provost (1998)

En esta matriz de orden 2, Kundu (2022) define los 4 elementos principales:

- Verdaderos positivos (VP): Se refiere a una muestra que pertenece a la clase positiva que se clasifica correctamente.
- Falsos positivos (FP): Se refiere a una muestra que pertenece a la clase negativa pero que se clasifica incorrectamente como perteneciente a la clase positiva. También es conocido como error tipo 1.

- Falsos negativos (FN): Se refiere a una muestra que pertenece a la clase positiva pero que se clasifica incorrectamente como perteneciente a la clase negativa. También es conocido como error tipo 2.
- Verdaderos negativos (VN): Se refiere a una muestra que pertenece a la clase negativa que se clasifica correctamente.

En el contexto específico de la variable dependiente, se pueden interpretar como:

- Verdaderos positivos (VP): Candidatos que se pronosticaron contratar, y realmente se contrataron.
- Falsos positivos (FP): Candidatos que se pronosticaron contratar, pero realmente no se contrataron.
- Falsos negativos (FN): Candidatos que se pronosticaron no contratar, pero realmente se contrataron.
- Verdaderos negativos (VN): Candidatos que se pronosticaron no contratar, y realmente no se contrataron.

Es en base a esa definición que se definen los siguientes 4 indicadores:

2.2.1.5.1 Indicador: Exactitud de selección de personal

Este indicador se define como el porcentaje de elementos clasificados correctamente sobre el conjunto total de elementos. El cálculo del indicador está descrito en la Ecuación 2.3:

$$ESP = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.3}$$

Dónde:

- ESP: Exactitud de selección de personal
- VP: Verdaderos positivos
- VN: Verdaderos negativos
- FP: Falsos positivos
- FN: Falsos negativos

2.2.1.5.2 Indicador: Precisión de selección de personal

Este indicador se define como el porcentaje de verdaderos positivos sobre el total de elementos pronosticados como positivos. El cálculo del indicador está descrito en la Ecuación 2.4:

$$PSP = \frac{VP}{VP + FP} \tag{2.4}$$

Dónde:

- PSP: Precisión de selección de personal
- VP: Verdaderos positivos
- FP: Falsos positivos

2.2.1.5.3 Indicador: Sensibilidad de selección de personal

Este indicador se define como el porcentaje de verdaderos positivos sobre el total de elementos con valor realmente positivo. El cálculo del indicador está descrito en la Ecuación 2.5:

$$SSP = \frac{VP}{VP + FN} \tag{2.5}$$

Dónde:

- SSP: Sensibilidad de selección de personal
- VP: Verdaderos positivos
- FN: Falsos negativos

2.2.1.5.4 Indicador: Robustez de selección de personal

Este indicador se define como el rendimiento combinado de los pronósticos considerando la precisión y la sensibilidad. Este indicador es calculado como la media armónica de esos dos indicadores. El cálculo del indicador está descrito en la Ecuación 2.6:

$$RSP = \frac{2 * PSP * SSP}{PSP + SSP} \tag{2.6}$$

Dónde:

- RSP: Robustez de selección de personal
- PSP: Precisión de selección de personal
- SSP: Sensibilidad de selección de personal

2.2.1.5.5 Indicador: Tiempo de filtrado de candidatos de selección de personal

Este indicador se define como la diferencia entre el tiempo final e inicial de la actividad de filtrado de candidatos (revisar el CV, verificar si se ajusta al puesto, corroborarlo con otros

candidatos y determinar si continúa en el proceso de selección). El cálculo del indicador está descrito en la Ecuación 2.7:

$$TFC = TFFC - TIFC \tag{2.7}$$

Dónde:

- TFC: Tiempo de filtrado de candidatos de selección de personal
- TFFC: Tiempo final de filtrado de candidatos de selección de personal
- TIFC: Tiempo inicial de filtrado de candidatos de selección de personal

2.2.1.5.6 Indicador: Tiempo de generación de reporte de selección de personal

Este indicador se define como la diferencia entre el tiempo final e inicial de la actividad de generación de reporte (filtrar los candidatos por convocatoria, corroborar los datos ingresados, darle formato al reporte y exportarlo en un archivo PDF). El cálculo del indicador está descrito en la Ecuación 2.8:

$$TGR = TFGR - TIGR \tag{2.8}$$

Dónde:

- TGR: Tiempo de generación de reporte de selección de personal
- TFGR: Tiempo final de generación de reporte de selección de personal
- TIGR: Tiempo inicial de generación de reporte de selección de personal

2.2.2 Variable independiente: Modelo predictivo

2.2.2.1 Modelo predictivo

Raschka (2015) lo define como una representación matemática que se ajusta automáticamente a los datos de entrenamiento y puede generalizar para hacer predicciones sobre datos no vistos previamente, esto enfocado principalmente en el campo de aprendizaje de máquina.

2.2.2.2 Aprendizaje de máquina

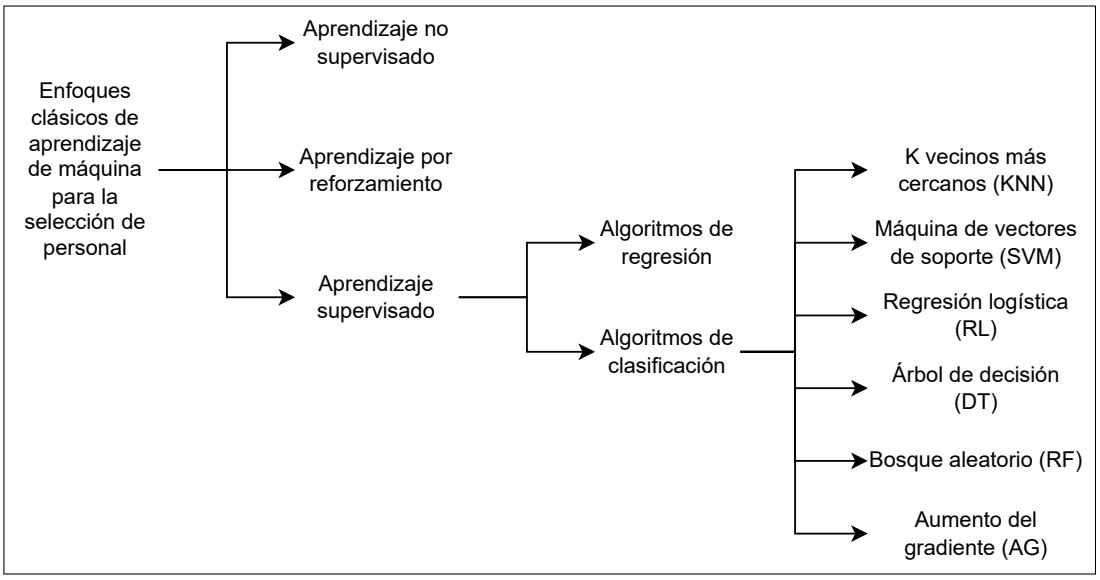
El aprendizaje de máquina es una rama de la inteligencia artificial que se centra en el estudio de algoritmos de computación que permiten que los programas informáticos mejoren automáticamente a través de la experiencia (Mitchell, 1997).

2.2.2.3 Modelo predictivo basado en aprendizaje de máquina

Goodfellow et al. (2016) lo define como sistema que pueden aprender de los datos y ajustar sus parámetros internos para realizar predicciones o tomar decisiones en nuevos datos sin necesidad de programación explícita.

Existen múltiples formas de interpretar la taxonomía de los modelos predictivos basados en aprendizaje de máquina. A continuación, en la Figura 2.14, se presenta una subdivisión en forma de árbol, la cual describe el campo de estudio de la presente investigación:

Figura 2.14: Taxonomía



Fuente: Hossain et al. (2020)

Algunos de los algoritmos de clasificación también pueden ser algoritmos de regresión, sin embargo, con fines explicativos solo se les ha categorizado como clasificación.

Existen 3 principales tipos de aprendizaje de máquina: no supervisado, por reforzamiento, y supervisado.

2.2.2.4 Tipos de aprendizaje de máquina

2.2.2.4.1 Aprendizaje no supervisado

Kohavi y Provost (1998) lo define como técnicas de aprendizaje que agrupan instancias sin un atributo dependiente preespecificado. Los algoritmos de agrupamiento generalmente no están supervisados.

En la actualidad no se han encontrado publicaciones destacadas que hagan uso del aprendizaje

no supervisado para resolver el problema de selección, sin embargo, si existen algunos estudios, como el de Eastwood (2020), el cual señala que mediante un análisis exploratorio de los atributos, permite aprender más de la calidad de los candidatos e inclusive, en los resultados descritos, señala que tuvo mejores resultados que un algoritmo de aprendizaje supervisado, por lo que es un campo en futuro desarrollo a tomar en cuenta.

2.2.2.4.2 Aprendizaje por reforzamiento

Sutton y Barto (2018) define el aprendizaje por refuerzo considerando 3 principales características:

- Es de ciclo cerrado de manera esencial: Las acciones del sistema influyen en sus entradas posteriores.
- No tener instrucciones directas: Sino que debe descubrir que acciones producen las mejores recompensas al probarlas.
- Cómo se desarrollan las secuencias de las acciones: En algunos casos, es posible que una acción no solo afecte a la recompensa inmediata, sino también a la siguiente, subsiguiente, y así consecutivamente.

En la actualidad no se han encontrado publicaciones destacadas que hagan uso del aprendizaje por reforzamiento para resolver el problema de selección, sin embargo, si existen algunas publicaciones que han hecho referencia a una ejemplificación de cómo se aplicaría para el problema, al menos de forma mínima. Jabbari et al. (2017) en su artículo ejemplifica cómo se podría dar este proceso:

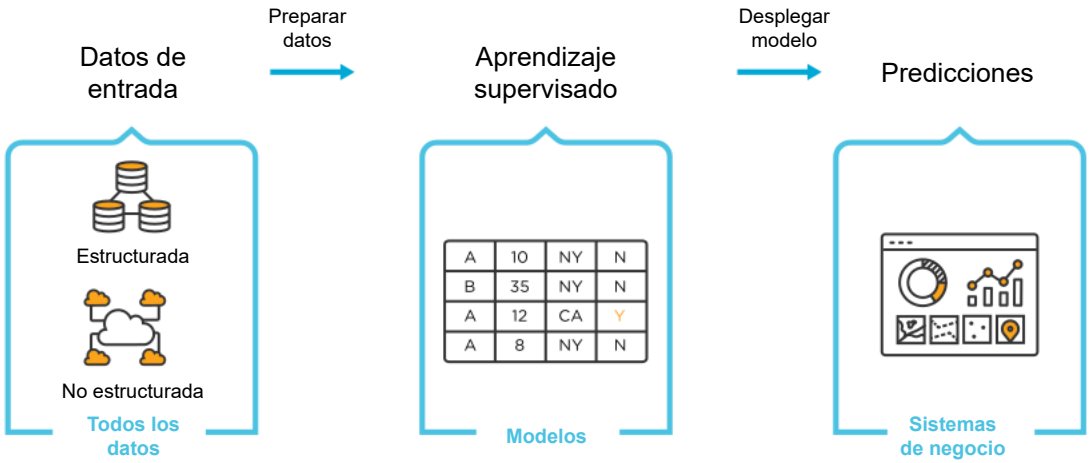
- Las acciones disponibles del sistema serían aquellas que afectan a las personas en sí (contratar o no contratar a un candidato).
- La recompensa por cada acción debe verse como el beneficio a corto plazo de tomar la decisión (la influencia a corto plazo en la productividad de la empresa una vez contratado).
- Las acciones que tome el algoritmo afectarán el estado futuro del sistema (la demografía de la empresa, el conjunto de candidatos aún disponible), el cual a su vez afectaría en las decisiones tomadas a futuro en el sistema.

2.2.2.4.3 Aprendizaje supervisado

Según IBM (2020), el aprendizaje supervisado es una subcategoría del aprendizaje de máquina, se define en usar un conjunto de datos etiquetados para entrenar algoritmos que clasifican datos o predicen resultados con precisión. El objetivo con esto es predecir una salida para nuevos

datos de entrada que no se hayan entrenado antes en el modelo. Un ejemplo del funcionamiento de estos algoritmos se describe en la Figura 2.15:

Figura 2.15: Aprendizaje supervisado



Fuente: Spotfire (2019b)

Tiene 2 principales subdivisiones, algoritmos de regresión y algoritmos de clasificación.

2.2.2.5 Tipos de algoritmos de aprendizaje supervisado

2.2.2.5.1 Algoritmos de regresión

Según Sharma (2021), los algoritmos de regresión son una técnica de aprendizaje de máquina en la que el modelo predice la salida como un valor numérico continuo. Este es el caso del artículo de Pessach et al. (2020), en el cual, por cada candidato de una convocatoria, obtiene un puntaje, en el intervalo de $[0, 1]$, el cual toma como referencia para su selección.

2.2.2.5.2 Algoritmos de clasificación

Aggarwal (2014) definió a los algoritmos de clasificación como un conjunto de pasos, los cuales tienen como tarea analizar los datos de entrada (compuestos por objetos de entrada y valores deseados de salida) y producir una función inferida discreta, que se puede usar para clasificar nuevas entradas o ejemplos del algoritmo, asignando una etiqueta de clase correcta a cada uno de ellos.

Este tipo de algoritmos solo predicen una etiqueta categórica de salida, el cual es una variable discreta en un número finito de valores posibles.

2.2.2.6 Tipos de algoritmos de clasificación

2.2.2.6.1 K vecinos más cercanos (KNN)

Este algoritmo se puede usar tanto como clasificación como para regresión.

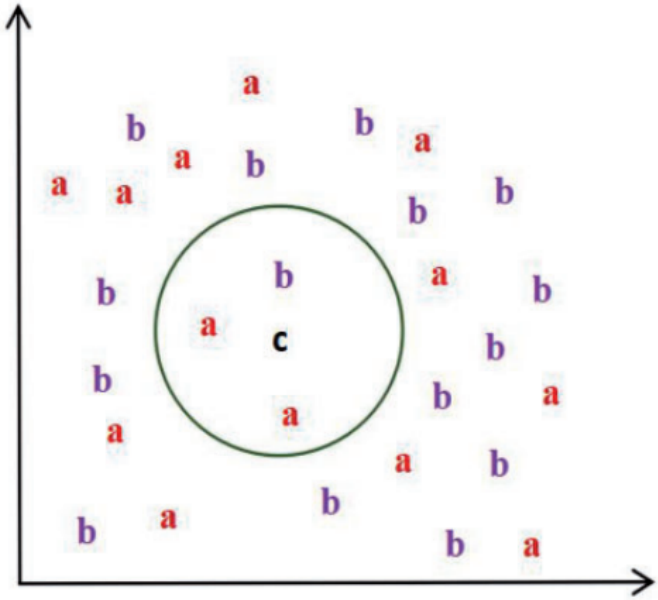
Los datos de entrada del algoritmo son aquel conjunto de datos de entrenamiento del modelo. Las salidas de los datos de entrada son las categorías, esas se colocan en un plano, y su posición depende de los valores de los atributos usados en esa ejecución del modelo. Una vez el modelo haya sido entrenado al colocar todas las etiquetas en el plano, se puede realizar las pruebas del modelo.

Por cada nuevo registro de los datos de prueba, lo que hace el algoritmo es, posicionar este dato dentro del plano, y calcular las etiquetas de los K vecinos (etiquetas en el plano) más cercanos a este. El valor de K es un parámetro del modelo, el cual debe ser un valor impar, ya que siempre genera un único ganador al aplicar mayoría simple a los K vecinos más cercanos.

Mishra et al. (2021) presenta un ejemplo descriptivo, ellos describen un ejemplo para modelo KNN con un K igual a 3.

En este caso, las 2 salidas del modelo son ’a’ y ’b’, que ya están posicionadas en el plano. Y en el ejemplo se visualiza un nuevo valor ’c’, que sería el dato por clasificar. En este caso, al tener dos vecinos con la etiqueta ’a’ y uno con la etiqueta ’b’, se determina que el nuevo dato se etiquetará el valor ’a’, debido a que ganó por mayoría simple. En la Figura 2.16 se describe este ejemplo:

Figura 2.16: Ejemplo del algoritmo KNN



Fuente: Mishra et al. (2021)

Algunos **casos de aplicación real** del algoritmo son: Sistemas de recomendación, búsqueda semántica y detección de anomalías.

Una de las principales **ventajas** que tiene el algoritmo es que es sencillo de aprender e implementar.

Una de las **desventajas** es que utiliza todo el conjunto de datos para poder entrenar cada punto del plano, lo que podría requerir mucha capacidad de procesamiento y memoria.

2.2.2.6.2 Máquina de vectores de soporte (SVM)

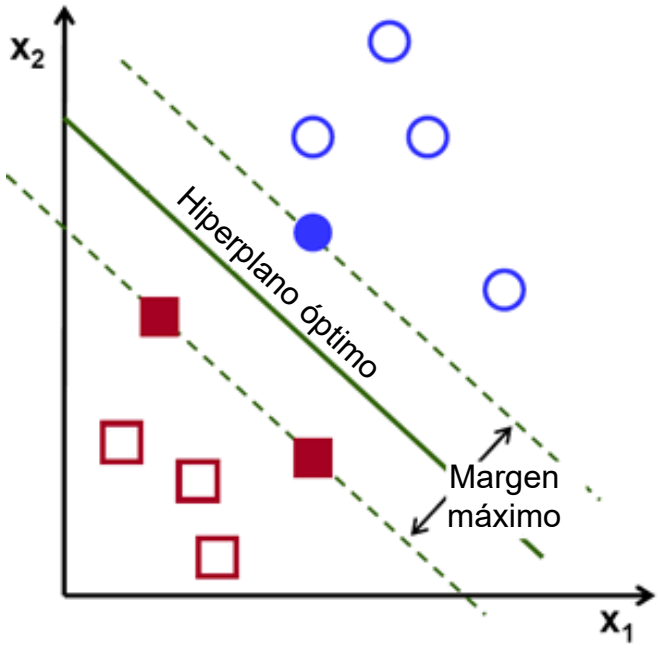
Este algoritmo se puede usar tanto como clasificación como para regresión.

Los datos de entrada del algoritmo son aquel conjunto de datos de entrenamiento del modelo. Las salidas de los datos de entrada son las categorías, esas se colocan en un espacio de N dimensiones (donde N es el número de atributos a usar), y su posición depende de los valores de los atributos usados en esa ejecución del modelo. Según Gandhi (2018), el objetivo principal del algoritmo es encontrar un hiperplano que pueda clasificar a las diferentes etiquetas de salida del modelo.

Esto se hace con el fin de, al ingresar una nueva entrada al modelo, se posiciona el dato en el espacio y, dependiendo en que zona delimitada por el hiperplano se encuentre, se clasificará en una etiqueta u otra.

Existen muchas formas de trazar este hiperplano, la prioridad es encontrar un plano que tenga un margen máximo, es decir, la distancia máxima entre 2 puntos de ambas clases, esos 2 puntos que se usan para trazar el plano son llamados vectores de soporte. Esto da cierto refuerzo a que los puntos de los datos futuros se clasifiquen con mayor confianza. Este hiperplano se describe en la Figura 2.17:

Figura 2.17: Hiperplano



Fuente: Gandhi (2018)

Algunos **casos de aplicación real** del algoritmo son: Aplicaciones médicas de procesamiento de señales, procesamiento del lenguaje natural y reconocimiento de imágenes y voz.

Algunas de las principales **ventajas** que tiene el algoritmo son: Elevada popularidad y buen rendimiento en clasificación y regresión, son más flexibles y permiten gestionar problemas no lineales.

Algunas de las principales **desventajas** que tiene el algoritmo son: Propenso al sobreajuste al tener más atributos que muestras, no son adecuados para grandes conjuntos de datos debido a su alto tiempo de formación.

2.2.2.6.3 Regresión logística (LR)

Este algoritmo se puede usar tanto como clasificación como para regresión.

Este algoritmo está basado en los modelos estadísticos clásicos de regresión, y se utiliza para estimar la probabilidad de una respuesta binaria (si o no) basada en una o más variables inde-

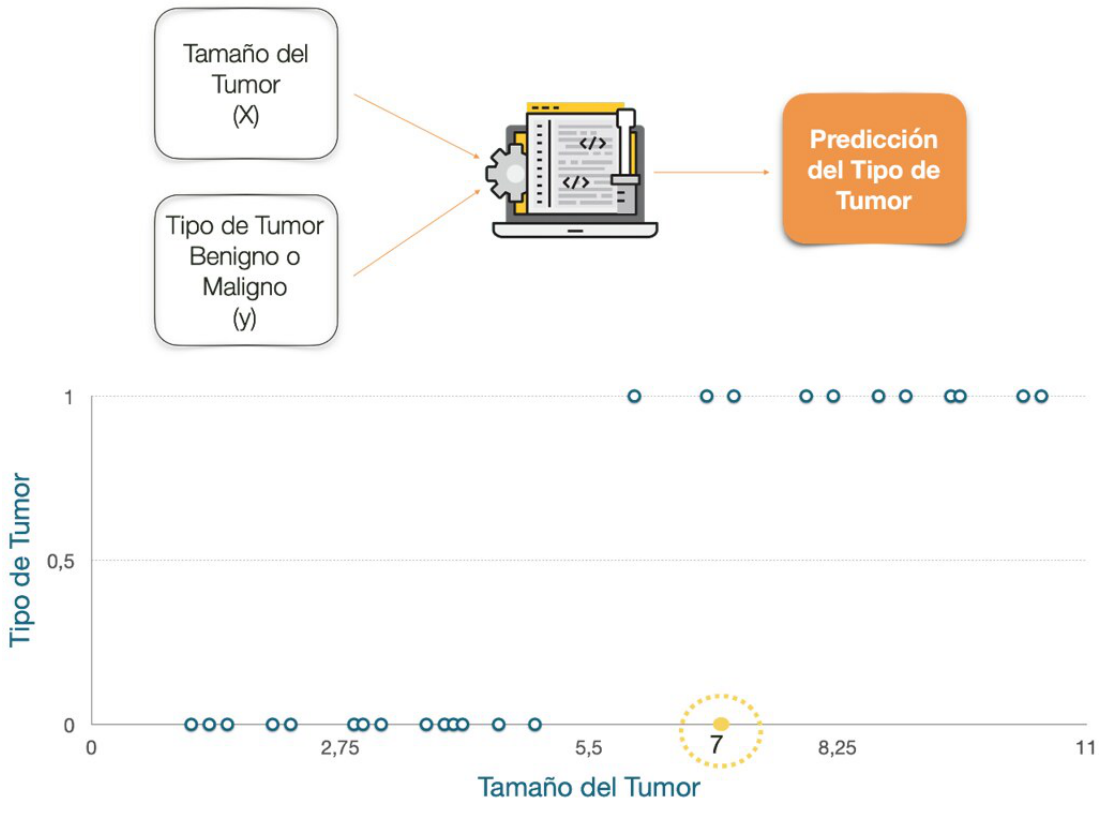
pendientes. Eso significa que la presencia de uno o más factores aumenta la probabilidad de un resultado u otro.

Gonzalez (2022) ejemplifica de manera más sencilla este algoritmo.

Supongamos que es necesario elaborar un modelo predictivo que determine el tipo de tumor, esto en base al tamaño del tumor. En la siguiente gráfica, vemos que los tamaños del tumor van desde 0cm hasta 11cm, y el punto crítico que determina si un tumor es benigno o maligno está alrededor de los 5.5cm.

En ese caso, es posible entrenar un modelo utilizando mediciones de la variable independiente (tamaño del tumor) y la dependiente (tipo del tumor), de tal forma que, cuando se haga una nueva medición del tamaño del tumor, sea posible predecir mediante regresión si el tumor es benigno o maligno. En la Figura 2.18 se describe este ejemplo:

Figura 2.18: Ejemplo de LR



Fuente: Gonzalez (2022)

Algunos **casos de aplicación real** del algoritmo son: Detección de spam, predicción de enfermedades, si un cliente comprará un producto o no.

Algunas de las principales **ventajas** que tiene el algoritmo son: Altamente eficiente, no necesita de gran recurso computacional, es fácilmente interpretable y fácil de entrenar.

Algunas de las principales **desventajas** que tiene el algoritmo son: No se puede utilizar para resolver problemas no lineales, no es el algoritmo más eficaz, además de tendencia al sobre-ajuste.

2.2.2.6.4 Árbol de decisión (DT)

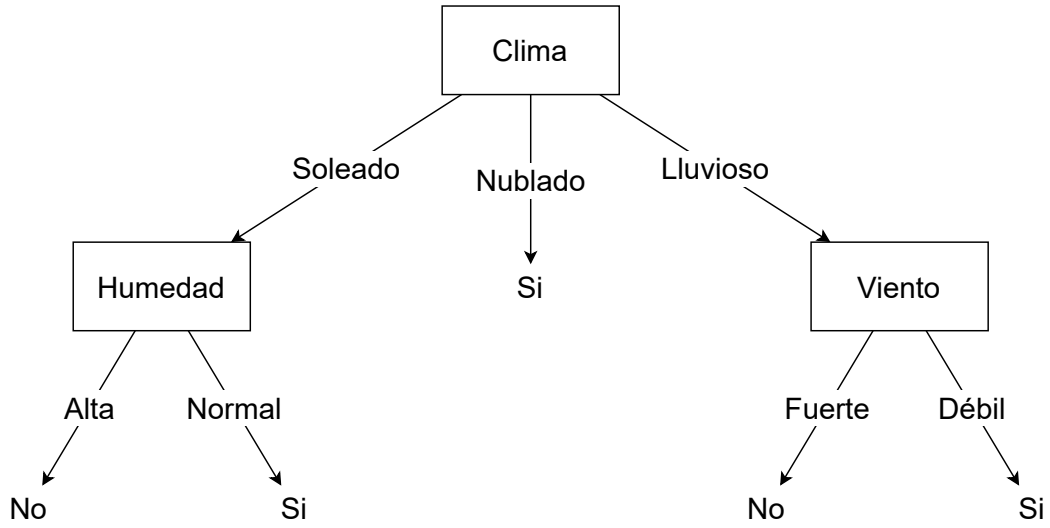
Este algoritmo se puede usar tanto como clasificación como para regresión.

Shubham (2018) define ampliamente el concepto. Este árbol de decisión parte de un nodo inicial llamado raíz, y a partir de él se descomponen el resto de los atributos de entrada, en dos o más ramas por cada uno (esto depende de si el tipo de dato es categórico o numérico), generando así más nodos. En cada uno de los nodos se plantea una condición que puede ser cierta o falsa, tantas ramas como subconjuntos de valores posibles de la condición se crearán a partir del mismo nodo.

Se itera de esa manera hasta llegar a las hojas, que son los nodos finales del árbol, es en estas hojas donde se toma la decisión final de la etiqueta a predecir.

En la Figura 2.19 se presenta un ejemplo de árbol de decisión, que determina si salir a jugar futbol, en base a cómo se encuentra el clima:

Figura 2.19: Ejemplo del algoritmo DT



Fuente: Shubham (2018)

Algunos **casos de aplicación real** del algoritmo son: Hacer estimaciones de la prima de seguros para cobrar a los asegurados, predecir si se debe ofrecer un producto a una determinada persona,

predecir la línea de crédito máxima de los clientes de un banco.

Algunas de las principales **ventajas** que tiene el algoritmo son: Sencillo de implementar, puede manejar data categórica y numérica, tolerante a valores atípicos.

Algunas de las principales **desventajas** que tiene el algoritmo son: Propenso al sobreajuste, posibilidad de crear árboles sesgados, medición constante para evaluar que tan bien lo está haciendo.

2.2.2.6.5 Bosque aleatorio (RF)

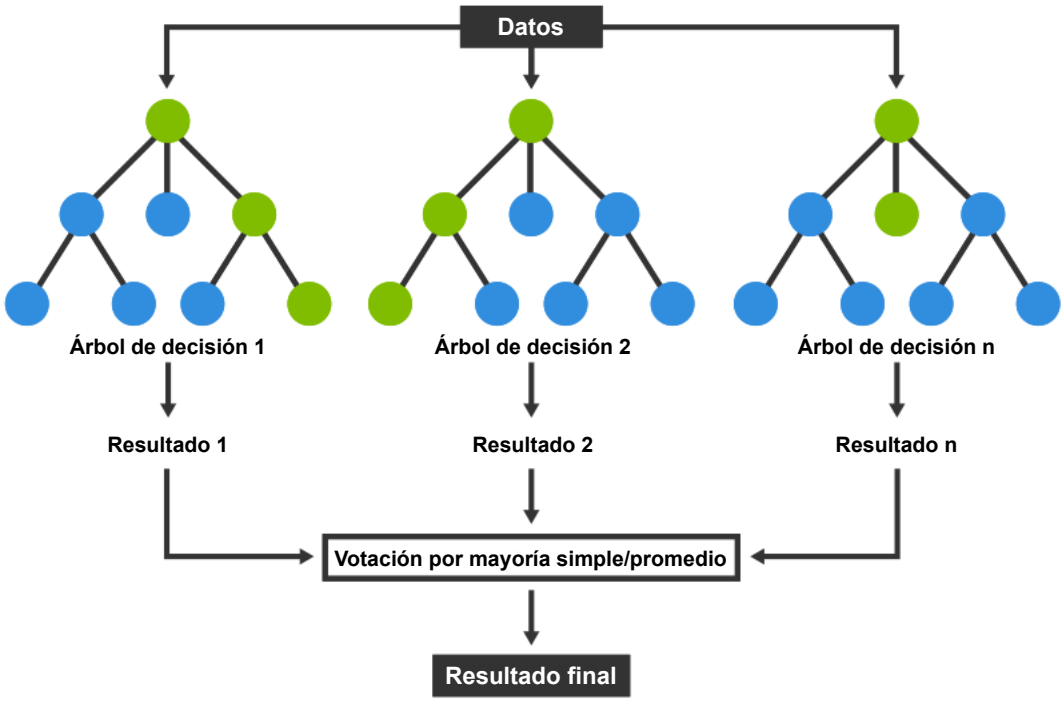
Este algoritmo se puede usar tanto como clasificación como para regresión.

Awujoola et al. (2021) lo define como un método en conjunto de un gran conjunto de árboles de decisión. En este caso, se inicia con la construcción de los árboles de decisión, su cantidad es definida por un parámetro N, en cada uno de los árboles, el algoritmo escoge un conjunto aleatorio de registros de los datos de entrada del algoritmo, para, a partir de ellos, armar el árbol de decisión de su iteración correspondiente, este paso se repite para los N árboles de decisión.

Posterior al entrenamiento de cada árbol, para poder conocer la etiqueta de salida de un nuevo registro ingresado, este registro nuevo se evalúa en cada uno de los árboles, por lo tanto, generará tantas salidas como árboles haya.

Finalmente, para elegir la etiqueta a predecir, se elige por mayoría simple según el número de ocurrencias tengan en las salidas de los árboles. Esto se describe en la Figura 2.20:

Figura 2.20: Algoritmo RF



Fuente: Spotfire (2019a)

Algunos **casos de aplicación real** del algoritmo son: Marketing telefónico para predecir el comportamiento de los clientes, o en finanzas para la gestión de riesgos

Algunas de las principales **ventajas** que tiene el algoritmo son: Riesgo reducido de sobreajuste, brinda flexibilidad, es fácil de determinar la importancia de los atributos.

Algunas de las principales **desventajas** que tiene el algoritmo son: Es un proceso que consume mucho tiempo, requiere muchos recursos computacionales, tiene una alta complejidad a comparación del árbol de decisión.

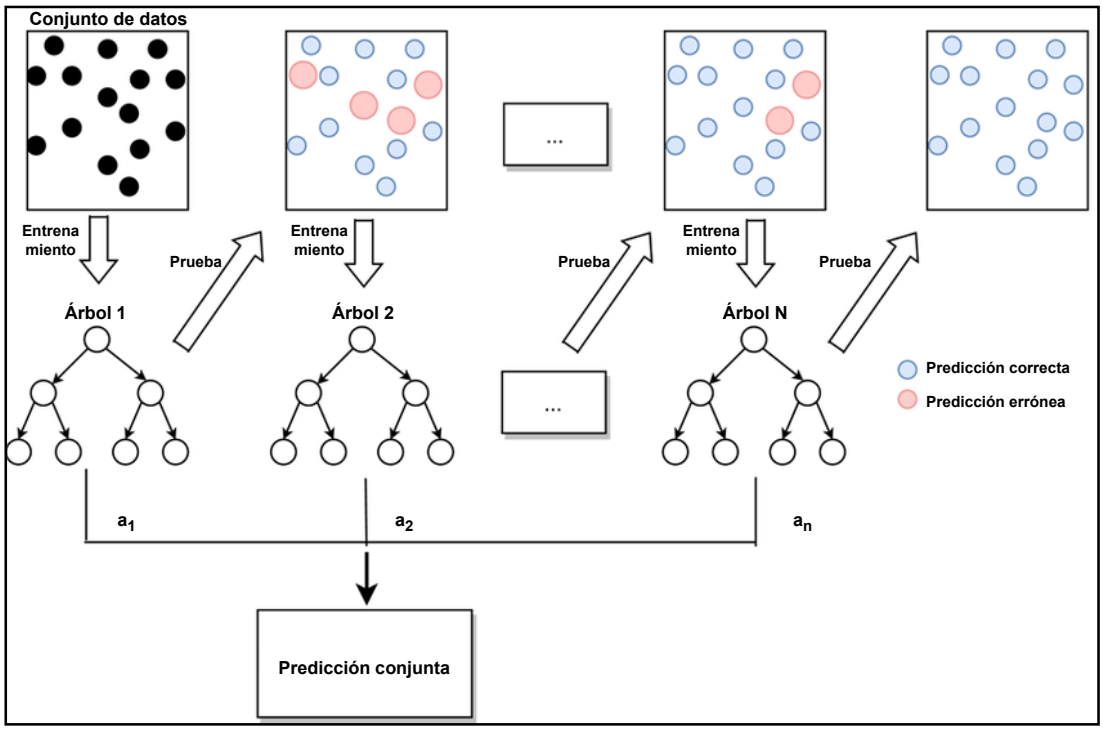
2.2.2.6.6 Aumento del gradiente (GBM)

Este algoritmo se puede usar tanto como clasificación como para regresión.

Este algoritmo está formado por un conjunto de árboles de decisión individuales, su cantidad es definida por un parámetro N. Estos árboles son entrenados de forma secuencial, uno detrás de otro, de tal forma que cada nuevo árbol trata de mejorar los errores de los árboles anteriores. Una vez entrenados los N árboles, al agregar un nuevo dato u observación, el resultado final se

obtiene agregando las predicciones de todos los árboles que forman el modelo (Amat, 2020). Se describe un ejemplo de este algoritmo en la Figura 2.21:

Figura 2.21: Algoritmo GBM



Fuente: Zhang et al. (2021)

Algunos **casos de aplicación real** del algoritmo son: Predicción atmosférica, diagnósticos médicos, detección de transacciones fraudulentas con tarjetas de crédito.

Algunas de las principales **ventajas** que tiene el algoritmo son: Capaces de seleccionar predictores de forma automática, maneja tanto predictores numéricos como categóricos, además que no se ven muy influenciados por valores atípicos.

Algunas de las principales **desventajas** que tiene el algoritmo son: Se pierde la interpretabilidad que tienen los modelos basados en un único árbol, además de que no son capaces de extrapolar fuera del rango de los predictores observado en los datos de entrenamiento.

2.2.2.7 Métricas de algoritmos de clasificación

Para definir las métricas de estos algoritmos, nos basaremos en la matriz de confusión:

2.2.2.7.1 Matriz de confusión

Kohavi y Provost (1998) la define como una matriz que muestra las clasificaciones pronosticadas y reales, de 2 o más clases. La matriz de confusión de orden 2 se presenta en la Tabla 2.13:

Tabla 2.13: Matriz de confusión

		Valor real	
		Positivo	**Negativo**
Valor pronosticado	**Positivo**	Verdaderos positivos (VP)	Falsos positivos (FP)
	Negativo	Falsos negativos (FN)	Verdaderos negativos (VN)

Fuente: Kohavi y Provost (1998)

En esta matriz de orden 2, Kundu (2022) define los 4 elementos principales:

- Verdaderos positivos (VP): Se refiere a una muestra que pertenece a la clase positiva que se clasifica correctamente.
- Falsos positivos (FP): Se refiere a una muestra que pertenece a la clase negativa pero que se clasifica incorrectamente como perteneciente a la clase positiva. También es conocido como error tipo 1.
- Falsos negativos (FN): Se refiere a una muestra que pertenece a la clase positiva pero que se clasifica incorrectamente como perteneciente a la clase negativa. También es conocido como error tipo 2.
- Verdaderos negativos (VN): Se refiere a una muestra que pertenece a la clase negativa que se clasifica correctamente.

Es a partir de esta matriz que se pueden definir los siguientes 4 indicadores:

2.2.2.7.2 Exactitud

La exactitud es la tasa de predicciones correctas hechas por el modelo sobre un conjunto de datos. Esta se evalúa generalmente sobre un conjunto de datos independiente que no se usó durante el proceso de aprendizaje del modelo (Kohavi y Provost, 1998). El cálculo del indicador está descrito en la Ecuación 2.9:

$$exactitud = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.9}$$

2.2.2.7.3 Precisión

La precisión es el grado de cercanía de los valores de varias medidas en un punto, con esta métrica es posible medir la calidad del modelo de aprendizaje de máquina en tareas de clasificación (Fernandes, 2020). El cálculo del indicador está descrito en la Ecuación 2.10:

$$precision = \frac{VP}{VP + FP} \tag{2.10}$$

2.2.2.7.4 Sensibilidad

Minaee (2019) lo define como la fracción de muestras de una clase que el modelo predice correctamente. También es complementado definiéndolo como la proporción de casos positivos reales que se identifican correctamente (Srivastava, 2023). El cálculo del indicador está descrito en la Ecuación 2.11:

$$sensibilidad = \frac{VP}{VP + FN} \tag{2.11}$$

2.2.2.7.5 Robustez

La métrica de robustez o valor F Se utiliza para evaluar los sistemas de clasificación binaria, que clasifican los ejemplos en positivos o negativos. El valor F es una forma de combinar la precisión y la recuperación del modelo, y se define como la media armónica de estos dos valores (Wood, 2019). El cálculo del indicador está descrito en la Ecuación 2.12:

$$robustez = \frac{2 * precision * sensibilidad}{precision + sensibilidad} \tag{2.12}$$

2.2.2.7.6 Tiempo

El tiempo está definido como la diferencia entre el tiempo final y tiempo inicial de algún hecho o evento. El cálculo del indicador está descrito en la Ecuación 2.13:

$$tiempo = tiempofinal - tiempoinicial \tag{2.13}$$

2.2.2.8 Técnicas de validación de algoritmos de clasificación

Estas técnicas de validación cruzada se usan para modelos ya entrenados, con el fin de validar la estabilidad del modelo, teniendo así una garantía de que el modelo tenga los patrones correctos,

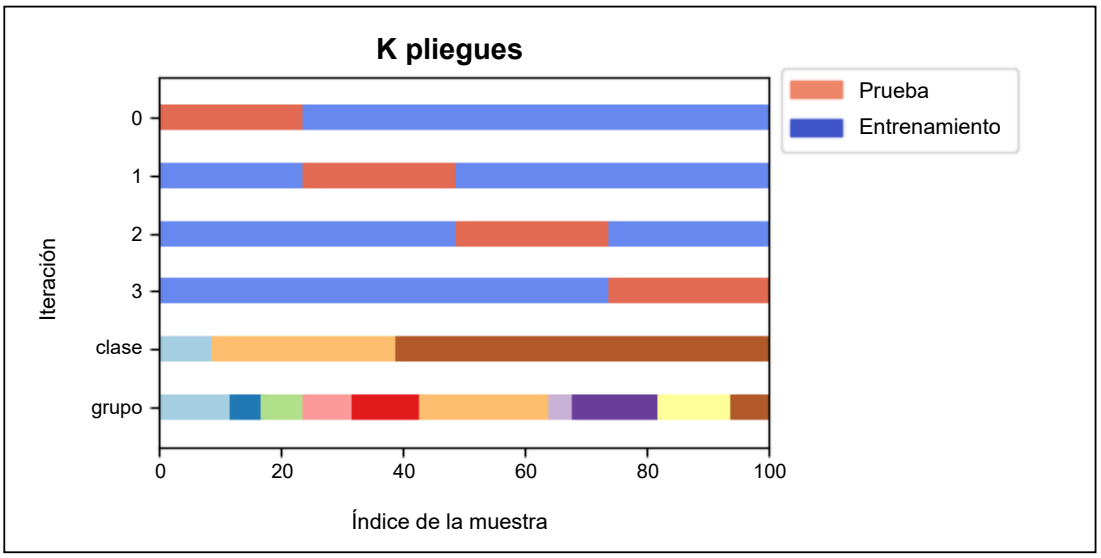
además de un sesgo y varianza bajos (Gupta, 2017).

Algunos ejemplos de técnicas son:

2.2.2.8.1 K pliegues (KP)

Consiste en dividir todas las muestras en K grupos de muestras, cada uno llamado pliegue, todos de igual tamaño (si es posible). Por cada iteración, el modelo evaluado es entrenado usando K-1 pliegues, y el pliegue restante es usado para las pruebas del modelo. Cabe recalcar que las propias clases de los datos no afectan la técnica (SciKitLearn, 2009). En la Figura 2.22 se muestra una descripción gráfica de la técnica:

Figura 2.22: K pliegues

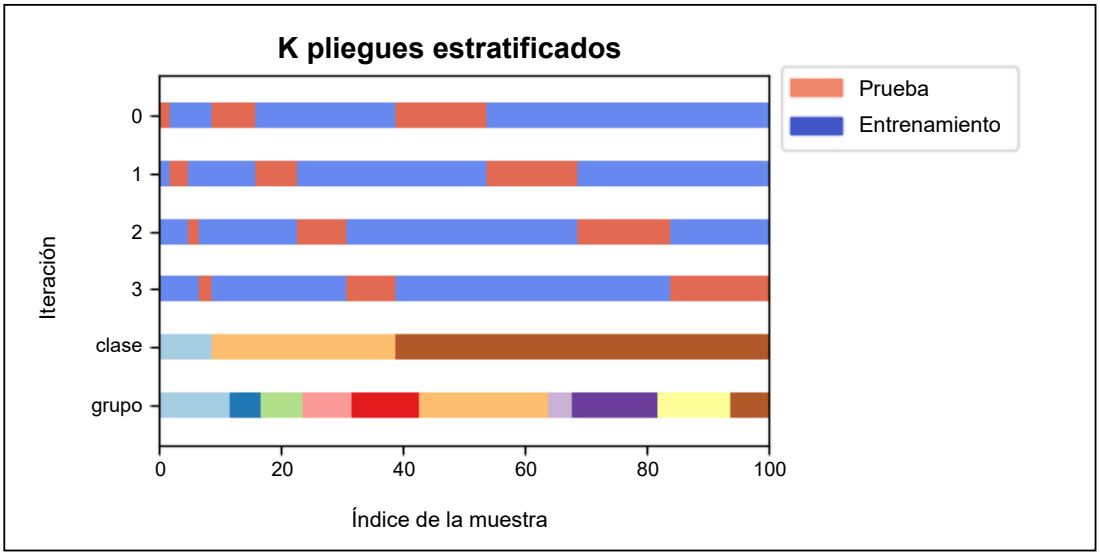


Fuente: SciKitLearn (2009)

2.2.2.8.2 K iteraciones estratificadas (KPE)

Es una variación de la técnica de K pliegues, en la cual se generan pliegues estratificados, cada uno de estos pliegues contiene aproximadamente el mismo porcentaje de muestras de cada clase objetivo del conjunto completo SciKitLearn (2009). En la Figura 2.23 se muestra una descripción gráfica de la técnica:

Figura 2.23: K pliegues estratificados



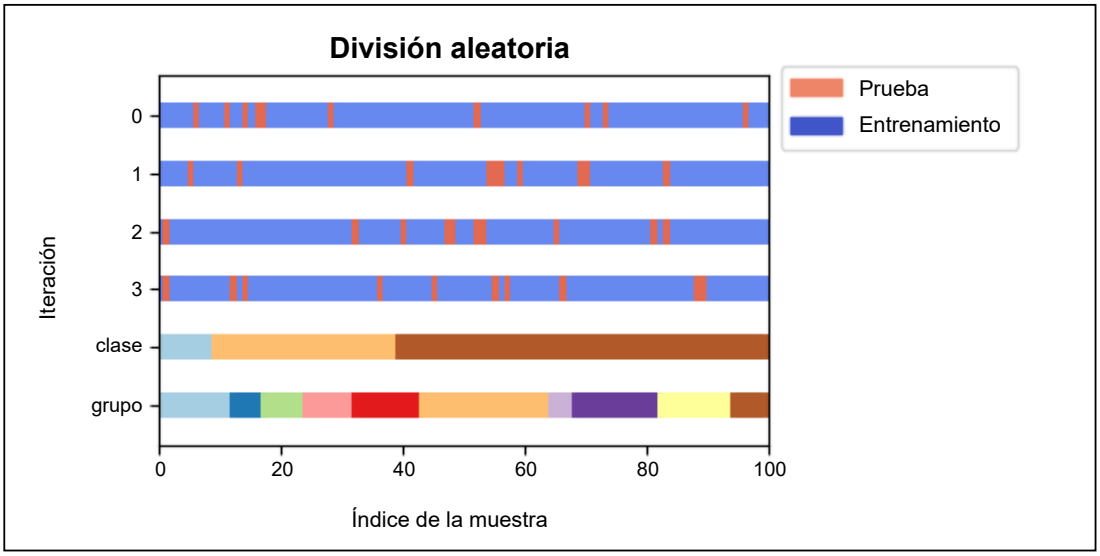
Fuente: SciKitLearn (2009)

2.2.2.8.3 División aleatoria (DA)

Esta técnica generará un número N de divisiones independientes del conjunto de datos inicial (no son excluyentes entre sí, a diferencia de la técnica de K pliegues), cada una de estas muestras tiene el mismo tamaño. Se entrena el modelo con N-1 divisiones y se prueba con la división restante.

