Escuela de Administración
Escuela de Ingeniería Ciencia y Tecnología

Maestría en Business Analytics

Implementación de un piloto analítico en una empresa de reclutamiento

Presentado por:

Manuel Felipe Vecino Martinez

Bogotá, D.C. 12 de mayo de 2025

Escuela de Administración
Escuela de Ingeniería Ciencia y Tecnología

Maestría en Business Analytics

Implementación de un piloto analítico en una empresa de reclutamiento

Presentado por:

Manuel Felipe Vecino Martinez

Bajo la dirección de:

Sergio Gutierrez Bonnet

Bogotá, D.C. 12 de mayo de 2025

Contenido

Preliminares

Agradecimientos

Gracias a mis padres por inspirarme durante todo mi camino educativo, a mi hermana por no dejar que me olvide de mis sueños y a Natalia Oviedo por su incondicional acompañamiento durante este proceso.

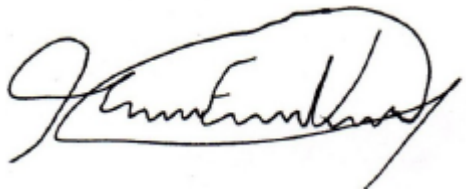
Dedicatoria

Dedicado a todas las personas que buscan una mejor vida por medio de un mejor empleo.

Declaración de originalidad y autonomía

Declaro bajo la gravedad del juramento, que he escrito el presente Proyecto Aplicado Empresarial (PAE), en la propuesta de solución a una problemática en el campo de conocimientos del programa de Maestría por mi propia cuenta y que, por lo tanto, su contenido es original.

Declaro que he indicado clara y precisamente todas las fuentes directas e indirectas de información y que este PAE no ha sido entregado a ninguna otra institución con fines de calificación o publicación.

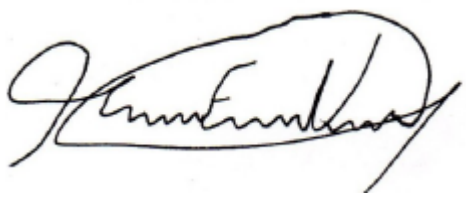


Manuel Felipe Vecino Martinez

Firmado en Bogotá, D.C. el 12 de mayo de 2025

Declaración de exoneración de responsabilidad

Declaro que la responsabilidad intelectual del presente trabajo es exclusivamente de su autor. La Universidad del Rosario no se hace responsable de contenidos, opiniones o ideologías expresadas total o parcialmente en él.



Manuel Felipe Vecino Martinez

Firmado en Bogotá, D.C. el 12 de mayo de 2025

Lista de ilustraciones

Lista de tablas

Abreviaturas

CV	Curriculum Vitae
NLP	Natural language processing (Procesamiento de lenguaje natural)
NER	Proceso de reconocimiento de entidades nombradas
ATS	Applicant tracking system (Sistema de seguimiento de candidatos)
RRHH	Recursos Humanos

Resumen Ejecutivo

Implementación de un piloto analítico en una empresa de reclutamiento

Este proyecto presenta el diseño e implementación de un piloto analítico en una empresa de reclutamiento, con el objetivo de optimizar su proceso de selección y, a largo plazo, fortalecer su madurez analítica. Esto bajo un contexto que presenta un entorno competitivo en el sector de reclutamiento, en donde la adopción de herramientas analíticas en procesos de recursos humanos está en crecimiento. Dado lo anterior, se propone lo siguiente para lograr el objetivo propuesto: por un lado, diagnosticar la calidad y uso de datos en el proceso de reclutamiento y, por otro lado, desarrollar una herramienta analítica basada en características de forma y estructura de los currículums vitae (CVs) de los candidatos. Esto, a través de técnicas de procesamiento de lenguaje natural y aprendizaje no supervisado (Segmentación con K-means). Así, se identificaron tres tipos de CVs con distintas probabilidades de avanzar en el embudo de reclutamiento, mostrando factores como la inclusión de una sección de cursos o de enlaces a sitios web externos como elementos diferenciadores. Los resultados permiten ofrecer recomendaciones particulares a los candidatos y sientan las bases para futuras iniciativas analíticas en la organización y a modo de estudios futuros. El estudio contribuye tanto al fortalecimiento estratégico de la empresa como al campo académico sobre analítica aplicada al reclutamiento.