Es posible controlar la aleatoriedad de la reproducibilidad de los resultados mediante un número generado. De igual forma, es una buena alternativa a la técnica K pliegues. SciKitLearn (2009). En la Figura 2.24 se muestra una descripción gráfica de la técnica:

Figura 2.24: División aleatoria

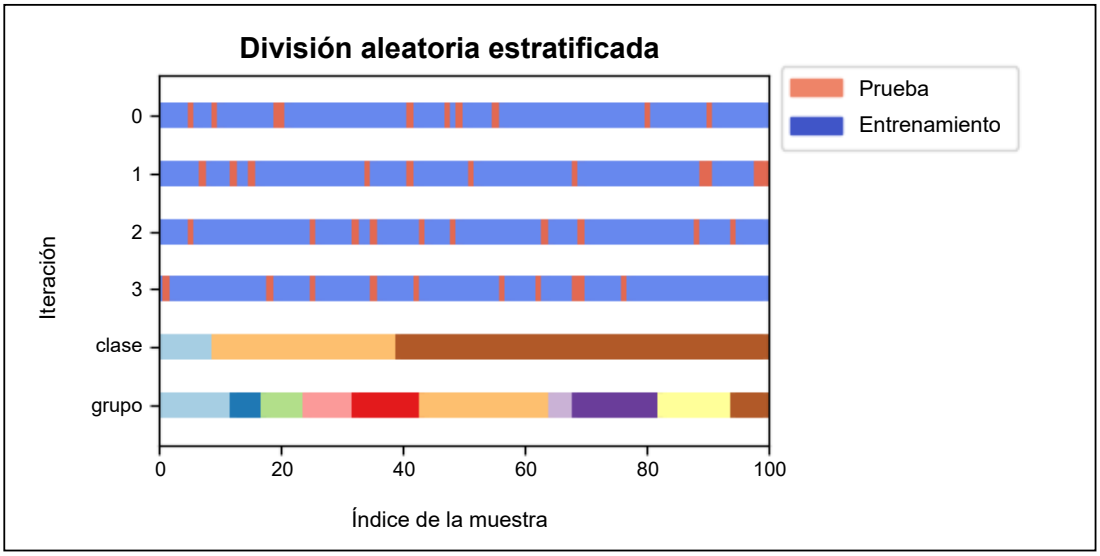


Fuente: SciKitLearn (2009)

2.2.2.8.4 División aleatoria estratificada (DAE)

Es una variación de la técnica de División aleatoria, en la cual se generan divisiones estratificadas, es decir, que crea divisiones conservando el mismo porcentaje para cada clase objetivo que en el conjunto completo SciKitLearn (2009). En la Figura 2.25 se muestra una descripción gráfica de la técnica:

Figura 2.25: División aleatoria estratificada



Fuente: SciKitLearn (2009)

2.2.2.9 Minería de datos

La minería de datos es el proceso de aplicar métodos computacionales a grandes cantidades de datos para revelar nueva información no trivial y relevante (Kallio y Tuimala, 2013). Hand (2007) complementa señalando que es el descubrimiento de estructuras interesantes, inesperadas o valiosas en grandes conjuntos de datos.

2.2.2.10 Metodologías de minería de datos

Si bien existen muchas metodologías en el ámbito de la minería de datos, existen 3 principales:

2.2.2.10.1 Metodología KDD

La metodología KDD (Descubrimiento de conocimiento en bases de datos) es un proceso que involucra la extracción de información útil, previamente desconocida y potencialmente valiosa de grandes conjuntos de datos. El proceso KDD es un proceso iterativo y requiere múltiples iteraciones de los pasos anteriores para extraer conocimiento preciso de los datos (Rajput, 2018).

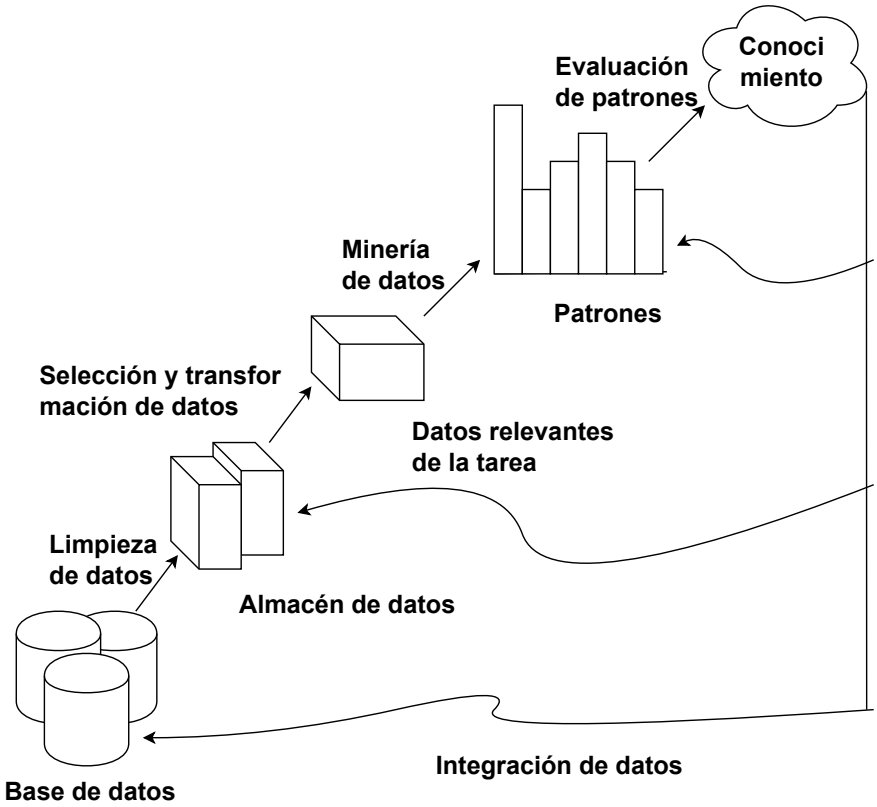
Las fases de esta metodología son:

- **Limpieza de datos:** Eliminación de datos ruidosos e irrelevantes de la recopilación.
- **Integración de datos:** Datos heterogéneos de múltiples fuentes combinados en una fuente común.

- **Selección de datos:** Se deciden los datos relevantes para el análisis y se recuperan de la recopilación de datos.
- **Transformación de datos:** Transformación de datos en la forma adecuada requerida por el procedimiento de minería.
- **Minería de datos:** Técnicas que se aplican para extraer patrones potencialmente útiles.
- **Evaluación de patrones:** Identificación de patrones estrictamente crecientes que representan el conocimiento basado en medidas dadas.
- **Representación del conocimiento:** Presentar los resultados de una manera que sea significativa y pueda usarse para tomar decisiones.

En la Figura 2.26, se describe la metodología de manera general:

Figura 2.26: Metodología KDD



Fuente: Rajput (2018)

2.2.2.10.2 Metodología SEMMA

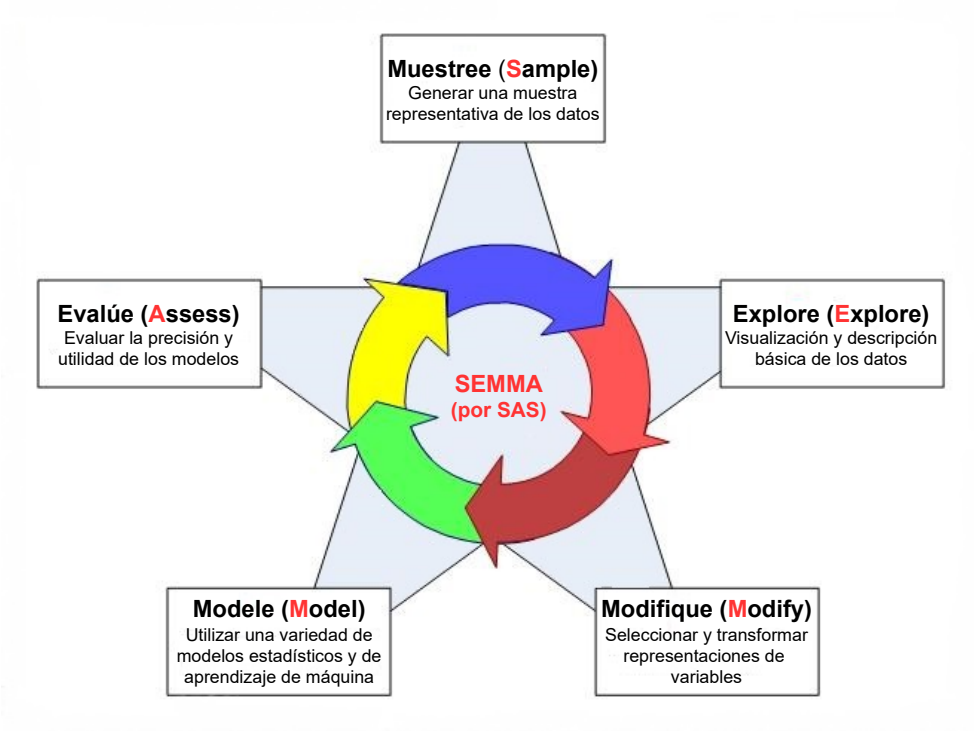
La metodología SEMMA son los métodos secuenciales para construir modelos de aprendizaje de máquina incorporados en SAS Enterprise Miner, un producto de SAS Institute Inc, uno de los mayores productores de software comercial de estadística e inteligencia empresarial. De igual manera, los pasos secuenciales guían el desarrollo de un sistema de aprendizaje de máquina (Jha, 2020).

Las fases de esta metodología son:

- **Muestree:** Seleccionar el subconjunto del conjunto de datos de volumen correcto de un gran conjunto de datos proporcionado para construir el modelo
- **Explore:** Se llevan a cabo actividades para comprender los vacíos de datos y la relación entre ellos
- **Modifique:** Las variables se limpian donde sea necesario. Las nuevas funciones derivadas se crean aplicando la lógica empresarial a las funciones existentes en función del requisito.
- **Modele:** Se aplican varias técnicas de modelado o minería de datos a los datos preprocesados para comparar su rendimiento con los resultados deseados
- **Evalúe:** El rendimiento del modelo se evalúa con los datos de prueba (no utilizados en el entrenamiento del modelo) para garantizar la confiabilidad y la utilidad empresarial

En la Figura 2.27, se describe la metodología de manera general:

Figura 2.27: Metodología SEMMA



Fuente: Samudra (2014)

2.2.2.10.3 Metodología CRISP-DM

La metodología CRISP-DM (Proceso estándar entre industrias para minería de datos) proporciona un enfoque estructurado para planificar un proyecto de minería de datos. Es una metodología robusta y bien probada, además de ser altamente práctica, flexible y útil.

Este modelo es una secuencia idealizada de eventos. En la práctica, muchas de las tareas se pueden realizar en un orden diferente y, a menudo, será necesario retroceder a tareas anteriores y repetir ciertas acciones. El modelo no intenta capturar todas las rutas posibles a través del proceso de minería de datos (Europe, 2017).

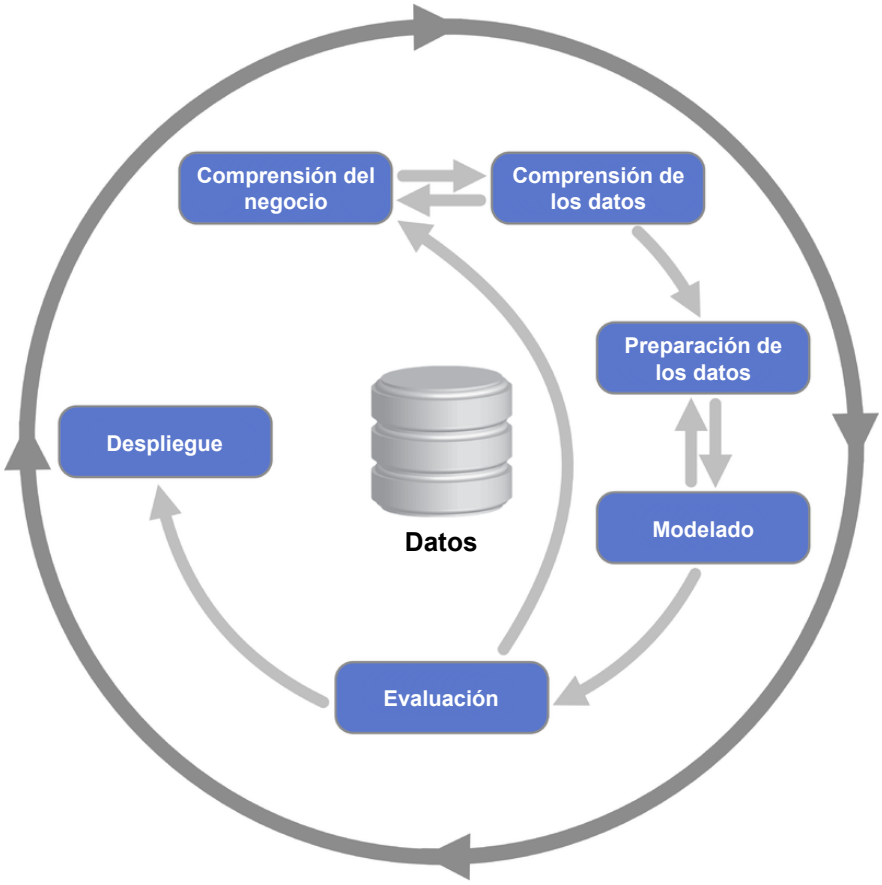
Las fases de esta metodología son:

- **Comprensión del negocio:** Comprender lo que desea lograr desde una perspectiva comercial y descubrir factores importantes que podrían influir en el resultado del proyecto.
- **Comprensión de los datos:** Recopilación inicial de los datos, pude incluir carga en datos en herramientas o integración de múltiples fuentes de datos.
- **Preparación de los datos:** Se deciden y procesan los datos que se van a utilizar para el análisis, además de considerar la calidad y el volumen total de datos.

- **Modelado:** Se seleccionan técnicas de modelado, se definen los procedimientos para validar el modelo y se usan herramientas de modelado para su construcción.
- **Evaluación:** Se evalúa el grado en que el modelo cumple con los objetivos comerciales y se busca determinar si existe alguna razón comercial por la cual este modelo es deficiente.
- **Despliegue:** Tomar los resultados de la evaluación y determinar una estrategia para su implementación.

En la Figura 2.28, se describe la metodología de manera general:

Figura 2.28: Metodología CRISP-DM



Fuente: Wikimedia Commons (2020a)

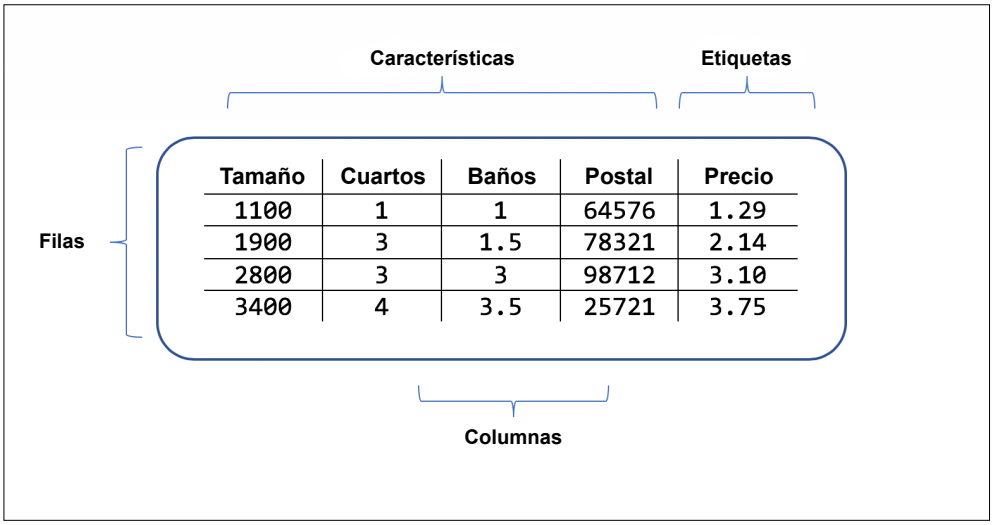
2.2.2.11 Otros conceptos

2.2.2.11.1 Conjunto de datos

Una colección de datos sin procesar, comúnmente organizados en formatos de hojas de cálculo o formatos estructurados, como el CSV (Google, 2023). Se ejemplifica un conjunto de datos en

la Figura 2.29:

Figura 2.29: Conjunto de datos



Fuente: Quintanilla (2022)

2.2.2.11.2 Columnas

Describe datos de un solo tipo. Todos los datos de una columna deben de tener la misma escala y tener un significado entre sí (Brownlee, 2019).

2.2.2.11.3 Filas

Una fila describe una sola instancia u observación dentro del conjunto de datos (Brownlee, 2019).

2.2.2.11.4 Características

También llamados atributos, son las variables de entrada para un modelo de aprendizaje de máquina, se usan para pronosticar la etiqueta de la fila en cuestión (Google, 2023).

2.2.2.11.5 Etiquetas

También llamadas clases, son las variables de salida para un modelo de aprendizaje de máquina, es la respuesta o resultado de una fila de un conjunto de datos (Google, 2023).

CAPÍTULO III
MÉTODO DE LA INVESTIGACIÓN

3.1 TIPO, NIVEL Y DISEÑO DE LA INVESTIGACIÓN

3.1.1 Tipo de la investigación

El tipo de esta investigación es aplicado, porque se resolverá un problema práctico, se aplicará un modelo predictivo de selección de personal para poder elegir siempre a empleados de alto rendimiento, que puedan aportar a la empresa.

3.1.2 Nivel de la investigación

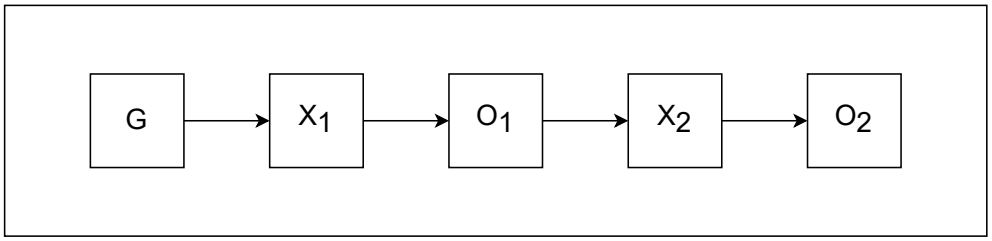
El nivel de esta investigación es explicativo, ya que tiene como objetivo comprender las relaciones causales entre variables. Busca identificar las causas y los efectos de un fenómeno y explicar cómo y por qué ocurren. Este nivel de investigación implica el uso de diseños experimentales y análisis estadísticos para establecer relaciones de causalidad.

3.1.3 Diseño de la investigación

El diseño de esta investigación es experimental porque se manejan una o más variables de estudio, controlando el aumento o disminución de estas variables para ver su efecto en las conductas observadas. Se requiere medir el efecto que tiene la variable independiente sobre la variable dependiente.

Se realizará un diseño de PreTest y PostTest con un solo grupo, el cual se presenta en la Figura 3.1

Figura 3.1: Diseño PreTest y PostTest



Fuente: La empresa
Elaboración: Propia

Dónde:

- Grupo experimental (G): Es el grupo al cual se le aplicó la medición para evaluar el proceso de selección de personal para medir la exactitud, precisión, sensibilidad, robustez, tiempo de filtrado de candidatos y tiempo de generación de reporte.
- Experimento 1 (X_1): Es la aplicación del modelo predictivo inicial sin validar en el proceso de selección de personal.
- PreTest (O_1): Medición del grupo experimental con la aplicación del modelo predictivo inicial sin validar, en el proceso de selección de personal de manera tradicional o manual. Esta medición será comparada con la medición del PostTest.
- Experimento 2 (X_2): Es la aplicación del modelo predictivo final validado en el proceso de selección de personal. Mediante dos mediciones (PreTest y PostTest) se podrá medir si el modelo predictivo final validado mejora el proceso de selección de personal.
- PostTest (O_2): Medición del grupo experimental después de la aplicación del modelo predictivo final validado, en el proceso de selección de personal de manera automatizada. Ambas mediciones serán comparadas y ayudará a determinar la exactitud, precisión, sensibilidad, robustez, tiempo de filtrado de candidatos y tiempo de generación de reporte, antes y después de la aplicación del modelo predictivo final validado.

El diseño de esta investigación también es cuantitativo, ya que se utilizará análisis estadístico para probar las hipótesis de la investigación.

3.2 VARIABLES Y OPERACIONALIZACIÓN

3.2.1 Variables

Variable independiente: Modelo predictivo.

Variable dependiente: Proceso de selección de personal.

3.2.2 Operacionalización de variables

En la Figura 3.1 se presenta la operacionalización de variables:

Tabla 3.1: Operacionalización de variables

Tipo	Variable	Definición conceptual	Definición operacional	Dimensión	Indicador	Unidad de medida
Independiente	Modelo predictivo	Raschka (2015) lo define como una representación matemática que se ajusta automáticamente a los datos de entrenamiento y puede generalizar para hacer predicciones sobre datos no vistos previamente, esto enfocado principalmente en el campo de aprendizaje de máquina	Realiza la limpieza de los datos de las convocatorias, preparación de datos y el entrenamiento del modelo de aprendizaje de máquina, para su posterior evaluación			
Dependiente	Proceso de selección de personal	Chiavenato (2009) define la selección de personal como el proceso que utiliza una organización para escoger, entre una lista de posibles candidatos, a la persona que mejor cumple con los criterios de selección	Realiza la publicación de convocatorias de trabajo, filtrado de candidatos, pruebas técnicas y entrevistas, con el fin de elegir finalmente, al mejor candidato para el puesto	Preselección	Exactitud	Porcentaje (%)
					Precisión	Porcentaje (%)
					Sensibilidad	Porcentaje (%)
					Robustez	Porcentaje (%)
					Tiempo de filtrado de candidatos	Segundo (s)
					Tiempo de generación de reporte	Segundo (s)

Fuente: La empresa
Elaboración: Propia

En la Figura 3.2 se presenta la operacionalización de indicadores:

Figura 3.2: Operacionalización de indicadores

Dimensión	Indicador	Descripción	Técnica	Instrumento	Unidad de medida	Fórmula
Preselección	Exactitud	Indica el porcentaje de elementos clasificados correctamente sobre el conjunto total de elementos	Observación	Registros o métricas del modelo	Porcentaje (%)	$\frac{VP+VN}{VP+VN+FP+FN}$
	Precisión	Indica el porcentaje de verdaderos positivos sobre el total de elementos pronosticados como positivos	Observación	Registros o métricas del modelo	Porcentaje (%)	$\frac{VP}{VP+FP}$
	Sensibilidad	Indica el porcentaje de verdaderos positivos sobre el total de elementos con valor realmente positivo	Observación	Registros o métricas del modelo	Porcentaje (%)	$\frac{VP}{VP+FN}$
	Robustez	Indica el rendimiento combinado de los pronósticos considerando la precisión y la sensibilidad	Observación	Registros o métricas del modelo	Porcentaje (%)	$\frac{2*PSP*SSP}{PSP+SSP}$
	Tiempo de filtrado de candidatos	Indica la diferencia entre el tiempo final e inicial de la actividad de filtrado de candidatos	Observación	Registros o métricas del modelo	Segundo (s)	$TFFC\ -TIFC$
	Tiempo de generación de reporte	Indica la diferencia entre el tiempo final e inicial de la actividad de generación de reporte	Observación	Registros o métricas del modelo	Segundo (s)	$TFGR\ -TIGR$

Fuente: La empresa
Elaboración: Propia

3.3 POBLACIÓN Y MUESTRA

3.3.1 Población

Se tomará como población a todos los postulantes a puestos de trabajo desde el 21 de junio del 2019 hasta el 22 de mayo del 2023 (1431 días o 3 años y 11 meses). Esto equivalen a 303 convocatorias de trabajo, el cual a su vez corresponde a **10562 postulaciones diferentes**.

Estas 10562 postulaciones estarán incluidas en 48 ejecuciones del modelo, de tal forma que, cada ejecución del modelo usará el total de postulaciones, pero las métricas resultantes por ejecución serán únicas. A estos **48 conjuntos de métricas** se le aplicará el análisis estadístico para las pruebas de normalidad e hipótesis.

3.3.2 Muestra

Para la presente investigación, se consideraron los límites y el tamaño de la muestra iguales a los de la población (10562 postulaciones), esto debido a:

1. Un modelo predictivo con mayor número de datos de entrenamiento y validación tiende a tener mejores resultados (exactitud, precisión, sensibilidad, robustez).
2. Los recursos disponibles son los suficientemente potentes para no ver comprometidos los tiempos (tiempo de generación de resultados, tiempo de generación de reporte).

3.4 TÉCNICAS E INSTRUMENTOS DE RECOLECCIÓN DE DATOS, VALIDEZ Y CONFIABILIDAD

3.4.1 Técnicas

Se usará la técnica de observación, debido a que se realizará una revisión manual de los indicadores.

3.4.2 Herramientas

Se usará la herramienta de registros o métricas en el modelo, ya que es la salida generada por parte del modelo, la cual tiene los valores finales de los indicadores.

3.5 MÉTODOS DE ANÁLISIS DE DATOS

Se usará el método de análisis cuantitativo, ya que se analizarán y manipularán los datos numéricos recopilados. En base a estos resultados se contrastarán las hipótesis y se formularán conclusiones del estudio (Río Sadornil, 2005).

Se utilizarán 2 diferentes pruebas para el análisis de datos:

3.5.1 Prueba de normalidad

Para la elección del tipo de prueba se tomó en cuenta el tamaño de la muestra. En ese sentido, se utilizará la prueba de Shapiro-Wilk para determinar la normalidad de los indicadores, ya que se trabajará con una muestra menor a 50.

3.5.1.1 Prueba de Shapiro-Wilk

Esta prueba se utiliza para contrastar la normalidad de un conjunto de datos y se considerada de las mejores pruebas para el contraste de normalidad.

De manera general, se definen las siguientes hipótesis estadísticas:

- H_0: El indicador presenta una distribución normal
- H_A: El indicador presenta una distribución no normal

Basadas en la siguiente toma de decisión:

- Valor P $< \alpha$: Se rechaza H_0
- Valor P $>= \alpha$: No se rechaza H_0

Dónde el estadístico de prueba es definido por la Ecuación 3.1:

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{3.1}$$

Dónde:

- $x_{(i)}$ (con el subíndice i entre paréntesis) es el número que ocupa la posición i en la muestra (con la muestra ordenada de menor a mayor).
- $\overline{x}$ es la media muestral.

3.5.2 Prueba de hipótesis

Para la elección del tipo de prueba se tomó en cuenta el tamaño de la muestra y los resultados de la prueba anterior (normalidad o no normalidad). En ese sentido, se utilizará la prueba de Wilcoxon, debido a la no normalidad de algunos indicadores, además del tamaño de la muestra (menor a 50).

3.5.2.1 Prueba de Wilcoxon

La prueba de Wilcoxon es una prueba no paramétrica para comparar el rango medio de dos muestras y determinar si existen diferencias entre ellas.

De manera general, se definirían las hipótesis de investigación:

- H_x : El modelo predictivo aumenta el indicador del proceso de selección de personal.

3.5.2.1.1 Hipótesis estadísticas

Definición de variables:

- I_a : Indicador antes de la aplicación del modelo predictivo final validado.
- I_d :: Indicador después de la aplicación del modelo predictivo final validado.

Definición de hipótesis estadísticas:

- H_0 : El modelo predictivo no aumenta el indicador del proceso de selección de personal, según la Ecuación 3.2.

$$H_0 = Ia >= Id \tag{3.2}$$

  El indicador sin el modelo predictivo es mejor que el indicador con el modelo predictivo.

- H_A : El modelo predictivo aumenta el indicador del proceso de selección de personal, según la Ecuación 3.3.

$$H_A = Ia < Id \tag{3.3}$$

  El indicador con el modelo predictivo es mejor que el indicador sin el modelo predictivo.

Dónde el estadístico de prueba es definido por la Ecuación 3.4:

$$z = \frac{W - \mu_w}{\sigma_w} \tag{3.4}$$

Dónde:

- W: Estadístico W de la prueba de Wilcoxon.
- μ_w: Valor esperado de W.
- σ_w: Desviación estándar de W.

El valor z con el cual comparar depende del nivel de confianza de la prueba, los valores comúnmente usados se describen en la Tabla 3.2:

Tabla 3.2: Niveles de confianza y Z_α para la prueba de Wilcoxon

Nivel de confianza	Z_α
99.75%	3
99%	2.58
98%	2.33
96%	2.05
95%	1.96
90%	1.645
80%	1.28
50%	0.674

Fuente: Statistics Kingdom (2023)

Para la presente investigación se ha definido un nivel de significancia del 5%. Pero al utilizar la prueba de Wilcoxon de 2 colas (ambos lados de la gráfica), el nivel de confianza es 90%, por lo tanto, el estadístico obtenido del indicador se comparará con $Z_\alpha = 1.645$

3.6 ASPECTOS LEGALES Y ÉTICOS

Yaranga (2022) señala que deben tenerse ciertos aspectos éticos en el proceso de selección, manteniendo una confidencialidad, confianza y seguridad de los datos en todo momento.

Dentro del proceso de selección de personal, existen diversas políticas que G&S debe tomar en cuenta al momento de poder completar el proceso sin incurrir en faltas legales contra leyes o decretos promulgados por el estado peruano, estas son:

- Ley N.º 26772: Igualdad de oportunidades y de trato.
    - La ley asegura que las ofertas de empleo y acceso a medios de formación educativa no podrán contener requisitos que constituyan discriminación, anulación o alteración de igualdad de oportunidades o de trato.
    - La ley define por discriminación a la anulación o alteración de la igualdad de oportunidades o de trato, en los **requerimientos de personal**, a los requisitos para acceder a centros de educación, formación técnica y profesional, que impliquen un trato diferenciado basado en motivos de raza, sexo, religión, opinión, origen social, condición económica, estado civil, edad o de cualquier índole.
    - La ley también señala que las personas individuales o jurídicas que incurran en discriminación serán sancionadas por el MTPE sin perjuicio de la indemnización a

que hubiere lugar.

- Ley N.º 30709: Ley que prohíbe la discriminación remunerativa entre varones y mujeres.
    - Esta ley prohíbe la **discriminación remunerativa entre varones y mujeres**, mediante la determinación de categorías, funciones y remuneraciones que permitan la ejecución del principio de igual remuneración por igual trabajo.
    - También asegura que en sus planes de formación profesional y de desarrollo de capacidades laborales de sus trabajadores se garantice la igualdad entre mujeres y hombres.
- Decreto Legislativo N.º 728: Ley de Productividad y Competitividad Laboral.
    - La ley prohíbe que el despido del empleado no se puede dar a través de la causa de discriminación por sexo, raza, religión, opinión o idioma.
- Ley N.º 29973: Ley general de la persona con discapacidad.
    - La ley establece que las personas con discapacidad tienen derecho a trabajar, en igualdad de condiciones que las demás, en un trabajo libremente elegido o aceptado, con igualdad de oportunidades y de remuneración por trabajo de igual valor, y con condiciones de trabajo justas, seguras y saludables.
- Ley N.º 30687: Ley de promoción de los derechos de las personas de talla baja
    - La ley establece que entidades públicas y privadas deberán fomentar entornos de trabajo que integren a las personas de talla baja, debiendo asegurar las modificaciones y/o adaptaciones necesarias y adecuadas, requeridas en un caso particular, en el espacio físico, en el mobiliario y las herramientas de trabajo.

La empresa se encuentra altamente comprometido con el cumplimiento de estas leyes para poder culminar el proceso de reclutamiento y selección de una manera segura, efectiva y conflictos legales y jurídicos con el MTPE, una entidad que vela por el cumplimiento de las disposiciones ya mencionadas.

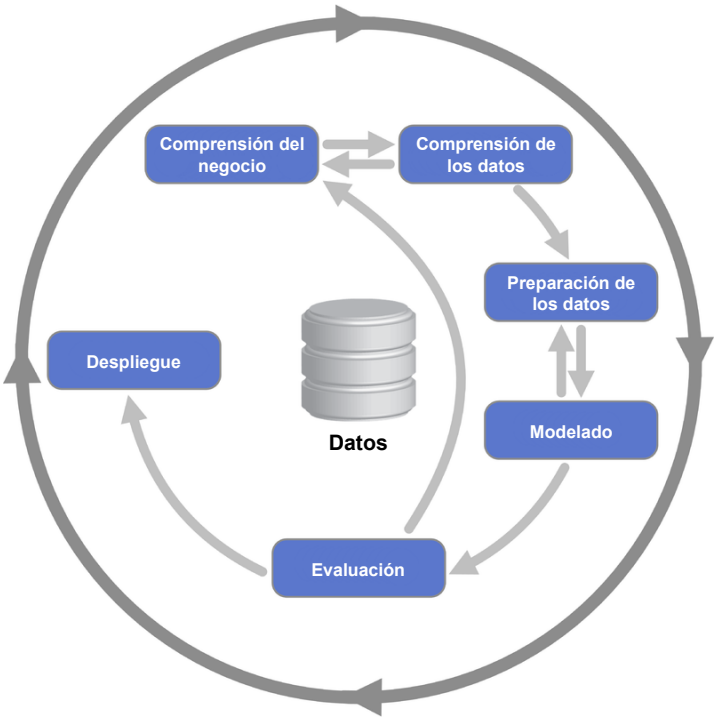
CAPÍTULO IV
DESARROLLO DE LA SOLUCIÓN

4.1 METODOLOGÍA DE DESARROLLO DE LA SOLUCIÓN

Para lograr el objetivo de la investigación se usará la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), la cual proporciona una descripción normalizada del ciclo de vida de un proyecto estándar de análisis de datos.

Esta metodología es un estándar de factor del mercado. Surgida en dos empresas que han sido pioneras en la aplicación de minería de datos a los procesos de negocio: DaimlerChrysler y SPSS (Vallalta, 2019). Las etapas de la metodología se muestran en la Figura 4.1:

Figura 4.1: Metodología CRISP-DM



Fuente: Wikimedia Commons (2020a)

La metodología CRISP-DM establece un proyecto de minería de datos como una secuencia de 6 fases, las cuales se describen a continuación:

4.1.1 Comprensión del negocio

El objetivo principal de esta fase es alinear los objetivos del proyecto de minería de datos con los objetivos del negocio. Tratando así de evitar embarcarnos en un proyecto de minería de datos que no produzca ningún efecto real en la organización (Vallalta, 2019).

En esta fase deberemos ser capaces de:

- Establecer los objetivos de negocio.
- Evaluar la situación actual.
- Fijar los objetivos a nivel de minería de datos.
- Obtener un plan de proyecto.

4.1.2 Comprensión de los datos

Dos puntos clave en esta fase: conocer los datos, estructura y distribución, y la calidad de los mismos (Vallalta, 2019).

En esta fase deberemos ser capaces de:

- Ejecutar procesos de captura de datos.
- Proporcionar una descripción del juego de datos.
- Realizar tareas de exploración de datos.
- Gestionar la calidad de los datos, identificando problemas y proporcionando soluciones.

4.1.3 Preparación de los datos

El objetivo final de esta fase es obtener los datos finales sobre los que aplicarán los modelos (Vallalta, 2019).

En esta fase deberemos ser capaces de:

- Establecer el universo de datos con los que trabajar.
- Realizar tareas de limpieza de datos.
- Construir un juego de datos apto para ser usado en modelos de minería de datos.
- Integrar datos de fuentes heterogéneas si es necesario.

4.1.4 Modelado

El objetivo último de esta fase es construir un modelo que nos permita alcanzar los objetivos del proyecto (Vallalta, 2019).

En esta fase deberemos ser capaces de:

- Seleccionar las técnicas de modelado más adecuadas para nuestro juego de datos y nuestros objetivos.
- Fijar una estrategia de verificación de la calidad del modelo.
- Construir un modelo a partir de la aplicación de las técnicas seleccionadas sobre el juego de datos.
- Ajustar el modelo evaluando su fiabilidad y su impacto en los objetivos anteriormente establecidos.

4.1.5 Evaluación

En esta fase nos centrarnos en evaluar el grado de acercamiento del modelo a los objetivos de negocio (Vallalta, 2019).

En esta fase deberemos ser capaces de:

- Evaluar el modelo o modelos generados hasta el momento.
- Revisar todo el proceso de minería de datos que nos ha llevado hasta este punto.
- Establecer los siguientes pasos a tomar, tanto si se trata de repetir fases anteriores como si se trata de abrir nuevas líneas de investigación.

4.1.6 Despliegue

El objetivo último de esta fase es realizar el despliegue de los resultados obtenidos de forma que sea propagado a los usuarios finales, así como el mantenimiento del mismo una vez el despliegue haya finalizado (Vallalta, 2019).

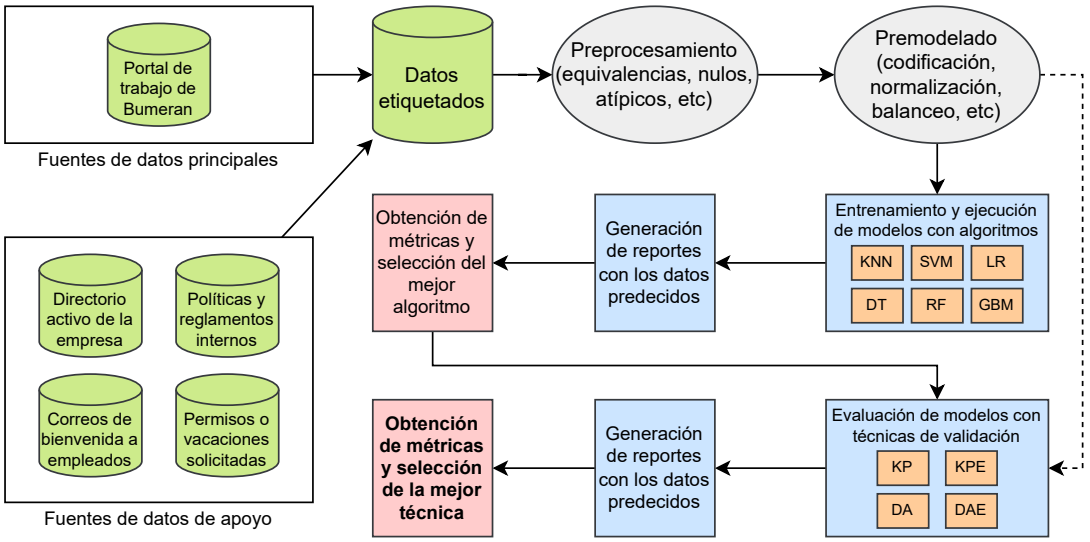
En esta fase deberemos ser capaces de:

- Diseñar un plan de despliegue de modelos y conocimiento sobre nuestra organización.
- Realizar seguimiento y mantenimiento de la parte más operativa del despliegue.
- Revisar el proyecto en su globalidad con el objetivo de identificar lecciones aprendidas.

4.2 APLICACIÓN DE LA METODOLOGÍA

Mediante la aplicación de la metodología CRISP-DM a las necesidades del proyecto, se definió un esquema de solución, presentado en la Figura 4.2:

Figura 4.2: Esquema de solución



Fuente: La empresa
Elaboración: Propia

4.2.1 Comprensión del negocio

Las actividades de comprensión del negocio que se contemplaron fueron:

- Identificación de necesidades
- Evaluación de la situación actual
- Establecer objetivos de minería de datos

4.2.1.1 Identificación de necesidades

Inicialmente, la necesidad fue identificada del Gerente de Operaciones de la empresa, el cual, al estar en contacto con la gerente de RRHH, se percató del problema que actualmente tienen para el proceso de selección de personal y esa información fue compartida hacia el investigador.

Posterior a ello, para poder ahondar en las necesidades, se realizaron sesiones continuas de análisis de requerimientos en conjunto con el Gerente de Operaciones, Rubén Parodi Guerrero. Estas sesiones se presentan en la Figura 4.3:

Figura 4.3: Reunión con Rubén Parodi Guerrero

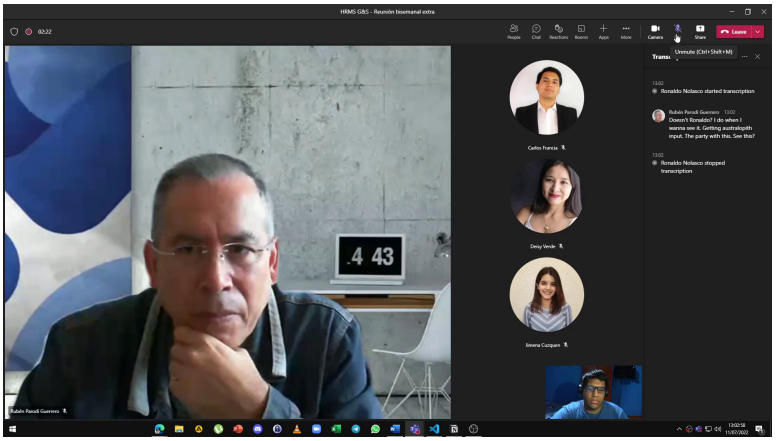
Fuente: La empresa

De igual manera, se tuvo como oportunidad realizar una sesión en conjunto con el equipo de RRHH de la empresa, conformado por:

- Deisy Verde (jefa de RRHH).
- Jimena Cuzquen (asistente de RRHH).
- Carlos Francia (asistente de RRHH).

Esta sesión es descrita en la Figura 4.4:

Figura 4.4: Reunión con equipo de RRHH



Fuente: La empresa

En base a ello, se pudo determinar que la principal necesidad está relacionada a poder elegir a un candidato que sea adecuado para la empresa, y pueda tener un buen rendimiento a corto y largo plazo.

4.2.1.2 Evaluación de la situación actual

Se solicitó información al área de RRHH acerca de cómo los procesos de selección se llevan actualmente.

Se identificó que todo el seguimiento de estas convocatorias se hace mediante Microsoft Excel, una herramienta que permite el manejo de hojas de cálculo para gestionar datos. Un ejemplo se encuentra en la Figura 4.5:

Figura 4.5: Seguimiento de las convocatorias

N° Solicitud	Línea de Negocio / Área	Cliente	Tipo	Fecha de solicitud	Mes	Perfil solicitado	Responsable	Fecha Requerida Según GL / Cliente	Impacto en el negocio
1	Analytics	Credicorp	Asignación fija	26-07-2021	Jul 21	Data Engineer	John Bautista		
1	Infra.Cloud	BCP	Asignación fija	10-08-2021	Aug 21	Consultor Cloud DevOps	José Cueva	02-11-2021	
73	MW.BS	BASC		17-08-2021	Aug 21	Desarrollador SharePoint 1	Walter Cristóbal	22-11-2021	Atraso en facturación
74	MW.BS	BASC		17-08-2021	Aug 21	Desarrollador SharePoint 2	Walter Cristóbal	22-11-2021	Atraso en facturación
24	Desarrollo	VUCE	Proyecto	07-09-2021	Sep 21	Lider Técnico	Percy Rojas	30-04-2022	
25	Desarrollo	VUCE	Proyecto	07-09-2021	Sep 21	Analista Programador Java Sr. 1	Percy Rojas	18-10-2021	
27	Desarrollo	VUCE	Proyecto	07-09-2021	Sep 21	Analista Programador Java Semi Sr. 3	Percy Rojas	03-01-2021	
29	Desarrollo	VUCE	Proyecto	07-09-2021	Sep 21	Analista Programador Java Jr. 1	Percy Rojas	15-01-2021	
30	PMO.Proyectos	VUCE	Proyecto	07-09-2021	Sep 21	Analista Programador Java Senior 4	Percy Rojas	15-02-2022	Atraso en facturación
31	PMO.Proyectos	MINCETUR - VUCE	Proyecto	07-09-2021	Sep 21	Analista Programador Java Jr.	Percy Rojas	06-06-2022	Atraso en facturación
32	PMO.Proyectos	VUCE	Proyecto	07-09-2021	Sep 21	Analista Programador Java Jr. 3	Percy Rojas	15-03-2022	Atraso en facturación
33	Desarrollo	VUCE	Proyecto	07-09-2021	Sep 21	Analista Programador Java Prac. 5	Percy Rojas	03-01-2021	
34	Desarrollo	VUCE	Proyecto	07-09-2021	Sep 21	Analista Programador Java Prac. 6	Percy Rojas	03-01-2021	
35	PMO.Proyectos	VUCE	Proyecto	07-09-2021	Sep 21	Analista Programador Java Jr. 4	Percy Rojas	15-03-2022	Atraso en facturación
36	PMO.Proyectos	MINCETUR - VUCE	Proyecto	07-09-2021	Sep 21	Analista de Calidad Senior	Percy Rojas	06-06-2022	Atraso en facturación
45	PMO.Proyectos	MINCETUR - VUCE	Proyecto	07-09-2021	Sep 21	Analista de Calidad Senior	Percy Rojas	15-06-2022	Atraso en facturación
37	Desarrollo	RIPLEY, VOLCAN		09-09-2021	Sep 21	Analista Programador .NET 1	Julio Lanza	15-02-2021	
38	Desarrollo	RIPLEY, VOLCAN		09-09-2021	Sep 21	Analista Programador .NET 2	Julio Lanza	15-02-2021	
39	Desarrollo	RIPLEY, VOLCAN		09-09-2021	Sep 21	Analista Programador .NET 3	Julio Lanza	03-01-2021	
40	Desarrollo	RIPLEY, VOLCAN		09-09-2021	Sep 21	Programador Frontend 1	Julio Lanza	03-01-2021	
41	Desarrollo	RIPLEY, VOLCAN		09-09-2021	Sep 21	Programador Frontend 2	Julio Lanza	01-12-2021	
42	Desarrollo	RIPLEY, VOLCAN		09-09-2021	Sep 21	Programador Frontend 3	Julio Lanza	01-12-2021	
43	Desarrollo	RIPLEY, VOLCAN		09-09-2021	Sep 21	Programador Frontend 4	Julio Lanza	15-11-2021	
44	Desarrollo	VUCE	Proyecto	09-09-2021	Sep 21	Analista Funcional 1	Percy Rojas	18-10-2021	
46	PMO.Proyectos	MINCETUR - VUCE	Proyecto	09-09-2021	Sep 21	Analista de Calidad	Percy Rojas	15-06-2022	Atraso en facturación
47	Desarrollo	VUCE	Proyecto	09-09-2021	Sep 21	Diseñador UX 2	Percy Rojas	03-01-2021	
52	PMO.Proyectos	VUCE	Proyecto	09-09-2021	Sep 21	Analista de Calidad Jr 1	Percy Rojas	15-03-2022	Atraso en facturación
83	MW.BS	LAP		09-09-2021	Sep 21	Desarrollador SharePoint 2	Javier Rojas	06-12-2021	
49	Desarrollo	VUCE	Proyecto	16-09-2021	Sep 21	Automatizador de procesos BPM 1	Percy Rojas		
50	Desarrollo	VUCE	Proyecto	16-09-2021	Sep 21	Automatizador de procesos BPM 2	Percy Rojas		
2	Analytics	Niubiz	Bolsa de horas	24-09-2021	Sep 21	Data Engineer	John Bautista	03-01-2021	
3	Analytics	BCP/Microsoft	Proyecto	24-09-2021	Sep 21	Data Engineer	John Bautista	15-12-2021	
4	Analytics	BCP/Microsoft	Proyecto	24-09-2021	Sep 21	Data Engineer	John Bautista	15-12-2021	
48	Desarrollo	VUCE	Proyecto	24-09-2021	Sep 21	Analista Funcional Semi Senior 3	Percy Rojas	03-01-2021	
53	Desarrollo	Credicorp	Asignación fija	28-09-2021	Sep 21	Especialista DevOps	Julio Lanza	06-12-2021	
55	Infra.Cloud	Credicorp	Asignación fija	28-09-2021	Sep 21	Cloud Specialist	José Cueva	18-10-2021	
97	Infra.Cloud	Credicorp	Asignación fija	28-09-2021	Sep 21	Especialista Oracle Cloud	Gabriel Quiroz	14-03-2022	
51	Desarrollo	VUCE	Proyecto	06-10-2021	Oct 21	Diseñador UX 1	Percy Rojas		

Fuente: La empresa

Pese a que sea una herramienta de fácil uso y se pueda utilizar para poder gestionar el proceso de selección, se pueden generar diversos errores en el registro de la información como errores de tipeo, insertar información de un candidato en la de otro, o no registrar correctamente el candidato que fue seleccionado en una convocatoria.

Es por ello que lo que se buscará es mantener la integridad de los datos, mientras se logra seleccionar al candidato más adecuado para el puesto.

4.2.1.3 Establecer objetivos de minería de datos

En el contexto del proyecto, se definieron objetivos de minería de datos, los cuales se usarán para corroborar que el modelo final cumpla con lo esperado.

Estos objetivos son:

- Predecir el candidato más adecuado para un puesto de trabajo en específico.
- Obtener el resultado de la predicción en el menor tiempo posible.

Dentro de los criterios de éxito de estos objetivos, se definieron de la siguiente manera:

- Obtener una exactitud del modelo de al menos 95%.
- Obtener una precisión del modelo de al menos 95%.
- Obtener una sensibilidad del modelo de al menos 95%.
- Obtener una robustez del modelo de al menos 95%.
- Obtener un tiempo de filtrado de candidatos de 5s como máximo.
- Obtener un tiempo de generación de reporte de 0.1s como máximo.

4.2.2 Comprensión de los datos

Las tareas de desarrollo propias de esta etapa se encuentran en un repositorio remoto de GitHub, el cual se encuentra en la referencia de Nolasco (2023).

Las actividades de comprensión de los datos que se contemplaron fueron:

- Lectura y procesamiento de los candidatos
- Lectura y procesamiento de los empleados
- Etiquetado de la variable objetivo
- Descripción general de los datos
- Diccionario de datos
- Visualización de datos

A continuación, se detallan cada una de estas actividades:

4.2.2.1 Lectura y procesamiento de los candidatos

Se revisaron las diversas fuentes de datos principales. Se determinó 1 fuente de datos principal, mostrada en la Figura 4.6:

1. Correos de candidatos a puestos publicados en Bumeran.

Figura 4.6: Bandeja de correo electrónico de ofertas laborales (Bumeran)

Fuente: La empresa

Estos correos son enviados automáticamente a la bandeja de entrada de RRHH, cuando un candidato realiza una postulación a una convocatoria.

De este correo, es posible obtener datos de los candidatos, como lo son:

- Datos de contacto (nombres, país de residencia, estado civil, medios de contacto, etc.)
- Objetivo laboral (objetivo laboral, sueldo pretendido)

- Experiencia laboral (Fecha de inicio y de fin, empresa, puesto desempeñado, detalle del puesto)
- Educación (Fecha de inicio y de fin, institución, nombre del estudio, estado del estudio, grado del estudio)
- Habilidades técnicas
- Idiomas
- Otros conocimientos

Toda la información de estos correos se leyó y se procesó en un único archivo. Finalmente, en la Figura 4.7 se tienen los datos principales integrados en un único archivo:

Figura 4.7: Datos de los candidatos

candidatePostulationDate	jobId	jobProfileName	candidateFullName	candidateResidenceCountry	candidateCivilStatus	salary	lastWorkCompany	lastWorkArea
30-09-2020 12:30	1114070449	Cloud Specialist - Ms Azure	Carlos Arbulu	Peru	Casado/A	6000	Tgestiona / Perupetro Sa	Infraestructura
30-09-2020 12:19	1114047602	Ejecutivo/A Comercial Ti	Alessio Uyema	Peru	Soltero/A	2500	Michael Page	Tecnologias De La Informacion
30-09-2020 23:39	1114087358	Analista Programador .Net	Joseth David Acosta Loayza	Peru	Soltero/A	4800	Canvia	Tecnologia / Sistemas
30-09-2020 22:40	1114087358	Analista Programador .Net	Jeanluigui Pierre Plasencia Arriaga	Peru	Soltero/A	3500	Xirect Software Solution	Programacion
30-09-2020 22:24	1114087358	Analista Programador .Net	Juilliand Rodolfo Damian Gomez	Peru	Soltero/A	2800	Aselera.S.A.C	Sistemas
30-09-2020 22:03	1114087358	Analista Programador .Net	Juan Carlos Sotelo Canales	Peru	Soltero/A	3000	Domiruth Travel Service Sac	Programacion
30-09-2020 21:48	1114087358	Analista Programador .Net	Edgar Alonso Camarena Nomberto	Peru	Soltero/A	3500	Seidor	Programacion
30-09-2020 21:30	1114087358	Analista Programador .Net	Heber Gonzalo Cachi Cruzado	Peru	Soltero/A	2300	Zegel Ipae	Programacion
30-09-2020 21:10	1114087358	Analista Programador .Net	Diego Gutierrez	Peru	Soltero/A	4000	Ardiles Import	Tecnologias De La Informacion
30-09-2020 20:14	1114023772	Ejecutivo(A) Comercial Ti	Zinthia Mansilla Bedoya	Peru	Casado/A	4500	Safetypay	Comercial
30-09-2020 3:09	1114084953	Especialista En Analitica	Carlos Dominguez	Peru	Casado/A	5000	Claro - Grupo America Movil	Gerencia / Direccion General
28-09-2022 11:53	1115406770	Programador .Net	Eduardo Passano	Peru	Soltero/A	6000	Minera Aurifera Retamas(Marsa)	Tecnologias De La Informacion
28-09-2022 9:08	1115407679	Ejecutivo Comercial Jr	Cesar Alexander Diaz Carbajal	Peru	Soltero/A	2000	Editorial Santillana (Santillana S.A.)	Comercial
28-09-2022 18:52	1115352072	Lider Tecnico	David Bryan Arias Castro	Peru	Soltero/A	6000	Win Net	Liderazgo De Proyecto
28-09-2022 16:45	1115402634	Analista De Calidad	Victor Garcia	Peru	Soltero/A	1000	Aerolinea American Airles	Comunicaciones Internas
25-09-2019 0:40	1113534187	Sql Server Consultant	Juan Kenny Munaque	Peru	Soltero/A	1500	Wind Telecomunicazioni	Programacion
25-09-2019 10:50	1113544623	Lider Tecnico	Carlos Alberto Casas Carbajal	Peru	Casado/A	5000	Sigma (Braedt, Otto Kunz Y Segoviana)	Produccion
25-09-2019 19:31	1113531950	Scrum Master	Robert Alonso Pareja Quispe	Peru	Soltero/A	7500	Everis Peru	Tecnologia / Sistemas
23-09-2020 10:57	1114047602	Ejecutivo/A Comercial Ti	Carlos Guillermo Flores Fernandez	Peru	Soltero/A	1600	Celeritech	Marketing
23-09-2020 21:45	1114070449	Cloud Specialist - Ms Azure	Luis Manuel Elizalde Villanueva	Peru	Soltero/A	1	Freelance	Sistemas
23-09-2020 4:59	1114070449	Cloud Specialist - Ms Azure	Jorge Raul Gallardo Ibarra	Peru	Soltero/A	5500	Tivit	Tecnologias De La Informacion
23-09-2020 14:52	1114069475	Analista De Ciberseguridad	Ximena Josely Mercedes Rabanal Lucho	Peru	Soltero/A	3000	Fractalia Peru S.A.	Telecomunicaciones
23-09-2020 13:57	1114047548	Jefe De Proyecto - Analytics & Ai	Anthony Alarcon	Peru	Soltero/A	10000	Ferreyros Sa	Exploracion Minera Y Petroquimica
21-09-2022 12:26	1115406770	Programador .Net	Ivan Cesar Ascencio Tiza	Peru	Soltero/A	4000	Programador Web Freelance	Programacion
21-09-2022 12:26	1115406770	Programador .Net	Ivan Cesar Ascencio Tiza	Peru	Soltero/A	4000	Programador Web Freelance	Programacion
21-09-2022 11:22	1115338594	Data Architect	Franco Octavio Gordillo Calagua	Peru	Soltero/A	1100	Seclim	Servicios
21-09-2022 11:19	1115406770	Programador .Net	Raul Emerson Barreto Hidalgo	Peru	Soltero/A	5000	Enotria Sac	Programacion
21-09-2022 23:08	1115407679	Ejecutivo Comercial Jr	Katherin Sandoval Cabello	Peru	Soltero/A	1600	Universidad Del Pacifico	Direccion
21-09-2022 10:41	1115352072	Lider Tecnico	Eduardo Gustavo Goche Penaloza	Peru	Soltero/A	10000	It Convergence	Tecnologias De La Informacion
21-09-2022 20:16	1115407679	Ejecutivo Comercial Jr	Manuel Antonio Santisteban Coronel	Peru	Soltero/A	1800	Jj Kalidad E.I.R.L.	Desarrollo De Negocios
21-09-2022 7:56	1115402634	Analista De Calidad	Jorge Diaz Bohorquez	Peru	Soltero/A	5000	Dayr Inversiones Multiples S.A.C.	Calidad
21-09-2022 18:58	1115407679	Ejecutivo Comercial Jr	Gianfranco Gonzales	Peru	Soltero/A	1200	Bungalows Mancora	Administracion
21-09-2022 18:15	1115407679	Ejecutivo Comercial Jr	Adriana Giorgina Arce Garcia	Peru	Soltero/A	0	Teleperformance	Call Center
21-09-2022 17:55	1115407679	Ejecutivo Comercial Jr	Marco Navarro Bustamante	Peru	Soltero/A	2500	Froizen	Desarrollo De Negocios
21-09-2022 17:44	1115407679	Ejecutivo Comercial Jr	Juan Carlos Seminario Atoche	Peru	Casado/A	5000	Tecnogas Sa (Praxair Peru)	Ventas
21-09-2022 16:56	1115407679	Ejecutivo Comercial Jr	Edinson Manuel Pena Ungaro	Peru	Casado/A	2500	Mapfre	Comercial

Fuente: La empresa
Elaboración: Propia

4.2.2.2 Lectura y procesamiento de los empleados

Se revisaron las diversas fuentes de datos de apoyo. Se determinaron 4 fuentes de datos>

1. Directorio activo de la empresa, descritas en la Figura 4.8 y Figura 4.9:

Figura 4.8: Lista de empleados en el directorio activo de la empresa

userPrincipalName	displayName	givenName	surname
Cazzia.rezza@gyscsp.onmicrosoft.com	Cazzia Rezza	Cazzia	Rezza
fpf_appmovil@gyscsp.onmicrosoft.com	FPF App Móvil	FPF	App Móvil
carlos.marcos@gyscsp.onmicrosoft.com	Carlos Marcos	Carlos	Marcos
jeremy.tornero@gyscsp.onmicrosoft.com	Jeremy Tornero	Jeremy	Tornero
proyecto.freebox02@gyscsp.onmicrosoft.com	Proyecto	FreeBox	02
credicorp.automatizaciones@gyscsp.onmicrosoft.com	Credicorp Automatizaciones	Credicorp	Automatizaciones
javier.gonzales_gestionysistemas.com#EXT#@gyscsp.onmicrosoft.com	Javier Gonzales	Javier	Gonzales
correos_backup@gyscsp.onmicrosoft.com	Backup PST	Backup	PST
Samir.ochochoque@gyscsp.onmicrosoft.com	Samir Ochochoque	Samir	Ochochoque
Luisa.Gavilan@gyscsp.onmicrosoft.com	Luisa Gavilan	Luisa	Gavilan
julio.alvarado@gyscsp.onmicrosoft.com	Julio Alvarado	Julio	Alvarado
osman.silvera@gyscsp.onmicrosoft.com	Osman Silvera	Osman	Silvera
andre.zegarra@gyscsp.onmicrosoft.com	Andre Zegarra	Andre	Zegarra
marco.limache@gyscsp.onmicrosoft.com	Marco Limache	Marco	Limache
testuser01@gyscsp.onmicrosoft.com	testuser01	test	user01
DemoGAF@gyscsp.onmicrosoft.com	Demo GAF	Demo GAF	
fiorella.galdos_gestionysistemas.com#EXT#@gyscsp.onmicrosoft.com	Fiorella Galdos	Fiorella	Galdos
Enrique.quiroz@gyscsp.onmicrosoft.com	Enrique Quiroz	Enrique	Quiroz
Julio.arguedas@gyscsp.onmicrosoft.com	Julio Arguedas	Julio	Arguedas
fiorella.galdos@gyscsp.onmicrosoft.com	Fiorella Galdos	Fiorella	Galdos
Raul.campos@gyscsp.onmicrosoft.com	Raul Campos	Raul	Campos
Renato.arrascue@gyscsp.onmicrosoft.com	Renato Arrascue	Renato	Arrascue
miguel.isa_gestionysistemas.com#EXT#@gyscsp.onmicrosoft.com	Miguel Isa	Miguel	Isa
credicorp.capital@gyscsp.onmicrosoft.com	Credicorp Capital	Credicorp	Capital
Dennis.Arteaga@gyscsp.onmicrosoft.com	Dennis Arteaga	Dennis	Arteaga
manuel.varillas_gestionysistemas.com#EXT#@gyscsp.onmicrosoft.com	Manuel Varillas	Manuel	Varillas
poc-banbif@gyscsp.onmicrosoft.com	Poc Banbif	Poc	Banbif
paraiso.demoppp2@gyscsp.onmicrosoft.com	Paraiso Demoppp2	Paraiso	Demoppp2
scom_azure@gyscsp.onmicrosoft.com	scom_azure	scom_azure	Pruebas_bs
Joan.ancajima@gyscsp.onmicrosoft.com	Joan Ancajima	Joan	Ancajima
marco.mesias@gyscsp.onmicrosoft.com	Marco Mesias	Marco	Mesias
jose.cueva@gyscsp.onmicrosoft.com	Jose Cueva	Jose	Cueva
Oscar.gensollen@gyscsp.onmicrosoft.com	Oscar Gensollen	Oscar	Gensollen
luis.pena@gyscsp.onmicrosoft.com	Luis Peña	Luis	Peña
Percy.rojas@gyscsp.onmicrosoft.com	Percy Rojas	Percy	Rojas

Fuente: La empresa

Figura 4.9: Lista de empleados en el directorio activo en Microsoft Azure

Display name	User principal name	User type	On-premises sy...	Identities	Company name	Creation type
Aaron Mejia	aaron.mejia_gestionysiste...	Guest	No	ExternalAzureAD		Invitation
Abel Quispe	aquispe_css.pe#EXT#@gy...	Guest	No	ExternalAzureAD		Invitation
abenites	abenites_luzdelsur.com.p...	Guest	No	gyscsp.onmicrosoft.com		Invitation
Acsel Alvarez	Acsel.alvarez_gestionysist...	Guest	No	ExternalAzureAD		Invitation
AD Connect	servidor.adconnect@gysc...	Member	No	gyscsp.onmicrosoft.com		
Admin BC	adminbc_theleafcompany...	Guest	No	ExternalAzureAD		Invitation
Admin Digesa	admin.digesa@gyscsp.on...	Member	No	gyscsp.onmicrosoft.com		
Admin G&S	AdminCSP@gyscsp.onmic...	Member	No	gyscsp.onmicrosoft.com		
Admin Support G&S	adminsupport@gyscsp.o...	Member	No	gyscsp.onmicrosoft.com		
Adrian Marquina	adrian.marquina_gestiony...	Guest	No	ExternalAzureAD		Invitation
AFP Habitat - Web Consulta	afphabitat_webconsulta@...	Member	No	gyscsp.onmicrosoft.com		
Aldo Carhuamaca	aldo.carhuamaca_gestion...	Guest	No	ExternalAzureAD		Invitation
Aldo Carhuamaca	Aldo.Carhuamaca@gyscs...	Member	No	gyscsp.onmicrosoft.com		
alecabrejos123	alecabrejos123_gmail.co...	Guest	No	gyscsp.onmicrosoft.com		Invitation
Alessandro Callirgos	alessandro.callirgos_gesti...	Guest	No	gyscsp.onmicrosoft.com		Invitation
Alexander Paisig	alexander.paisig_gestiony...	Guest	No	ExternalAzureAD		Invitation
Alexandra Ccolqque	alexandra.ccolqque_gesti...	Guest	No	ExternalAzureAD		Invitation
alexis.cayo	alexis.cayo_repsol.com#E...	Guest	No	gyscsp.onmicrosoft.com		Invitation
Alfred Aylas	Alfred.aylas_gestionysiste...	Guest	No	ExternalAzureAD		Invitation
AlfredAylas	alfred.aylas@gyscsp.onmi...	Member	No	gyscsp.onmicrosoft.com		
Alfredo Benaute	alfredo.benaute_dasser.co...	Guest	No	MicrosoftAccount		Invitation

Fuente: La empresa

2. Políticas y reglamentos internos, descritas en la Figura 4.10 y Figura 4.11:

Figura 4.10: Lista de empleados en reglamentos y políticas internas

N°	NOMBRE
21	Echegaray Medina Marily Lizeth
22	Eneque Pisfil Juan
23	Escobar Estrada Cesar Alberto
24	Espinoza Alarcon Shyla
25	Espinoza Pereda Juan Carlos
26	Galvez Mayo Sebastian Alberto Saul
27	Gamarra López Jannette Johana
28	Gómez García Richard Robinson
29	Gonzales Sandoval Javier Guillermo
30	Gonzales Sandoval José Federico
31	Guere Contreras Paulo Cesar
32	Guillén Huarcaya Diego Joaquín
33	Huachopoma Cruz Arthur Tonny
34	Huaman Ramos Moises
35	Huertas Zanabria Raul Gianmarco
36	Jara Carpio Carlos Francisco
37	La Rosa Zavalla Luis Miguel
38	Lanza Romero Julio Cesar
39	Lara Rabanal Shirley Isabella
40	León Orué Marco Orlando
41	López Figueroa Stephanie Almendra
42	Luna Vigo Miguel Angel
43	Mazzoni De Oliveira Bruno
44	Medina Manay Alberto Antonio
45	Mejía Alfaro Aarón Yamír
46	Mendez Soto Alexandra Dayana
47	Mendoza Yucra Solange Alexandra
48	Mescco Herrera Percy Mariano
49	Minaya Vidal Sussan Paola
50	Mosquera Villalobos Fiorella Del Carmen
51	Nevado Talledo Jose Eduardo
52	Niño Pazos Jason Joel
53	Ochochoque Chambi Samir Jann
54	Olortegui Abanto Segundo Manuel
55	Orellana Benites Aurelio Johnny
56	Ormeño Vera Wilber Javier
57	Parodi Guerrero Ruben
58	Perales Castillo Jhon Antonio
59	Perez Quispe Scott Edinson
60	Portal Zapata Priscilla Del Pilar

N°	NOMBRE
61	Pretell Kishimoto Joaquin Alonso
62	Quiroz Villamonte Gabriel Eduardo
63	Ramirez Laimito Jose Daniel
64	Ramirez Limo Fernando Manuel
65	Remigio Romàn Milagros Andrea
66	Rezza Lauz Cazzia Ivonne
67	Rivera Livaque Jhon Francis Jr.
68	Rocha Angeles Evelyn Grace
69	Rodriguez Alomías Jose Luis
70	Rojas Chavez Javier Fernando
71	Rojas Cuadros Víctor Jean Franz
72	Rojas Espinoza Anibal Gonzalo
73	Rojas Osnayo Jery William
74	Rojas Paredes Percy Luis
75	Rojas Trujillo Rosario
76	Rosales Siche Héctor Junior
77	Saavedra Canessa Segundo Nicolas
78	Sanchez Moran, Zarela Esther
79	Salvador Calzada Carlos Eduardo
80	Segura Vilca Gilberto Franquito
81	Silva Ortiz Brando Antonio
82	Silvera Herrera Osman Emilio
83	Silverio Rivera Vladimir Marcos
84	Soto Gutiérrez Sthephany Jemina
85	Suarez Huamán Marlon Jheral
86	Takia Rios Ana Claudia
87	Tineo Quispe Christian Victor
88	Tito Cruzado Fabio Andres
89	Toratto Olortegui Jully
90	Tornero Landeo Jeremy Walter
91	Torres Solórzano Gean Carlos
92	Valdez Landa Max Enrique
93	Valer Oroncoy Jhon Henry
94	Vargas Arredondo Fernando Andres
95	Vasquez Gratelli Carlos
96	Verde Jara Deisy Yohana
97	Villalobos Pulache Omar Jose
98	Villanueva Nicho Christian Leonardo
99	Wong Sato José Luis
100	Zamora Noreña Luis Fernando
101	Zavaleta Roldan Edward Luis
102	Zegarra Ore Emmanuel Andre

Fuente: La empresa

Figura 4.11: Lista de empleados capacitados para el teletrabajo

Este cuestionario de autoevaluación se considera preventivamente aceptable, únicamente, cuando todas las respuestas sean "sí".

Este cuestionario debes enviarlo firmado y escaneado al área de recursos humanos (rr.hh@gestionysistemas.com) ya que será la declaración responsable a la que hace referencia al art 26.4 del DS 002-2023-TR para la gestión del teletrabajo.

Conserva este documento, que servirá como guía de información ya que recoge las recomendaciones más importantes a tener en cuenta para un puesto de teletrabajo seguro.

FORMULARIO DE AUTOEVALUACIÓN DE RIESGOS PARA LA SEGURIDAD Y SALUD EN EL TELETRABAJO				
Nombre del/la trabajador/a: Ronaldo Farid Nolasco Chavez				
Lugar donde se aplica la autoevaluación (dirección): Urb. Husares de Junin Mz F Lote 26, S.M.P.				
Puesto de trabajo: Analista programador			Empresa: Gestion y Sistemas SA.C	
Presenta alguna discapacidad (señale de qué tipo): No			Fecha: 19/05/2023	
N°	ENUNCIADO	SI	NO	OBSERVACIONES
1	Dispone de espacio suficiente en su lugar para el desarrollo del teletrabajo, para levantarse y sentarse sin dificultad.	X		

Fuente: La empresa

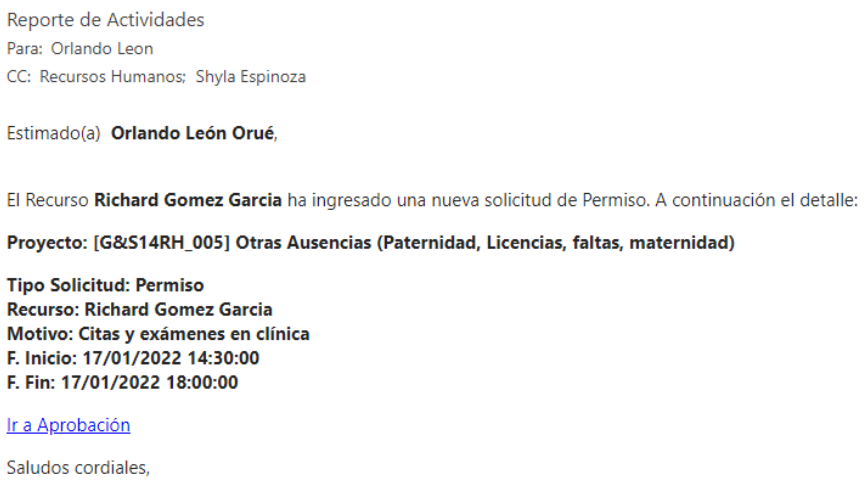
3. Correos de bienvenida a empleados, descritas en la Figura 4.12:

Figura 4.12: Correos de bienvenida a empleados en su ingreso a la empresa

Fuente: La empresa

4. Permisos o vacaciones solicitadas, descritas en la Figura 4.13, Figura 4.14 y Figura 4.15:

Figura 4.13: Solicitud de permiso



Reporte de Actividades
Para: Orlando Leon
CC: Recursos Humanos; Shyla Espinoza

Estimado(a) **Orlando León Orué**,

El Recurso **Richard Gomez Garcia** ha ingresado una nueva solicitud de Permiso. A continuación el detalle:

Proyecto: [G&S14RH_005] Otras Ausencias (Paternidad, Licencias, faltas, maternidad)

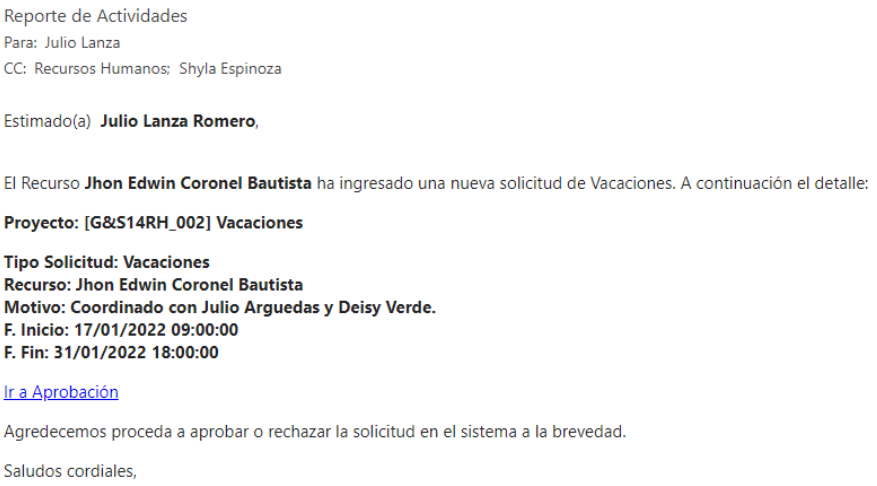
Tipo Solicitud: Permiso
Recurso: Richard Gomez Garcia
Motivo: Citas y exámenes en clínica
F. Inicio: 17/01/2022 14:30:00
F. Fin: 17/01/2022 18:00:00

Ir a Aprobación

Saludos cordiales,

Fuente: La empresa

Figura 4.14: Solicitud de vacaciones



Reporte de Actividades
Para: Julio Lanza
CC: Recursos Humanos; Shyla Espinoza

Estimado(a) **Julio Lanza Romero**,

El Recurso **Jhon Edwin Coronel Bautista** ha ingresado una nueva solicitud de Vacaciones. A continuación el detalle:

Proyecto: [G&S14RH_002] Vacaciones

Tipo Solicitud: Vacaciones
Recurso: Jhon Edwin Coronel Bautista
Motivo: Coordinado con Julio Arguedas y Deisy Verde.
F. Inicio: 17/01/2022 09:00:00
F. Fin: 31/01/2022 18:00:00

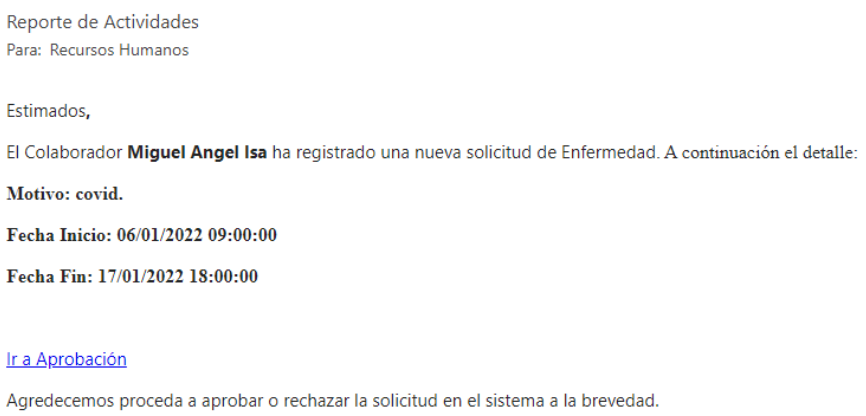
Ir a Aprobación

Agredecemos proceda a aprobar o rechazar la solicitud en el sistema a la brevedad.

Saludos cordiales,

Fuente: La empresa

Figura 4.15: Solicitud de enfermedad



Reporte de Actividades
Para: Recursos Humanos

Estimados,

El Colaborador **Miguel Angel Isa** ha registrado una nueva solicitud de Enfermedad. A continuación el detalle:

Motivo: covid.

Fecha Inicio: 06/01/2022 09:00:00

Fecha Fin: 17/01/2022 18:00:00

Ir a Aprobación

Agredecemos proceda a aprobar o rechazar la solicitud en el sistema a la brevedad.

Fuente: La empresa

Cada una de estas fuentes se leyó y procesó, sin embargo, no se unieron en un único archivo, esto debido a las diferencias de estructura de las fuentes de datos.

4.2.2.3 Etiquetado de la variable objetivo

Debido a que en los datos de los candidatos no se tenía un atributo que diferenciara a que candidatos se contrataron y a cuáles no, se utilizaron los datos de los empleados.

Se analizo a cada candidato y, si este se encontraba en al menos una de las 4 fuentes de datos de apoyo, se etiquetaba como contratado, caso contrario se etiquetaba como no contratado.

Finalmente, se formó el conjunto de datos final, expuesto en la Figura 4.16:

Figura 4.16: Conjunto de datos final

fechaPostulacion	idConvocatoria	nombrePerfilConvocatoria	empresaUltimoTrabajo	areaUltimoTrabajo	institucionUltimoEstudio	areaUltimoEstudio	contratado
30-09-2020 12:30	1114070449	Cloud Specialist - Ms Azure	Tgestiona / Perupetro Sa	Infraestructura	Universidad Nacional Pedro Ruiz Gallo - Lambayeque	Computacion / Informatica	0
30-09-2020 12:19	1114047602	Ejecutivo/A Comercial Ti	Michael Page	Tecnologias De La Informacion	Universidad De Lima	Ing. Industrial	0
30-09-2020 23:39	1114087358	Analista Programador .Net	Canvia	Tecnologia / Sistemas	Icpna	Traduccion	0
30-09-2020 22:40	1114087358	Analista Programador .Net	Xirect Software Solution	Programacion	Instituto Tecnologico Del Norte	Computacion / Informatica	0
30-09-2020 22:24	1114087358	Analista Programador .Net	Aselera.S.A.C	Sistemas	Usil	Ing. En Sistemas	0
30-09-2020 22:03	1114087358	Analista Programador .Net	Domiruth Travel Service Sac	Programacion	Sistemasuni	Tecnologias De La Informacion	0
30-09-2020 21:48	1114087358	Analista Programador .Net	Seidor	Programacion	Univeridad Autonoma Del Peru	Ing. En Sistemas	0
30-09-2020 21:30	1114087358	Analista Programador .Net	Zegel Ipae	Programacion	Universidad Nacional De Cajamarca	Ing. En Sistemas	0
30-09-2020 21:10	1114087358	Analista Programador .Net	Ardiles Import	Tecnologias De La Informacion	Universidad Cientifica Del Sur	Ing. En Sistemas	0
30-09-2020 20:14	1114023772	Ejecutivo(A) Comercial Ti	Safetypay	Comercial	Universidad San Ignacio De Loyola	Marketing / Comercializacion	0
30-09-2020 3:09	1114084953	Especialista En Analitica	Claro - Grupo America Movil	Gerencia / Direccion General	Universidad Inca Garcilaso De La Vega	Ing. En Sistemas	1
28-09-2022 11:53	1115406770	Programador .Net	Minera Aurifera Retamas(Marsa)	Tecnologias De La Informacion	Universidad Privada De Cienascias Aplicadas	Tecnologias De La Informacion	0
28-09-2022 9:08	1115407679	Ejecutivo Comercial Jr	Editorial Santillana (Santillana S.A.)	Comercial	Universidad Ricardo Palma	Adm. De Empresas	0
28-09-2022 18:52	1115352072	Lider Tecnico	Win Net	Liderazgo De Proyecto	Utel	Ing. En Sistemas	0
28-09-2022 16:45	1115402634	Analista De Calidad	Aerolinea American Airles	Comunicaciones Internas	Instituto De Formacion Bancaria Ifb		0
25-09-2019 0:40	1113534187	Sql Server Consultant	Wind Telecomunicazioni	Programacion	Pontificia Universidad Catolica Del Peru	Ing. Informatica	0
25-09-2019 10:50	1113544623	Lider Tecnico	Sigma (Braedt, Otto Kunz Y Segoviana)	Produccion	Capacity Academy Mx	Tecnologias De La Informacion	0
25-09-2019 19:31	1113531950	Scrum Master	Everis Peru	Tecnologia / Sistemas	Universidad Peruana De Ciencias Aplicadas	Ing. En Sistemas	0
23-09-2020 10:57	1114047602	Ejecutivo/A Comercial Ti	Celeritech	Marketing	Universidad De Lima	Adm. De Empresas	0
23-09-2020 21:45	1114070449	Cloud Specialist - Ms Azure	Freelance	Sistemas	Universidad San Ignacio De Loyola	Tecnologias De La Informacion	0
23-09-2020 4:59	1114070449	Cloud Specialist - Ms Azure	Tivit	Tecnologias De La Informacion	Rioja	Ing. Informatica	0
23-09-2020 14:52	1114069475	Analista De Ciberseguridad	Fractalia Peru S.A.	Telecomunicaciones	Cisco	Telecomunicaciones	0
23-09-2020 13:57	1114047548	Jefe De Proyecto - Analytics & Ai	Ferreyros Sa	Exploracion Minera Y Petroquimica	Ccance	Tecnologias De La Informacion	0
21-09-2022 12:26	1115406770	Programador .Net	Programador Web Freelance	Programacion	Universidad De Barcelona / Online	Finanzas	0
21-09-2022 12:26	1115406770	Programador .Net	Programador Web Freelance	Programacion	Universidad De Barcelona / Online	Finanzas	0
21-09-2022 11:22	1115338594	Data Architect	Seclim	Servicios	Universidad Nacional De Moquegua	Computacion / Informatica	0
21-09-2022 11:19	1115406770	Programador .Net	Enotria Sac	Programacion	Instituto Tecnologico Del Norte	Tecnologias De La Informacion	0
21-09-2022 23:08	1115407679	Ejecutivo Comercial Jr	Universidad Del Pacifico	Direccion	Universidad San Ignacio Del Loyola	Adm. De Empresas	0
21-09-2022 10:41	1115352072	Lider Tecnico	It Convergence	Tecnologias De La Informacion	Universidad Rey Juan Carlos	Adm. De Empresas	0
21-09-2022 20:16	1115407679	Ejecutivo Comercial Jr	Jj Kalidad E.I.R.L.	Desarrollo De Negocios	Universidad Privada Del Norte	Ing. En Sistemas	0
21-09-2022 7:56	1115402634	Analista De Calidad	Dayr Inversiones Multiples S.A.C.	Calidad	Tecsup	Transporte	0
21-09-2022 18:58	1115407679	Ejecutivo Comercial Jr	Bungalows Mancora	Administracion	Upc	Adm. De Empresas	1
21-09-2022 18:15	1115407679	Ejecutivo Comercial Jr	Teleperformance	Call Center	Universidad Nacional Mayor De San Marcos	Contabilidad / Auditoria	0
21-09-2022 17:55	1115407679	Ejecutivo Comercial Jr	Froizen	Desarrollo De Negocios	Camara De Comercio Del Peru	Adm. Y Gestion Publica	0
21-09-2022 17:44	1115407679	Ejecutivo Comercial Jr	Tecnogas Sa (Praxair Peru)	Ventas	Universidad Inca Garcilaso De La Vega	Economia	0
21-09-2022 16:56	1115407679	Ejecutivo Comercial Jr	Mapfre	Comercial	Universidad Catolica Sedes Sapientiae	Adm. De Empresas	0

Fuente: La empresa
Elaboración: Propia

4.2.2.4 Descripción general de los datos

En base a estos datos recopilados, se pueden detallar algunas características sobre ellos, los cuales se presentan en la Tabla 4.1:

Tabla 4.1: Características de los datos - Comprensión de datos

Número de filas	10597
Fila con fecha más antigua	15-08-2017 12:52
Fila con fecha más reciente	22-05-2023 19:31
Número de columnas	36
Número de columnas con valores categóricos	19
Número de columnas con valores numéricos	17
Número de columnas con al menos un valor nulo	21

Fuente: La empresa
Elaboración: Propia

4.2.2.5 Diccionario de datos

También se presenta el diccionario de los datos recopilados, en la Figura 4.2 y Figura 4.3:

Tabla 4.2: Diccionario de datos (primera parte) - Comprensión de datos

ID	Campo	Origen	Tipo	Descripción	Valores posibles
1	fechaPostulacion	Candidato	Numérico	Fecha y hora de postulación del candidato	Formato DD-MM-YYYY HH:MM, fecha válida posterior al año 1900
2	idConvocatoria	Convocatoria	Numérico	Código de la convocatoria de trabajo, autogenerado por el portal de empleo	Número entero de 10 dígitos
3	nombrePerfilConvocatoria	Convocatoria	Categórico	Nombre del perfil del puesto solicitado	Texto
4	nombreCompleto	Candidato	Categórico	Nombre completo del candidato	Texto
5	paisResidencia	Candidato	Categórico	País de residencia del candidato	195 países diferentes
6	estadoCivil	Candidato	Categórico	Estado civil del candidato	Casado/A, Divorciado/A, Pareja De Hecho, Soltero/A, Unión Libre, Viudo/A
7	numeroDocumento	Candidato	Numérico	Número de documento de identidad del candidato	Número entero, de 8 a 12 dígitos
8	fechaNacimiento	Candidato	Numérico	Fecha de nacimiento del candidato	Formato DD-MM-YYYY, fecha válida posterior al año 1900
9	paisNacimiento	Candidato	Categórico	País de nacimiento del candidato	195 países diferentes
10	direccion	Candidato	Categórico	Dirección del domicilio del candidato	Texto
11	numeroCasa	Candidato	Numérico	Número de teléfono de casa del candidato	Número entero, de 9 dígitos
12	numeroCelular	Candidato	Numérico	Número de teléfono celular del candidato	Número entero, de 9 dígitos
13	correoElectronico	Candidato	Categórico	Correo electrónico del candidato	Texto
14	objetivoLaboral	Candidato	Categórico	Objetivo laboral del candidato	Texto
15	sueldoPretendido	Candidato	Numérico	Sueldo pretendido, solicitado por el candidato en la convocatoria	Número decimal mayor a 0
16	diasUltimoTrabajo	Candidato	Numérico	Número de días de permanencia del último trabajo del candidato	Número entero, mayor o igual a 0
17	empresaUltimoTrabajo	Candidato	Categórico	Nombre de la empresa del último trabajo del candidato	Texto
18	paisUltimoTrabajo	Candidato	Categórico	País del último trabajo del candidato	195 países diferentes

Fuente: La empresa
Elaboración: Propia

Tabla 4.3: Diccionario de datos (segunda parte) - Comprensión de datos

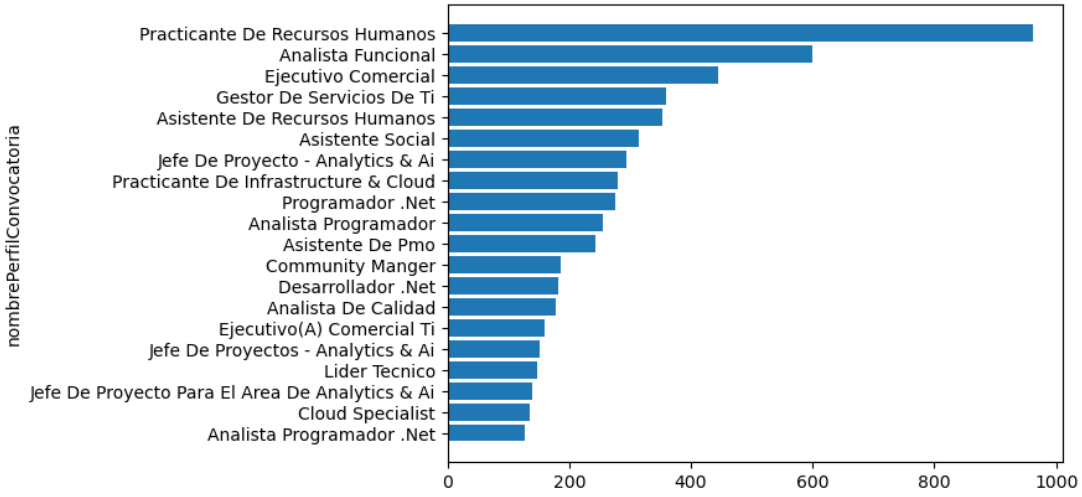
ID	Campo	Origen	Tipo	Descripción	Valores posibles
19	areaUltimoTrabajo	Candidato	Categórico	Área del último trabajo del candidato	Texto
20	nombreUltimoTrabajo	Candidato	Categórico	Nombre del último trabajo del candidato	Texto
21	descripcionUltimoTrabajo	Candidato	Categórico	Descripción del último trabajo del candidato	Texto
22	aniosExperiencia	Candidato	Numérico	Años de experiencia del candidato	Número decimal mayor o igual a 0
23	numeroTrabajos	Candidato	Numérico	Número de trabajos del candidato	Número entero, mayor o igual a 0
24	diasUltimoEstudio	Candidato	Numérico	Número de días de permanencia del último estudio del candidato	Número entero, mayor o igual a 0
25	institucionUltimoEstudio	Candidato	Categórico	Institución del último estudio del candidato	Texto
26	paisUltimoEstudio	Candidato	Categórico	País del último estudio del candidato	195 países diferentes
27	areaUltimoEstudio	Candidato	Categórico	Área del último estudio del candidato	Texto
28	nombreUltimoEstudio	Candidato	Categórico	Nombre del último estudio del candidato	Texto
29	estadoUltimoEstudio	Candidato	Categórico	Estado del último estudio del candidato	Abandonado, En Curso, Graduado
30	gradoUltimoEstudio	Candidato	Categórico	Grado alcanzado del último estudio del candidato	Otro, Secundario, Terciario/Técnico, Universitario, Posgrado, Master, Doctorado
31	aniosEstudio	Candidato	Numérico	Años de estudio del candidato	Número entero, mayor o igual a 0
32	numeroEstudios	Candidato	Numérico	Número de estudios del candidato	Número entero, mayor o igual a 0
33	habilidadesTecnicas	Candidato	Numérico	Número de habilidades técnicas del candidato	Número entero, mayor o igual a 0
34	idiomas	Candidato	Numérico	Número de idiomas del candidato	Número entero, mayor o igual a 0
35	otrasHabilidades	Candidato	Numérico	Número de otras habilidades (blandas, informáticas) del candidato	Número entero, mayor o igual a 0
36	contratado	Convocatoria	Numérico	El candidato fue rechazado o fue contratado	0 o 1

Fuente: La empresa
Elaboración: Propia

4.2.2.6 Visualización de datos

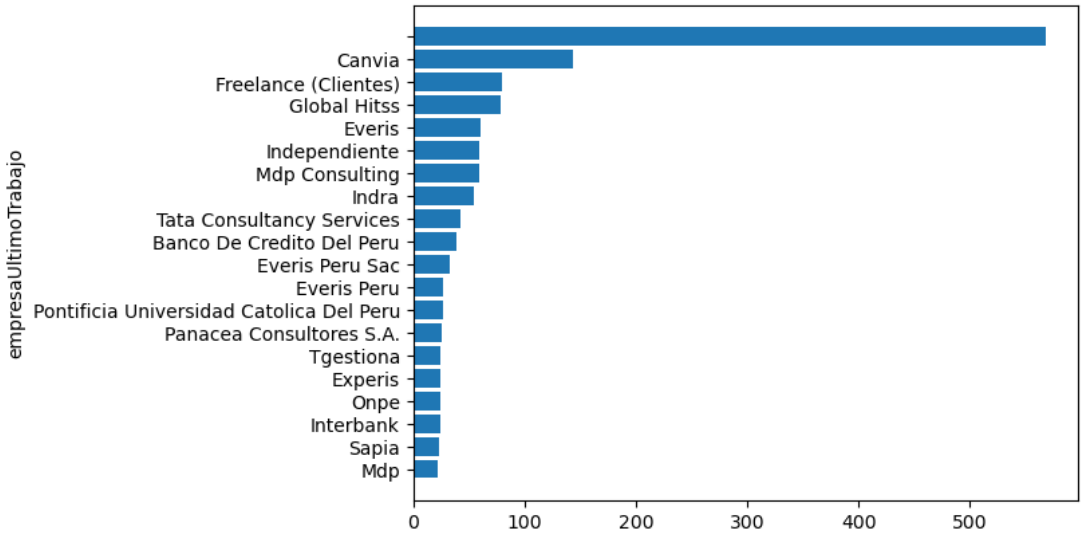
A continuación, se muestran algunas gráficas descriptivas de los datos recopilados, en la Figura 4.17, Figura 4.18, Figura 4.19, Figura 4.20, Figura 4.21, Figura 4.22, Figura 4.23, Figura 4.24, Figura 4.25 y Figura 4.26:

Figura 4.17: Nombre del perfil de la convocatoria



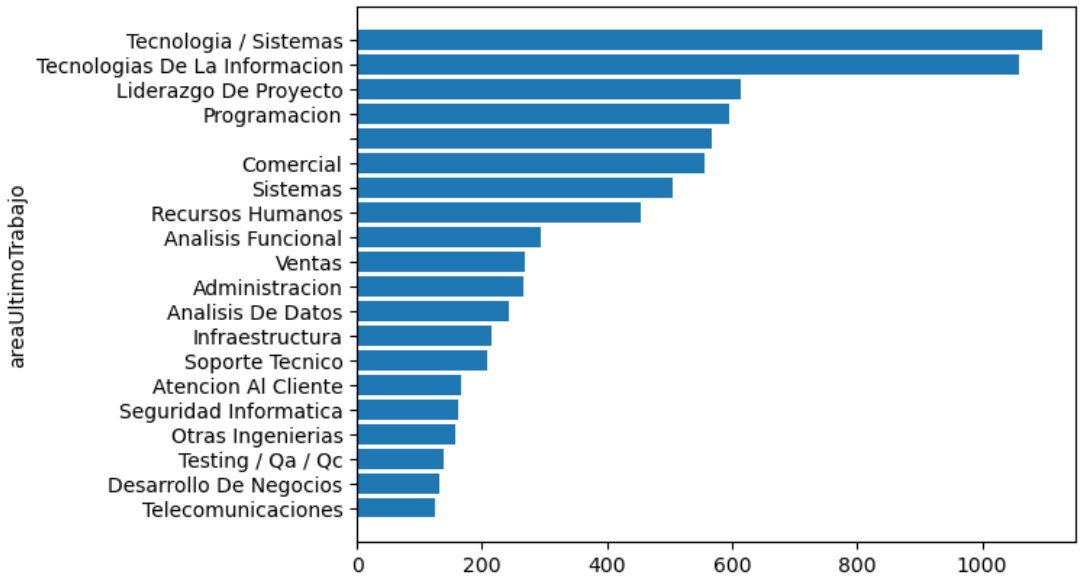
Fuente: La empresa
Elaboración: Propia

Figura 4.18: Empresa del último trabajo



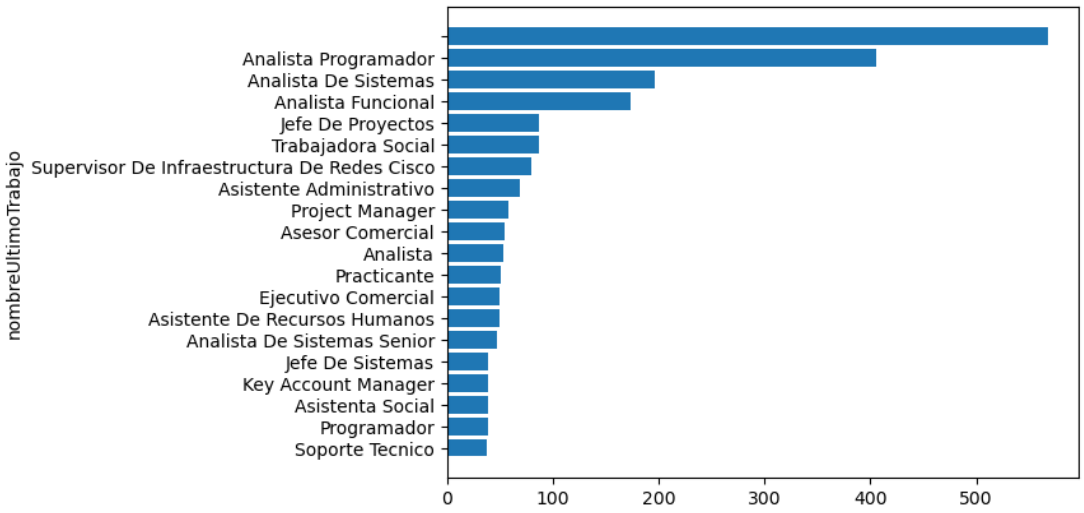
Fuente: La empresa
Elaboración: Propia

Figura 4.19: Área del último trabajo



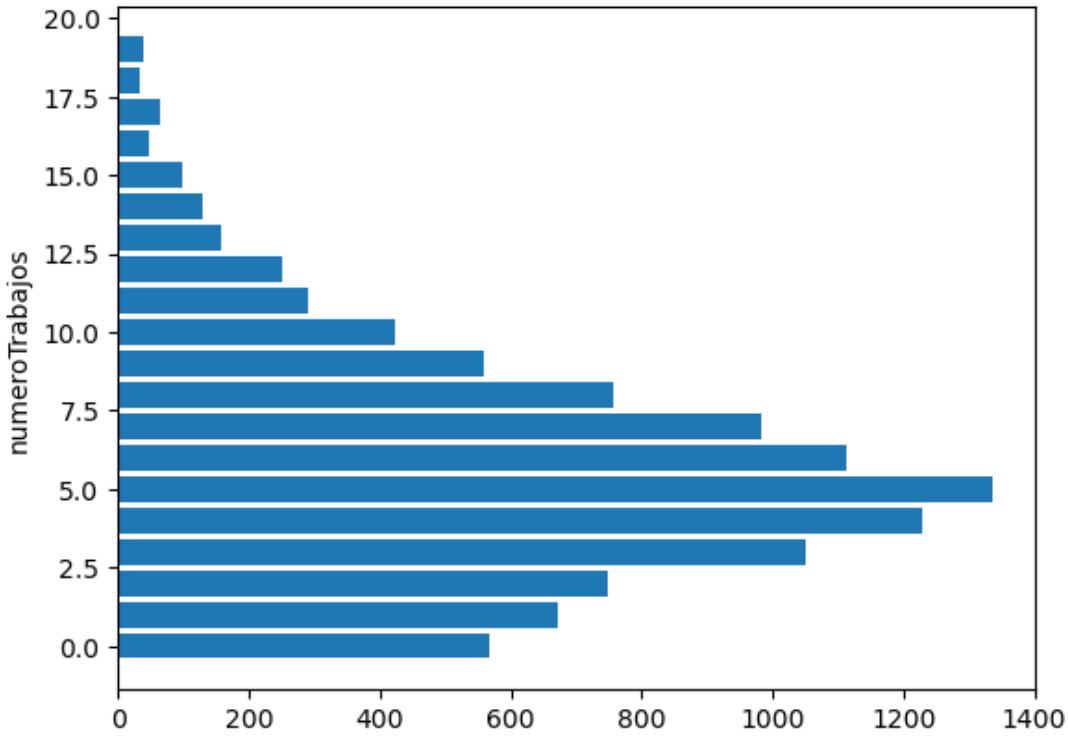
Fuente: La empresa
Elaboración: Propia

Figura 4.20: Nombre del último trabajo



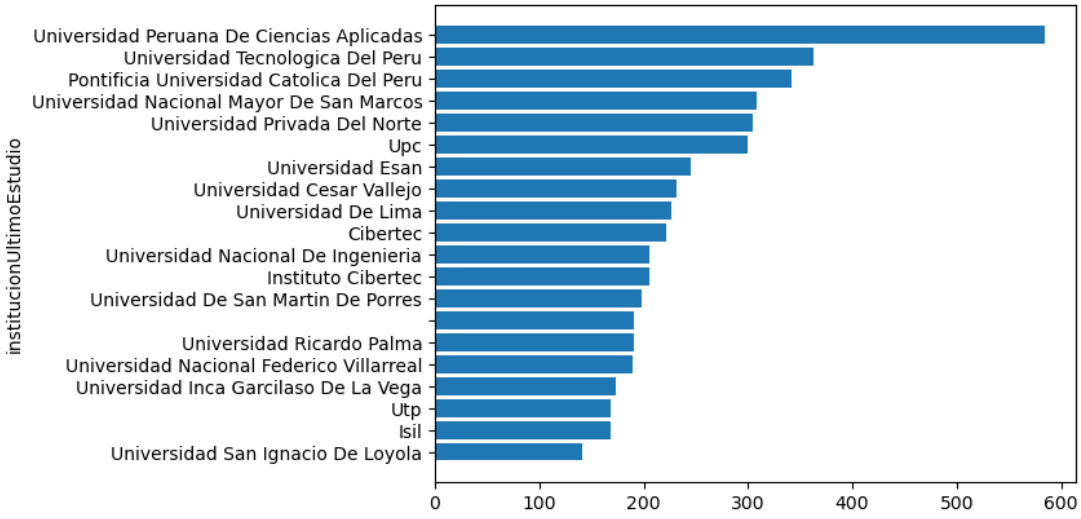
Fuente: La empresa
Elaboración: Propia

Figura 4.21: Número de trabajos



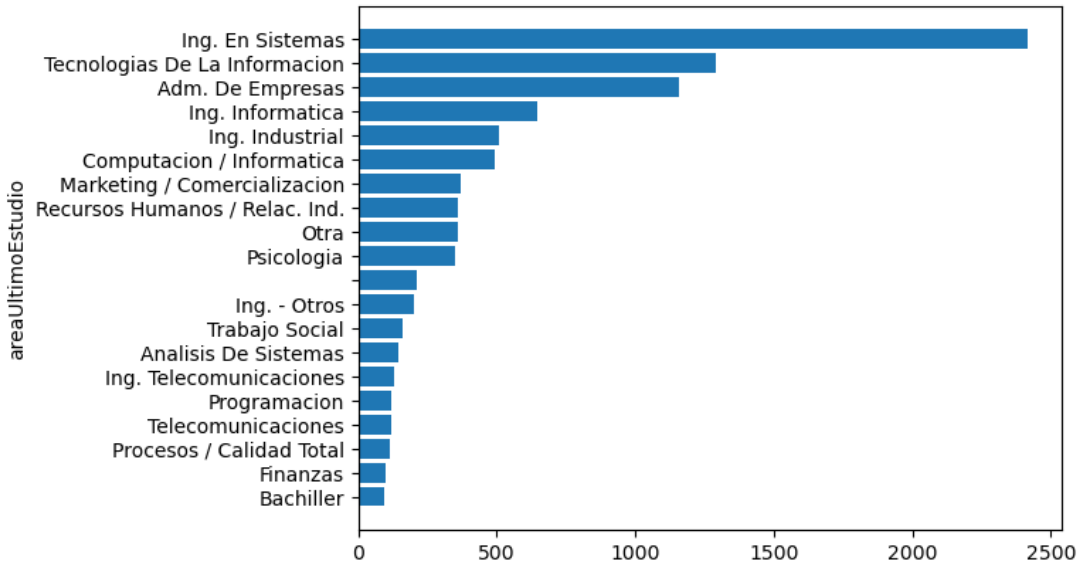
Fuente: La empresa
Elaboración: Propia

Figura 4.22: Institución del último estudio



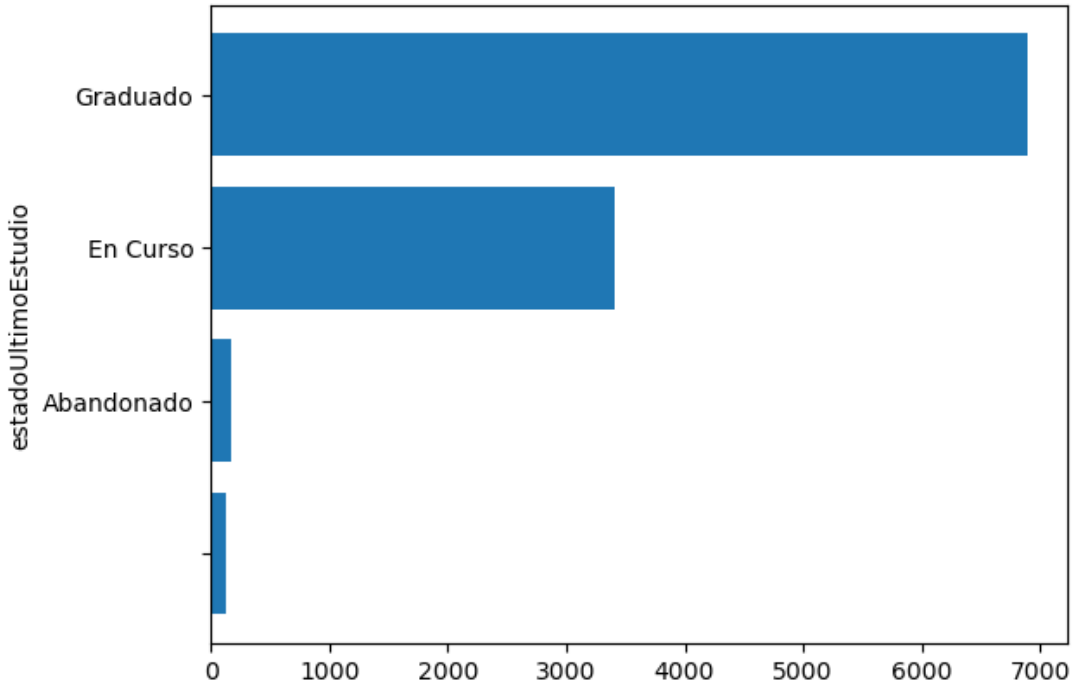
Fuente: La empresa
Elaboración: Propia

Figura 4.23: Área del último estudio



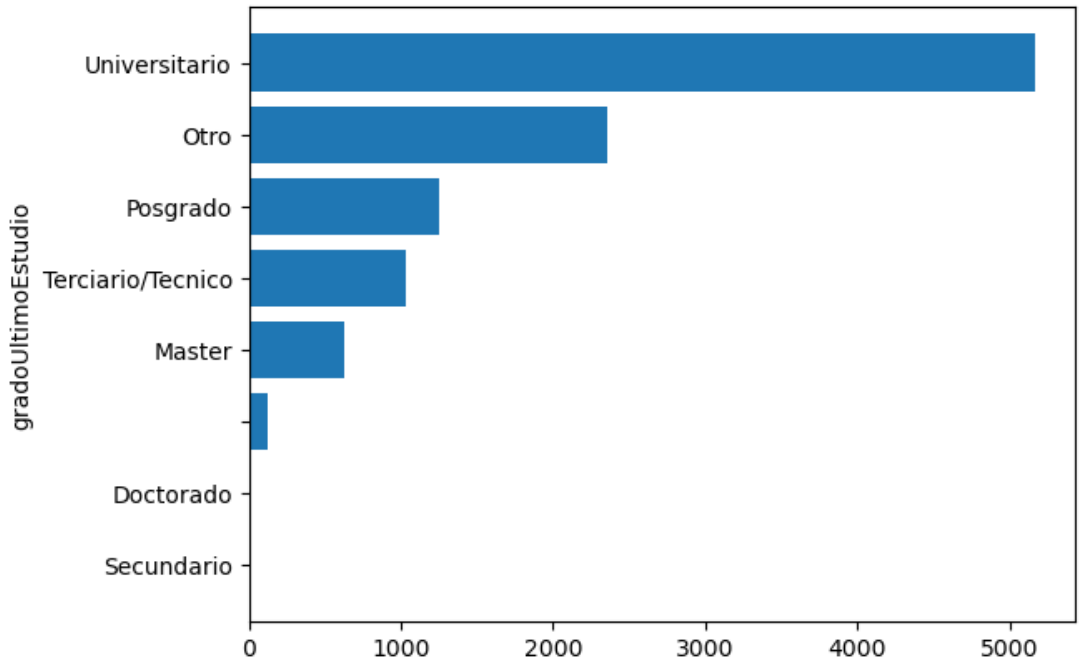
Fuente: La empresa
Elaboración: Propia

Figura 4.24: Estado del último estudio



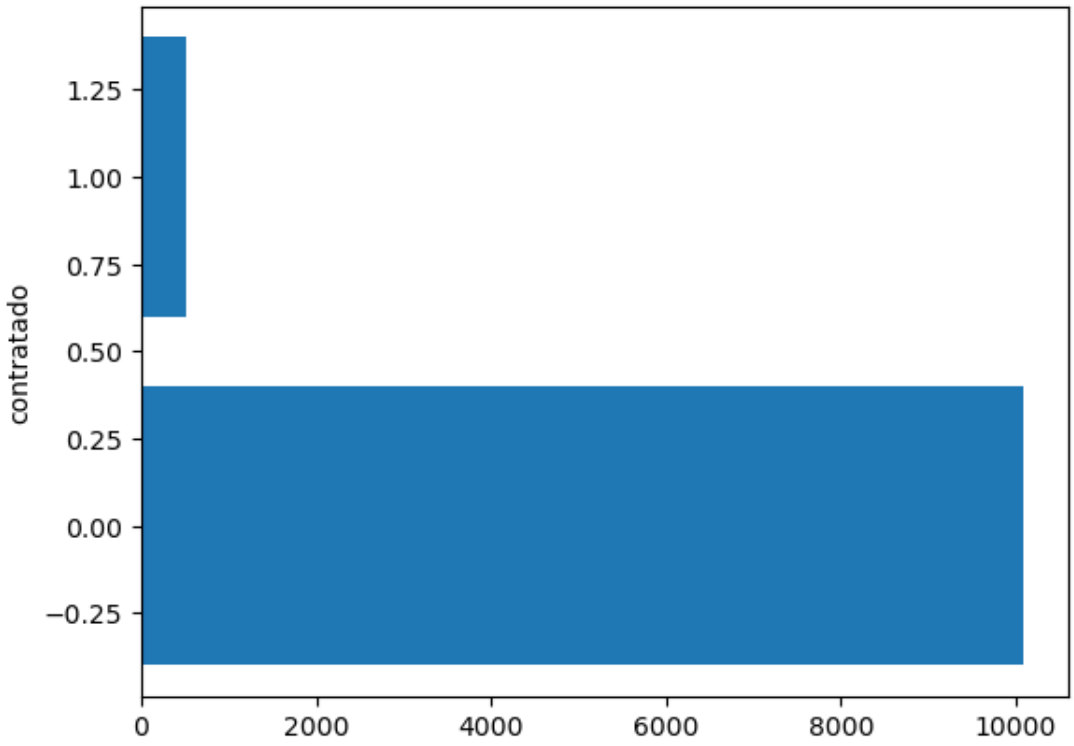
Fuente: La empresa
Elaboración: Propia

Figura 4.25: Grado del último estudio



Fuente: La empresa
Elaboración: Propia

Figura 4.26: Contratado



Fuente: La empresa
Elaboración: Propia

4.2.3 Preparación de los datos

Las tareas de desarrollo propias de esta etapa se encuentran en un repositorio remoto de GitHub, el cual se encuentra en la referencia de Nolasco (2023).

Las actividades de preparación de los datos que se contemplaron fueron:

- Filtrado al rango de fechas delimitado
- Eliminación de columnas
- Aplicación de equivalencias a datos de columnas
- Reemplazo de valores nulos
- Reemplazo de valores atípicos
- Nueva descripción general de los datos
- Nuevo diccionario de datos

- Nueva visualización de datos

A continuación, se detallan cada una de estas actividades:

4.2.3.1 Filtrado al rango de fechas delimitado

En la comprensión de datos, se extrajeron todos los datos de los candidatos posibles, dentro de la bandeja del correo del área de RRHH. Sin embargo, se acotó que la fecha de postulación debería ser entre el 21 de junio del 2019 y el 22 de mayo del 2023. Se eliminaron las filas que no cumplan con la condición, y el número total de registros disminuyó de 10597 a 10562.

4.2.3.2 Eliminación de columnas

Se determinaron 3 criterios para la eliminación de columnas del conjunto de datos:

4.2.3.2.1 Columnas con datos sensibles

Se contemplan diversas columnas con datos sensibles de los candidatos, las cuales son:

- Nombre completo del candidato.
- Número de documento del candidato.
- Fecha de nacimiento del candidato.
- Dirección del candidato.
- Número de casa del candidato.
- Número celular del candidato.
- Correo electrónico del candidato.

Se eliminaron las columnas que cumplen con este criterio, y el número total de atributos disminuyó de 36 a 29.

4.2.3.2.2 Columnas no relevantes

Se contemplan diversas columnas con datos que no aportan al modelo predictivo, las cuales son:

- Fecha de postulación.
- Id de convocatoria.
- Objetivo laboral.

- Descripción del último trabajo.

Se eliminaron las columnas que cumplen con este criterio, y el número total de atributos disminuyó de 29 a 25.

4.2.3.2.3 Columnas normadas por ley

Se contemplan diversas columnas con datos que están normados por la Ley N.º 26772: Igualdad de oportunidades y de trato, las cuales son:

- Estado civil del candidato.
- País de nacimiento del candidato.

Se eliminaron las columnas que cumplen con este criterio, y el número total de atributos disminuyó de 25 a 23.

4.2.3.3 Aplicación de equivalencias a datos de columnas

Existen algunas columnas con alto número de datos diferentes, pero que, existen un subconjunto de estos datos, que tienen equivalencias con otros datos pertenecientes a la misma columna.

Por ejemplo, dentro de las instituciones de último estudio, existe un valor *Universidad Peruana de Ciencias Aplicadas*, mientras que también hay otro dato con valor *Upc*. Pese a que ambos sean datos diferentes, realmente hacen referencia al mismo centro de estudios, mientras uno es el nombre completo, otro es el acrónimo.

Esto aparte de ocurrir con acrónimos, también ocurre con errores ortográficos propios de la fuente de datos de Bumeran. Por ejemplo, confundir *Universidad* con *Univeridad*, o con *Universidade*.

Estas equivalencias se aplicaron para las siguientes 5 columnas:

- Nombre del perfil de la convocatoria.
- Empresa de último trabajo.
- Nombre de último trabajo.
- Institución de último estudio.
- Nombre de último estudio.

Se muestra un resumen del número de equivalencias aplicadas en la Tabla 4.4:

Tabla 4.4: Equivalencias - Comprensión de datos

Columna	Número de datos diferentes antes de las equivalencias	Número de equivalencias	Número de datos diferentes después de las equivalencias
nombrePerfilConvocatoria	172	72	100
empresaUltimoTrabajo	5533	1299	4234
nombreUltimoTrabajo	4389	884	3505
institucionUltimoEstudio	1512	505	1007
nombreUltimoEstudio	3489	552	2937

Fuente: La empresa
Elaboración: Propia

Al solo reemplazarse los datos en sí, el número de filas y columnas no fue alterado.

4.2.3.4 Reemplazo de valores nulos

La presencia de valores nulos dentro de un conjunto de datos puede afectar la calidad y la utilidad de los datos, generando sesgos, e interpretación errónea de estos datos.

En la Figura 4.5 se muestra la proporción de registros nulos dentro del conjunto de datos, por cada columna:

Tabla 4.5: Porcentaje de nulos por columna

Orden	Columna	Tipo	Porcentaje nulos
1	otrasHabilidades	Numérico	22.85%
2	habilidadesTecnicas	Numérico	11.83%
3	idiomas	Numérico	7.45%
4	diasUltimoEstudio	Numérico	6.87%
5	diasUltimoTrabajo	Numérico	5.72%
6	empresaUltimoTrabajo	Categórico	5.45%
7	aniosExperiencia	Numérico	5.44%
8	areaUltimoTrabajo	Categórico	5.38%
9	nombreUltimoTrabajo	Categórico	5.38%
10	numeroTrabajos	Numérico	5.38%
11	paisUltimoTrabajo	Categórico	5.38%
12	sueldoPretendido	Numérico	3.83%
13	areaUltimoEstudio	Categórico	2.00%
14	institucionUltimoEstudio	Categórico	1.84%
15	nombreUltimoEstudio	Categórico	1.40%
16	aniosEstudio	Numérico	1.25%
17	estadoUltimoEstudio	Categórico	1.22%
18	gradoUltimoEstudio	Categórico	1.21%
19	numeroEstudios	Numérico	1.21%
20	paisUltimoEstudio	Categórico	1.21%
21	paisResidencia	Categórico	0.60%
22	nombrePerfilConvocatoria	Categórico	0.00%
23	contratado	Numérico	0.00%

Fuente: La empresa
Elaboración: Propia

Debido a que existen valores nulos presentes en las columnas de los datos, se optó por aplicar diversas estrategias para tratar estos valores nulos:

4.2.3.4.1 Para columnas mayores o iguales a 30% de valores nulos

En este caso, si la columna tiene mayor o igual al 30% de valores nulos, la columna se eliminaría directamente del modelo y no se consideraría para las siguientes fases.

En este caso, al ser la columna *otrasHabilidades* la columna con mayores valores nulos, y siendo esta de 22.85%, no se eliminará ninguna columna bajo este criterio.

Por lo cual, el porcentaje de nulos por columnas no se vio afectado.

4.2.3.4.2 Para columnas menores a 30% de valores nulos, y de tipo categórico

En este caso, si la columna tiene menor al 30% de valores nulos, y es del tipo categórico, lo que se emplea es la aleatorización de valores no nulos.

Es decir, para esos valores nulos, se reemplazará aleatoriamente por algún valor no nulo dentro de la misma columna, esto con el fin de mantener homogeneidad en la data para esas columnas.

En la Figura 4.6 se muestra la nueva proporción de valores nulos una vez aplicada esta estrategia:

Tabla 4.6: Porcentaje de nulos por columna 2

Orden	Columna	Tipo	Porcentaje nulos
1	otrasHabilidades	Numérico	22.85%
2	habilidadesTecnicas	Numérico	11.83%
3	idiomas	Numérico	7.45%
4	diasUltimoEstudio	Numérico	6.87%
5	diasUltimoTrabajo	Numérico	5.72%
6	aniosExperiencia	Numérico	5.44%
7	numeroTrabajos	Numérico	5.38%
8	sueldoPretendido	Numérico	3.83%
9	aniosEstudio	Numérico	1.25%
10	numeroEstudios	Numérico	1.21%
11	nombrePerfilConvocatoria	Categórico	0.00%
12	nombreUltimoEstudio	Categórico	0.00%
13	gradoUltimoEstudio	Categórico	0.00%
14	estadoUltimoEstudio	Categórico	0.00%
15	institucionUltimoEstudio	Categórico	0.00%
16	areaUltimoEstudio	Categórico	0.00%
17	paisUltimoEstudio	Categórico	0.00%
18	paisResidencia	Categórico	0.00%
19	nombreUltimoTrabajo	Categórico	0.00%
20	areaUltimoTrabajo	Categórico	0.00%
21	paisUltimoTrabajo	Categórico	0.00%
22	empresaUltimoTrabajo	Categórico	0.00%
23	contratado	Numérico	0.00%

Fuente: La empresa
Elaboración: Propia

4.2.3.4.3 Para columnas menores a 30% de valores nulos, y de tipo numérico

En este caso, si la columna tiene menor al 30% de valores nulos, y es del tipo numérico, lo que se emplea es el reemplazo por promedio.

Es decir, para esos valores nulos, se reemplazará a todos los valores con el promedio de los valores no nulos para esa columna, esto con el fin de mantener homogeneidad en la data para esas columnas.

En la Figura 4.7 se muestra la nueva proporción de valores nulos una vez aplicada esta estrategia:

Tabla 4.7: Porcentaje de nulos por columna 3

Orden	Columna	Tipo	Porcentaje nulos
1	nombrePerfilConvocatoria	Categórico	0.00%
2	paisUltimoEstudio	Categórico	0.00%
3	otrasHabilidades	Numérico	0.00%
4	idiomas	Numérico	0.00%
5	habilidadesTecnicas	Numérico	0.00%
6	numeroEstudios	Numérico	0.00%
7	aniosEstudio	Numérico	0.00%
8	gradoUltimoEstudio	Categórico	0.00%
9	estadoUltimoEstudio	Categórico	0.00%
10	nombreUltimoEstudio	Categórico	0.00%
11	areaUltimoEstudio	Categórico	0.00%
12	institucionUltimoEstudio	Categórico	0.00%
13	paisResidencia	Categórico	0.00%
14	diasUltimoEstudio	Numérico	0.00%
15	numeroTrabajos	Numérico	0.00%
16	aniosExperiencia	Numérico	0.00%
17	nombreUltimoTrabajo	Categórico	0.00%
18	areaUltimoTrabajo	Categórico	0.00%
19	paisUltimoTrabajo	Categórico	0.00%
20	empresaUltimoTrabajo	Categórico	0.00%
21	diasUltimoTrabajo	Numérico	0.00%
22	sueldoPretendido	Numérico	0.00%
23	contratado	Numérico	0.00%

Fuente: La empresa
Elaboración: Propia

4.2.3.5 Reemplazo de valores atípicos

El reemplazo de valores atípicos solo aplica para columnas numéricas, esto debido a que, para detectar valores atípicos, se requiere que haya una distancia numérica entre los elementos, o que puedan ser colocados en una escala ascendente o descendente, esto es posible para los valores numéricos, pero no para los categóricos.

Existen dos tipos de valores atípicos, atípicos leves y atípicos extremos.

Se define el rango intercuartílico mediante la Ecuación 4.1:

$$IQR = Q_3 - Q_1 \tag{4.1}$$

Dónde:

- IQR: Rango intercuartílico.
- Q_3: Tercer cuartil del conjunto de datos.

- Q_1: Primer cuartil del conjunto de datos.

4.2.3.5.1 Valor atípico leve

Un valor atípico leve será aquel que cumpla con la condición descrita en la Ecuación 4.2:

$$q < Q_1 - 1.5 \cdot IQR \quad \vee \quad q > Q_3 + 1.5 \cdot IQR \tag{4.2}$$

Dónde:

- q: Valor de la muestra a evaluar.
- Q_1: Primer cuartil del conjunto de datos.
- IQR: Rango intercuartílico.
- Q_3: Tercer cuartil del conjunto de datos.

Cabe recalcar que, estos valores atípicos leves no se reemplazarán.

4.2.3.5.2 Valor atípico extremo

Un valor atípico extremo será aquel que cumpla con la condición descrita en la Ecuación 4.3:

$$q < Q_1 - 3 \cdot IQR \quad \vee \quad q > Q_3 + 3 \cdot IQR \tag{4.3}$$

Dónde:

- q: Valor de la muestra a evaluar.
- Q_1: Primer cuartil del conjunto de datos.
- IQR: Rango intercuartílico.
- Q_3: Tercer cuartil del conjunto de datos.

Cabe recalcar que, estos valores atípicos extremos si se reemplazarán.

4.2.3.5.3 Reemplazo de valores atípicos extremos

En la Figura 4.8 se muestra la nueva proporción de valores atípicos extremos, con respecto al total:

Tabla 4.8: Porcentaje de atípicos por columna

Orden	Columna	Tipo	Porcentaje atípicos
1	idiomas	Numérico	3.55%
2	habilidadesTecnicas	Numérico	2.66%
3	otrasHabilidades	Numérico	2.41%
4	diasUltimoTrabajo	Numérico	2.27%
5	numeroEstudios	Numérico	2.25%
6	aniosEstudio	Numérico	1.63%
7	aniosExperiencia	Numérico	0.80%
8	sueldoPretendido	Numérico	0.42%
9	numeroTrabajos	Numérico	0.34%
10	diasUltimoEstudio	Numérico	0.12%
11	nombrePerfilConvocatoria	Categórico	0.00%
12	nombreUltimoEstudio	Categórico	0.00%
13	gradoUltimoEstudio	Categórico	0.00%
14	estadoUltimoEstudio	Categórico	0.00%
15	institucionUltimoEstudio	Categórico	0.00%
16	areaUltimoEstudio	Categórico	0.00%
17	paisUltimoEstudio	Categórico	0.00%
18	paisResidencia	Categórico	0.00%
19	nombreUltimoTrabajo	Categórico	0.00%
20	areaUltimoTrabajo	Categórico	0.00%
21	paisUltimoTrabajo	Categórico	0.00%
22	empresaUltimoTrabajo	Categórico	0.00%
23	contratado	Numérico	0.00%

Fuente: La empresa
Elaboración: Propia

Se aplicó el reemplazo de los valores atípicos, aplicando las siguientes reglas:

- Si el valor atípico se encuentra más cercano al límite inferior $(Q_1 - 3 \cdot IQR)$, será reemplazado por este límite inferior.
- Caso contrario, será reemplazado por el límite superior $(Q_3 + 3 \cdot IQR)$

Se realizó el reemplazo de estos valores en el conjunto de datos.

En la Figura 4.9 se muestra la nueva proporción de valores atípicos una vez aplicada esta estrategia:

Tabla 4.9: Porcentaje de atípicos por columna 2

Orden	Columna	Tipo	Porcentaje atípicos
1	nombrePerfilConvocatoria	Categórico	0.00%
2	paisUltimoEstudio	Categórico	0.00%
3	otrasHabilidades	Numérico	0.00%
4	idiomas	Numérico	0.00%
5	habilidadesTecnicas	Numérico	0.00%
6	numeroEstudios	Numérico	0.00%
7	aniosEstudio	Numérico	0.00%
8	gradoUltimoEstudio	Categórico	0.00%
9	estadoUltimoEstudio	Categórico	0.00%
10	nombreUltimoEstudio	Categórico	0.00%
11	areaUltimoEstudio	Categórico	0.00%
12	institucionUltimoEstudio	Categórico	0.00%
13	paisResidencia	Categórico	0.00%
14	diasUltimoEstudio	Numérico	0.00%
15	numeroTrabajos	Numérico	0.00%
16	aniosExperiencia	Numérico	0.00%
17	nombreUltimoTrabajo	Categórico	0.00%
18	areaUltimoTrabajo	Categórico	0.00%
19	paisUltimoTrabajo	Categórico	0.00%
20	empresaUltimoTrabajo	Categórico	0.00%
21	diasUltimoTrabajo	Numérico	0.00%
22	sueldoPretendido	Numérico	0.00%
23	contratado	Numérico	0.00%

Fuente: La empresa
Elaboración: Propia

Con esto concluyen las actividades de preparación de datos.

4.2.3.6 Nueva descripción general de los datos

En base a estos datos preparados, se pueden detallar algunas características sobre ellos, los cuales se presentan en la Tabla 4.10:

Tabla 4.10: Características de los datos - Preparación de datos

Número de filas	10562
Fila con fecha más antigua	21-06-2019 14:09
Fila con fecha más reciente	22-05-2023 19:31
Número de columnas	23
Número de columnas con valores categóricos	12
Número de columnas con valores numéricos	11
Número de columnas con al menos un valor nulo	0

Fuente: La empresa
Elaboración: Propia

4.2.3.7 Nuevo diccionario de datos

También se presenta el diccionario de los datos preparados, en la Figura 4.11:

Tabla 4.11: Diccionario de datos - Preparación de datos

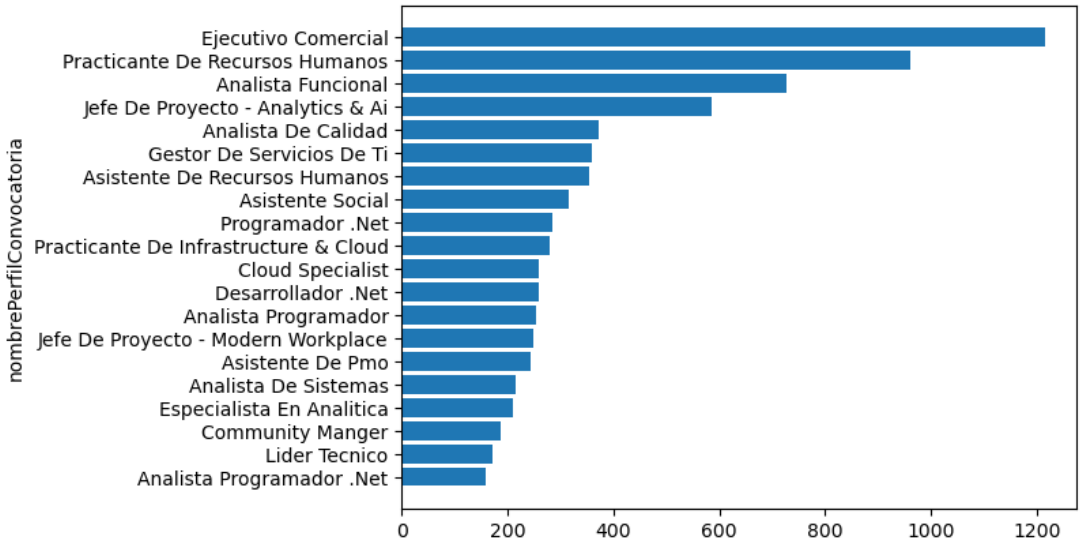
ID	Campo	Origen	Tipo	Descripción	Valores posibles
1	nombrePerfilConvocatoria	Convocatoria	Categórico	Nombre del perfil del puesto solicitado	Texto
2	paisResidencia	Candidato	Categórico	País de residencia del candidato	195 países diferentes
3	sueldoPretendido	Candidato	Numérico	Sueldo pretendido, solicitado por el candidato en la convocatoria	Número decimal mayor a 0
4	diasUltimoTrabajo	Candidato	Numérico	Número de días de permanencia del último trabajo del candidato	Número entero, mayor o igual a 0
5	empresaUltimoTrabajo	Candidato	Categórico	Nombre de la empresa del último trabajo del candidato	Texto
6	paisUltimoTrabajo	Candidato	Categórico	País del último trabajo del candidato	195 países diferentes
7	areaUltimoTrabajo	Candidato	Categórico	Área del último trabajo del candidato	Texto
8	nombreUltimoTrabajo	Candidato	Categórico	Nombre del último trabajo del candidato	Texto
9	aniosExperiencia	Candidato	Numérico	Años de experiencia del candidato	Número decimal mayor o igual a 0
10	numeroTrabajos	Candidato	Numérico	Número de trabajos del candidato	Número entero, mayor o igual a 0
11	diasUltimoEstudio	Candidato	Numérico	Número de días de permanencia del último estudio del candidato	Número entero, mayor o igual a 0
12	institucionUltimoEstudio	Candidato	Categórico	Institución del último estudio del candidato	Texto
13	paisUltimoEstudio	Candidato	Categórico	País del último estudio del candidato	195 países diferentes
14	areaUltimoEstudio	Candidato	Categórico	Área del último estudio del candidato	Texto
15	nombreUltimoEstudio	Candidato	Categórico	Nombre del último estudio del candidato	Texto
16	estadoUltimoEstudio	Candidato	Categórico	Estado del último estudio del candidato	Abandonado, En Curso, Graduado
17	gradoUltimoEstudio	Candidato	Categórico	Grado alcanzado del último estudio del candidato	Otro, Secundario, Terciario/Técnico, Universitario, Posgrado, Master, Doctorado
18	aniosEstudio	Candidato	Numérico	Años de estudio del candidato	Número entero, mayor o igual a 0
19	numeroEstudios	Candidato	Numérico	Número de estudios del candidato	Número entero, mayor o igual a 0
20	habilidadesTecnicas	Candidato	Numérico	Número de habilidades técnicas del candidato	Número entero, mayor o igual a 0
21	idiomas	Candidato	Numérico	Número de idiomas del candidato	Número entero, mayor o igual a 0
22	otrasHabilidades	Candidato	Numérico	Número de otras habilidades (blandas, informáticas) del candidato	Número entero, mayor o igual a 0
23	contratado	Convocatoria	Numérico	El candidato fue rechazado o fue contratado	0 o 1

Fuente: La empresa
Elaboración: Propia

4.2.3.8 Nueva visualización de datos

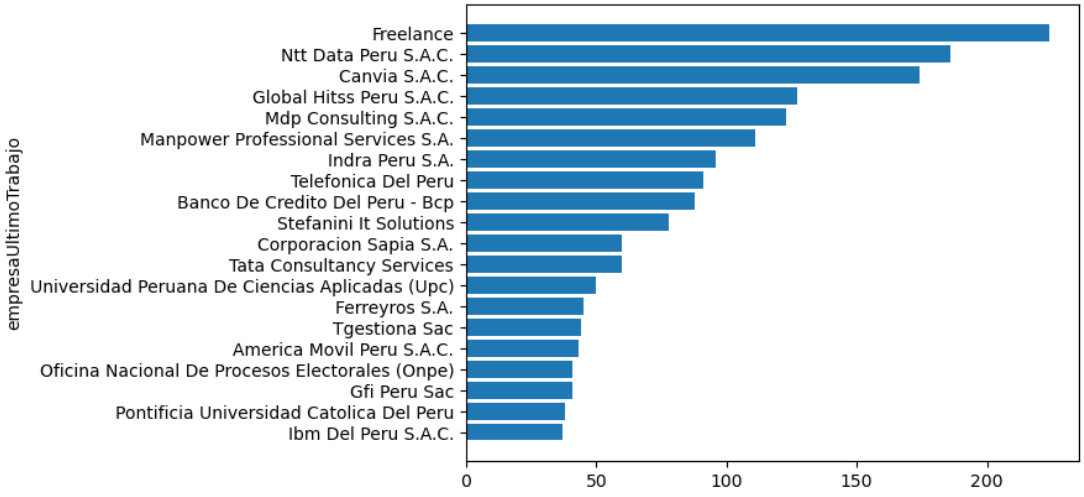
A continuación, se muestran algunas gráficas descriptivas de los datos preparados, en la Figura 4.27, Figura 4.28, Figura 4.29, Figura 4.30, Figura 4.31, Figura 4.32, Figura 4.33, Figura 4.34, Figura 4.35 y Figura 4.36:

Figura 4.27: Nombre del perfil de la convocatoria



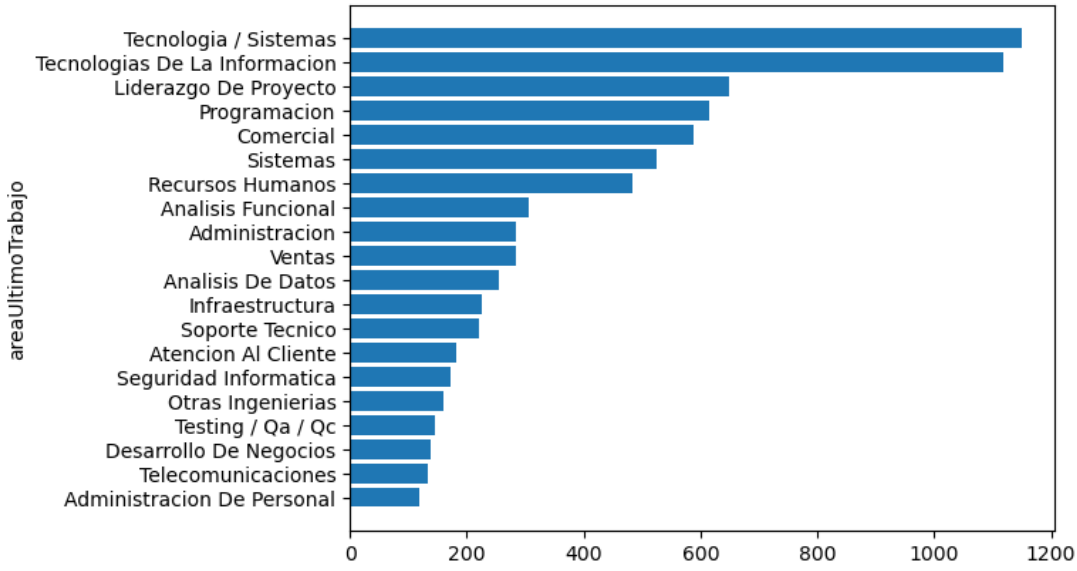
Fuente: La empresa
Elaboración: Propia

Figura 4.28: Empresa del último trabajo



Fuente: La empresa
Elaboración: Propia

Figura 4.29: Área del último trabajo

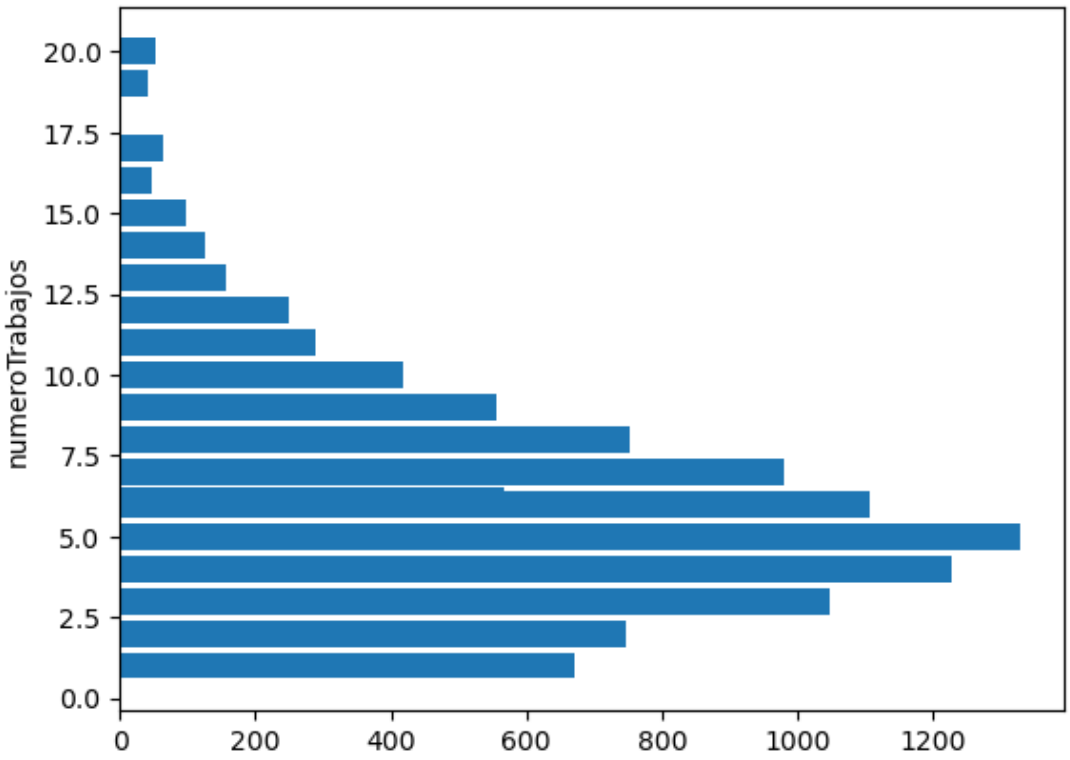


Fuente: La empresa
Elaboración: Propia

Figura 4.30: Nombre del último trabajo

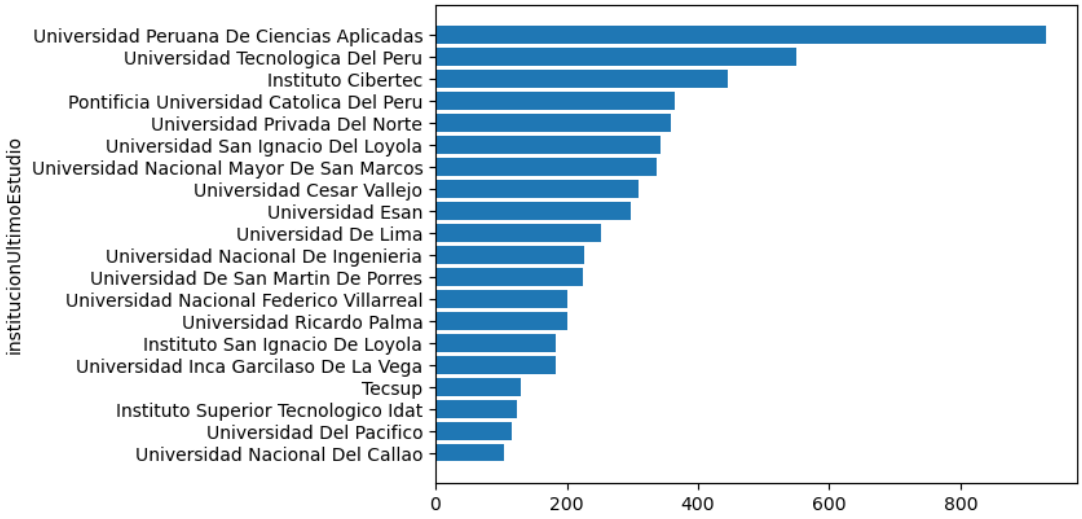
Fuente: La empresa
Elaboración: Propia

Figura 4.31: Número de trabajos



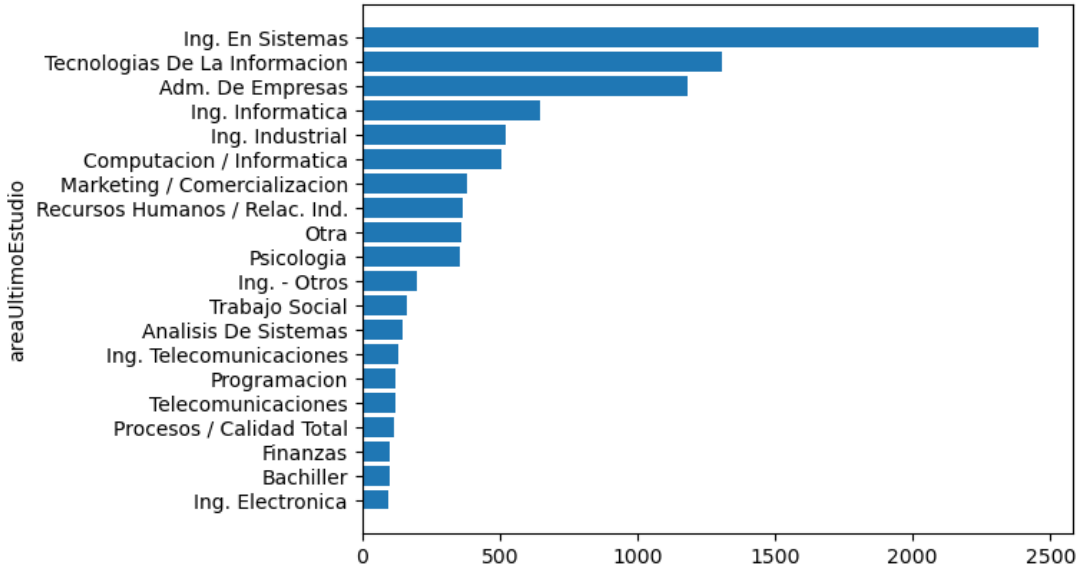
Fuente: La empresa
Elaboración: Propia

Figura 4.32: Institución del último estudio



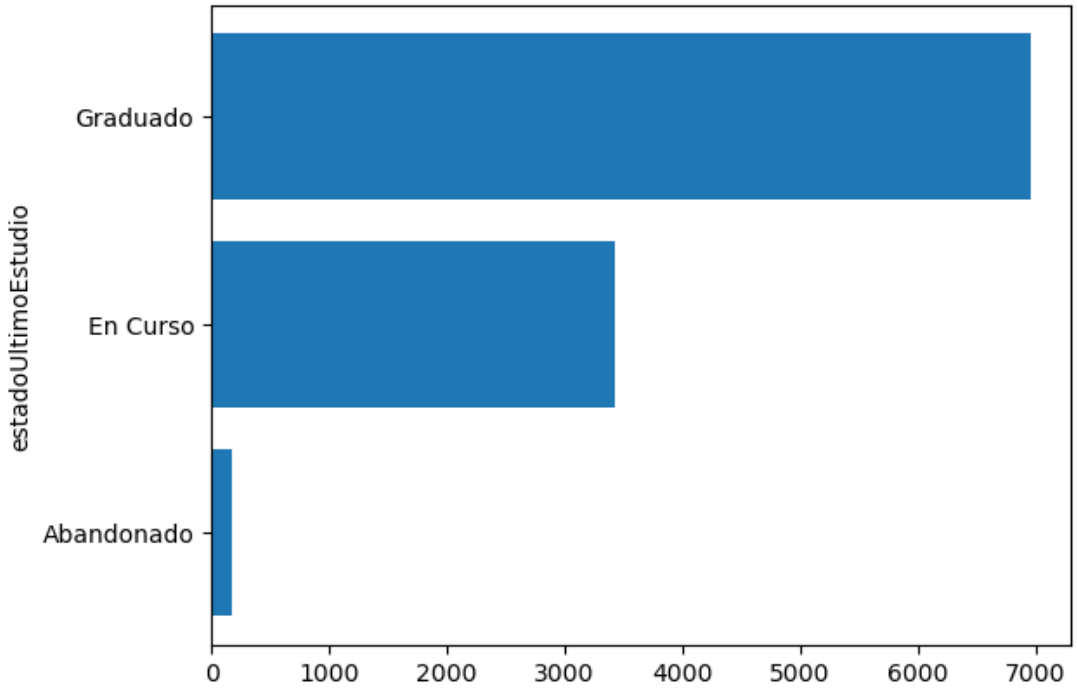
Fuente: La empresa
Elaboración: Propia

Figura 4.33: Área del último estudio



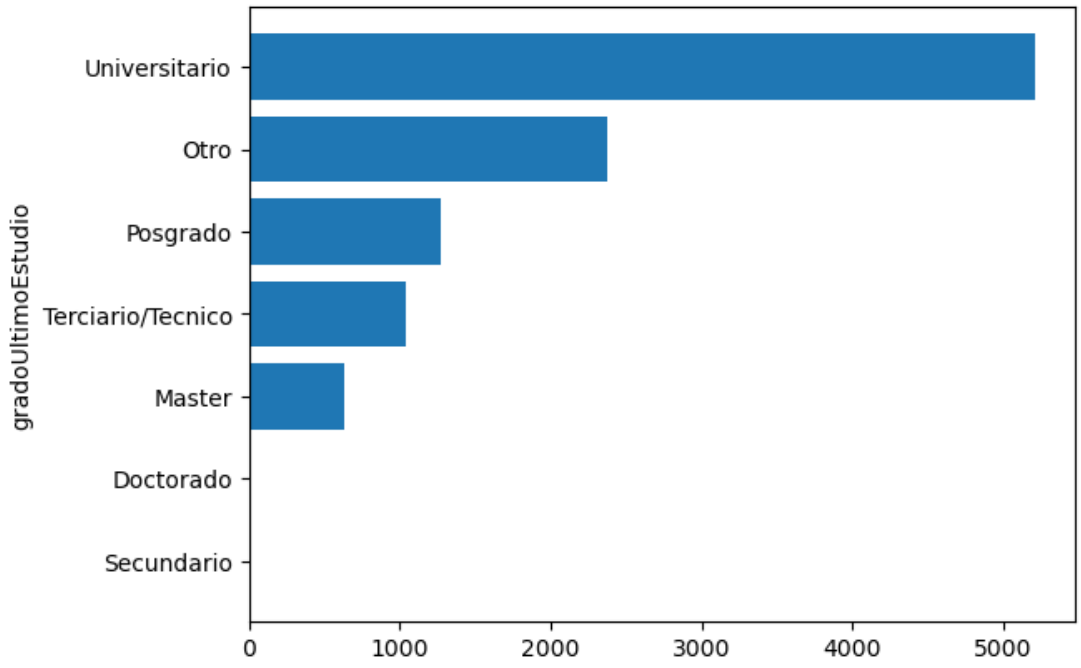
Fuente: La empresa
Elaboración: Propia

Figura 4.34: Estado del último estudio



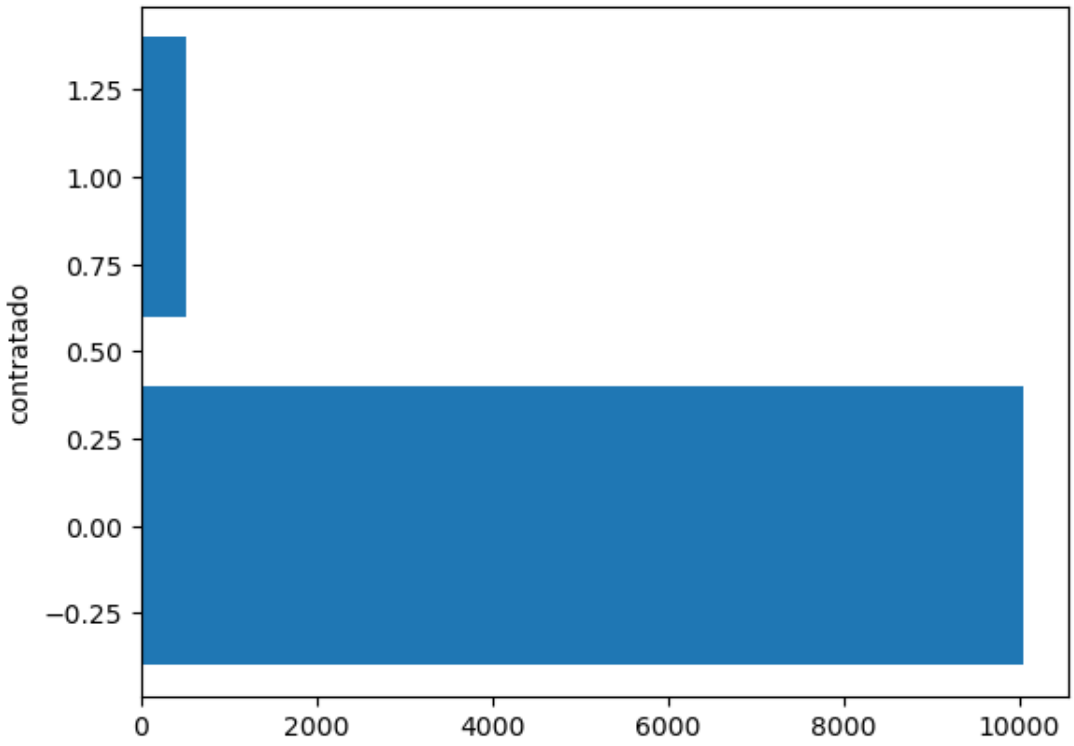
Fuente: La empresa
Elaboración: Propia

Figura 4.35: Grado del último estudio



Fuente: La empresa
Elaboración: Propia

Figura 4.36: Contratado



Fuente: La empresa
Elaboración: Propia

4.2.4 Modelado

Las tareas de desarrollo propias de esta etapa se encuentran en un repositorio remoto de GitHub, el cual se encuentra en la referencia de Nolasco (2023).

Las actividades de modelado que se contemplaron fueron:

- Balanceo de datos
- Codificación de las variables categóricas ordinales
- Codificación de las variables categóricas cardinales
- Normalización de las variables numéricas
- Eliminación de columnas con varianza cercana a cero
- Eliminación de columnas con correlación cercana a uno
- Separación de datos de entrenamiento y validación

- Entrenamiento de los modelos
- Obtención de métricas de cada modelo
- Selección del mejor algoritmo

A continuación, se detallan cada una de estas actividades:

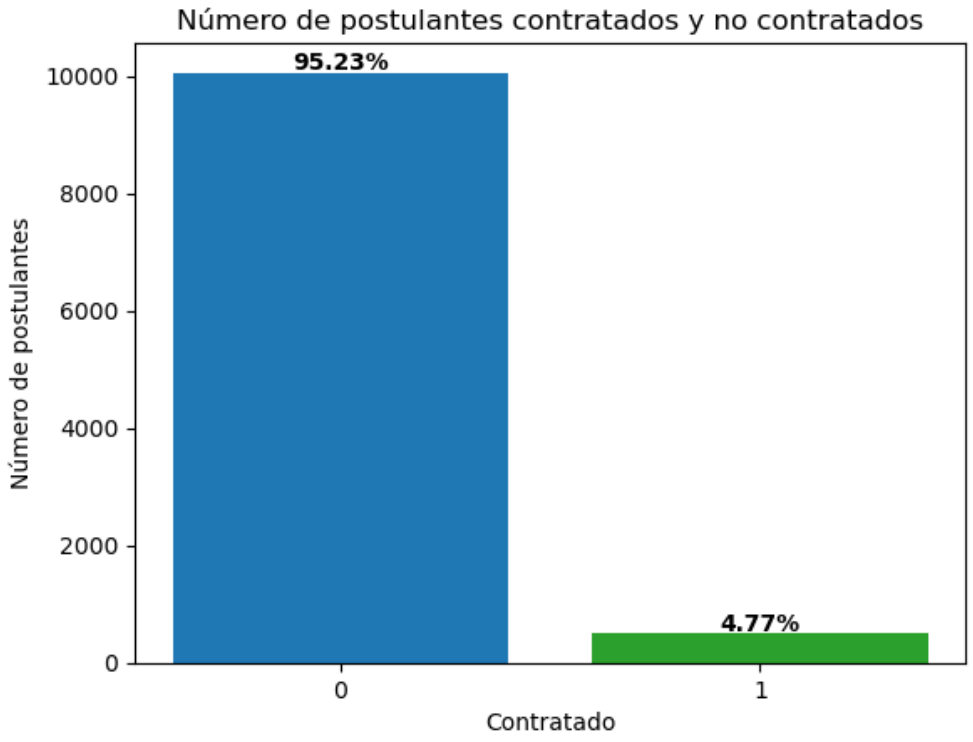
4.2.4.1 Balanceo de datos

Dentro de un contexto donde existan 2 clases a predecir, se define clase mayoritaria como aquella que tiene más observaciones y minoritaria como aquella que menos tiene.

El balanceo de datos se define como métodos para poder tratar datos desbalanceados, con el fin de predecir correctamente las clases minoritarias poco representadas en nuestro conjunto de datos.

En la Figura 4.37, se muestra la proporción de ambas clases antes del sobre muestreo aleatorio:

Figura 4.37: Antes del sobre muestreo aleatorio



Fuente: La empresa
Elaboración: Propia

La proporción inicial de datos se presenta en la Tabla 4.12:

Tabla 4.12: Antes del sobre muestreo - Modelado

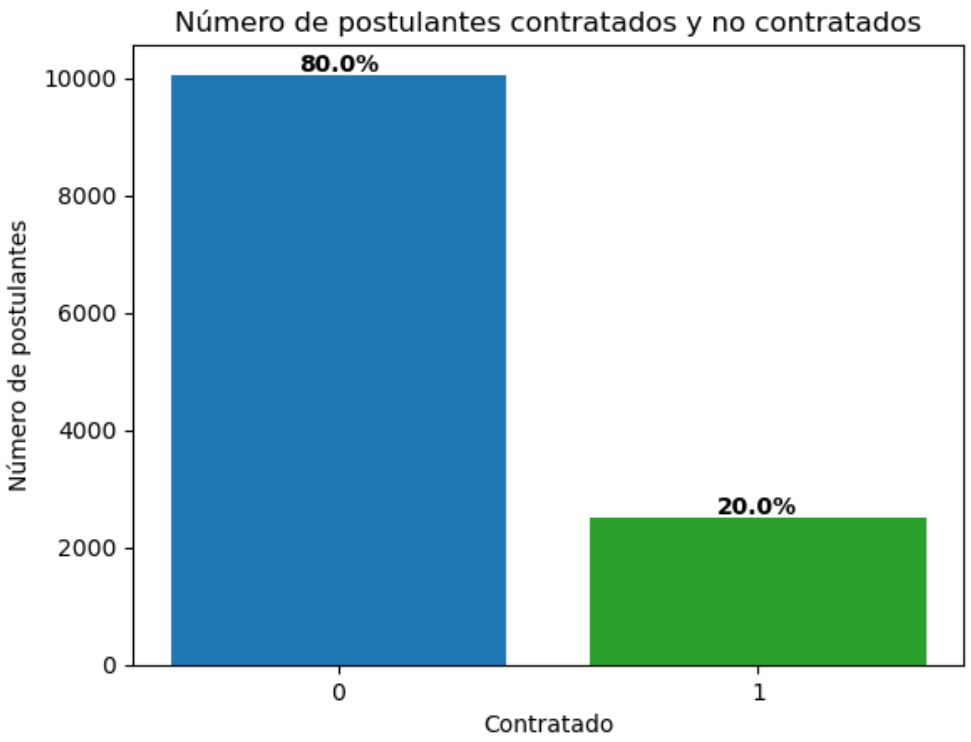
Número de filas con clase 0	10058
Número de filas con clase 1	504
Número total de filas	10562

Fuente: La empresa
Elaboración: Propia

En este caso, se empleará la técnica del sobre muestreo aleatorio, la cual se aplica a la clase minoritaria, y consiste en añadir copias de esta clase para aumentar su peso total.

En la Figura 4.38, se muestra la proporción de ambas clases después del sobre muestreo aleatorio:

Figura 4.38: Después del sobre muestreo aleatorio



Fuente: La empresa
Elaboración: Propia

La proporción final de datos se presenta en la Tabla 4.13:

Tabla 4.13: Después del sobre muestreo - Modelado

Número de filas con clase 0	10058
Número de filas con clase 1	2514
Número total de filas	12572

Fuente: La empresa
Elaboración: Propia

Esto hizo que el número total de filas se elevara de 10562 a 12572.

Además, al ser ahora la proporción de clases de 1 a 4, se podría afirmar que, para el conjunto de datos propuesto, por cada 5 candidatos postulantes a un puesto de trabajo, existe 1 candidato contratado.

4.2.4.2 Codificación de las variables categóricas ordinales

Dentro de los algoritmos de aprendizaje de máquina, es necesario que todos los atributos o variables a usar hayan sido codificados, ya que las entradas de estos algoritmos son solo numéricas.

Debido a que las variables categóricas ordinales tienen un orden y jerarquía lógica, es posible hallar una codificación equivalente para este tipo de variables.

Es por ello que se aplicará la codificación ordinal para estas variables.

Dentro del diccionario de datos final, se identifican 2 variables categóricas ordinales, las cuales se presentan en la Figura 4.14:

Tabla 4.14: Variables categóricas ordinales

ID	Campo	Origen	Tipo	Descripción	Valores posibles
16	estadoUltimoEstudio	Candidato	Categórico	Estado del último estudio del candidato	Abandonado, En Curso, Graduado
17	gradoUltimoEstudio	Candidato	Categórico	Grado alcanzado del último estudio del candidato	Otro, Secundario, Terciario/Técnico, Universitario, Posgrado, Master, Doctorado

Fuente: La empresa
Elaboración: Propia

De estas variables, se ordenó su jerarquía de menor a mayor, asignando números en orden ascendentes desde el número 0. Esta jerarquía se muestra en la Tabla 4.15 y la Tabla 4.16:

Tabla 4.15: Jerarquías del estado de último estudio - Modelado

Categoría	Jerarquía
Graduado	2
En Curso	1
Abandonado	0

Fuente: La empresa
Elaboración: Propia

Tabla 4.16: Jerarquías del grado de último estudio - Modelado

Categoría	Jerarquía
Doctorado	6
Master	5
Posgrado	4
Universitario	3
Terciario/Técnico	2
Secundario	1
Otro	0

Fuente: La empresa
Elaboración: Propia

Estas jerarquías fueron reemplazadas por los valores propios de cada variable dentro del conjunto de datos.

El número de filas del conjunto de datos no cambió, pero esas 2 variables fueron cambiadas de categóricas a numéricas.

4.2.4.3 Codificación de las variables categóricas cardinales

A diferencia de las categóricas ordinales, las cardinales no tienen ningún orden o jerarquía entre sí, por lo cual cada valor tiene el mismo peso dentro de su columna.

Para este tipo de variable se aplicará la codificación en caliente, la cual consiste en hacer que cada valor diferente de una variable se transforme en una columna, haciendo que estas nuevas columnas tengan 0 o 1, dependiendo si cierta observación tiene ese valor de ese atributo o no.

Esto permite manejar estas variables de manera numérica, pero la complejidad se traslada a las columnas, haciendo que haya un gran número de columnas.

En la Tabla 4.17, se describen los números de valores diferentes que tienen las variables categóricas cardinales:

Tabla 4.17: Número de valores diferentes - Modelado

Orden	Columna	Número de valores diferentes
1	nombrePerfilConvocatoria	100
2	paisResidencia	14
3	empresaUltimoTrabajo	4229
4	paisUltimoTrabajo	31
5	areaUltimoTrabajo	152
6	nombreUltimoTrabajo	3154
7	institucionUltimoEstudio	1006
8	paisUltimoEstudio	30
9	areaUltimoEstudio	105
10	nombreUltimoEstudio	2478

Fuente: La empresa
Elaboración: Propia

Siendo así un total de 11299 valores totales diferentes.

Cada uno de estos valores se trasformará en una columna dentro del conjunto de datos. Siendo así un total de $23 + 11299 = 11322$ columnas.

Pero ya que esas 10 columnas cardinales originales ya se han codificado y agregado, se deben eliminar las columnas originales, quedando así un total de $11322 - 10 = 11312$ columnas al finalizar esta codificación.

4.2.4.4 Normalización de las variables numéricas

Cada una de las variables numéricas presentes dentro del conjunto de datos presentan una escala diferente, es decir, tienen valores mínimos y máximos permitidos diferentes para cada uno.

Esto puede causar que, para fines interpretativos del modelo, algunas variables tengan más peso que otras, debido a que tendrán un valor numérico mayor.

Es por ello que todas las escalas se normalizarán a un intervalo $[0, 1]$, donde 0 es el valor mínimo

y 1 el valor máximo.

Esta normalización se aplicará a las columnas presentes en la Tabla 4.18 (debido a que las variables estadoUltimoEstudio y gradoUltimoEstudio fueron transformadas previamente a numéricas, también están incluidas aquí):

Tabla 4.18: Máximos y mínimos de variables numéricas - Modelado

Orden	Columna	Mínimo valor	Máximo valor
1	sueldoPretendido	1	14600
2	diasUltimoTrabajo	5	2619
3	aniosExperiencia	0.1	32.9
4	numeroTrabajos	1	20
5	diasUltimoEstudio	1	6653.5
6	estadoUltimoEstudio	0	2
7	gradoUltimoEstudio	0	6
8	aniosEstudio	0.1	21
9	numeroEstudios	1	13
10	habilidadesTecnicas	1	35
11	idiomas	1	2.2
12	otrasHabilidades	1	18

Fuente: La empresa
Elaboración: Propia

El número de filas y estructura del conjunto de datos no cambió.

4.2.4.5 Eliminación de columnas con varianza cercana a cero

La baja varianza dentro de los valores de una columna determina que todos los valores presentes son valores muy cercanos o que no tienen mucha variación, lo cual podría hacer que esas columnas no sean de alto valor o representativas para el modelo.

Es por ello que, a forma de optimizar el modelo, y utilizar las columnas que más aporte tengan a la variable objetivo, dejaremos el 10% de columnas que tengan mayor varianza que las demás.

Se calculó la varianza para todas las columnas, y se determinó que el umbral para ese 10% de columnas es 0.025, es decir, que solo se mantendrán las columnas que tengan una varianza mayor a 0.025.

En ese sentido, el número total de columnas se redujo de 11312 a 1111.

4.2.4.6 Eliminación de columnas con correlación cercana a uno

La correlación entre columnas también representa un punto a analizar dentro del conjunto de datos, ya que, si existen 2 o más columnas altamente relacionadas, se podría afirmar que ambas columnas representan lo mismo, por lo que sería posible solo mantener una de esas columnas.

Es por ello que, a forma de optimizar el modelo, y utilizar las columnas que más aporte tengan a la variable objetivo, dejaremos el 90% de columnas que tengan menor correlación que las demás.

Se calculó la correlación promedio para todas las columnas, y se determinó que el umbral para ese 90% de columnas es 0.766, es decir, que solo se mantendrán las columnas que tengan una correlación menor o igual a 0.766.

En ese sentido, el número total de columnas se redujo de 1111 a 1001.

4.2.4.7 Separación de datos de entrenamiento y validación

Para el entrenamiento de los modelos, es necesario crear dos conjuntos de datos, en base a los datos iniciales:

- Conjunto de datos de entrenamiento, el cual se usarán como entradas al modelo, las cuales estarán etiquetadas y el modelo aprenderá de estos datos
- Conjunto de datos de validación, el cual se usarán como datos para evaluar la exactitud y precisión del modelo entrenado, con el fin de que pronostique la misma etiqueta o clase del dato real.

La proporción elegida para esta investigación es 75 a 25, es decir:

- 75% de datos para el entrenamiento del modelo
- 25% de datos para la validación del modelo

4.2.4.8 Entrenamiento de los modelos

Para esta etapa, se entrenan los modelos predictivos, utilizando los 2 conjuntos de datos previamente creados.

Se utilizarán los 6 algoritmos descritos en el marco teórico:

1. K vecinos más cercanos (KNN)
2. Máquina de vectores de soporte (SVM)

3. Regresión logística (RL)
4. Árbol de decisión (DT)
5. Bosque aleatorio (RF)
6. Aumento de gradiente (GBM)

4.2.4.9 Obtención de métricas de cada modelo

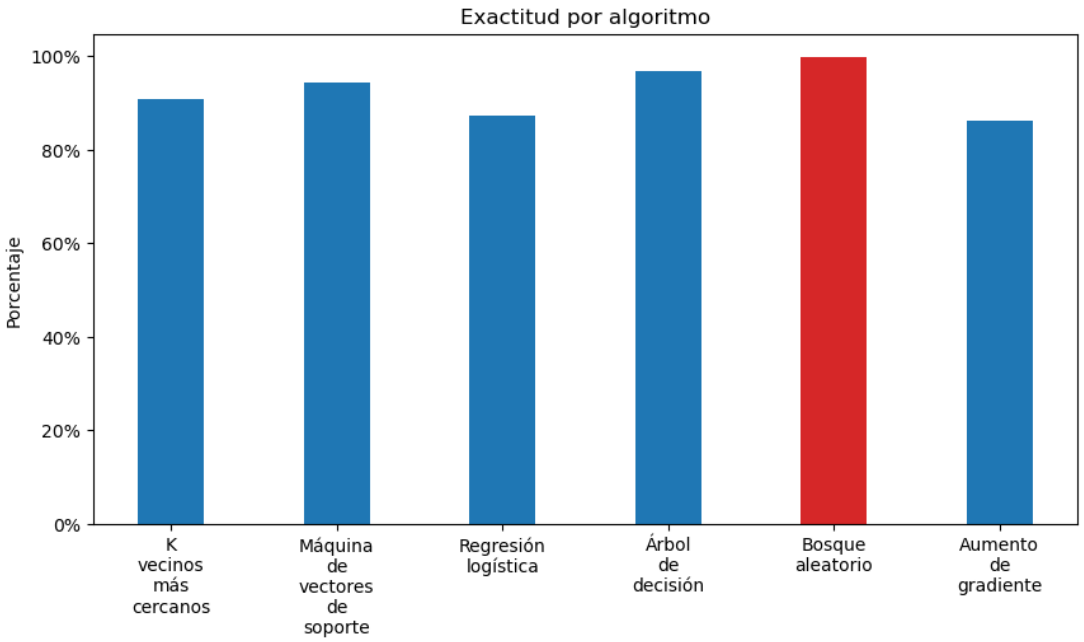
Tomando como base 8 iteraciones con 6 algoritmos diferentes, se realizaron un total de 48 modelos diferentes.

Se hallaron las siguientes métricas por cada algoritmo:

- Exactitud
- Precisión
- Sensibilidad
- Robustez
- Tiempo de filtrado de candidatos
- Tiempo de generación de reporte

4.2.4.9.1 Exactitud

Figura 4.39: Exactitud - Modelado

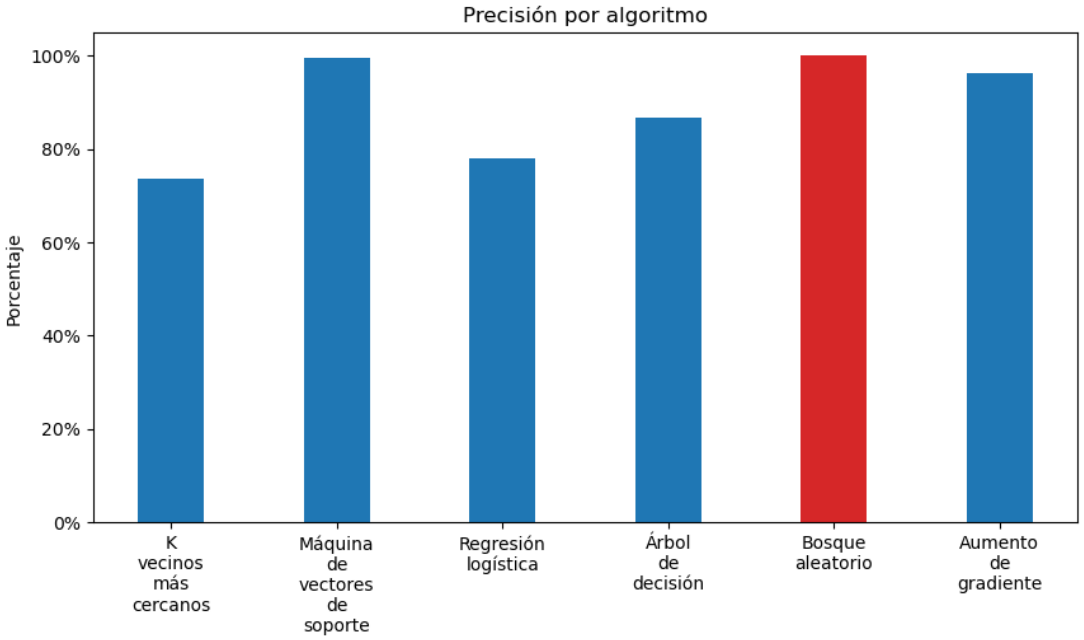


Fuente: La empresa
Elaboración: Propia

De la Figura 4.39, se identifica que el algoritmo que tiene mayor exactitud es el algoritmo Bosque aleatorio (99.70%), seguido de Árbol de decisión (96.74%) y Máquina de vectores de soporte (94.47%).

4.2.4.9.2 Precisión

Figura 4.40: Precisión - Modelado

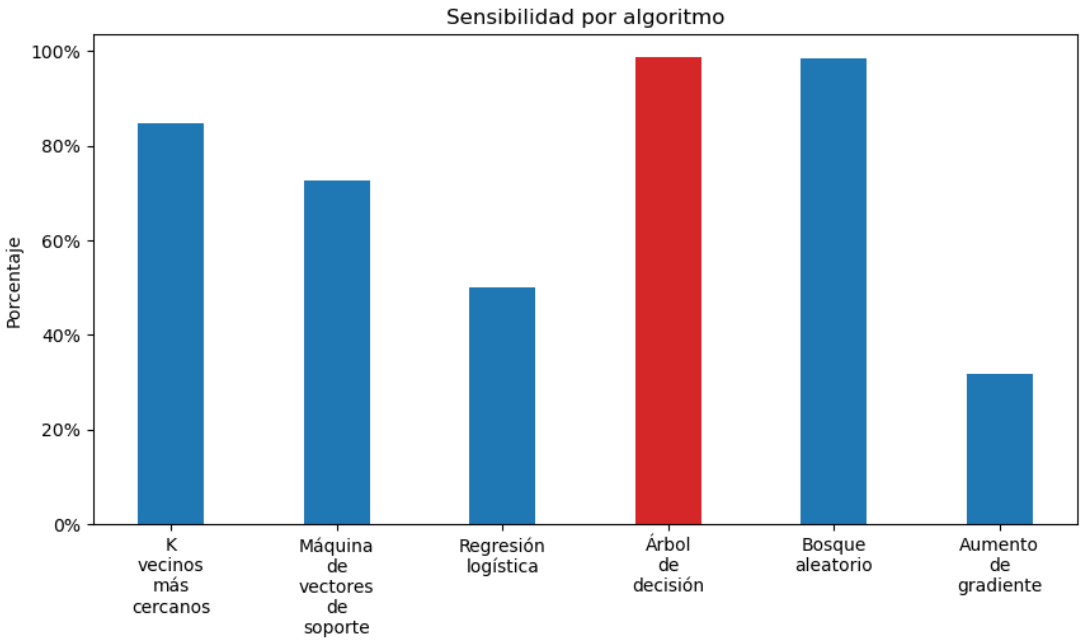


Fuente: La empresa
Elaboración: Propia

De la Figura 4.40, se identifica que el algoritmo que tiene mayor precisión es el algoritmo Bosque aleatorio (100.00 %), seguido de Máquina de vectores de soporte (99.43 %) y Aumento de gradiente (96.25 %).

4.2.4.9.3 Sensibilidad

Figura 4.41: Sensibilidad - Modelado

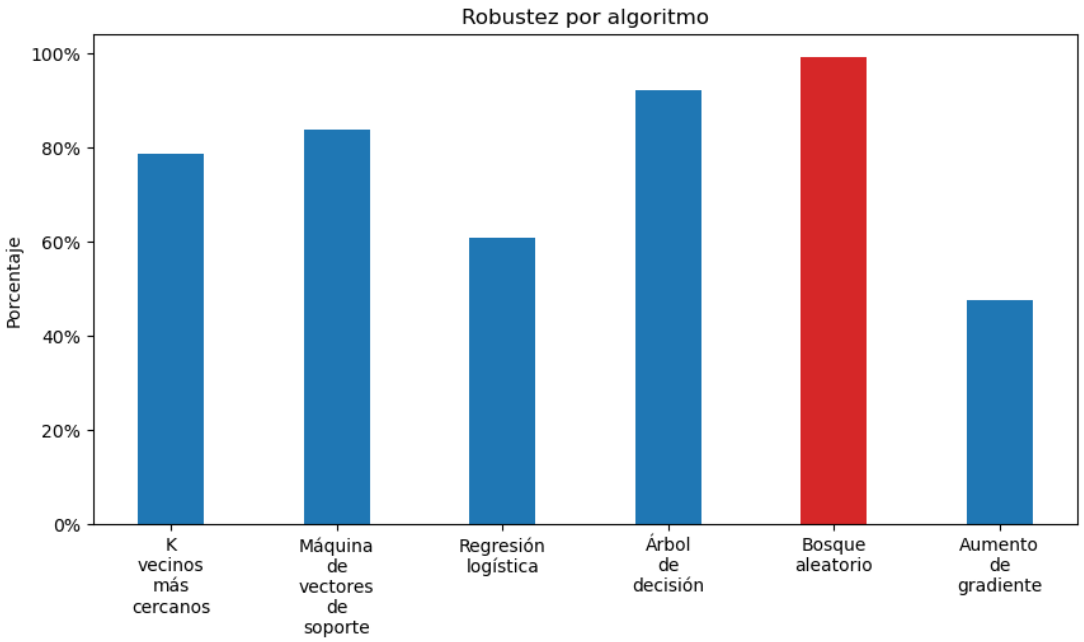


Fuente: La empresa
Elaboración: Propia

De la Figura 4.41, se identifica que el algoritmo que tiene mayor sensibilidad es el algoritmo Árbol de decisión (98.64%), seguido de Bosque Aleatorio (98.47%) y K vecinos más cercanos (84.65%).

4.2.4.9.4 Robustez

Figura 4.42: Robustez - Modelado

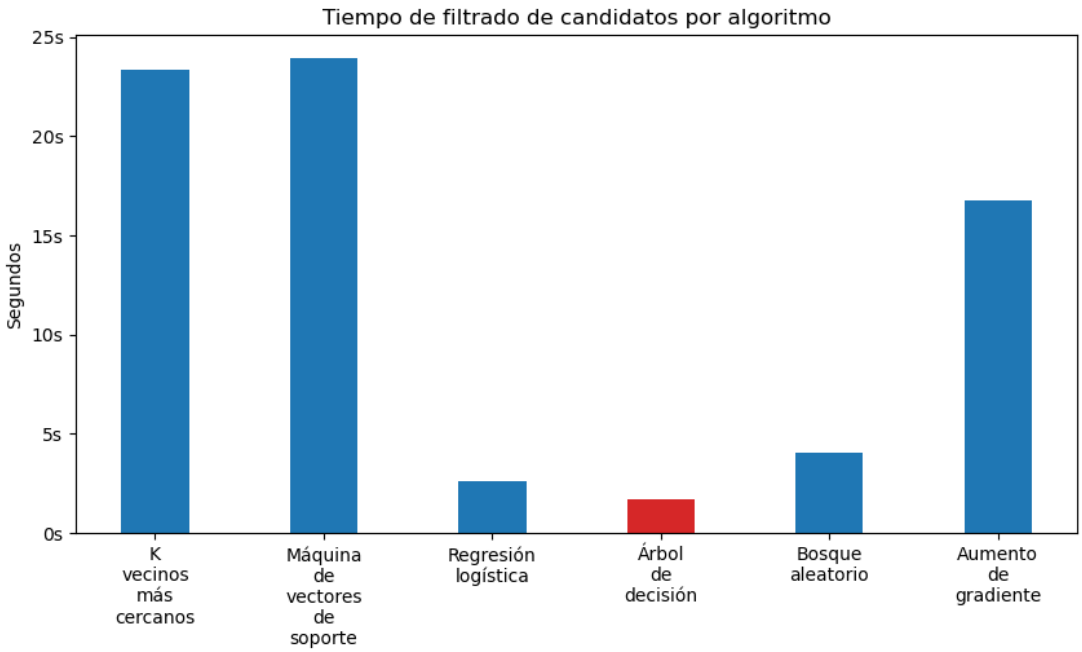


Fuente: La empresa
Elaboración: Propia

De la Figura 4.42, se identifica que el algoritmo que tiene mayor robustez es el algoritmo Bosque aleatorio (99.23%), seguido de Árbol de decisión (92.32%) y Máquina de vectores de soporte (83.88%).

4.2.4.9.5 Tiempo de filtrado de candidatos

Figura 4.43: Tiempo de filtrado de candidatos - Modelado

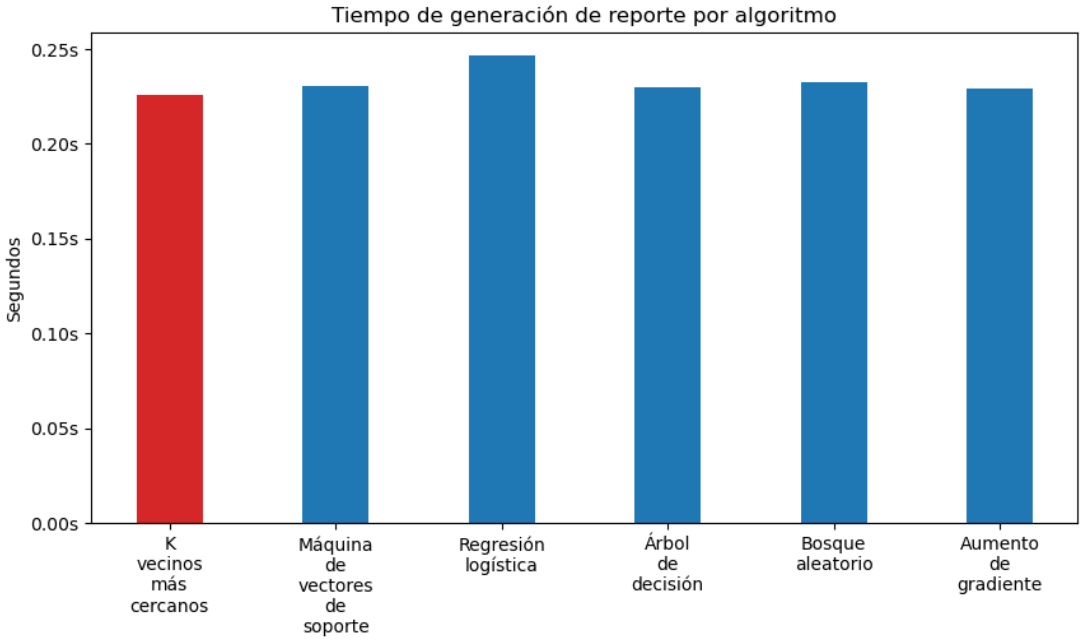


Fuente: La empresa
Elaboración: Propia

De la Figura 4.43, se identifica que el algoritmo que tiene menor tiempo de filtrado de candidatos es el algoritmo Árbol de decisión (1.68s), seguido de Regresión logística (2.62s) y Bosque aleatorio (4.04s).

4.2.4.9.6 Tiempo de generación de reporte

Figura 4.44: Tiempo de generación de reporte - Modelado



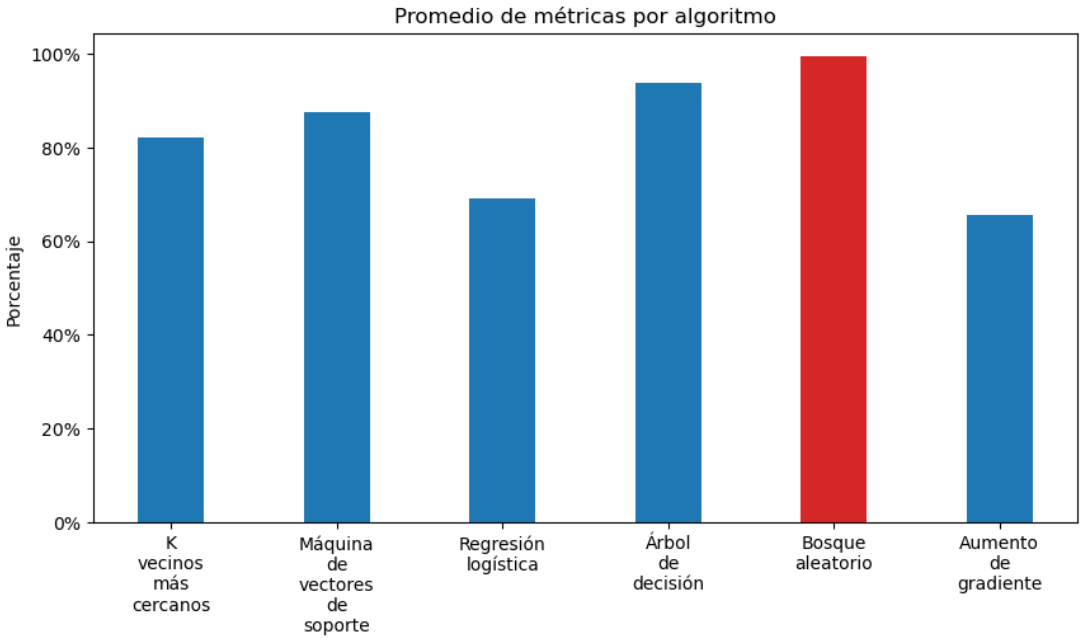
Fuente: La empresa
Elaboración: Propia

De la Figura 4.44, se identifica que el algoritmo que tiene menor tiempo de generación de reporte es el algoritmo K vecinos más cercanos (0.2259s), seguido de Aumento de gradiente (0.2289s) y Árbol de decisión (0.2298s).

4.2.4.9.7 Promedio de métricas

Debido a que las métricas de tiempo de filtrado de candidatos y tiempo de generación de reporte cuentan con una unidad de medida diferente a la de las 4 otras métricas, es pertinente no considerar esa métrica para el promedio de métricas.

Figura 4.45: Promedio de métricas - Modelado



Fuente: La empresa
Elaboración: Propia

Finalmente, hallando un promedio de las 4 principales métricas (exactitud, precisión, sensibilidad y robustez), de la Figura 4.45 se determina que el algoritmo con mayor promedio es el algoritmo Bosque aleatorio (99.35%), superando a otros algoritmos como Árbol de decisión (93.62%), Máquina de vectores de soporte (87.59%) y K vecinos más cercanos (82.02%).

En la Tabla 4.19, se muestra un resumen de las métricas de los 6 modelos entrenados:

Tabla 4.19: Resumen de métricas - Modelado

Siglas	KNN	SVM	LR	DT	RF	GBM
Algoritmo	K vecinos más cercanos	Máquina de vectores de soporte	Regresión logística	Árbol de decisión	Bosque aleatorio	Aumento de gradiente
Registros entrenamiento	9429	9429	9429	9429	9429	9429
Registros prueba	3143	3143	3143	3143	3143	3143
Exactitud	90.94%	94.47%	87.25%	96.74%	99.70%	86.19%
Precisión	73.70%	99.43%	77.95%	86.77%	100.00%	96.25%
Sensibilidad	84.65%	72.56%	49.95%	98.64%	98.47%	31.71%
Robustez	78.78%	83.88%	60.85%	92.32%	99.23%	47.68%
Tiempo de filtrado de candidatos (s)	23.37	23.91	2.62	1.69	4.04	16.78
Tiempo de generación de reporte (s)	0.2259	0.2306	0.2462	0.2298	0.2321	0.2289
Promedio de métricas	82.02%	87.59%	69.00%	93.62%	**99.35%**	65.46%

Fuente: La empresa
Elaboración: Propia

4.2.4.10 Selección del mejor algoritmo

Como se describió en la sección previa, el algoritmo de Bosque aleatorio es el que cuenta con mayor promedio de métricas que los demás, es por ello que el algoritmo seleccionado para el modelo predictivo final es:

- Bosque aleatorio

Teniendo así finalmente las siguientes métricas de este modelo, presentados en la Tabla 4.20:

Tabla 4.20: Métricas finales - Modelado

Algoritmo	Bosque aleatorio
Registros entrenamiento	9429
Registros prueba	3143
Exactitud	99.70%
Precisión	100.00%
Sensibilidad	98.47%
Robustez	99.23%
Tiempo de filtrado de candidatos (s)	4.04
Tiempo de generación de reporte (s)	0.2321
Promedio de métricas	**99.35%**

Fuente: La empresa
Elaboración: Propia

Este modelo será evaluado en mayor profundidad en la siguiente etapa de la metodología, mediante otras técnicas de validación.

4.2.5 Evaluación

Las tareas de desarrollo propias de esta etapa se encuentran en un repositorio remoto de GitHub, el cual se encuentra en la referencia de Nolasco (2023).

Las actividades de evaluación que se contemplaron fueron:

- Aplicación de las técnicas de validación
- Comparación de resultados de las técnicas de validación
- Selección de la mejor técnica de validación

4.2.5.1 Aplicación de las técnicas de validación

Dentro de las técnicas de validación de modelos predictivos, existe la técnica de validación cruzada, la cual consiste en el entrenamiento de varios modelos, utilizando múltiples subconjuntos de los datos de entrada, y evaluarlo con un subconjunto complementario de datos.

Esto se realiza con el fin de determinar si existen errores de pronóstico para datos nuevos, también conocido como sobreajuste del modelo. Además, esto también sirve para validar el

rendimiento real del modelo, entrenándolo con múltiples escenarios o situaciones de los datos iniciales.

Para esta etapa, se evalúa cada modelo, utilizando una técnica de validación cruzada en concreto.

Se utilizarán las 4 técnicas de validación descritas en el marco teórico:

1. K pliegues (KP)
2. K pliegues estratificados (KPE)
3. División aleatoria (DA)
4. División aleatoria estratificada (DAE)

Todas las técnicas usarán $N = 4$, donde N es el número de particiones o pliegues a usar para la validación cruzada.

4.2.5.2 Comparación de resultados de las técnicas de validación

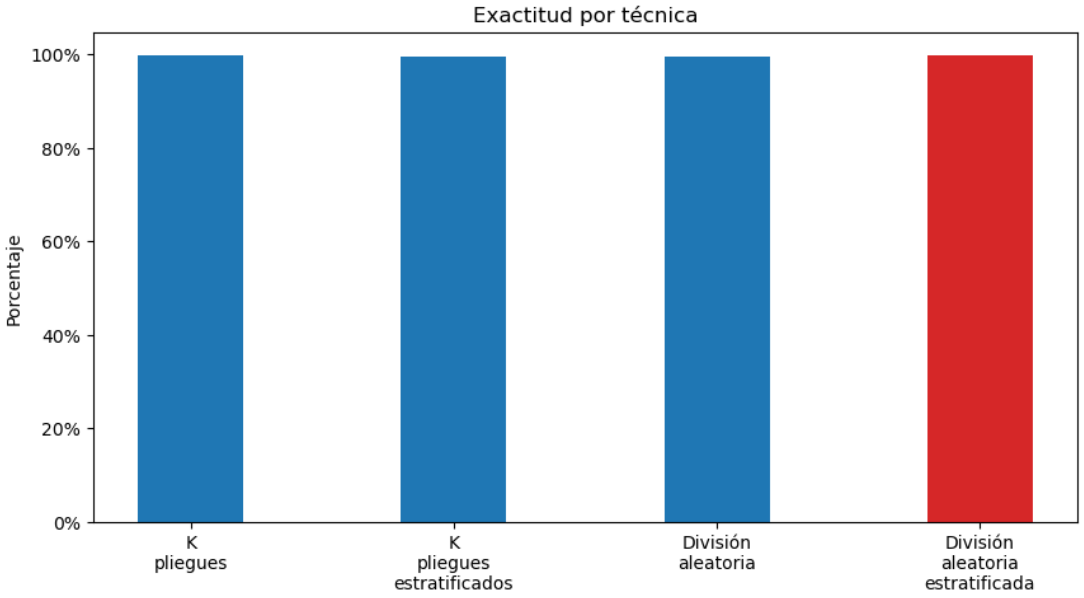
Tomando como base 4 particiones por iteración, donde cada técnica tiene 3 iteraciones, y hay 4 técnicas diferentes, se realizaron un total de 48 modelos diferentes.