Palabras clave

Reclutamiento, Curriculum Vitae, Procesamiento de Lenguaje Natural, Segmentación.

Abstract

Implementation of an Analytical Pilot in a Recruitment Company

This project presents the design and implementation of an analytical pilot in a recruitment company. This with the aim of optimizing its selection process and, in the long term, strengthening its analytical maturity. This project takes place within a competitive environment in the recruitment sector where the adoption of analytical tools in human resources processes is on the rise. To achieve the objective it was proposed to diagnose the quality and usage of data in the recruitment process and the development of an analytical tool based on the structure and formatting features of candidates resumes (CVs). This action is carried out using natural language processing techniques and unsupervised learning (clustering with K-means). As a result, three types of CVs were identified, each with different probabilities of progressing through the recruitment funnel. The process allowed to highlight factors such as the inclusion of a "Courses" section or links to external websites as differentiating elements. The results provide tailored recommendations for candidates and lay the foundation for future analytical initiatives within the organization and further academic studies. The study contributes both to the strategic strengthening of the company and to the academic field of analytics applied to recruitment.

Keywords

Recruitment, Curriculum Vitae, Natural Language Processing, Clustering.

1. Introducción

1.1. Contexto del problema

El sector de talento viene digitalizándose desde hace varios años. Herramientas digitales como LinkedIn se han convertido en la principal manera para conseguir empleo, así como para las empresas, conseguir el talento que necesitan. Para ponerlo en perspectiva, solo LinkedIn como plataforma de empleo cuenta con más de 900mil millones de usuarios (Toxigon, 2025) con 8,7 millones de aplicaciones diarias y 61 millones de personas semanalmente buscando empleo activamente en la plataforma (Amplitude Marketing, 2024). El 2024 continúa la tendencia post-pandemia, en dónde existen más ofertas de empleo que personas buscando trabajo (appcast, 2025). Habiendo además 50 millones de empresas registradas (Toxigon, 2025) con la necesidad de encontrar candidatos calificados, la competencia para conseguir candidatos calificados es una realidad para las compañías en la actualidad. Con esto en mente, existen empresas como la cual se estudiará en este trabajo con el objetivo de conectar el talento adecuado con las empresas buscándolo, las cuales se benefician del estado del mercado laboral actual, pero también se encuentran en esta feroz competencia, habiendo más de 27,000 compañías de dedicadas al Reclutamiento nada más en los Estados Unidos (American Staffing Association, s.f.).

De acuerdo con LinkedIn (2025), alrededor de 34% de empresas en el rubro se encuentran ya integrando herramientas analíticas en sus procesos de reclutamiento y selección. Siendo la inteligencia artificial la principal, donde el ritmo de adopción de esta es rápido, con 76% de personas usando IA en su trabajo, las cuales el 46% adoptó estas herramientas en tan solo 6 meses (Microsoft and LinkedIn, 2024). Siendo aún una tecnología en desarrollo, en donde apenas está siendo integrada en la industria, ser un adoptante

temprano puede traer beneficios tan importantes como lo son medir la calidad de las contrataciones y aumentar la eficiencia de los procesos (LinkedIn, 2025).

Como una muestra de lo anterior, en un estudio de Deloitte (Mazor, Goretsky, Moss, Cleary, & Fineman, 2018) se compara el nivel de madurez analítica, particularmente enfocado en analítica para HR y sus indicadores financieros. Dicho estudio revela que las empresas con un alto nivel de madurez analítica prácticamente duplican en flujo de caja y rentabilidad a las empresas con un bajo nivel de madurez. Entendiendo esto como una señal de la importancia de empezar la experimentación e integración de herramientas analíticas dentro de los procesos de reclutamiento y selección en este tipo de compañías de Reclutamiento.

Siendo así, se buscará entender la manera en la cual mejorar los procesos actuales de uso y manejo de datos en la organización e implementar un primer piloto del uso de una herramienta analítica que incremente la efectividad y asertividad dentro del proceso de reclutamiento, mejorando la calidad percibida frente a los clientes, aumentando la probabilidad de éxito de los candidatos en el proceso y generando una ventaja frente a la competencia.