Se hallaron las siguientes métricas por cada técnica:

- Exactitud
- Precisión
- Sensibilidad
- Robustez
- Tiempo de filtrado de candidatos
- Tiempo de generación de reporte

4.2.5.2.1 Exactitud

Figura 4.46: Exactitud - Evaluación

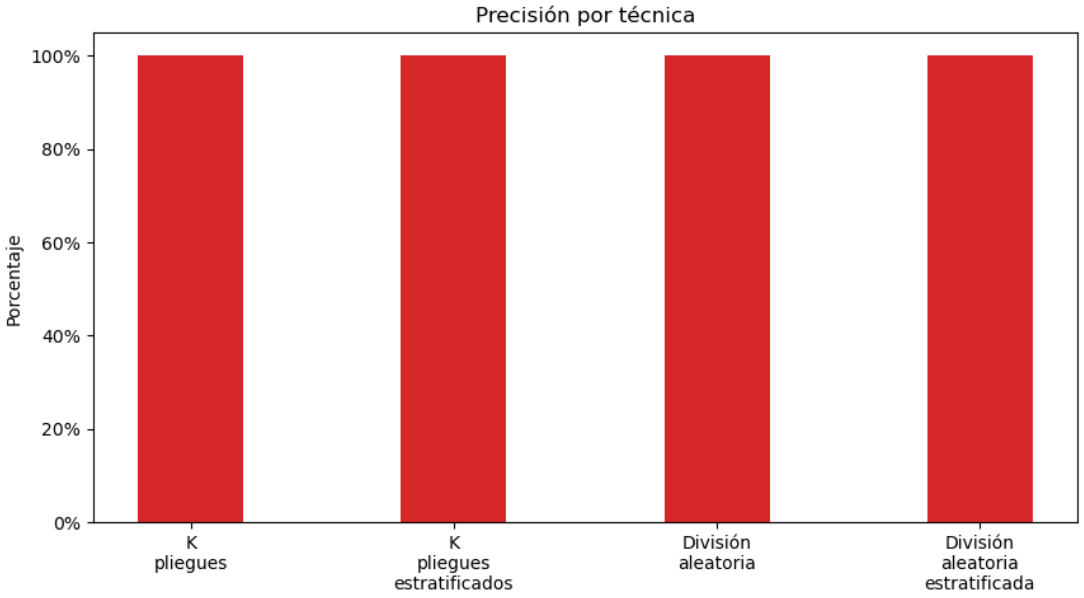


Fuente: La empresa
Elaboración: Propia

De la Figura 4.46, se identifica que la técnica que tiene mayor exactitud es la técnica División aleatoria estratificada (99.78%), seguida de K pliegues (99.71%).

4.2.5.2.2 Precisión

Figura 4.47: Precisión - Evaluación

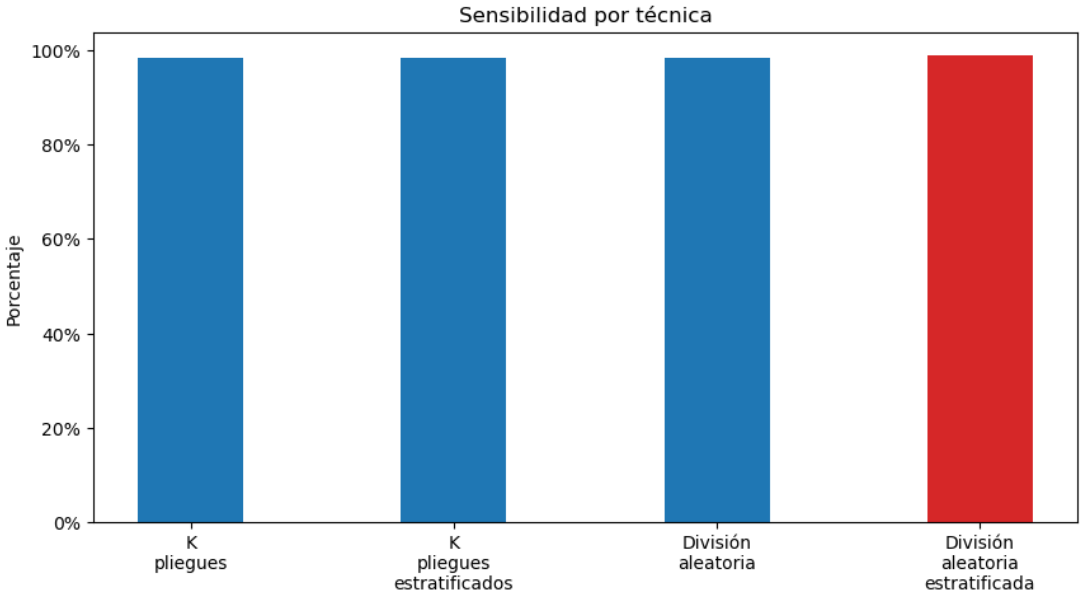


Fuente: La empresa
Elaboración: Propia

De la Figura 4.47, se identifica que todas las técnicas presentan un 100 % de precisión.

4.2.5.2.3 Sensibilidad

Figura 4.48: Sensibilidad - Evaluación

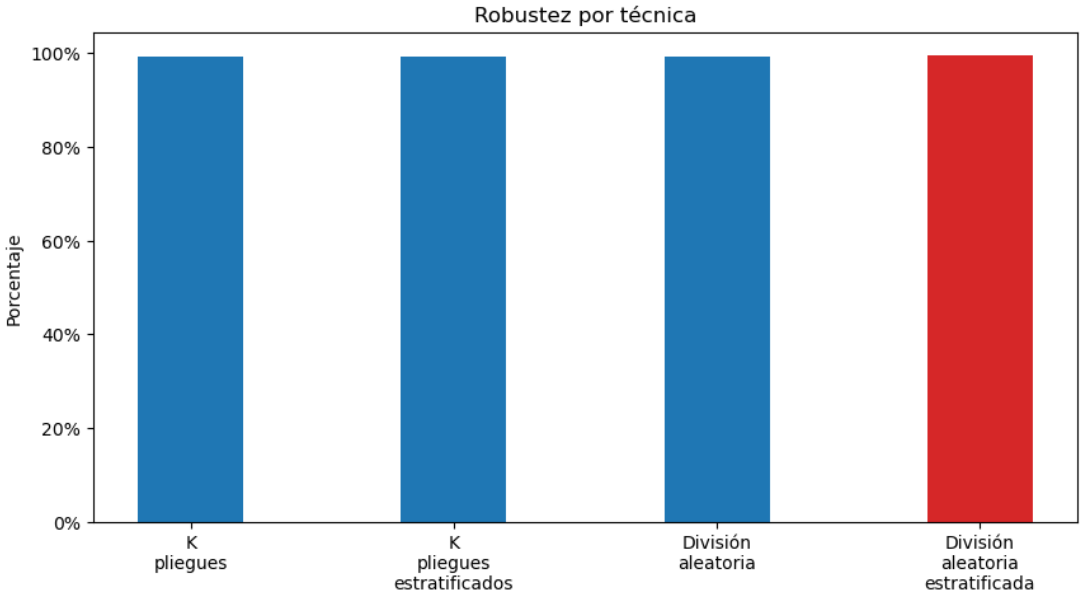


Fuente: La empresa
Elaboración: Propia

De la Figura 4.48, se identifica que la técnica que tiene mayor sensibilidad es la técnica División aleatoria estratificada (98.90%), seguida de K pliegues (98.52%).

4.2.5.2.4 Robustez

Figura 4.49: Robustez - Evaluación

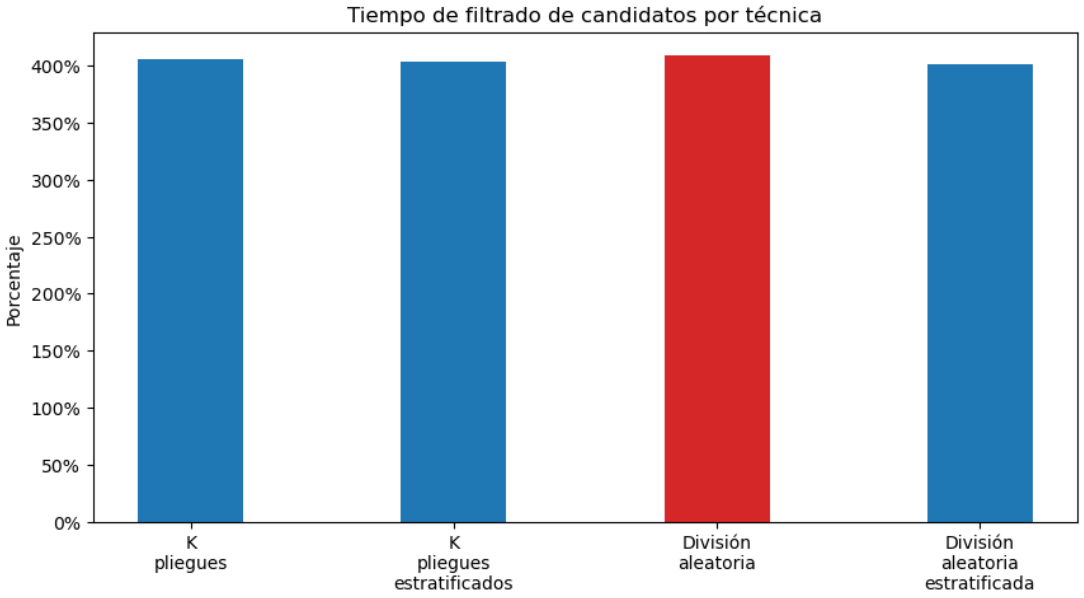


Fuente: La empresa
Elaboración: Propia

De la Figura 4.49, se identifica que la técnica que tiene mayor robustez es la técnica División aleatoria estratificada (99.45 %), seguida de K pliegues (99.25 %).

4.2.5.2.5 Tiempo de filtrado de candidatos

Figura 4.50: Tiempo de filtrado de candidatos - Evaluación

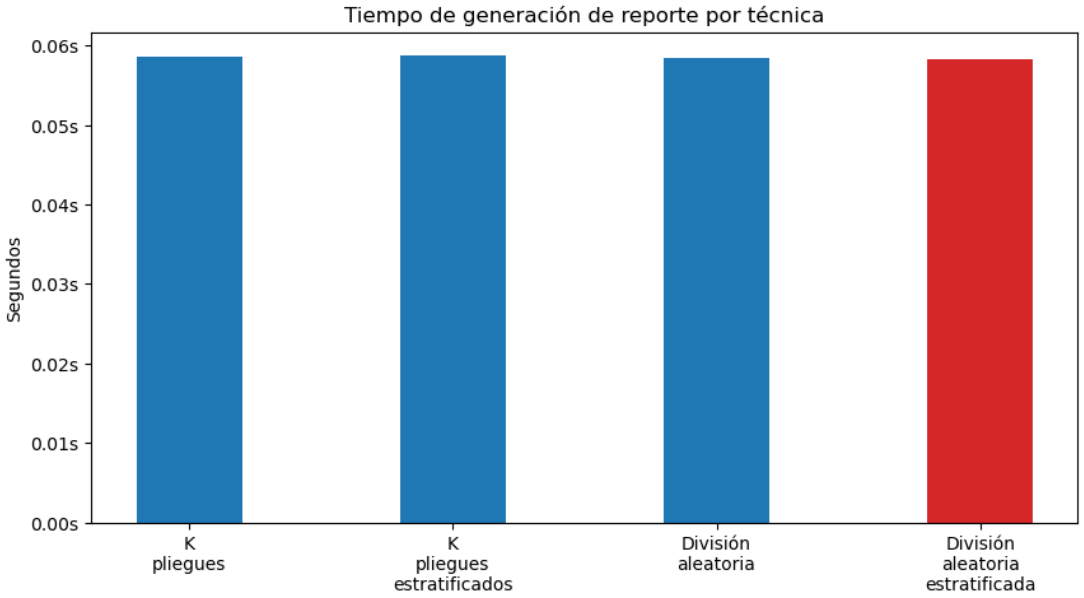


Fuente: La empresa
Elaboración: Propia

De la Figura 4.50, se identifica que la técnica que tiene menor tiempo de filtrado de candidatos es la técnica División aleatoria estratificada (4.00s), seguida de K pliegues estratificados (4.03s).

4.2.5.2.6 Tiempo de generación de reporte

Figura 4.51: Tiempo de generación de reporte - Evaluación



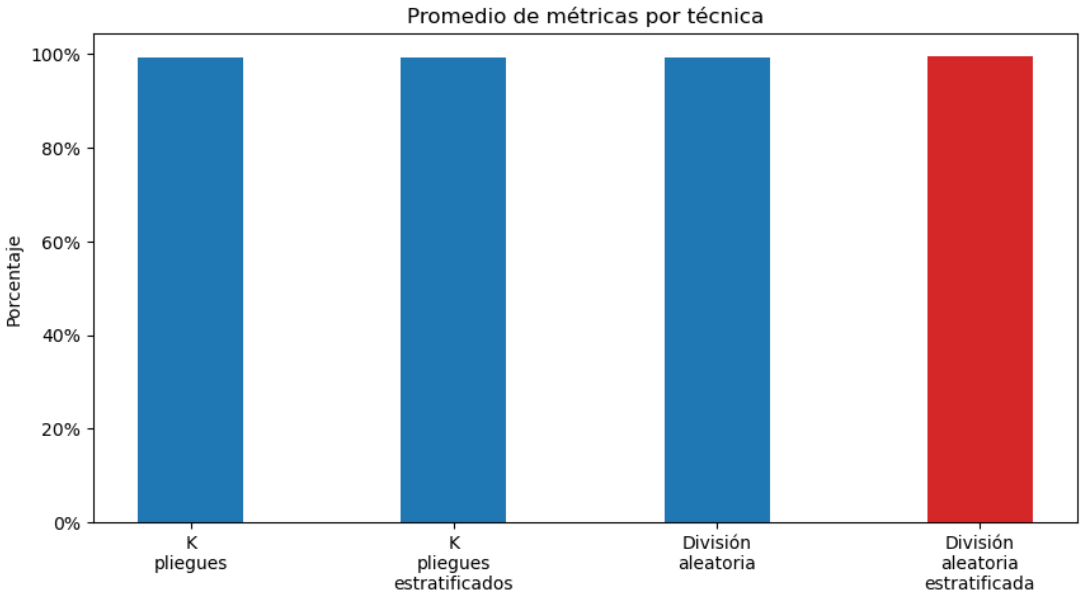
Fuente: La empresa
Elaboración: Propia

De la Figura 4.51, se identifica que la técnica que tiene menor tiempo de generación de reporte es la técnica División aleatoria estratificada (0.0582s), seguida de División aleatoria (0.0584s).

4.2.5.2.7 Promedio de métricas

Debido a que las métricas de tiempo de filtrado de candidatos y tiempo de generación de reporte cuentan con una unidad de medida diferente a la de las 4 otras métricas, es pertinente no considerar esa métrica para el promedio de métricas.

Figura 4.52: Promedio de métricas - Evaluación



Fuente: La empresa
Elaboración: Propia

Finalmente, hallando un promedio de las 4 principales métricas (exactitud, precisión, sensibilidad y robustez), de la Figura 4.52 se determina que la técnica con mayor promedio es la técnica de División aleatoria estratificada (99.53%), superando a otras técnicas como K pliegues (99.37%), División aleatoria (99.31%) y K pliegues estratificados (99.27%).

En la Tabla 4.21, se muestra un resumen de las métricas de las 4 técnicas utilizadas:

Tabla 4.21: Resumen de métricas - Evaluación

Sigla	KP	KPE	DA	DAE
Técnica	K pliegues	K pliegues estratificados	División aleatoria	División aleatoria estratificada
Algoritmo	Bosque aleatorio	Bosque aleatorio	Bosque aleatorio	Bosque aleatorio
Registros entrenamiento	9429	9429	9429	9429
Registros prueba	3143	3143	3143	3143
Exactitud	99.71%	99.66%	99.68%	99.78%
Precisión	100.00%	100.00%	100.00%	100.00%
Sensibilidad	98.52%	98.29%	98.38%	98.90%
Robustez	99.25%	99.13%	99.18%	99.45%
Tiempo de filtrado de candidatos (s)	4.06	4.04	4.09	4.01
Tiempo de generación de reporte (s)	0.0585	0.0587	0.0584	0.0582
Promedio de métricas	99.37%	99.27%	99.31%	**99.53%**

Fuente: La empresa
Elaboración: Propia

4.2.5.3 Selección de la mejor técnica de validación

Como se describió en la sección previa, la técnica de División aleatoria estratificada es la que cuenta con mayor promedio de métricas que los demás, es por ello que la técnica de validación seleccionada para el modelo predictivo final es:

- División aleatoria estratificada

Teniendo así finalmente las siguientes métricas de esta técnica, presentados en la Tabla 4.22:

Tabla 4.22: Métricas finales - Evaluación

Algoritmo	Bosque aleatorio
Técnica	División aleatoria estratificada
Registros entrenamiento	9429
Registros prueba	3143
Exactitud	99.78%
Precisión	100.00%
Sensibilidad	98.90%
Robustez	99.45%
Tiempo de filtrado de candidatos (s)	4.01
Tiempo de generación de reporte (s)	0.0582
Promedio de métricas	**99.53%**

Fuente: La empresa
Elaboración: Propia

De igual manera, al comparar el modelo final obtenido con los objetivos y criterios de éxito, se corrobora que cumple con los 5 criterios, por lo que se cumplen los objetivos.

4.2.6 Despliegue

Las actividades de despliegue que se contemplaron fueron:

- Revisión de la infraestructura a usar
- Justificación de los programas a usar
- Realización del despliegue local

4.2.6.1 Revisión de la infraestructura a usar

Se realizará un despliegue local en la computadora del tesista, con el fin de probar los desarrollos realizados en las 4 secciones anteriores.

Esta computadora cuenta con las características descritas en la Tabla 4.23:

Tabla 4.23: Infraestructura, periféricos y programas - Despliegue

Infraestructura	**Tarjeta madre**	Gigabyte H610M H DDR4
	Procesador	Intel(R) Core(TM) i5-12400F
	Memoria RAM	HyperX Fury Black DDR4 16GB X 2
	Tarjeta de video	NVIDIA GeForce RTX 3050 8GB
	Almacenamiento	1 SSD 1TB WD GREEN SN350 1 HDD 2TB Western Digital Blue 1 HDD 2TB Seagate Barracuda Green
Periféricos	**Pantalla**	ASUS VG248QG
	Teclado	HyperX Alloy Origins Core
	Ratón	Logitech G203 LIGHTSYNC
	Audífonos	Sony WH-CH510
	Micrófono	iBlue 46169-BK-V2
Programas	**Sistema operativo**	Windows 10 Pro 22H2 19045.3086
	Anaconda	Anaconda3 2023.03-1
	Python	3.10.9 64-bit
	Visual Studio Code	1.79.2
	Git	2.40.1.windows.1

Fuente: La empresa
Elaboración: Propia

Cabe recalcar que ninguno de estos componentes fue modificado desde el inicio hasta el final del desarrollo de la metodología.

4.2.6.2 Justificación de los programas a usar

El sistema operativo a usar es Windows 10 Pro, en su versión 22H2. Este es un sistema operativo desarrollado por Microsoft, y es el principal sistema operativo utilizado en la actualidad a nivel mundial (StatCounter, 2023).

Este sistema operativo fue lanzado oficialmente el 15 de julio del 2015, y actualmente aún cuenta con soporte activo, sin embargo, Microsoft ya ha confirmado que este cesará el 14 de octubre de 2025 (Microsoft Learn, 2023b).

Es debido a ser el sistema más estable de Windows que cuenta con vigencia, que se eligió y se instaló para el despliegue. El logo se presenta en la Figura 4.53:

Figura 4.53: Logo de Windows 10

Fuente: Wikimedia Commons (2020b)

De igual manera, se instaló el entorno de trabajo Anaconda, en su versión 2023.03-1. Es un entorno de trabajo desarrollado por Continuum Analytics, y en los últimos años se ha vuelto un estándar para aplicaciones de la computación científica (ciencia de datos, desarrollo de modelos predictivos, visualización de datos, manejo de grandes volúmenes de datos, etc.).

Esta versión fue lanzada el 24 de abril del 2023, como una versión oficial de soporte a largo plazo (LTS) y según palabras de los mismos desarrolladores, se da continuidad a las versiones hasta por 10 años, una vez lanzadas (Nolan, 2022).

Es debido a ser la versión más reciente estable, que se eligió e instaló para el despliegue. El logo se presenta en la Figura 4.54:

Figura 4.54: Logo de Anaconda

Fuente: Anaconda (2023)

Junto con la instalación de Anaconda, el lenguaje de programación Python también viene incluido, en su versión 3.10.9, para los sistemas operativos de 64 bits. Es un lenguaje de programación de propósito general, ampliamente conocido y utilizado en múltiples sectores (desarrollo web, ciencia de datos, automatización de procesos, visión por computadora, videojuegos).

Esta versión 3.10.9 fue lanzada el 6 de diciembre del 2022, como una versión oficial, y según el portal de Python, tienen estimado el soporte a esta versión hasta octubre del 2026 (Galindo, 2020).

Es debido a ser una versión estable, y que viene instalada con el entorno de Anaconda, que se eligió e instaló para el despliegue. El logo se presenta en la Figura 4.55:

Figura 4.55: Logo de Python

Fuente: Python (2023)

Todos los desarrollos fueron creados uno a uno mediante el uso del editor de código Visual Studio Code, desarrollado por Microsoft. Este es un editor muy utilizado en el ámbito de programación, y conocido por la posibilidad de compilación y ejecución de múltiples lenguajes de programación (C++, Java, Python, Javascript, SQL, etc.).

La versión a utilizar es la 1.79.2, lanzada el 14 de junio del 2023. La primera versión fue lanzada el 14 de abril del 2016, y Microsoft aún tiene planeado mantener este editor de código con soporte activo, para todas sus versiones, es decir, aún no tiene fecha de fin de su ciclo de vida (Microsoft Learn, 2023a).

Es debido a la alta integración con el lenguaje de Python, además de la versatilidad y facilidad de uso, que se eligió e instaló para el despliegue. El logo se presenta en la Figura 4.56:

Figura 4.56: Logo de Visual Studio Code

Fuente: Microsoft (2021)

Finalmente, para la gestión de versiones de cada etapa del desarrollo, además del manejo de múltiples ramas o características del proyecto, se empleó el sistema de control de versiones Git. Es un sistema de versiones ampliamente utilizado en proyectos de desarrollo de software, pero incluso es posible usarlo para proyectos de manera general.

La versión a utilizar es la 2.40.1.windows.1, lanzada el 13 de marzo del 2023, siendo la última versión oficial publicada. Actualmente no se ha definido una fecha para el fin del soporte, pero existen versiones como la 2.30, lanzada el 27 de diciembre del 2020, que aun cuentan con soporte activo (Wikipedia, 2023). Por lo que se espera que el soporte para esta versión dure al menos unos 2 años más, desde la fecha.

Es debido a su uso en el mercado, y fácil integración y uso, que se eligió e instaló para el despliegue. El logo se presenta en la Figura 4.57:

Figura 4.57: Logo de Git

Fuente: Git (2023)

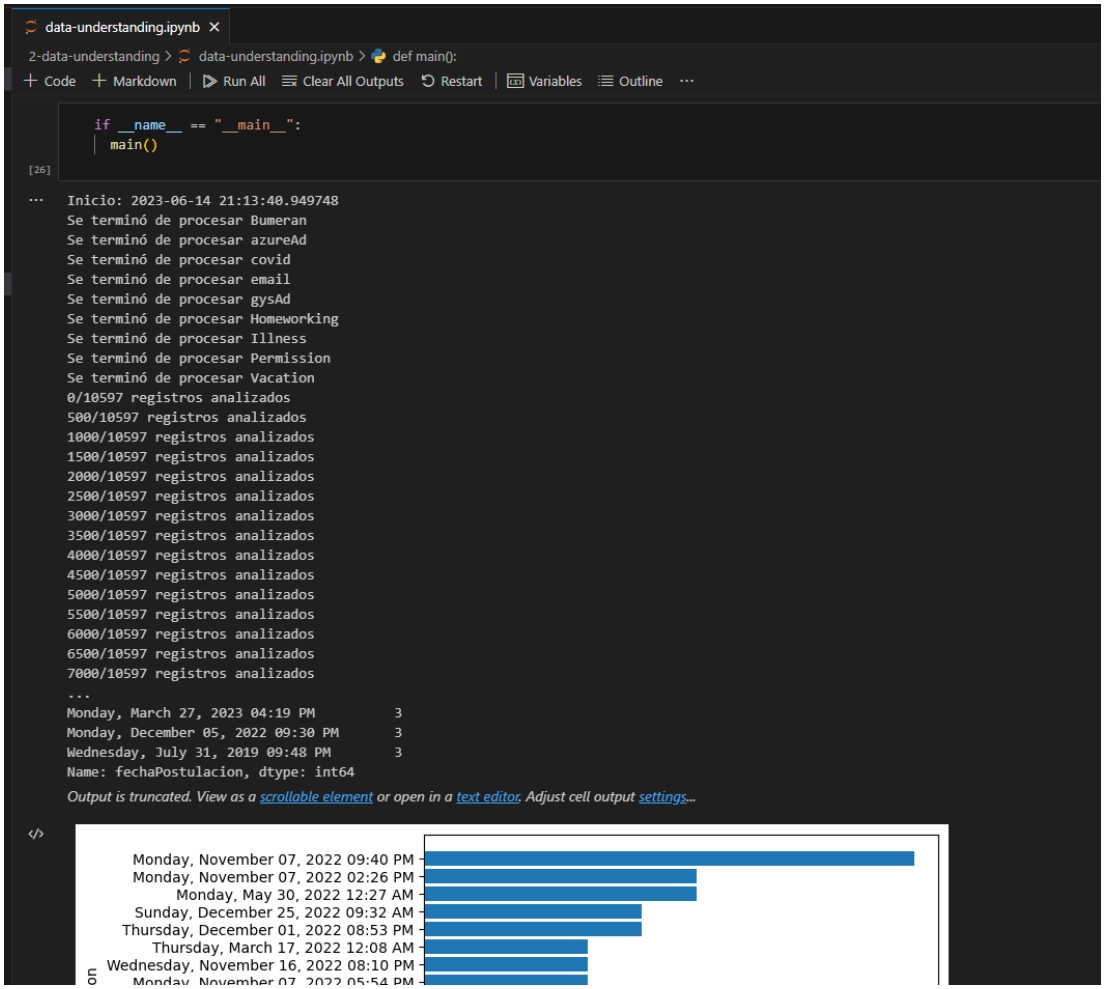
Una vez determinado este análisis, se procede con el despliegue.

4.2.6.3 Realización del despliegue local

Se ejecutaron estos desarrollos de los 4 capítulos previos, uno a uno, desde el editor de código Visual Studio Code.

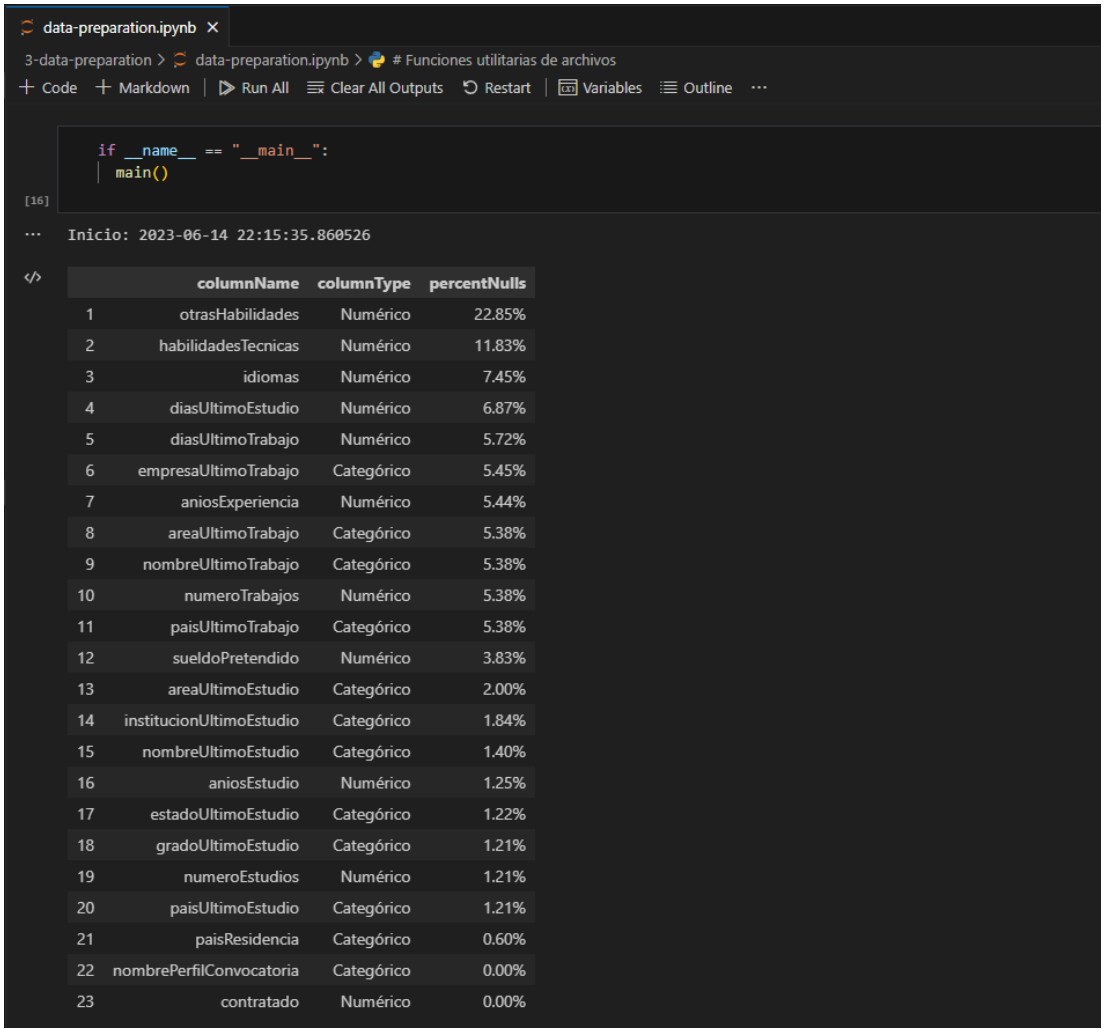
En la Figura 4.58, Figura 4.59, Figura 4.60, Figura 4.61 y Figura 4.62, se muestran las evidencias de las ejecuciones de las 4 etapas previas, además del reporte final:

Figura 4.58: Comprensión de datos - evidencia



Fuente: La empresa
Elaboración: Propia

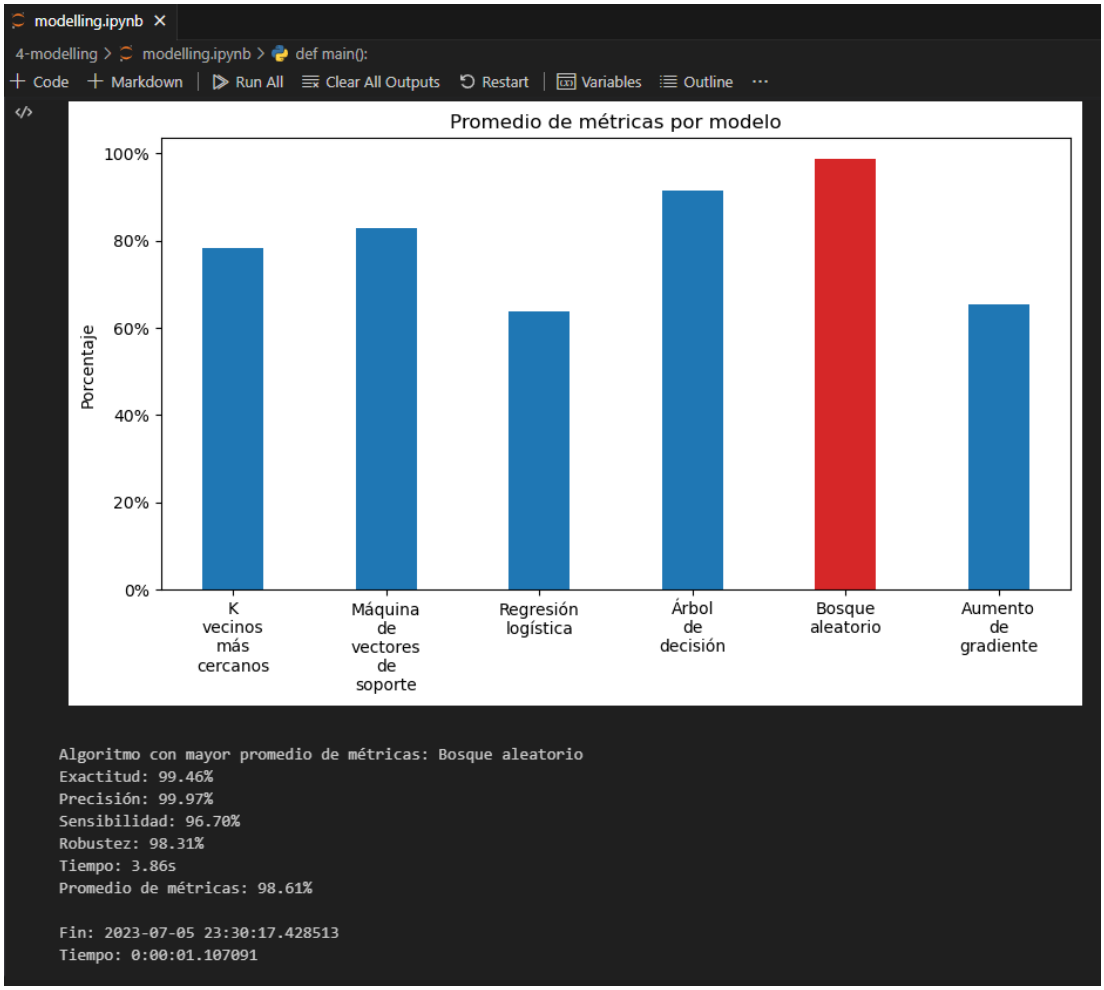
Figura 4.59: Preparación de datos - evidencia



	columnName	columnType	percentNulls
1	otrasHabilidades	Numérico	22.85%
2	habilidadesTecnicas	Numérico	11.83%
3	idiomas	Numérico	7.45%
4	diasUltimoEstudio	Numérico	6.87%
5	diasUltimoTrabajo	Numérico	5.72%
6	empresaUltimoTrabajo	Categórico	5.45%
7	aniosExperiencia	Numérico	5.44%
8	areaUltimoTrabajo	Categórico	5.38%
9	nombreUltimoTrabajo	Categórico	5.38%
10	numeroTrabajos	Numérico	5.38%
11	paisUltimoTrabajo	Categórico	5.38%
12	sueldoPretendido	Numérico	3.83%
13	areaUltimoEstudio	Categórico	2.00%
14	institucionUltimoEstudio	Categórico	1.84%
15	nombreUltimoEstudio	Categórico	1.40%
16	aniosEstudio	Numérico	1.25%
17	estadoUltimoEstudio	Categórico	1.22%
18	gradoUltimoEstudio	Categórico	1.21%
19	numeroEstudios	Numérico	1.21%
20	paisUltimoEstudio	Categórico	1.21%
21	paisResidencia	Categórico	0.60%
22	nombrePerfilConvocatoria	Categórico	0.00%
23	contratado	Numérico	0.00%

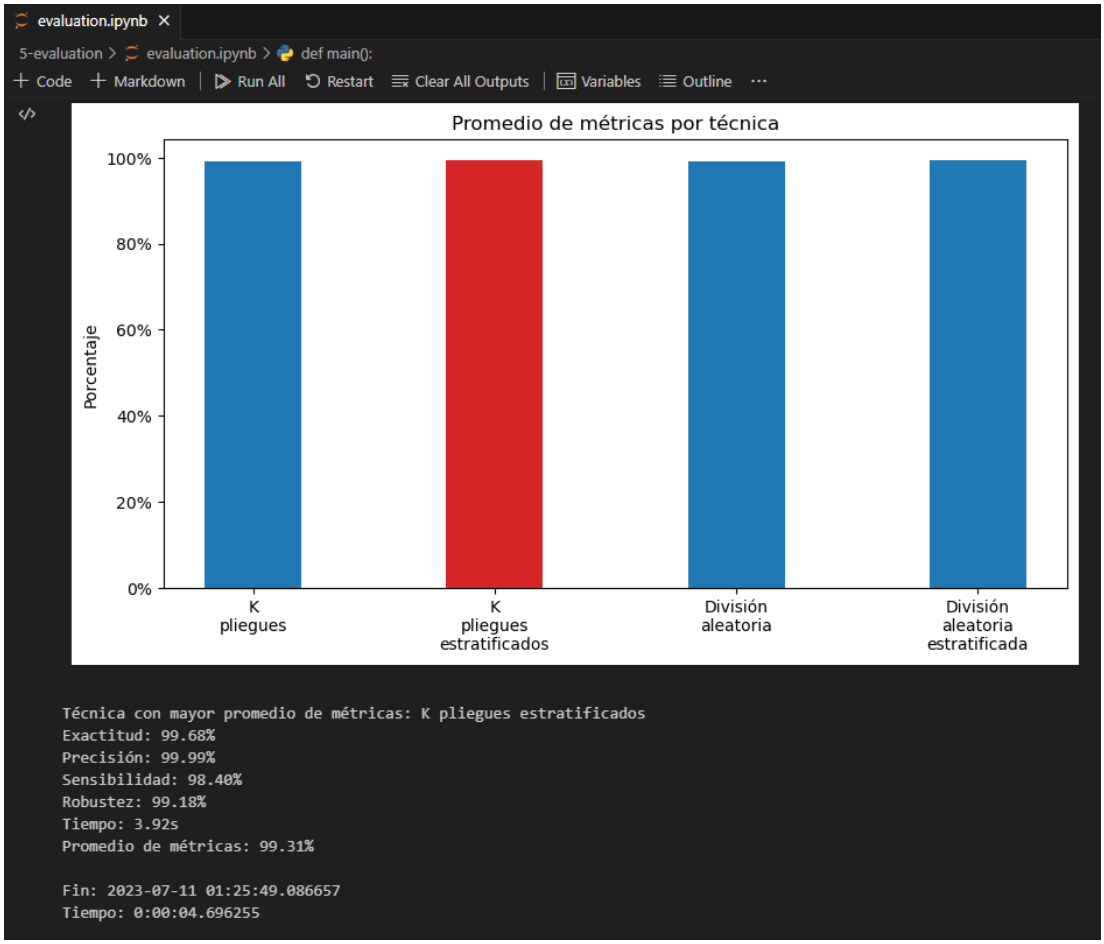
Fuente: La empresa
Elaboración: Propia

Figura 4.60: Modelado - evidencia



Fuente: La empresa
Elaboración: Propia

Figura 4.61: Evaluación - evidencia



Fuente: La empresa
Elaboración: Propia

Figura 4.62: Reporte generado - evidencia

fechaPostulacion	idConvocatoria	nombrePerfilConvocatoria	nombreCompleto	numeroCelular	sueldoPretendido	contratado
30-09-2020 22:03	1114087358	Analista Programador .Net	Juan Carlos Sotelo Canales	51 938589577	3000	Si
30-09-2020 21:48	1114087358	Analista Programador .Net	Edgar Alonso Camarena Nomberto	51 934611558	3500	Si
30-09-2020 21:30	1114087358	Analista Programador .Net	Heber Gonzalo Cachi Cruzado	51 961828127	2300	Si
30-09-2020 21:10	1114087358	Analista Programador .Net	Diego Gutierrez	051 994886846	4000	No
30-09-2020 3:09	1114084953	Especialista En Analitica	Carlos Dominguez	51 952951269	5000	No
28-09-2022 9:08	1115407679	Ejecutivo Comercial Jr	Cesar Alexander Diaz Carbajal	51 947190913	2000	No
28-09-2022 16:45	1115402634	Analista De Calidad	Victor Garcia	+51 916237071	1000	Si
21-09-2022 12:26	1115406770	Programador .Net	Ivan Cesar Ascencio Tiza	51 920802197	4000	Si
21-09-2022 11:22	1115338594	Data Architect	Franco Octavio Gordillo Calagua	51 918626232	1100	No
21-09-2022 16:38	1115407679	Ejecutivo Comercial Jr	Andres Valcarcel Minness	51 973946181	1500	Si
16-09-2020 23:27	1114047548	Jefe De Proyecto - Analytics & Ai	Sandro Flavio Angulo Ccahuana	01 946094254	12000	Si
16-09-2020 21:17	1114067993	Especialista Devops	Carlos Godoy	51 926893401	8000	Si
16-09-2020 18:59	1114069475	Analista De Ciberseguridad	Angel Junior Ruiz Caldas	01 991439332	3700	Si
14-09-2022 12:46	1115362010	Ejecutivo Comercial Ti	Luis Huamani Aliaga	51 989099197	12000	Si
14-09-2022 20:11	1115396754	Especialista Ciberseguridad	Dennis Henry Carrera Travezano	01 967699337	2500	Si
14-09-2022 14:25	1115319638	Especialista Desarrollador De Software	Juan Jose Briceno Ochoa	51 994762302	2000	Si
11-09-2019 22:54	1113534187	Sql Server Consultant	Juan Manuel Rafael Fabian	51 993452316	8300	Si
11-09-2019 18:13	1113563516	Cloud Specialist	Renzo Portilla	051 941468147	1	Si
11-09-2019 17:18	1113563516	Cloud Specialist	George Felix Huayna Montes	051 954799320	6000	Si
11-09-2019 14:29	1113563516	Cloud Specialist	Angelo Guillermo Minaya Montes	51 998378949	1399	No
09-09-2020 9:43	1114058833	Lider Tecnico	Henry Fernando Terrones Cortez	511 949346218	1	Si
08-09-2021 21:28	1114632424	Practicante De Recursos Humanos	Hugo Luis Perez Ibarra	51 983504610	2000	Si
08-09-2021 14:33	1114632432	Asistente De Recursos Humanos	Edgar Andree Quinto Porras	51 943094289	1000	Si
07-09-2022 22:08	1115352072	Lider Tecnico	Jesus Abel Mejia Ramirez	511 959332129	6000	Si
07-09-2022 17:36	1115371204	Data Engineer Sr.	Joel Gutierrez Espiritu	+ 913414177	8000	Si
07-09-2022 15:55	1115338587	Jefe De Proyecto Para El Area De Analytics & Ai	Dennis Henry Carrera Travezano	01 967699337	3500	Si
04-09-2019 0:51	1113393261	Cloud Specialist	Jorge Monzon Nanez	01 938512855	7000	Si
04-09-2019 12:35	1113532025	Desarrollador Java	Jesus Jhonatan Cahuana Auquipuma	051 992257680	4500	Si
04-09-2019 10:55	1113393261	Cloud Specialist	Rodrigo Pinedo Diaz	01 953237774	2500	Si
04-09-2019 7:27	1113532025	Desarrollador Java	Fernando Bustamante	01 965713193	4500	Si
04-09-2019 15:35	1113534187	Sql Server Consultant	Fernando Pinto Polar	0051 997488448	3500	Si
02-09-2020 23:03	1114047602	Ejecutivo/A Comercial Ti	Carlos Chavez	51 987385839	3000	Si
02-09-2020 9:05	1114005544	Especialista Dynamics 365 Business Central	Ehytel Celestino Vilca Mosquera	51 962971387	3500	Si
02-09-2020 20:11	1114019175	Especialista En Analitica	Juan Bernardo Marca Andia	51 933963971	3000	Si
02-09-2020 19:42	1114047602	Ejecutivo/A Comercial Ti	Jose Martin Barriga Huaman	51 987097807	2500	Si

Fuente: La empresa
Elaboración: Propia

Este reporte generado de manera automatizada contiene los datos del candidato, del puesto, y una columna llamada *contratado* que determina si se pronostica contratarlo o no al candidato para ese puesto.

Este reporte se compartirá al personal de RRHH, el cual verá la recomendación por parte del modelo, para posteriormente tomar la decisión final.

CAPÍTULO V
RESULTADOS

5.1 RESULTADOS DESCRIPTIVOS

En el estudio se aplicó un modelo predictivo para evaluar la exactitud, precisión, sensibilidad, robustez, tiempo de filtrado de candidatos y tiempo de generación de reporte del proceso de selección de personal. Para ello se aplicó un PreTest que permite conocer las condiciones iniciales del indicador, posteriormente se aplicó el modelo predictivo final validado, y luego se aplicó un PostTest, registrando nuevamente la exactitud, precisión, sensibilidad, robustez, tiempo de filtrado de candidatos y tiempo de generación de reporte del proceso de selección de personal.

Las tareas de desarrollo propias de esta etapa se encuentran en un repositorio remoto de GitHub, el cual se encuentra en la referencia de Nolasco (2023).

5.1.1 Medidas descriptivas

5.1.1.1 Exactitud

Los resultados de las medidas descriptivas se describen en la Tabla 5.1:

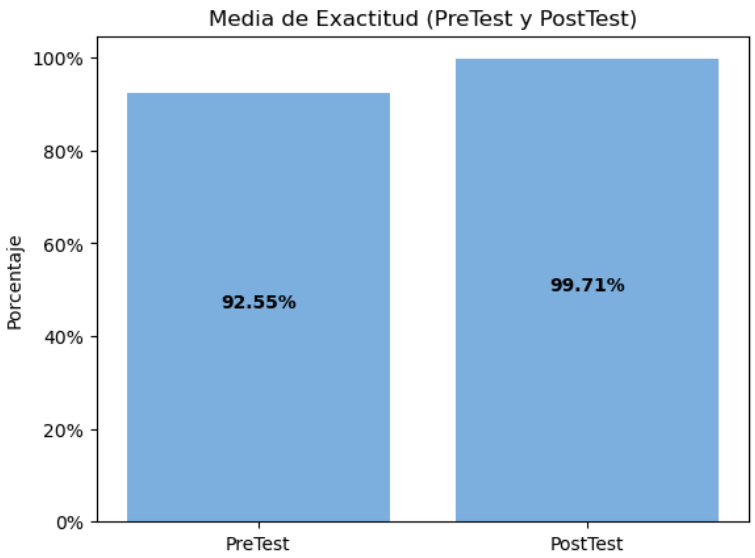
Tabla 5.1: Medidas descriptivas - Exactitud

Nombre	Muestra	Mínimo	Máximo	Media	Desviación estándar
PreTest_Exactitud	48	0.8514	0.9990	0.9255	0.0490
PostTest_Exactitud	48	0.9898	1.0000	0.9971	0.0017

Fuente: La empresa
Elaboración: Propia

En el caso de la exactitud del proceso de selección de personal, en el PreTest se obtuvo un valor de 92.55%, mientras que en el PostTest es de 99.71%, lo cual denota una leve diferencia entre el antes y después. Esta comparación se puede visualizar en la Figura 5.1:

Figura 5.1: Comparación de medias - Exactitud



Fuente: La empresa
Elaboración: Propia

5.1.1.2 Precisión

Los resultados de las medidas descriptivas se describen en la Tabla 5.2:

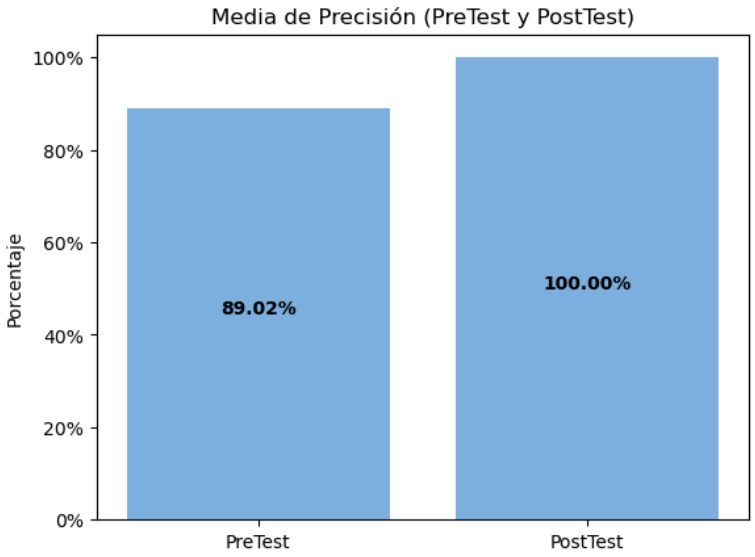
Tabla 5.2: Medidas descriptivas - Precisión

Nombre	Muestra	Mínimo	Máximo	Media	Desviación estándar
PreTest_Precision	48	0.7012	1.0000	0.8902	0.1046
PostTest_Precision	48	1.0000	1.0000	1.0000	0.0000

Fuente: La empresa
Elaboración: Propia

En el caso de la precisión del proceso de selección de personal, en el PreTest se obtuvo un valor de 89.02%, mientras que en el PostTest es de 100.00%, lo cual denota una significativa diferencia entre el antes y después. Esta comparación se puede visualizar en la Figura 5.2:

Figura 5.2: Comparación de medias - Precisión



Fuente: La empresa
Elaboración: Propia

5.1.1.3 Sensibilidad

Los resultados de las medidas descriptivas se describen en la Tabla 5.3:

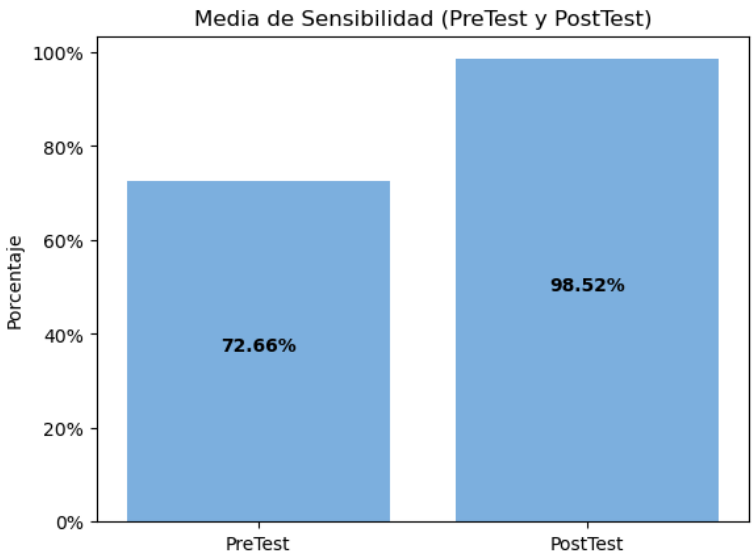
Tabla 5.3: Medidas descriptivas - Sensibilidad

Nombre	Muestra	Mínimo	Máximo	Media	Desviación estándar
PreTest_Sensibilidad	48	0.2850	0.9952	0.7266	0.2481
PostTest_Sensibilidad	48	0.9490	1.0000	0.9852	0.0087

Fuente: La empresa
Elaboración: Propia

En el caso de la sensibilidad del proceso de selección de personal, en el PreTest se obtuvo un valor de 72.66%, mientras que en el PostTest es de 98.52%, lo cual denota una considerable diferencia entre el antes y después. Esta comparación se puede visualizar en la Figura 5.3:

Figura 5.3: Comparación de medias - Sensibilidad



Fuente: La empresa
Elaboración: Propia

5.1.1.4 Robustez

Los resultados de las medidas descriptivas se describen en la Tabla 5.4:

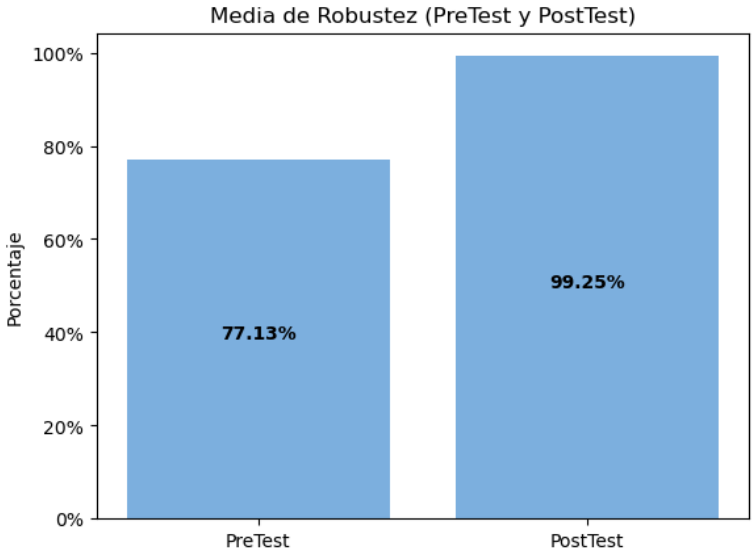
Tabla 5.4: Medidas descriptivas - Robustez

Nombre	Muestra	Mínimo	Máximo	Media	Desviación estándar
PreTest_Robustez	48	0.4339	0.9976	0.7713	0.1785
PostTest_Robustez	48	0.9739	1.0000	0.9925	0.0045

Fuente: La empresa
Elaboración: Propia

En el caso de la robustez del proceso de selección de personal, en el PreTest se obtuvo un valor de 77.13 %, mientras que en el PostTest es de 99.25 %, lo cual denota una significativa diferencia entre el antes y después. Esta comparación se puede visualizar en la Figura 5.4:

Figura 5.4: Comparación de medias - Robustez



Fuente: La empresa
Elaboración: Propia

5.1.1.5 Tiempo de filtrado de candidatos

Los resultados de las medidas descriptivas se describen en la Tabla 5.5:

Tabla 5.5: Medidas descriptivas - Tiempo de filtrado de candidatos

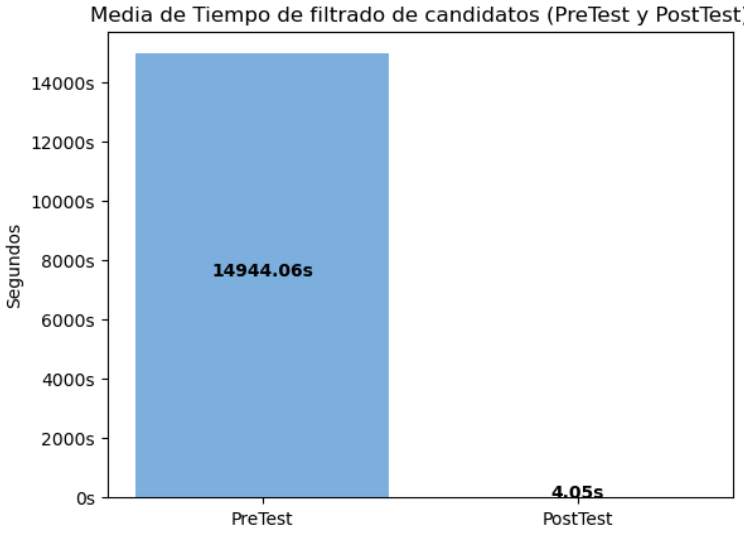
Nombre	Muestra	Mínimo	Máximo	Media	Desviación estándar
PreTest_Tiempo_filtrado_candidatos_manual	48	13756	16081	14944.06	750.79
PostTest_Tiempo_filtrado_candidatos_automatizado	48	3.8351	4.3077	4.0493	0.1114

Fuente: La empresa
Elaboración: Propia

En el caso del tiempo de filtrado de candidatos del proceso de selección de personal, en el

PreTest se obtuvo un valor de 14944.06s, mientras que en el PostTest es de 4.04s, lo cual denota una gigantesca diferencia entre el antes y después. Esta comparación se puede visualizar en la Figura 5.5:

Figura 5.5: Comparación de medias - Tiempo de filtrado de candidatos



Fuente: La empresa
Elaboración: Propia

5.1.1.6 Tiempo de generación de reporte

Los resultados de las medidas descriptivas se describen en la Tabla 5.6:

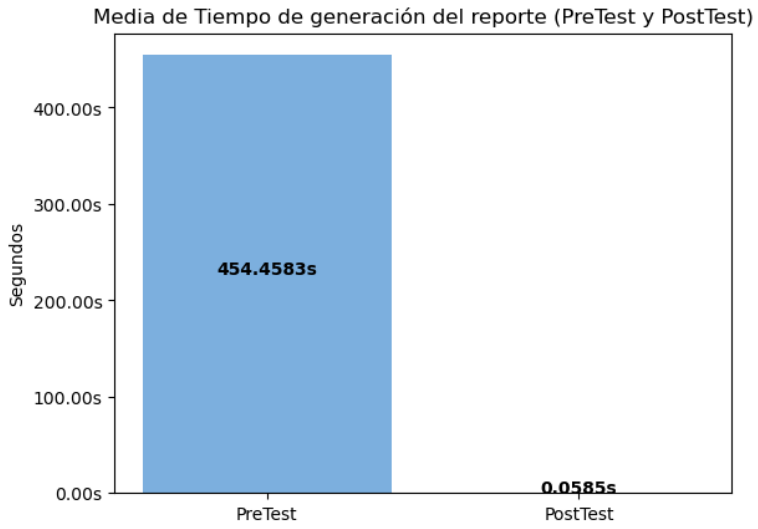
Tabla 5.6: Medidas descriptivas - Tiempo de generación de reporte

Nombre	Muestra	Mínimo	Máximo	Media	Desviación estándar
PreTest_Tiempo_ generacion_reporte_ manual	48	302	615	454.46	104.93
PostTest_Tiempo_ generacion_reporte_ automatizado	48	0.0549	0.0773	0.0585	0.0045

Fuente: La empresa
Elaboración: Propia

En el caso del tiempo de generación de reporte del proceso de selección de personal, en el PreTest se obtuvo un valor de 454.46s, mientras que en el PostTest es de 0.05s, lo cual denota una gigantesca diferencia entre el antes y después. Esta comparación se puede visualizar en la Figura 5.6:

Figura 5.6: Comparación de medias - Tiempo de generación de reporte



Fuente: La empresa
Elaboración: Propia

5.2 RESULTADOS INFERENCIALES

Las tareas de desarrollo propias de esta etapa se encuentran en un repositorio remoto de GitHub, el cual se encuentra en la referencia de Nolasco (2023).

5.2.1 Prueba de normalidad

Se procedió a hacer las pruebas de normalidad para los indicadores de exactitud, precisión, sensibilidad, robustez, tiempo de filtrado de candidatos y tiempo de generación de reporte a través del método de Shapiro-Wilk, debido a que el tamaño de la muestra está conformado por 48 modelos, el cual es menor a 50.

Se definen las siguientes hipótesis estadísticas:

- H_0: El indicador presenta una distribución normal
- H_A: El indicador presenta una distribución no normal

Basadas en la siguiente toma de decisión:

- Valor P $< \alpha$: Se rechaza H_0
- Valor P $>= \alpha$: No se rechaza H_0

Dónde:

- Valor P: Probabilidad de haber obtenido un estadístico de prueba suponiendo que la hipótesis nula es cierta, es el nivel crítico del contraste de la hipótesis.
- α: Probabilidad de rechazar la hipótesis nula cuando es verdadera, también llamado nivel de significancia.

Para todas las pruebas, se utilizó un $\alpha = 5\,\%$.

5.2.1.1 Exactitud

Los resultados de la prueba de normalidad se describen en la Figura 5.7:

Tabla 5.7: Prueba de normalidad - Exactitud

Nombre	Grados de libertad	Estadístico de prueba	Valor P
PreTest_Exactitud	48	0.9024	0.0008
PostTest_Exactitud	48	0.9002	0.0006

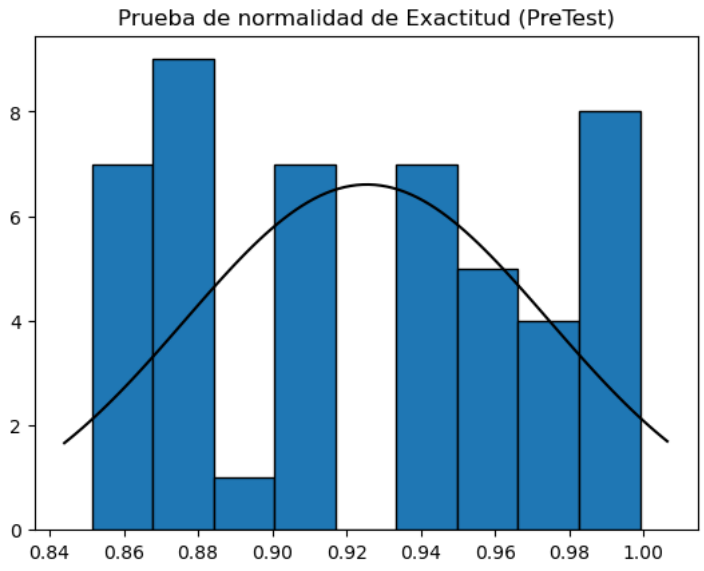
Fuente: La empresa
Elaboración: Propia

Los resultados de la prueba indican que:

- PreTest: El Valor P es 0.0008, el cual es menor a 0.05. Por lo tanto, la exactitud en el PreTest se distribuye **no normalmente**.
- PostTest: El Valor P es 0.0006, el cual es menor a 0.05. Por lo tanto, la exactitud en el PostTest se distribuye **no normalmente**.

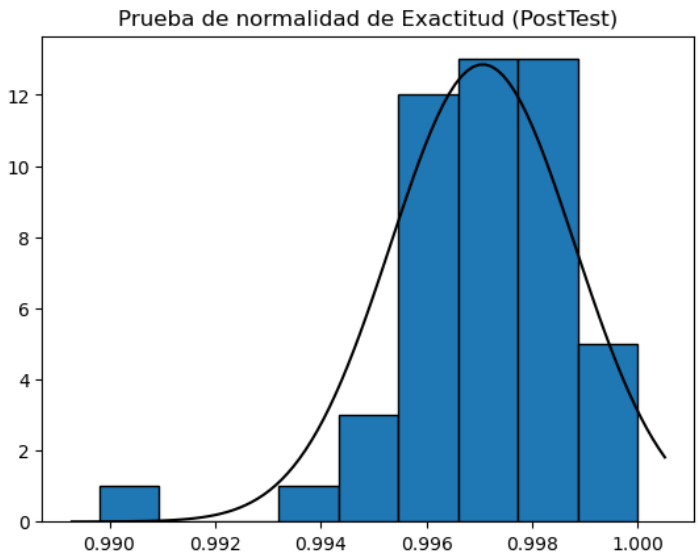
Ambas pruebas de normalidad se describen en la Figura 5.7 y Figura 5.8:

Figura 5.7: Prueba de normalidad - Exactitud (PreTest)



Fuente: La empresa
Elaboración: Propia

Figura 5.8: Prueba de normalidad - Exactitud (PostTest)



Fuente: La empresa
Elaboración: Propia

5.2.1.2 Precisión

Los resultados de la prueba de normalidad se describen en la Figura 5.8:

Tabla 5.8: Prueba de normalidad - Precisión

Nombre	Grados de libertad	Estadístico de prueba	Valor P
PreTest_Precision	48	0.8410	1.27×10^{-5}
PostTest_Precision	48	1.0000	1.0000

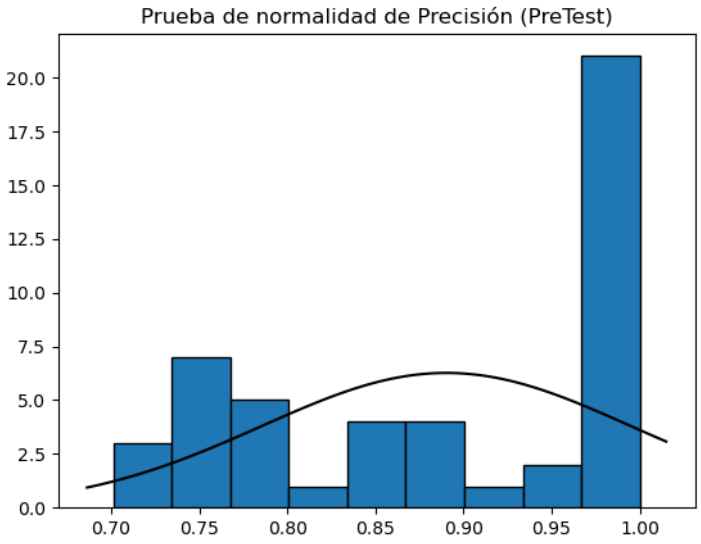
Fuente: La empresa
Elaboración: Propia

Los resultados de la prueba indican que:

- PreTest: El Valor P es 1.27×10^{-5}, el cual es menor a 0.05. Por lo tanto, la precisión en el PreTest se distribuye **no normalmente**.
- PostTest: El Valor P es 1.0000, el cual es mayor a 0.05. Por lo tanto, la precisión en el PostTest se distribuye **normalmente**.

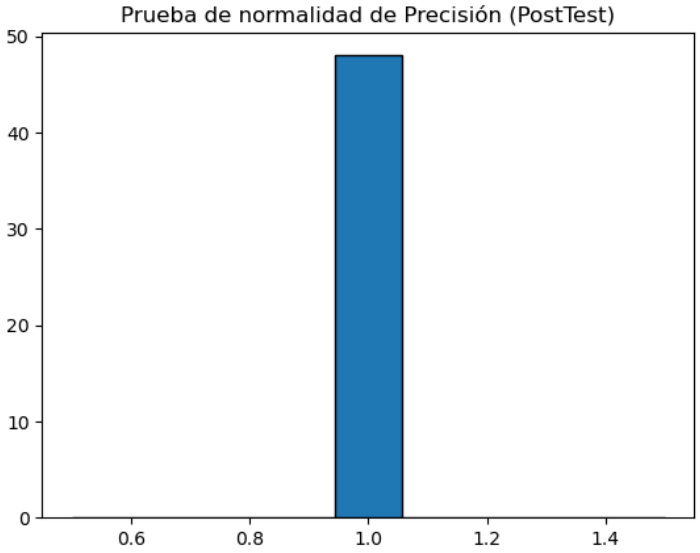
Ambas pruebas de normalidad se describen en la Figura 5.9 y Figura 5.10:

Figura 5.9: Prueba de normalidad - Precisión (PreTest)



Fuente: La empresa
Elaboración: Propia

Figura 5.10: Prueba de normalidad - Precisión (PostTest)



Fuente: La empresa
Elaboración: Propia

5.2.1.3 Sensibilidad

Los resultados de la prueba de normalidad se describen en la Figura 5.9:

Tabla 5.9: Prueba de normalidad - Sensibilidad

Nombre	Grados de libertad	Estadístico de prueba	Valor P
PreTest_Sensibilidad	48	0.8599	4.02×10^{-5}
PostTest_Sensibilidad	48	0.9068	0.0011

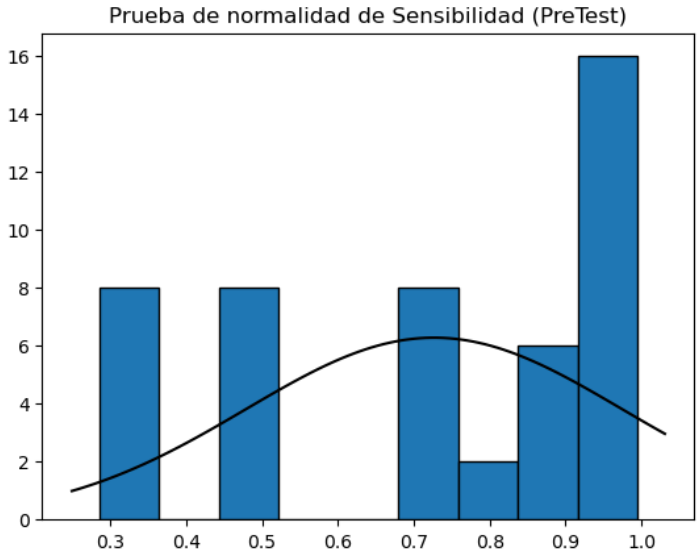
Fuente: La empresa
Elaboración: Propia

Los resultados de la prueba indican que:

- PreTest: El Valor P es 4.02×10^{-5}, el cual es menor a 0.05. Por lo tanto, la sensibilidad en el PreTest se distribuye **no normalmente**.
- PostTest: El Valor P es 0.0011, el cual es menor a 0.05. Por lo tanto, la sensibilidad en el PostTest se distribuye **no normalmente**.

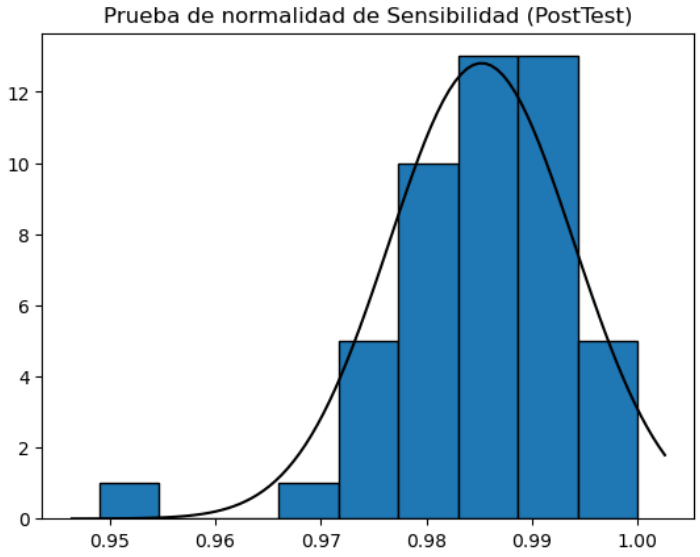
Ambas pruebas de normalidad se describen en la Figura 5.11 y Figura 5.12:

Figura 5.11: Prueba de normalidad - Sensibilidad (PreTest)



Fuente: La empresa
Elaboración: Propia

Figura 5.12: Prueba de normalidad - Sensibilidad (PostTest)



Fuente: La empresa
Elaboración: Propia

5.2.1.4 Robustez

Los resultados de la prueba de normalidad se describen en la Figura 5.10:

Tabla 5.10: Prueba de normalidad - Robustez

Nombre	Grados de libertad	Estadístico de prueba	Valor P
PreTest_Robustez	48	0.9007	0.0007
PostTest_Robustez	48	0.9027	0.0008

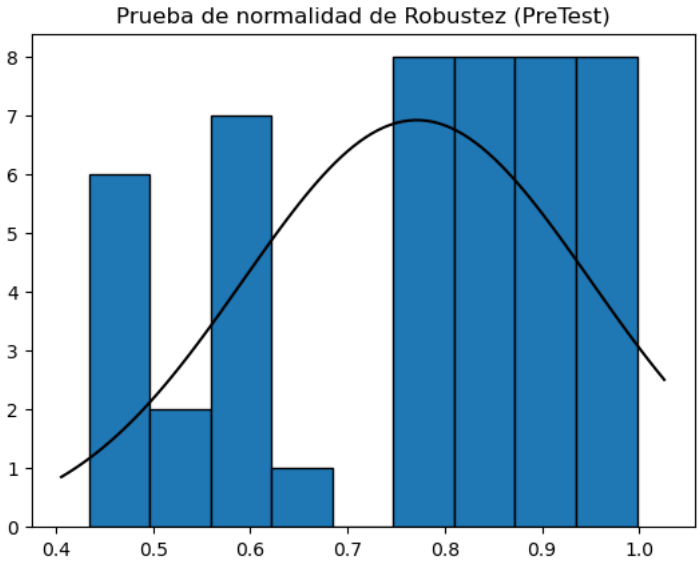
Fuente: La empresa
Elaboración: Propia

Los resultados de la prueba indican que:

- PreTest: El Valor P es 0.0007, el cual es menor a 0.05. Por lo tanto, la robustez en el PreTest se distribuye **no normalmente**.
- PostTest: El Valor P es 0.0008, el cual es menor a 0.05. Por lo tanto, la robustez en el PostTest se distribuye **no normalmente**.

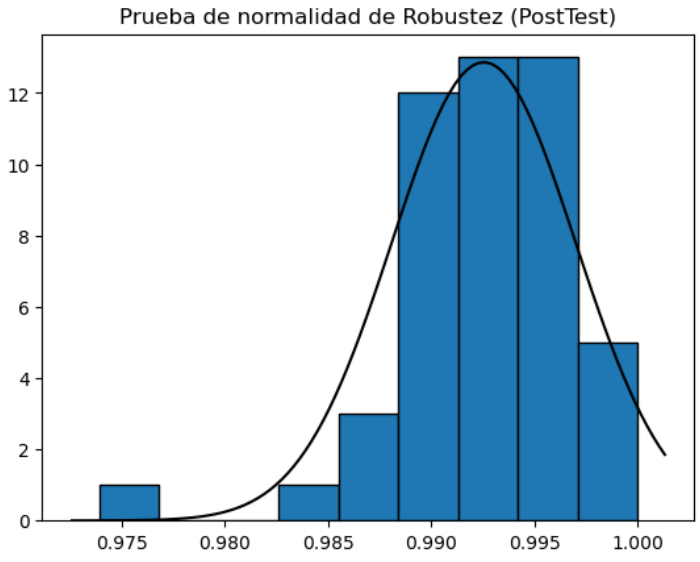
Ambas pruebas de normalidad se describen en la Figura 5.13 y Figura 5.14:

Figura 5.13: Prueba de normalidad - Robustez (PreTest)



Fuente: La empresa
Elaboración: Propia

Figura 5.14: Prueba de normalidad - Robustez (PreTest)



Fuente: La empresa
Elaboración: Propia

5.2.1.5 Tiempo de filtrado de candidatos

Los resultados de la prueba de normalidad se describen en la Figura 5.11:

Tabla 5.11: Prueba de normalidad - Tiempo de filtrado de candidatos

Nombre	Grados de libertad	Estadístico de prueba	Valor P
PreTest_Tiempo_filtrado_candidatos_manual	48	0.9089	0.0012
PostTest_Tiempo_filtrado_candidatos_automatizado	48	0.9808	0.6119

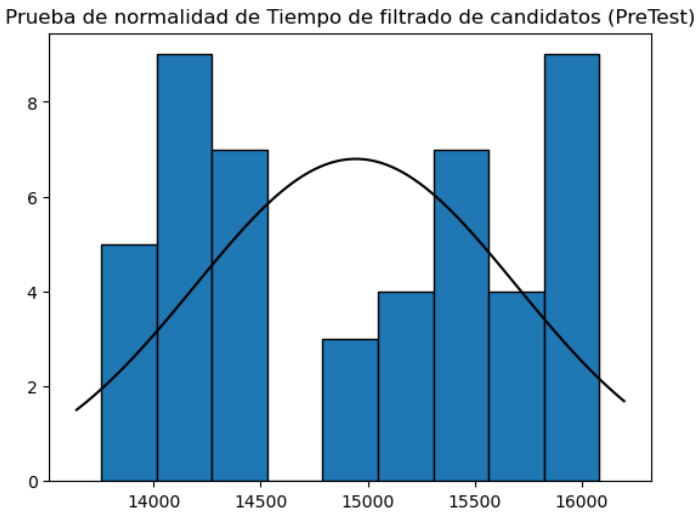
Fuente: La empresa
Elaboración: Propia

Los resultados de la prueba indican que:

- PreTest: El Valor P es 0.0012, el cual es menor a 0.05. Por lo tanto, el tiempo de filtrado de candidatos en el PreTest se distribuye **no normalmente**.
- PostTest: El Valor P es 0.6119, el cual es mayor a 0.05. Por lo tanto, el tiempo de filtrado de candidatos en el PostTest se distribuye **normalmente**.

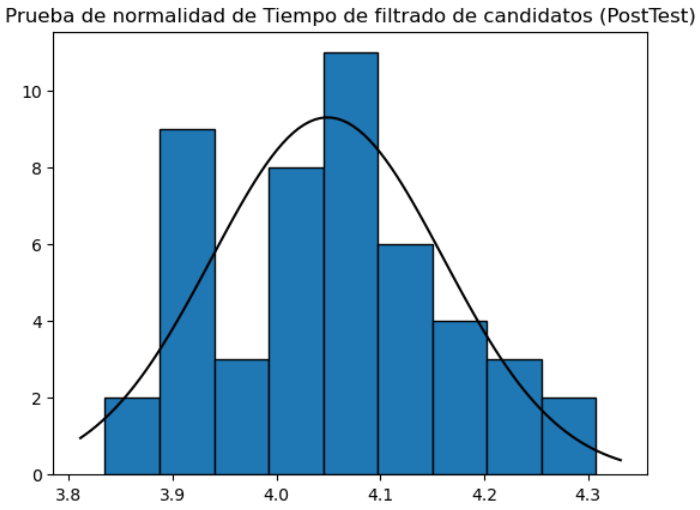
Ambas pruebas de normalidad se describen en la Figura 5.15 y Figura 5.16:

Figura 5.15: Prueba de normalidad - Tiempo de filtrado de candidatos (PreTest)



Fuente: La empresa
Elaboración: Propia

Figura 5.16: Prueba de normalidad - Tiempo de filtrado de candidatos (PostTest)



Fuente: La empresa
Elaboración: Propia

5.2.1.6 Tiempo de generación de reporte

Los resultados de la prueba de normalidad se describen en la Figura 5.12:

Tabla 5.12: Prueba de normalidad - Tiempo de generación de reporte

Nombre	Grados de libertad	Estadístico de prueba	Valor P
PreTest_Tiempo_generacion_reporte_manual	48	0.9087	0.0012
PostTest_Tiempo_generacion_reporte_automatizado	48	0.5398	4.80×10^{-11}

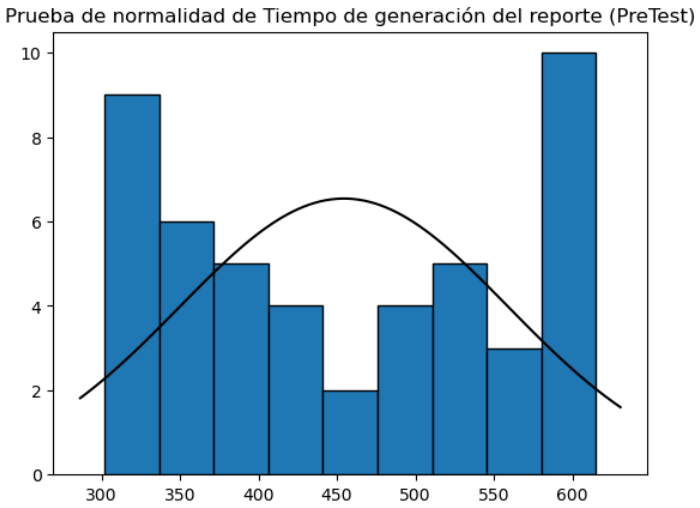
Fuente: La empresa
Elaboración: Propia

Los resultados de la prueba indican que:

- PreTest: El Valor P es 0.0012, el cual es menor a 0.05. Por lo tanto, el tiempo de generación de reporte en el PreTest se distribuye **no normalmente**.
- PostTest: El Valor P es 4.80×10^{-11}, el cual es menor a 0.05. Por lo tanto, el tiempo de generación de reporte en el PostTest se distribuye **no normalmente**.

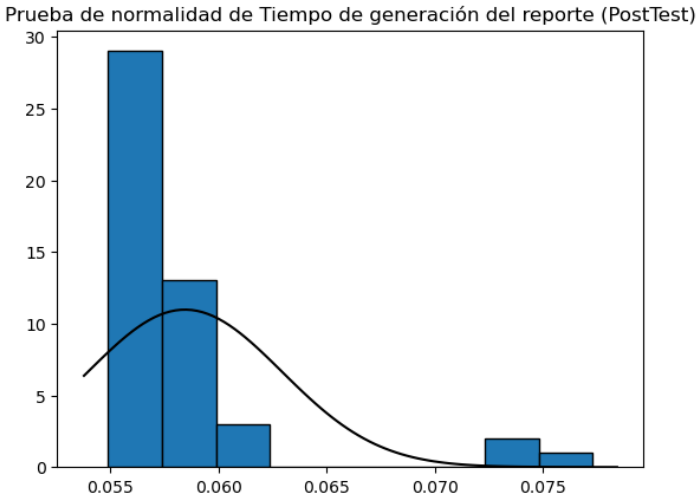
Ambas pruebas de normalidad se describen en la Figura 5.17 y Figura 5.18:

Figura 5.17: Prueba de normalidad - Tiempo de generación de reporte (PreTest)



Fuente: La empresa
Elaboración: Propia

Figura 5.18: Prueba de normalidad - Tiempo de generación de reporte (PostTest)



Fuente: La empresa
Elaboración: Propia

5.2.2 Prueba de hipótesis

5.2.2.1 Exactitud

5.2.2.1.1 Hipótesis de investigación 1

- H_1 : El modelo predictivo aumenta la exactitud del proceso de selección de personal.
- Indicador: Exactitud

5.2.2.1.2 Hipótesis estadísticas

Definición de variables:

- IE_a : Exactitud antes de la aplicación del modelo predictivo final validado.
- IE_d : Exactitud después de la aplicación del modelo predictivo final validado.

Definición de hipótesis:

- H_0 : El modelo predictivo no aumenta la exactitud del proceso de selección de personal, según la Ecuación 5.1.

$$H_0 = IEa >= IEd \tag{5.1}$$

  El indicador sin el modelo predictivo es mejor que el indicador con el modelo predictivo.

- H_A : El modelo predictivo aumenta la exactitud del proceso de selección de personal, según la Ecuación 5.2.

$$H_A = IEa < IEd \tag{5.2}$$

  El indicador con el modelo predictivo es mejor que el indicador sin el modelo predictivo.

En cuanto al resultado del contraste de hipótesis se aplicó la prueba de Wilcoxon, debido a que al menos una de las distribuciones de los datos obtenidos durante la investigación (PreTest y PostTest) es no normal. Es en base a ello que se puede afirmar que el Valor P (6.25×10^{-13}) es menor a α (0.05). En la Tabla 5.13, se muestran los resultados de la prueba de Wilcoxon:

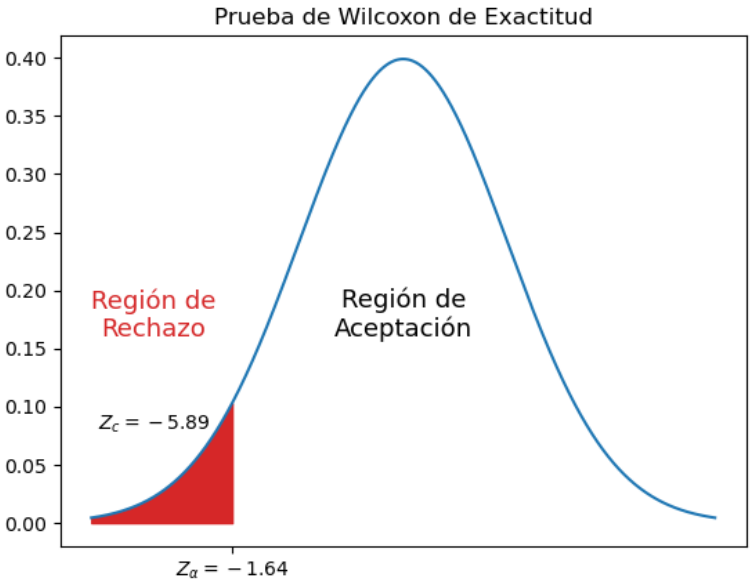
Tabla 5.13: Prueba de hipótesis - Exactitud

Nombre	Grados de libertad	Estadístico de prueba	Valor P
PostTest_PreTest_ Exactitud	48	-5.89	6.25×10^{-13}

Fuente: La empresa
Elaboración: Propia

De igual manera, el valor del estadístico de prueba Z_c (-5.89) es menor a Z_α (-1.64), esto también se puede corroborar graficándolo en una distribución normal con μ = 0 y σ = 1, donde nos percatamos que se encuentra en la zona de rechazo. Este contraste se presenta en la Figura 5.19:

Figura 5.19: Prueba de hipótesis - Exactitud



Fuente: La empresa
Elaboración: Propia

Es debido a ello, que se rechaza la hipótesis nula, y se acepta la hipótesis alterna con un 90% de confianza. Por lo tanto, **el modelo predictivo aumenta la exactitud del proceso de selección de personal**.

5.2.2.2 Precisión

5.2.2.2.1 Hipótesis de investigación 2

- H_2 : El modelo predictivo aumenta la precisión del proceso de selección de personal.
- Indicador: Precisión

5.2.2.2.2 Hipótesis estadísticas

Definición de variables:

- IP_a : Precisión antes de la aplicación del modelo predictivo final validado.
- IP_d : Precisión después de la aplicación del modelo predictivo final validado.

Definición de hipótesis:

- H_0 : El modelo predictivo no aumenta la precisión del proceso de selección de personal, según la Ecuación 5.3.

$$H_0 = IPa >= IPd \tag{5.3}$$

  El indicador sin el modelo predictivo es mejor que el indicador con el modelo predictivo.

- H_A : El modelo predictivo aumenta la precisión del proceso de selección de personal, según la Ecuación 5.4.

$$H_A = IPa < IPd \tag{5.4}$$

  El indicador con el modelo predictivo es mejor que el indicador sin el modelo predictivo.

En cuanto al resultado del contraste de hipótesis se aplicó la prueba de Wilcoxon, debido a que al menos una de las distribuciones de los datos obtenidos durante la investigación (PreTest y PostTest) es no normal. Es en base a ello que se puede afirmar que el Valor P (5.26×10^{-8}) es menor a α (0.05). En la Tabla 5.14, se muestran los resultados de la prueba de Wilcoxon:

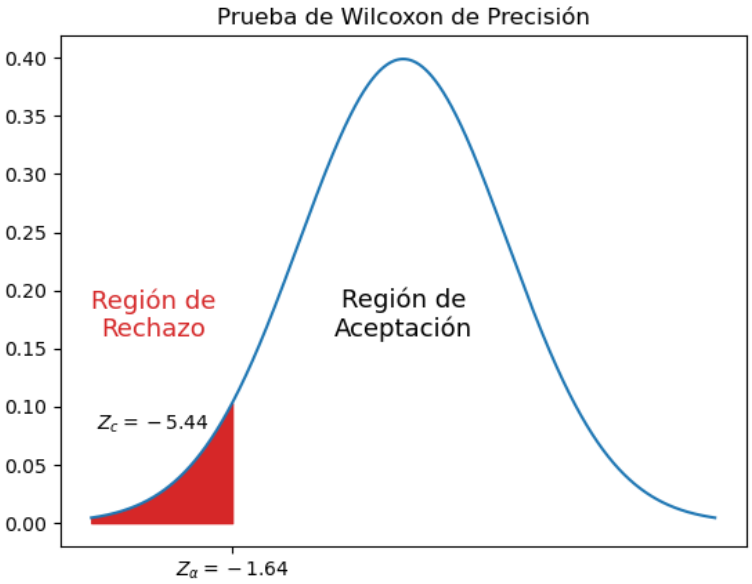
Tabla 5.14: Prueba de hipótesis - Precisión

Nombre	Grados de libertad	Estadístico de prueba	Valor P
PostTest_PreTest_ Precision	48	-5.44	5.26×10^{-8}

Fuente: La empresa
Elaboración: Propia

De igual manera, el valor del estadístico de prueba Z_c (-5.44) es menor a Z_α (-1.64), esto también se puede corroborar graficándolo en una distribución normal con μ = 0 y σ = 1, donde nos percatamos que se encuentra en la zona de rechazo. Este contraste se presenta en la Figura 5.20:

Figura 5.20: Prueba de hipótesis - Precisión



Fuente: La empresa
Elaboración: Propia

Es debido a ello, que se rechaza la hipótesis nula, y se acepta la hipótesis alterna con un 90% de confianza. Por lo tanto, **el modelo predictivo aumenta la precisión del proceso de selección de personal**.

5.2.2.3 Sensibilidad

5.2.2.3.1 Hipótesis de investigación 3

- H_3 : El modelo predictivo aumenta la sensibilidad del proceso de selección de personal.
- Indicador: Sensibilidad

5.2.2.3.2 Hipótesis estadísticas

Definición de variables:

- IS_a : Sensibilidad antes de la aplicación del modelo predictivo final validado.
- IS_d : Sensibilidad después de la aplicación del modelo predictivo final validado.

Definición de hipótesis:

- H_0 : El modelo predictivo no aumenta la sensibilidad del proceso de selección de personal, según la Ecuación 5.5.

$$H_0 = ISa >= ISd \tag{5.5}$$

  El indicador sin el modelo predictivo es mejor que el indicador con el modelo predictivo.

- H_A : El modelo predictivo aumenta la sensibilidad del proceso de selección de personal, según la Ecuación 5.6.

$$H_A = ISa < ISd \tag{5.6}$$

  El indicador con el modelo predictivo es mejor que el indicador sin el modelo predictivo.

En cuanto al resultado del contraste de hipótesis se aplicó la prueba de Wilcoxon, debido a que al menos una de las distribuciones de los datos obtenidos durante la investigación (PreTest y PostTest) es no normal. Es en base a ello que se puede afirmar que el Valor P (1.55×10^{-9}) es menor a α (0.05). En la Tabla 5.15, se muestran los resultados de la prueba de Wilcoxon:

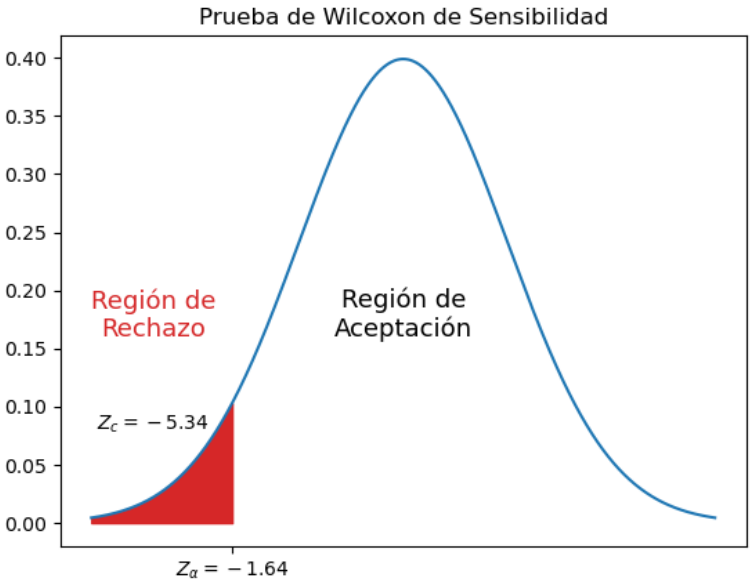
Tabla 5.15: Prueba de hipótesis - Sensibilidad

Nombre	Grados de libertad	Estadístico de prueba	Valor P
PostTest_PreTest_ Sensibilidad	48	-5.34	1.55×10^{-9}

Fuente: La empresa
Elaboración: Propia

De igual manera, el valor del estadístico de prueba Z_c (-5.34) es menor a Z_α (-1.64), esto también se puede corroborar graficándolo en una distribución normal con μ = 0 y σ = 1, donde nos percatamos que se encuentra en la zona de rechazo. Este contraste se presenta en la Figura 5.21:

Figura 5.21: Prueba de hipótesis - Sensibilidad



Fuente: La empresa
Elaboración: Propia

Es debido a ello, que se rechaza la hipótesis nula, y se acepta la hipótesis alterna con un 90% de confianza. Por lo tanto, **el modelo predictivo aumenta la sensibilidad del proceso de selección de personal**.

5.2.2.4 Robustez

5.2.2.4.1 Hipótesis de investigación 4

- H_4 : El modelo predictivo aumenta la robustez del proceso de selección de personal.
- Indicador: Robustez

5.2.2.4.2 Hipótesis estadísticas

Definición de variables:

- IP_a : Robustez antes de la aplicación del modelo predictivo final validado.
- IP_d : Robustez después de la aplicación del modelo predictivo final validado.

Definición de hipótesis:

- H_0 : El modelo predictivo no aumenta la robustez del proceso de selección de personal, según la Ecuación 5.7.

$$H_0 = IRa >= IRd \tag{5.7}$$

  El indicador sin el modelo predictivo es mejor que el indicador con el modelo predictivo.

- H_A : El modelo predictivo aumenta la robustez del proceso de selección de personal, según la Ecuación 5.8.

$$H_A = IRa < IRd \tag{5.8}$$

  El indicador con el modelo predictivo es mejor que el indicador sin el modelo predictivo.

En cuanto al resultado del contraste de hipótesis se aplicó la prueba de Wilcoxon, debido a que al menos una de las distribuciones de los datos obtenidos durante la investigación (PreTest y PostTest) es no normal. Es en base a ello que se puede afirmar que el Valor P (4.97×10^{-13}) es menor a α (0.05). En la Tabla 5.16, se muestran los resultados de la prueba de Wilcoxon:

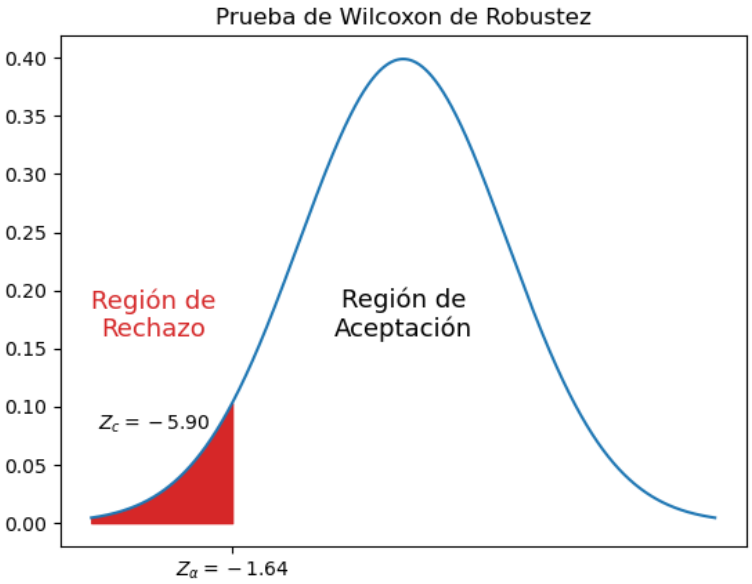
Tabla 5.16: Prueba de hipótesis - Robustez

Nombre	Grados de libertad	Estadístico de prueba	Valor P
PostTest_PreTest_ Robustez	48	-5.9	4.97×10^{-13}

Fuente: La empresa
Elaboración: Propia

De igual manera, el valor del estadístico de prueba Z_c (-5.9) es menor a Z_α (-1.64), esto también se puede corroborar graficándolo en una distribución normal con μ = 0 y σ = 1, donde nos percatamos que se encuentra en la zona de rechazo. Este contraste se presenta en la Figura 5.22:

Figura 5.22: Prueba de hipótesis - Robustez



Fuente: La empresa
Elaboración: Propia

Es debido a ello, que se rechaza la hipótesis nula, y se acepta la hipótesis alterna con un 90% de confianza. Por lo tanto, **el modelo predictivo aumenta la robustez del proceso de selección de personal**.