1.2. Presentación de la empresa

La empresa en la que se basará el estudio es una compañía de Servicios de Reclutamiento fundada en el año 2016. Desde sus inicios su misión principal ha sido conectar empresas con talento a nivel global, estando especializada en procesos de reclutamiento y selección de personal para clientes en diversas locaciones, incluyendo Latinoamérica, Estados Unidos y Europa. La principal base de operaciones se encuentra en Colombia,

aunque cuenta con presencia en el Reino Unido y EE.UU, facilitando la consecución de clientes y candidatos a nivel internacional. Su operación se centra significativamente en el búsqueda y contratación de candidatos para vacantes de tecnología, siendo la gran mayoría de las vacantes con las que esta trabaja relacionadas con roles que intervienen en el ciclo de desarrollo de software, el cual comprenden la planeación, diseño, desarrollo, pruebas, despliegue y mantenimiento (AWS, 2024) de aplicaciones web y móviles. Los roles a buscar se concentran en la fase de desarrollo, siendo estos principalmente desarrolladores e ingenieros de software.

La empresa tiene un proceso ante todo simple en su concepción con el objetivo de lograr la mejor calidad posible en cada proceso. Dicho proceso comienza con la recepción del requerimiento del cliente, para posteriormente iniciar la búsqueda especializada del perfil, entrevistar los candidatos encontrados y enviar los Curriculum Vitae (de ahora en adelante CVs) más adecuados para la vacante, para posterior entrevista del cliente y contratación. En la figura a continuación se puede apreciar de forma visual el embudo de reclutamiento.

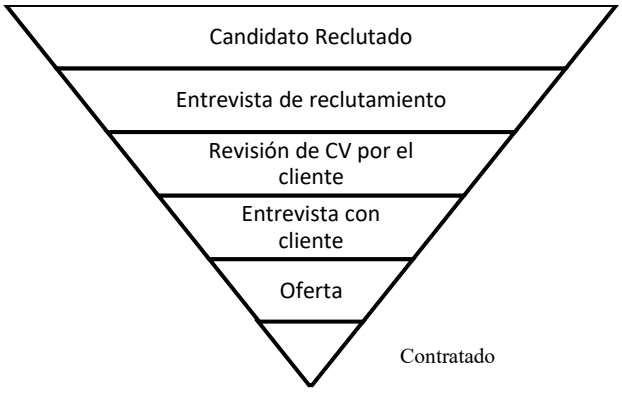


Ilustración 1: Embudo de reclutamiento

Entendiendo cómo opera la compañía y de acuerdo con el contexto competitivo explicado en la sección anterior, de la mano con la creciente adopción de la analítica en el sector de Reclutamiento, la empresa tiene el reto de adoptar herramientas analíticas que le permitan obtener una ventaja competitiva frente al resto de la industria. Particularmente, teniendo en cuenta el embudo existente y siendo este el centro de su operación, una de las oportunidades inmediatas sería la oportunidad de mejorar la tasa de conversión en cada una de las etapas del embudo con el objetivo de incrementar las contrataciones al final de este, que es desde donde vienen los ingresos, además, mejorar la calidad percibida desde el cliente e incrementar la satisfacción de los candidatos para con el proceso.

Para abordar este reto, y como parte de su objetivo de ser una organización enfocada a datos, este proyecto se enfoca en una doble estrategia: primero, sentar las bases para un manejo de datos más robusto y estandarizado, mejorando la calidad y gobernanza de la información; y segundo, desarrollar e implementar una herramienta analítica a modo de piloto que muestre las oportunidades que existen en esta área para la compañía y deje un camino a seguir para la implementación de nuevas herramientas en un futuro cercano, logrando aventajarse a la competencia.

1.3. Justificación

La realización de este proyecto es relevante tanto para la empresa, al alinearse con sus necesidades estratégicas, como por su enfoque académico e investigativo, al sentar la base para nuevos estudios de analítica en reclutamiento.