5.2.2.5 Tiempo de filtrado de candidatos

5.2.2.5.1 Hipótesis de investigación 5

- H_5 : El modelo predictivo reduce el tiempo de filtrado de candidatos del proceso de selección de personal.
- Indicador: Tiempo de filtrado de candidatos

5.2.2.5.2 Hipótesis estadísticas

Definición de variables:

- $ITFC_a$: Tiempo de filtrado de candidatos antes de la aplicación del modelo predictivo final validado.
- $ITFC_d$: Tiempo de filtrado de candidatos después de la aplicación del modelo predictivo final validado.

Definición de hipótesis:

- H_0 : El modelo predictivo no reduce el tiempo de filtrado de candidatos del proceso de selección de personal, según la Ecuación 5.9.

$$H_0 = ITFCa <= ITFCd \tag{5.9}$$

  El indicador sin el modelo predictivo es mejor que el indicador con el modelo predictivo.

- H_A : El modelo predictivo reduce el tiempo de filtrado de candidatos del proceso de selección de personal, según la Ecuación 5.10.

$$H_A = ITFCa < ITFCd \tag{5.10}$$

  El indicador con el modelo predictivo es mejor que el indicador sin el modelo predictivo.

En cuanto al resultado del contraste de hipótesis se aplicó la prueba de Wilcoxon, debido a que al menos una de las distribuciones de los datos obtenidos durante la investigación (PreTest y PostTest) es no normal. Es en base a ello que se puede afirmar que el Valor P (7.11×10^{-15}) es menor a α (0.05). En la Tabla 5.17, se muestran los resultados de la prueba de Wilcoxon:

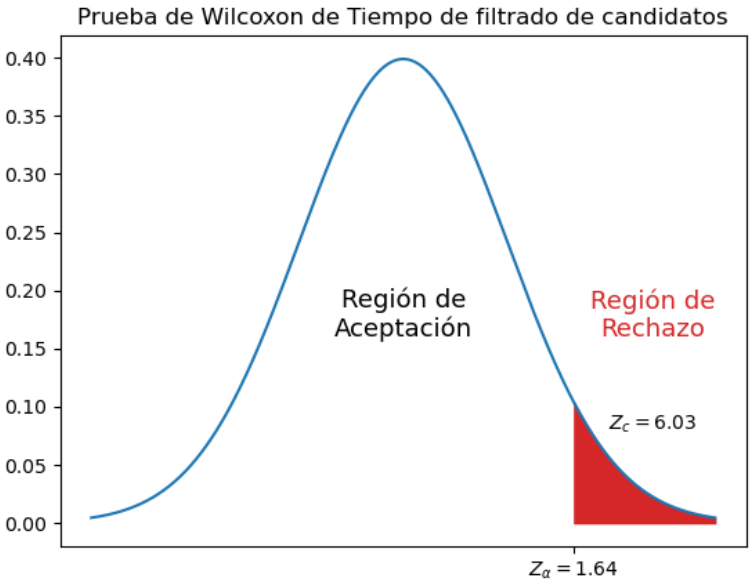
Tabla 5.17: Prueba de hipótesis - Tiempo de filtrado de candidatos

Nombre	Grados de libertad	Estadístico de prueba	Valor P
PostTest_PreTest_ Tiempo_filtrado_ candidatos	48	6.03	7.11×10^{-15}

Fuente: La empresa
Elaboración: Propia

De igual manera, el valor del estadístico de prueba Z_c (6.03) es mayor a Z_α (1.64), esto también se puede corroborar graficándolo en una distribución normal con μ = 0 y σ = 1, donde nos percatamos que se encuentra en la zona de rechazo. Este contraste se presenta en la Figura 5.23:

Figura 5.23: Prueba de hipótesis - Tiempo de filtrado de candidatos



Fuente: La empresa
Elaboración: Propia

Es debido a ello, que se rechaza la hipótesis nula, y se acepta la hipótesis alterna con un 90% de confianza. Por lo tanto, **el modelo predictivo reduce el tiempo de filtrado de candidatos del proceso de selección de personal**.

5.2.2.6 Tiempo de generación de reporte

5.2.2.6.1 Hipótesis de investigación 6

- H_6 : El modelo predictivo reduce el tiempo de generación de reporte del proceso de selección de personal.
- Indicador: Tiempo de generación de reporte

5.2.2.6.2 Hipótesis estadísticas

Definición de variables:

- $ITGR_a$: Tiempo de generación de reporte antes de la aplicación del modelo predictivo final validado.
- $ITGR_d$: Tiempo de generación de reporte después de la aplicación del modelo predictivo final validado.

Definición de hipótesis:

- H_0 : El modelo predictivo no reduce el tiempo de generación de reporte del proceso de selección de personal, según la Ecuación 5.11.

$$H_0 = ITGRa <= ITGRd \tag{5.11}$$

  El indicador sin el modelo predictivo es mejor que el indicador con el modelo predictivo.

- H_A : El modelo predictivo reduce el tiempo de generación de reporte del proceso de selección de personal, según la Ecuación 5.12.

$$H_A = ITGRa < ITGRd \tag{5.12}$$

  El indicador con el modelo predictivo es mejor que el indicador sin el modelo predictivo.

En cuanto al resultado del contraste de hipótesis se aplicó la prueba de Wilcoxon, debido a que al menos una de las distribuciones de los datos obtenidos durante la investigación (PreTest y PostTest) es no normal. Es en base a ello que se puede afirmar que el Valor P (7.11×10^{-15}) es menor a α (0.05). En la Tabla 5.18, se muestran los resultados de la prueba de Wilcoxon:

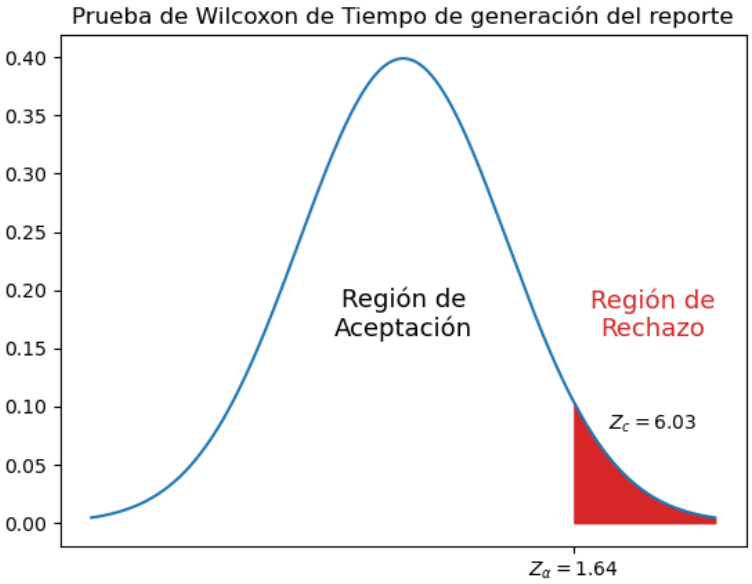
Tabla 5.18: Prueba de hipótesis - Tiempo de generación de reporte

Nombre	Grados de libertad	Estadístico de prueba	Valor P
PostTest_PreTest_ Tiempo_generacion_ reporte	48	6.03	7.11×10^{-15}

Fuente: La empresa
Elaboración: Propia

De igual manera, el valor del estadístico de prueba Z_c (6.03) es mayor a Z_α (1.64), esto también se puede corroborar graficándolo en una distribución normal con μ = 0 y σ = 1, donde nos percatamos que se encuentra en la zona de rechazo. Este contraste se presenta en la Figura 5.24:

Figura 5.24: Prueba de hipótesis - Tiempo de generación de reporte



Fuente: La empresa
Elaboración: Propia

Es debido a ello, que se rechaza la hipótesis nula, y se acepta la hipótesis alterna con un 90% de confianza. Por lo tanto, **el modelo predictivo reduce el tiempo de generación de reporte del proceso de selección de personal**.

CAPÍTULO VI
DISCUSIÓN DE LOS RESULTADOS

6.1 CONTRASTACIÓN DE LA HIPÓTESIS

En base a los resultados la presente investigación se analiza una comparación sobre la exactitud, precisión, sensibilidad, robustez, tiempo de filtrado de candidatos y tiempo de generación de reporte obtenidos en el PreTest y PostTest.

Finalmente, luego del análisis de medidas descriptivas de los cinco principales indicadores, se encontró que:

- La exactitud aumentó de un 92.55 % a un 99.71 %, lo que equivale a un aumento promedio de 7.16 %, un aumento leve.
- La precisión aumentó de un 89.02 % a un 100.00 %, lo que equivale a un aumento promedio de 10.98 %, un aumento leve pero significativo.
- La sensibilidad aumentó de un 72.66 % a un 98.52 %, lo que equivale a un aumento promedio de 25.86 %, un aumento muy considerable.
- La robustez aumentó de un 77.13 % a un 99.25 %, lo que equivale a un aumento promedio de 22.12 %, un aumento considerable.
- El tiempo de filtrado de candidatos se redujo de 14944.06s a 4.04s, lo que equivale a una reducción promedio de 14940.02s, la cual de manera porcentual representa una reducción del 99.97 %, una reducción gigantesca.
- El tiempo de generación de reporte se redujo de 454.46s a 0.06s, lo que equivale a una reducción promedio de 454.40s, la cual de manera porcentual representa una reducción del 99.98 %, una reducción gigantesca.

De igual forma, luego de las pruebas de normalidad y las pruebas de hipótesis, se puede afirmar que, a un 90 % de confianza:

- El modelo predictivo aumenta la exactitud del proceso de selección de personal.

- El modelo predictivo aumenta la precisión del proceso de selección de personal.
- El modelo predictivo aumenta la sensibilidad del proceso de selección de personal.
- El modelo predictivo aumenta la robustez del proceso de selección de personal.
- El modelo predictivo reduce el tiempo de filtrado de candidatos del proceso de selección de personal.
- El modelo predictivo reduce el tiempo de generación de reporte del proceso de selección de personal.

Finalmente, debido a que se aceptan las 6 hipótesis específicas de la presente investigación, también se puede afirmar que:

- El modelo predictivo mejora el proceso de selección de personal.

6.2 CONTRASTACIÓN DE LA HIPÓTESIS CON RESULTADOS SIMILARES

Comparando los resultados obtenidos con los resultados de los antecedentes de investigación, se tiene lo siguiente:

- El antecedente de Jagan Mohan Reddy et al. (2020) se puede relacionar con los valores obtenidos. Este antecedente tiene una precisión del 100.00%, a comparación de la presente investigación que tiene una precisión del 100%, denotando así una diferencia del 0.00%.
- El antecedente de Mishra et al. (2021) se puede relacionar con los valores obtenidos. Este antecedente tiene una sensibilidad del 96.02% y robustez de 96.26%, a comparación de la presente investigación que tiene una sensibilidad del 98.52% y robustez de 99.25%, denotando así una leve mejora del 2.50% y 2.99%, respectivamente.
- El antecedente de Roy et al. (2018) se puede relacionar con los valores obtenidos. Este antecedente tiene una exactitud del 90.30%, a comparación de la presente investigación que tiene una exactitud del 99.71%, denotando así una gran mejora del 9.41%.

Finalmente, los resultados obtenidos en la presente investigación comprueban el aumento de la exactitud, precisión, sensibilidad, junto con la reducción del tiempo de filtrado de candidatos y tiempo de generación de reporte, tomando en cuenta el modelo predictivo. Es en base a ello que se concluye que el modelo predictivo mejora el proceso de selección de personal.

CONCLUSIONES

1. La aplicación del modelo predictivo final, validado mediante técnicas de validación cruzada, permitió elevar las métricas porcentuales por encima del 98%, lo cual permite concluir que el modelo predictivo es excelente, con respecto a la exactitud, precisión, sensibilidad y robustez.

2. La aplicación del modelo predictivo final, validado mediante técnicas de validación cruzada, permitió reducir el tiempo de filtrado de candidatos por debajo de los 5 segundos y el tiempo de generación de reporte por debajo de los 0.1 segundos, lo cual permite concluir que el modelo es bueno con respecto al tiempo, considerando los valores iniciales de 14944.06 segundos y 454.46 segundos, respectivamente.

3. La tarea de balanceo de datos, junto a otras tareas, permitió obtener mejores resultados del modelo, debido a la gran diferencia en la proporción inicial entre clases o etiquetas en el modelo (95.23% y 4.77%) comparada con la proporción final (80% y 20%).

4. Las tareas de eliminación de variables debido a baja varianza y alta correlación, junto a otras tareas, permitió reducir ampliamente el tiempo de filtrado de candidatos y el tiempo de generación de reporte, eliminando gran cantidad de atributos que no tenían alto impacto en la predicción de la clase final.

5. El algoritmo a utilizar fue el de Bosque aleatorio, debido a la gran diferencia en resultados que tuvo con respecto a los demás algoritmos.

6. Si bien es cierto las 4 técnicas de validación tuvieron muy buenos resultados, la mejor técnica que se utilizó fue la División aleatoria estratificada.

RECOMENDACIONES

1. Se recomienda incluir otro tipo de indicadores, como el Área bajo la curva (AUC), Error medio absoluto (MAE), Raíz del error cuadrático medio (RMSE) o R-cuadrado (R^2), con el fin de tener mayores criterios para evaluar los modelos y así poder seleccionar el modelo más adecuado.

2. Se recomienda aplicar la técnica de ajuste de hiper parámetros del modelo, para así, de manera iterativa, encontrar la mejor configuración posible para los modelos y, por ende, conseguir los mejores resultados posibles.

3. Se recomienda extraer mayor cantidad de datos del puesto en cuestión (responsabilidades, prerrequisitos, beneficios) a fin de poder hacer una predicción más ajustada a las capacidades, necesidades y limitaciones del candidato en cuestión.

4. Se recomienda aplicar más fases para el filtrado de candidatos, en las cuales sean las posteriores a la preselección (pruebas psicométricas, pruebas técnicas, entrevistas), de tal forma de poder no solo predecir a los candidatos a contratar, sino en qué fase ellos se encontrarían al final del proceso de selección.

5. Se recomienda utilizar algoritmos de regresión, considerando los datos disponibles. Con ello no se definiría si contratar o no contratar, sino se le daría un puntaje al candidato, haciendo posible ordenarlos en una escala numérica de compatibilidad con el puesto, lo cual haría mucha más precisa y enriquecedora la selección de personal.

6. Se recomienda profundizar más en la etapa de despliegue, realizándolo no solo a nivel local, sino también dentro de la red interna de la empresa, o también en algún servicio de la nube (Azure, Amazon, Google Cloud).

REFERENCIAS BIBLIOGRÁFICAS

Aggarwal, C. C. (2014). *Data classification: Algorithms and applications* (1.ª ed.). Florida: Chapman & Hall/CRC.

Amat, J. (2020, Oct). *Gradient boosting con python.* Descargado de https://www.cienciadedatos.net/documentos/py09_gradient_boosting_python

Anaconda. (2023, Sep). *Anaconda | the world's most popular data science platform.* Descargado de https://www.anaconda.com/

Andrés, A. (2023, Mar). *Fuentes de reclutamiento o cómo buscar candidatos de forma eficiente.* Descargado de https://www.bizneo.com/blog/fuentes-de-reclutamiento/

Awujoola, O., Odion, P. O., Irhebhude, M. E., y Aminu, H. (2021, Apr.). Performance evaluation of machine learning predictive analytical model for determining the job applicants employment status. *Malaysian Journal of Applied Sciences*, *6*(1), 67-79. Descargado de https://journal.unisza.edu.my/myjas/index.php/myjas/article/view/276 doi: 10.37231/myjas.2021.6.1.276

Bock, L. (2015). *Work rules!: Insights from inside google that will transform how you live and lead.* John Murray. Descargado de https://books.google.com.pe/books?id=YbPCoAEACAAJ

Brownlee, J. (2019, Aug). *Machine learning terminology from statistics and computer science.* Descargado de https://machinelearningmastery.com/data-terminology-in-machine-learning/

Chiavenato, I. (2011). *Administración de recursos humanos* (8.ª ed.). Mexico, DF: McGraw Hill.

Coronel, E. (2021). *Machine learning en la mejora del proceso de selección del personal administrativo de la corte superior de justicia de lima, 2020* (Tesis de maestría, Universidad César Vallejo). Descargado de https://hdl.handle.net/20.500.12692/61903

Eastwood, B. (2020, 09). *Exploration-based algorithms can improve hiring quality and diversity.* Descargado de https://mitsloan.mit.edu/ideas-made-to-matter/exploration-based-algorithms-can-improve-hiring-quality-and-diversity

Europe, S. V. (2017). *Building and applying predictive models in ibm spss modeler training webinar.* Descargado de https://www.sv-europe.com/crisp-dm-methodology/

Fernandes, A. (2020). *Exactitud y precisión.* Descargado de https://www.diferenciador.com/diferencia-entre-exactitud-y-precision/

Ford, M. (2016). *Rise of the robots: Technology and the threat of a jobless future*. USA: Basic Books, Inc.

Galindo, P. (2020). *Pep 619 - python 3.10 release schedule.* Descargado de https://peps.python.org/pep-0619/

Gandhi, R. (2018). *Support vector machine — introduction to machine learning algorithms.* Descargado de https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Gartner Inc. (Ed.). (2021). *HR Glossary: HR Analytics.* Descargado de https://www.gartner.com/en/human-resources/glossary/hr-analytics

Git. (2023). *Git - logo downloads.* Descargado de https://git-scm.com/downloads/logos

Gonzalez, L. (2022, Sep). *Regresión logística - teoría.* Descargado de https://aprendeia.com/algoritmo-regresion-logistica-machine-learning-teoria/

Goodfellow, I., Bengio, Y., y Courville, A. (2016). *Deep learning*. MIT Press. Descargado de http://www.deeplearningbook.org (Book in preparation for MIT Press)

Google. (2023, Aug). *Machine learning glossary | google for developers.* Autor. Descargado de https://developers.google.com/machine-learning/glossary

Gupta, P. (2017, Jun). *Cross-validation in machine learning.* Towards Data Science. Descargado de https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f

Hand, D. J. (2007, 01 de Jul). Principles of data mining. *Drug Safety*, *30*(7), 621-622. Descargado de https://doi.org/10.2165/00002018-200730070-00010 doi: 10.2165/00002018-200730070-00010

Hossain, M. A., Md. Noor, R., Yau, K.-L., Razalli, S., Zraba, M., y Ahmedy, I. (2020, 04). Comprehensive survey of machine learning approaches in cognitive radio-based vehicular ad hoc networks. *IEEE Access*, *8*, 78054-78108. Descargado de https://doi.org/10.1109/ACCESS.2020.2989870 doi: 10.1109/ACCESS.2020.2989870

IBM (Ed.). (2020). *IBM Cloud Education.* Descargado de https://www.ibm.com/cloud/learn/supervised-learning

INEI. (2021). *En el perú existen más de 2 millones 838 mil empresas.* Descargado 2022-05-03, de https://www.inei.gob.pe/prensa/noticias/en-el-peru-existen-mas-de-2-millones-838-mil-empresas-12937/

Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., y Roth, A. (2017, 06–11 Aug). Fairness in reinforcement learning. En D. Precup y Y. W. Teh (Eds.), *Proceedings of the 34th international conference on machine learning* (Vol. 70, pp. 1617–1626). PMLR. Descargado de https://proceedings.mlr.press/v70/jabbari17a.html

Jagan Mohan Reddy, D., Regella, S., y Seelam, S. R. (2020). Recruitment prediction using machine learning. En *2020 5th international conference on computing, communication and security (icccs)* (p. 1-4). Descargado de https://doi.org/10.1109/ICCCS49678.2020.9276955 doi: 10.1109/ICCCS49678.2020.9276955

Jaime, H. (2023, May). *Proceso de reclutamiento: Interno, externo y mixto.* Descargado de https://www.pandape.com/blog/proceso-de-reclutamiento-interno-y-externo/

Jantawan, B., y Tsai, C.-F. (2013). *The application of data mining to build classification model for predicting graduate employment.* arXiv. Descargado de https://arxiv.org/abs/1312.7123 doi: 10.48550/ARXIV.1312.7123

Jha, V. (2020, Sep). *Semma model.* Descargado de https://www.geeksforgeeks.org/semma-model/

Kallio, A., y Tuimala, J. (2013). Data mining. En W. Dubitzky, O. Wolkenhauer, K.-H. Cho, y H. Yokota (Eds.), *Encyclopedia of systems biology* (pp. 525–528). New York, NY: Springer New York. Descargado de https://doi.org/10.1007/978-1-4419-9863-7_599 doi: 10.1007/978-1-4419-9863-7_599

Kohavi, R., y Provost, F. (1998, 01). Glossary of terms. special issue of applications of machine learning and the knowledge discovery process. *Mach. Learn.*, *30*.

Kundu, R. (2022, Sep). *Confusion matrix: How to use it & interpret results examples.* Descargado de https://www.v7labs.com/blog/confusion-matrix-guide

Lather, A. S., Malhotra, R., Saloni, P., Singh, P., y Mittal, S. (2019). Prediction of employee performance using machine learning techniques. En *Proceedings of the international conference on advanced information science and system.* New York, NY, USA: Association for Computing Machinery. Descargado de https://doi.org/10.1145/3373477.3373696 doi: 10.1145/3373477.3373696

Microsoft. (2021, Nov). *Visual studio code and vs code icons and names usage guidelines.* Autor. Descargado de https://code.visualstudio.com/brand

Microsoft Learn (Ed.). (2023a). *Visual studio code - microsoft lifecycle | microsoft learn.* Descargado de https://learn.microsoft.com/en-us/lifecycle/products/visual-studio-code

Microsoft Learn (Ed.). (2023b). *Windows 10 home and pro - microsoft lifecycle | microsoft learn.* Descargado de https://learn.microsoft.com/en-us/lifecycle/products/windows-10-home-and-pro

Minaee, S. (2019, Oct). *20 popular machine learning metrics. part 1: Classification & regression evaluation metrics.* Towards Data Science. Descargado de https://towardsdatascience.com/20-popular-machine-learning-metrics-part-1-classification-regression-evaluation-metrics-1ca3e282a2ce

Mishra, S., Mallick, P. K., Tripathy, H. K., Jena, L., y Chae, G.-S. (2021). Stacked knn with hard voting predictive approach to assist hiring process in it organizations. *The International Journal of Electrical Engineering & Education*, 0020720921989015. Descargado de https://doi.org/10.1177/0020720921989015 doi: 10.1177/0020720921989015

Mitchell, T. (1997). *Machine learning*. New York: McGraw Hill.

Nolan, S. (2022). *Anaconda is launching long-term support.* Descargado de https://www.anaconda.com/blog/anaconda-is-launching-long-term-support

Nolasco, R. (2023, Jun). *Modelo predictivo basado en aprendizaje de máquina para mejorar el proceso de selección de personal en una empresa de consultoría tecnológica.* Descargado de https://github.com/RonaldoNolasco/personal-thesis-development

Pessach, D., Singer, G., Avrahami, D., Chalutz Ben-Gal, H., Shmueli, E., y Ben-Gal, I. (2020). Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming. *Decision Support Systems*, *134*, 113290. Descargado de https://doi.org/10.1016/j.dss.2020.113290 doi: 10.1016/j.dss.2020.113290

Python. (2023). *The python logo | python software foundation.* Descargado de https://www.python.org/community/logos/

Quintanilla, L. (2022, Nov). *What is model builder and how does it work? - ml.net.* Descargado de https://learn.microsoft.com/en-us/dotnet/machine-learning/automate-training-with-model-builder

Rajput, A. (2018, Jun). *Kdd process in data mining - geeksforgeeks.* Descargado de https://www.geeksforgeeks.org/kdd-process-in-data-mining/

Raschka, S. (2015). *Python machine learning*. Packt Publishing.

Rosales, M., y Parrales, E. (2022). *Utilización de machine learning para el proceso de selección de personal en una microempresa* (Tesis de bachiller, Universidad Politécnica Salesiana). Descargado de https://dspace.ups.edu.ec/handle/123456789/23335

Roy, K. S., Roopkanth, K., Teja, V. U., Bhavana, V., y Priyanka, J. (2018). Student career prediction using advanced machine learning techniques. *International Journal of Engineering & Technology*, *7*(2.20), 26–29. Descargado de https://www.sciencepubco.com/index.php/ijet/article/view/11738 doi: 10.14419/ijet.v7i2.20.11738

Río Sadornil, D. d. (2005). *Diccionario-glosario de metodología de la investigación social / dionisio del río sadornil* (1a ed. ed.). Madrid: UNED.

Saaty, T. (1980). *The analytic hierarchy process: Planning, priority setting, resource allocation*. McGraw-Hill International Book Company. Descargado de https://books.google.com.pe/books?id=Xxi7AAAAIAAJ

Sadler, M. (2021, Sep). *Applicant tracking systems (ats) market update 2021.* Descargado de https://www.trustradius.com/vendor-blog/ats-applicant-tracking-systems-software-market

Samudra, H. (2014, Nov). *Processes in data mining.* Descargado de https://sisbinus.blogspot.com/2014/11/processes-in-data-mining.html

SciKitLearn. (2009). *3.1. cross-validation: evaluating estimator performance — scikit-learn 0.21.3 documentation.* Descargado de https://scikit-learn.org/stable/modules/cross_validation.html

Sharma, G. (2021, 05). *Regression algorithms | 5 regression algorithms you should know.* Descargado de https://www.analyticsvidhya.com/blog/2021/05/5-regression-algorithms-you-should-know-introductory-guide/

Shubham. (2018, May). *Decision tree.* Descargado de https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/ml-decision-tree/tutorial/

Spotfire. (2019a). *Demystifying the random forest algorithm for accurate predictions.* Descargado de https://www.tibco.com/reference-center/what-is-a-random-forest

Spotfire. (2019b). *Supervised learning: From binary classification to polynomial regression".* Descargado de https://www.spotfire.com/glossary/what-is-supervised-learning

Srivastava, T. (2023, May). *12 important model evaluation metrics for machine learning everyone should know (updated 2023).* Descargado de https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/#Confusion_Matrix

StatCounter (Ed.). (2023). *Desktop windows version market share worldwide.* Descargado de https://gs.statcounter.com/windows-version-market-share/desktop/worldwide/2022

Statistics Kingdom. (2023). *Z table.* Descargado de https://www.statskingdom.com/z_table.html

Sutton, R. S., y Barto, A. G. (2018). *Reinforcement learning: An introduction* (Second ed.). The MIT Press. Descargado de http://incompleteideas.net/book/the-book-2nd.html

Vallalta, J. (2019, Nov). *Crisp-dm: Una metodología para minería de datos en salud.* Health Data Miner. Descargado de https://healthdataminer.com/data-mining/crisp-dm-una-metodologia-para-mineria-de-datos-en-salud/

Wahid, Z., Satter, A. K. M. Z., Al Imran, A., y Bhuiyan, T. (2019). Predicting absenteeism at work using tree-based learners. En *Proceedings of the 3rd international conference on machine learning and soft computing* (p. 7–11). New York, NY, USA: Association for Computing Machinery. Descargado de https://doi.org/10.1145/3310986.3310994 doi: 10.1145/3310986.3310994

Wikimedia Commons. (2020a). *File:crisp-dm process diagram.png — wikimedia commons, the free media repository.* Descargado de https://commons.wikimedia.org/w/index.php?title=File:CRISP-DM_Process_Diagram.png&oldid=506972775 ([Online; accessed 6-October-2023])

Wikimedia Commons. (2020b). *File:windows 10 logo.svg — wikimedia commons, the free media repository.* Descargado de https://commons.wikimedia.org/w/index.php?title=File:Windows_10_Logo.svg&oldid=784872472 ([Online; accessed 6-October-2023])

Wikipedia (Ed.). (2023). *Git - wikipedia.* Descargado de https://en.wikipedia.org/wiki/Git

Wood, T. (2019, May). *F-score.* Descargado de https://deepai.org/machine-learning-glossary-and-terms/f-score

Yaranga, I. (2022). *Machine learning en la mejora del proceso de selección del personal docente en una universidad nacional, lima 2021* (Tesis de maestría, Universidad César Vallejo). Descargado de https://hdl.handle.net/20.500.12692/85185

Zhang, T., Lin, W., Vogelmann, A., Zhang, M., Xie, S., Qin, Y., y Golaz, J. (2021, 05). Improving convection trigger functions in deep convective parameterization schemes using machine learning. *Journal of Advances in Modeling Earth Systems*, *13*. doi: 10.1029/2020MS002365

Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., y Zhu, X. (2019). Employee turnover prediction with machine learning: A reliable approach. En K. Arai, S. Kapoor, y R. Bhatia (Eds.), *Intelligent systems and applications* (pp. 737–758). Cham: Springer International Publishing. Descargado de https://doi.org/10.1007/978-3-030-01057-7_56 doi: 10.1007/978-3-030-01057-7_56

ANEXOS

ANEXO A: Árbol de problemas

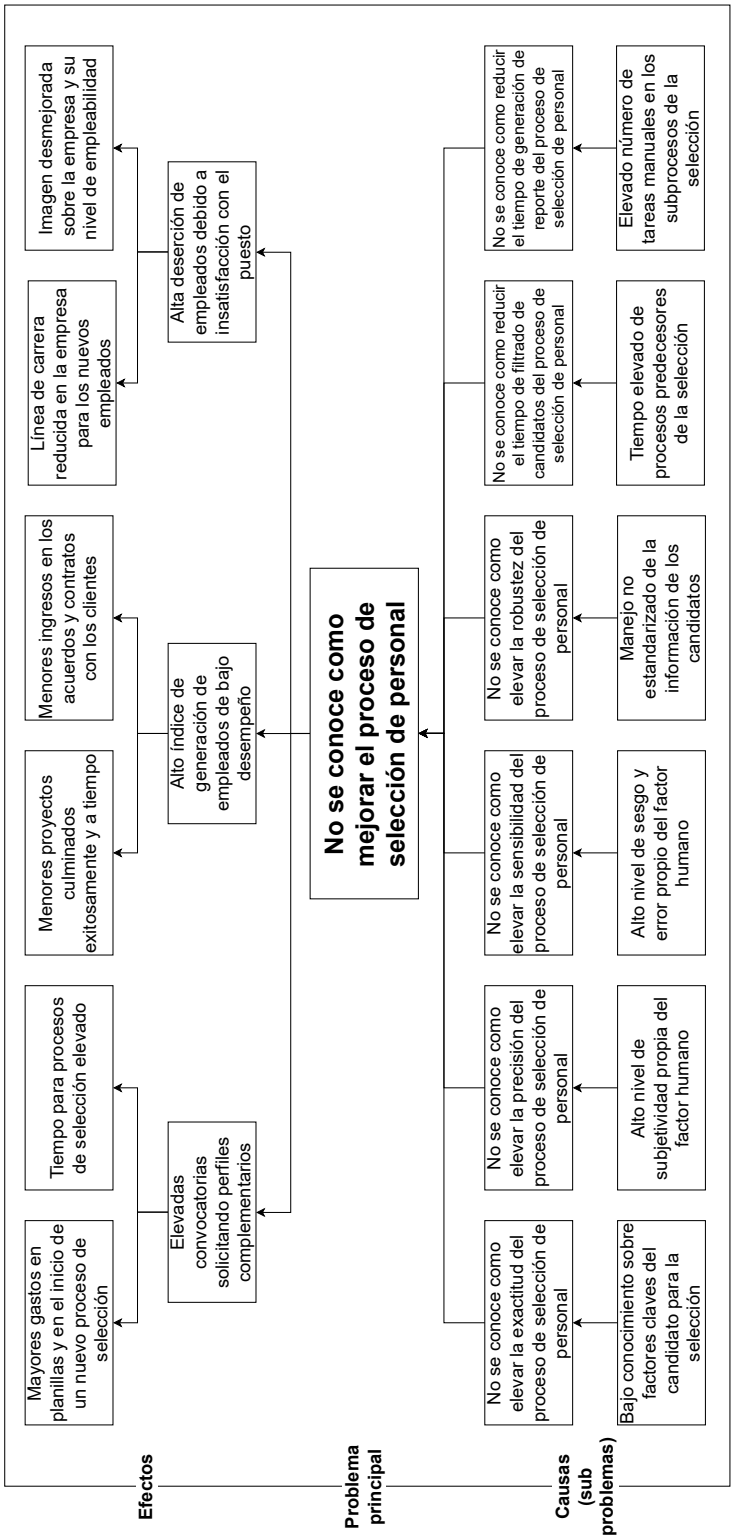


Figura A1: Árbol de problemas

Fuente: La empresa
Elaboración: Propia

ANEXO B: Árbol de objetivos

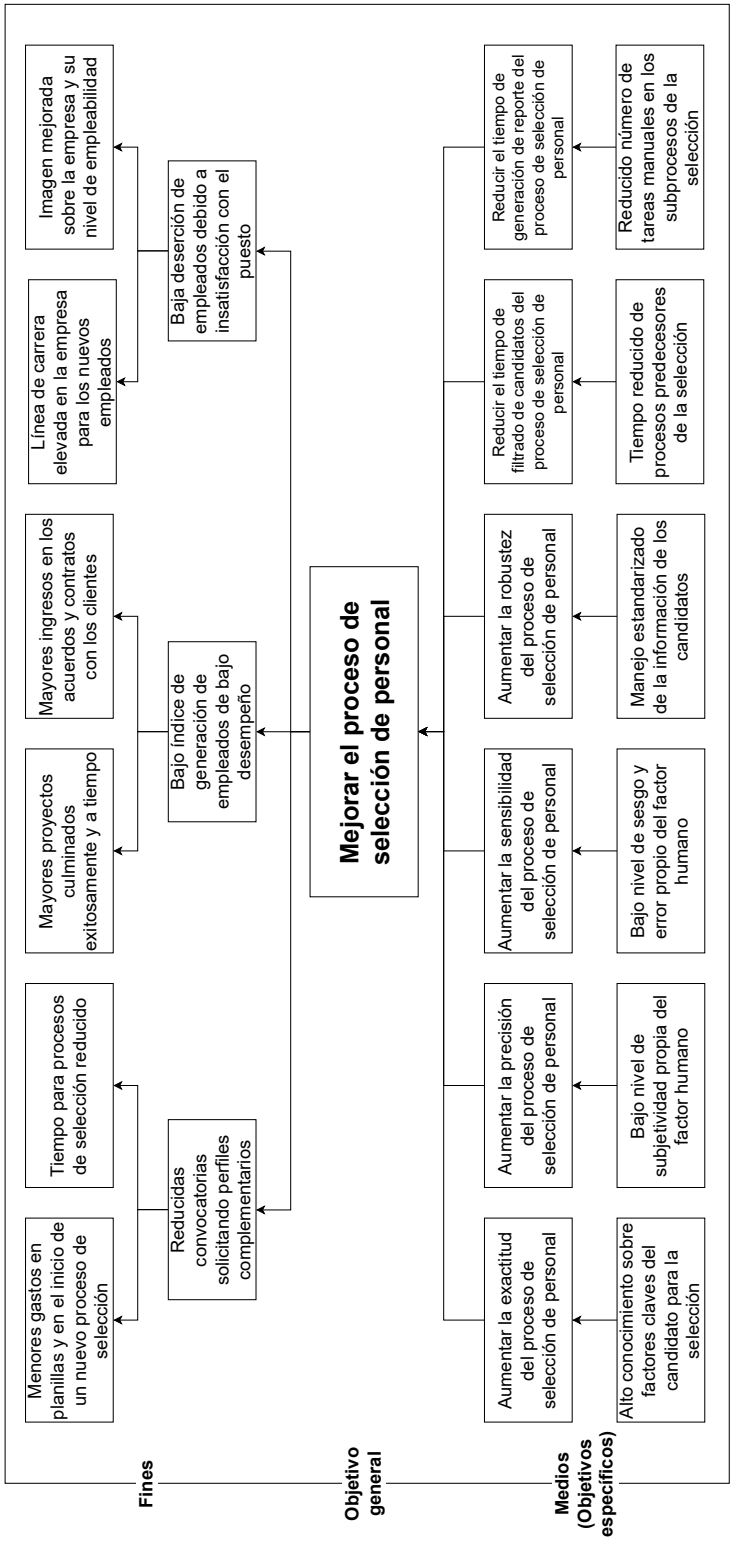


Figura A2: Árbol de objetivos

Fuente: La empresa
Elaboración: Propia

ANEXO C: Matriz de consistencia

Figura A3: Matriz de consistencia

Problema	Objetivos	Hipótesis	Variables	Indicadores	Medición
Problema principal: ¿Cómo influye el modelo predictivo en el proceso de selección de personal?	**Objetivo general:** Determinar cómo influye el modelo predictivo en el proceso de selección de personal.	**Hipótesis general:** El modelo predictivo mejora el proceso de selección de personal.	**VI:** Modelo predictivo **VD:** Proceso de selección de personal		
Problema específico: ¿Cómo influye el modelo predictivo en la exactitud del proceso de selección de personal?	**Objetivo específico:** Determinar cómo influye el modelo predictivo en la exactitud del proceso de selección de personal.	**Hipótesis específica:** El modelo predictivo aumenta la exactitud del proceso de selección de personal.	**VI:** Modelo predictivo **VD:** Exactitud	**VI:** ¿Está desarrollado el modelo predictivo? **VD:** Exactitud mayor a 95%	**VI:** ¿Está desarrollado el modelo predictivo? **VD:** Revisión de las métricas del modelo
Problema específico: ¿Cómo influye el modelo predictivo en la precisión del proceso de selección de personal?	**Objetivo específico:** Determinar cómo influye el modelo predictivo en la precisión del proceso de selección de personal.	**Hipótesis específica:** El modelo predictivo aumenta la precisión del proceso de selección de personal.	**VI:** Modelo predictivo **VD:** Precisión	**VI:** ¿Está desarrollado el modelo predictivo? **VD:** Precisión mayor a 95%	**VI:** ¿Está desarrollado el modelo predictivo? **VD:** Revisión de las métricas del modelo
Problema específico: ¿Cómo influye el modelo predictivo en la sensibilidad del proceso de selección de personal?	**Objetivo específico:** Determinar cómo influye el modelo predictivo en la sensibilidad del proceso de selección de personal.	**Hipótesis específica:** El modelo predictivo aumenta la sensibilidad del proceso de selección de personal.	**VI:** Modelo predictivo **VD:** Sensibilidad	**VI:** ¿Está desarrollado el modelo predictivo? **VD:** Sensibilidad mayor a 95%	**VI:** ¿Está desarrollado el modelo predictivo? **VD:** Revisión de las métricas del modelo
Problema específico: ¿Cómo influye el modelo predictivo en la robustez del proceso de selección de personal?	**Objetivo específico:** Determinar cómo influye el modelo predictivo en la robustez del proceso de selección de personal.	**Hipótesis específica:** El modelo predictivo aumenta la robustez del proceso de selección de personal.	**VI:** Modelo predictivo **VD:** Robustez	**VI:** ¿Está desarrollado el modelo predictivo? **VD:** Robustez mayor a 95%	**VI:** ¿Está desarrollado el modelo predictivo? **VD:** Revisión de las métricas del modelo
Problema específico: ¿Cómo influye el modelo predictivo en el tiempo de filtrado de candidatos del proceso de selección de personal?	**Objetivo específico:** Determinar cómo influye el modelo predictivo en el tiempo de filtrado de candidatos del proceso de selección de personal.	**Hipótesis específica:** El modelo predictivo reduce el tiempo de filtrado de candidatos del proceso de selección de personal.	**VI:** Modelo predictivo **VD:** Tiempo de filtrado de candidatos	**VI:** ¿Está desarrollado el modelo predictivo? **VD:** Tiempo de filtrado de candidatos menor a 5s	**VI:** ¿Está desarrollado el modelo predictivo? **VD:** Revisión de las métricas del modelo
Problema específico: ¿Cómo influye el modelo predictivo en el tiempo de generación de reporte del proceso de selección de personal?	**Objetivo específico:** Determinar cómo influye el modelo predictivo en el tiempo de generación de reporte del proceso de selección de personal.	**Hipótesis específica:** El modelo predictivo reduce el tiempo de generación de reporte del proceso de selección de personal.	**VI:** Modelo predictivo **VD:** Tiempo de generación de reporte	**VI:** ¿Está desarrollado el modelo predictivo? **VD:** Tiempo de generación de reporte menor a 0.1s	**VI:** ¿Está desarrollado el modelo predictivo? **VD:** Revisión de las métricas del modelo

Fuente: La empresa
Elaboración: Propia

ANEXO D: Cronograma

Figura A.4: Cronograma

Nº	Actividad	Duración (días)	Inicio	Fin
1	Reunión con stakeholders	14	2022-04-11	2022-04-25
2	Tener borrador de propuesta de investigación	7	2022-04-25	2022-05-02
3	Revisión y retroalimentación del borrador	7	2022-05-02	2022-05-09
4	Revisión de la literatura	49	2022-05-09	2022-06-27
5	Segundo borrador de propuesta de investigación	7	2022-06-27	2022-07-04
6	Revisión y retroalimentación del borrador	7	2022-07-04	2022-07-11
7	Tercer borrador de propuesta de investigación	7	2022-07-11	2022-07-18
8	Revisión y retroalimentación del borrador	7	2022-07-18	2022-07-25
9	Propuesta final de investigación	7	2022-07-25	2022-08-01
	Hito 1: Definición del proyecto de investigación			
10	Identificación de necesidades	14	2022-08-01	2022-08-15
11	Evaluación de la situación actual	14	2022-08-15	2022-08-29
12	Establecer objetivos de minería de datos	7	2022-08-29	2022-09-05
	Hito 2: Comprensión del negocio			
13	Lectura y procesamiento de los candidatos	14	2022-09-05	2022-09-19
14	Lectura y procesamiento de los empleados	14	2022-09-19	2022-10-03
15	Etiquetado de la variable objetivo	21	2022-10-03	2022-10-24
16	Descripción general de los datos	7	2022-10-24	2022-10-31
17	Diccionario de datos	7	2022-10-24	2022-10-31
18	Visualización de datos	7	2022-10-24	2022-10-31
	Hito 3: Comprensión de los datos			
19	Filtrado al rango de fechas delimitado	7	2022-10-31	2022-11-07
20	Eliminación de columnas con datos sensibles	7	2022-11-07	2022-11-14
21	Eliminación de columnas no relevantes	7	2022-11-07	2022-11-14
22	Eliminación de columnas normadas por Ley	7	2022-11-07	2022-11-14
23	Aplicación de equivalencias a datos de columnas	14	2022-11-14	2022-11-28
24	Reemplazo de valores nulos	14	2022-11-28	2022-12-12
25	Reemplazo de valores atípicos	14	2022-11-28	2022-12-12
26	Nuevo descubrimiento en los datos	7	2022-12-26	2023-01-02
27	Nuevo diccionario de datos	7	2022-12-26	2023-01-02
28	Nueva visualización de datos	7	2022-12-26	2023-01-02
	Hito 4: Preparación de datos			
29	Balanceo de datos	7	2023-01-02	2023-01-09
30	Codificación de las variables categóricas ordinales	7	2023-01-09	2023-01-16
31	Codificación de las variables categóricas cardinales	7	2023-01-16	2023-01-23
32	Normalización de las variables numéricas	7	2023-01-23	2023-01-30
33	Eliminación de columnas con varianza cercana a cero	7	2023-01-30	2023-02-06
34	Eliminación de columnas con correlación cercana a uno	7	2023-01-30	2023-02-06
35	Separación de datos de entrenamiento y validación	7	2023-02-06	2023-02-13
36	Entrenamiento de los modelos	28	2023-02-13	2023-03-13
37	Obtención de métricas de cada modelo	14	2023-03-13	2023-03-27
38	Selección del mejor algoritmo	21	2023-03-27	2023-04-17
	Hito 5: Modelado			
39	Aplicación de las técnicas de validación	21	2023-04-17	2023-05-08
40	Comparación de resultados de las técnicas de validación	14	2023-05-08	2023-05-22
41	Selección de la mejor técnica de validación	14	2023-05-22	2023-06-05
	Hito 6: Evaluación			
42	Revisión de la infraestructura a usar	7	2023-06-05	2023-06-12
43	Justificación de los programas a usar	7	2023-06-12	2023-06-19
44	Realización del despliegue local	14	2023-06-19	2023-07-03
	Hito 7: Despliegue			
45	Resultados	14	2023-07-03	2023-07-17
46	Discusión de los resultados	14	2023-07-17	2023-07-31
	Hito 8: Fin del proyecto de tesis			

Fuente: La empresa
Elaboración: Propia

ANEXO E: Constancia emitida por la empresa

Figura A5: Constancia emitida por la empresa

Miraflores, 20 de setiembre de 2023

CONSTANCIA

Por medio de la presente dejamos constancia que el Sr. Ronaldo Farid Nolasco Chavez, identificado con DNI N.º 76146602, tiene autorización y se encuentra desarrollando su trabajo de investigación de pregrado titulado "DESARROLLO DE UN MODELO PREDICTIVO BASADO EN APRENDIZAJE DE MÁQUINA PARA MEJORAR EL PROCESO DE SELECCIÓN DE PERSONAL EN UNA EMPRESA DE CONSULTORÍA TECNOLÓGICA" en nuestra empresa.

Dicha investigación lleva siendo realizada desde el mes de abril del 2022 hasta la fecha.

Se otorga la presente constancia para los fines que el interesado considere conveniente.

Rubén Parodi Guerrero
Gerente de Recursos Humanos
G&S Gestión y Sistemas S.A.C
Av. Del Ejército Nro. 250 interior 204
Urb. Santa cruz, Miraflores
Telefax: (511) – 715- 00 77
www.gestionysistemas.com

Telf: 51(1) 715-0077 Telefax: 51(1) 717-1763 I Av. Del Ejercito Nro. 250 Ofic. 204, Urb. Santa Cruz – Lima – Miraflores
WWW.GESTIONYSISTEMAS.COM

Fuente: La empresa