Particularmente para la empresa, el proyecto está directamente relacionado con su objetivo de ser Data-Driven, así como con la oportunidad de optimizar procesos clave en su

operación del día a día. El proyecto busca sentar las bases para que futuros proyectos analíticos impacten directamente en el núcleo de la operación. Aunque en este caso el enfoque está en el embudo de reclutamiento, se espera que a su vez impulse a un mayor nivel la madurez analítica de la compañía. A corto plazo, se busca que la optimización de la conversión en el embudo de reclutamiento pueda traducirse en un aumento de las contrataciones exitosas, una mejora en la percepción de calidad por parte de los clientes y una mayor eficiencia en la operación. Lo anterior, a mediano plazo, debe convertirse en una ventaja competitiva que le permita a la empresa distinguirse efectivamente de la competencia.

Con respecto a su relevancia académica, el estudio contribuirá con descubrimientos importantes sobre el entendimiento de factores particulares que pueden afectar un proceso de selección y sobre la aplicación de modelos supervisados y no supervisados dentro de estos procesos para la generación de herramientas analíticas basadas en ellos. Aportará, además, al entendimiento de las dificultades inherentes a la aplicación de dichas herramientas en el contexto estudiado

1.4. Pregunta de investigación

Con lo anterior en mente y teniendo en cuenta el alcance mismo del proyecto (en el que se profundizará más adelante), es necesario acotar la investigación a una única fase del embudo de reclutamiento. Esto, principalmente, para facilitar su entendimiento como proceso individual y obtener los datos de conversión y del funcionamiento interno de la fase de forma más adecuada y completa. De la misma forma, esto facilita su posterior implementación y uso.

Siendo así, al considerar nuevamente el embudo de reclutamiento y sus 5 fases, se observa que 2 de ellas (“Candidato Reclutado” y “Entrevista de Reclutador”) dependen del trabajo interno de los expertos de la compañía para encontrar al candidato adecuado y filtrarlo efectivamente en una entrevista donde se busca garantizar que cumpla los requisitos mínimos exigidos por el cliente. Asimismo, las 2 fases posteriores dependen de la decisión del cliente de avanzar con un candidato, en principio, por cómo este se caracteriza en el CV y posteriormente, por su desempeño en la entrevista con ellos. La fase de contratación simplemente depende de que el candidato acepte o no la oferta laboral.

En cuanto a las fases “Revisión del CV por parte del cliente” y “Entrevista con el cliente”, se menciona que el cliente es quien decide con quién avanzar o no, con criterios principalmente subjetivos por su parte. Sin embargo, ahondando en la fase 3 (“Revisión del CV por parte del cliente”), se encontró uno de los pocos factores que no implican interacción de 2 seres humanos durante proceso: el CV mismo.

Al ser una empresa de talento, se está todo el tiempo sujeto a caer en juicios subjetivos. Estos, como de acaba de ver de ver, se encuentran presentes en todas las etapas del proceso y difícilmente dejarán de estarlo, ya que muchas veces los criterios mismos para crear una vacante provienen de necesidades subjetivas, como, por ejemplo, las habilidades blandas de un candidato.

Dentro de su responsabilidad de garantizar los mejores candidatos para los clientes, los expertos en reclutamiento de la empresa buscan identificar estos factores subjetivos mediante su expertise en procesos de búsqueda en plataformas de empleabilidad y en entrevistas situacionales y por competencias. Lo anterior buscando enviar candidatos con las mejores habilidades posibles para dicha vacante. Pero ¿qué pasa si el CV del candidato no es

suficientemente bueno? O, más bien, ¿qué factores (más allá del contenido explícito sobre habilidades y experiencia) hacen que un CV no sea lo suficientemente bueno para el cliente y que este decida avanzar o no con un candidato? Esto se refiere a factores principalmente formales del CV que a simple vista pueden ser pasados por alto por un reclutador, a diferencia de las habilidades, estudios o experiencia del candidato, que es lo que el reclutador usa para comparar al candidato frente a los requisitos del puesto y tomar una decisión.

Dado lo anterior, habiendo entendido las fases del embudo y lo que sucede en cada una de ellas, se puede dar a entender que se profundizará en la fase de "Revisión del CV por parte del cliente" para realizar el proceso de modelado y crear la herramienta analítica.

Con este contexto en mente, las herramientas analíticas pueden ayudar a entender qué tipos de CV, en cuanto a forma, avanzan más frente a otros la revisión por parte del cliente. Además, qué características de estas hacen que tenga más probabilidad de avanzar en dicha fase. De esta manera, se tendría la oportunidad de sugerir a los candidatos agregar o modificar dichas características en sus CVs para, potencialmente, aumentar su posibilidad de éxito frente al cliente.

Con esto en mente, la pregunta de investigación a contestar es la siguiente:

¿Qué tipos de CVs, definidos principalmente por sus características estructurales y formales, se asocian con una mayor probabilidad de superar la fase de revisión por parte del cliente, y qué características específicas definen a estos tipos?

1.5. Objetivos

1.5.1. Objetivo General

Diagnosticar la gestión de datos del proceso de reclutamiento en la empresa generando recomendaciones de uso y manejo de datos, desarrollando además una herramienta analítica piloto que siente las bases junto con las recomendaciones dadas para futuras implementaciones que incrementen a largo plazo la madurez analítica de la organización.

1.5.2. Objetivos Específicos:

- Entender el estado actual de la gestión y calidad de los datos asociados al proceso de selección en la empresa.
- Diseñar un conjunto de recomendaciones para estandarizar y mejorar la recolección, calidad y uso de la información.
- Desarrollar e implementar una herramienta analítica a modo de piloto con el fin de identificar distintos segmentos o perfiles de CVs basados principalmente en aspectos formales y estructurales.

1.6. Alcance y Limitaciones

El proyecto se centra en el contexto de la empresa estudiada, abordando específicamente su proceso de reclutamiento para los roles de desarrollo de software.

Los ejes bajo los cuales se desarrollará el proyecto serán: el diagnóstico del manejo y uso de datos en la organización, particularmente los relacionados con el proceso de selección. A partir de dicho diagnóstico, se formularán recomendaciones orientadas a mejorar el manejo de datos, específicamente en el proceso de selección. Aunque la compañía utiliza datos en

otras áreas y procesos de su funcionamiento, el proyecto se limita a este proceso, ya que es el núcleo de su operación y permite realizar una investigación más profunda, dadas las restricciones temporales para el desarrollo del proyecto. Adicionalmente, las recomendaciones propuestas buscan sentar las bases para realizar este mismo análisis a posteriori en otras áreas y procesos de la organización.

Por otro lado, el segundo eje importante para el desarrollo del proyecto es la creación de una herramienta analítica piloto, que se implementará en una fase del proceso núcleo de la compañía: el proceso de reclutamiento y selección. Esta limitación a una única fase, explicada anteriormente, conduce a realizar un análisis inicial de la manera en que la compañía puede, desde su operación actual, empezar a explotar su potencial analítico para, posteriormente, crear e implementar herramientas analíticas; inicialmente, en otras fases del proceso de selección y, luego, en otros procesos de la organización. El piloto busca, inicialmente, explorar la viabilidad de identificar perfiles de CVs y analizar su posible relación con el avance de los candidatos en la fase de revisión por parte del cliente, sirviendo como una prueba de concepto y con la expectativa de continuar su desarrollo en el corto plazo.

Es importante subrayar que este proyecto tiene un carácter exploratorio y fundacional, ya que la empresa no cuenta actualmente con herramientas analíticas avanzadas dentro de sus procesos, más allá de análisis descriptivos generales. Se busca, por tanto, establecer un punto de partida para la adopción de prácticas analíticas más avanzadas en la empresa, y no lograr la implementación de una solución productiva completa.

Asimismo, el proyecto está sujeto a diferentes limitaciones inherentes a su naturaleza y contexto. En primer lugar, la disponibilidad de datos estuvo limitada, contándose

únicamente con las CVs procesadas por el autor durante el periodo de ejecución del proyecto. Esto pudo afectar la capacidad de los diferentes modelos probados para brindar resultados más precisos o generalizables.

Igualmente, es importante recordar que el modelo se enfoca en un análisis particular de las características formales y estructurales de los CVs, sin incorporar otras variables utilizadas frecuentemente en estos estudios (como se explorará más adelante en el estado del arte). La omisión de dichas variables pueden afectar el comportamiento de los modelos probados respecto al objetivo de identificar qué factores hacen que un CV avance en el proceso.

Del mismo modo, las limitaciones de tiempo y recursos computacionales restringieron el acceso a herramientas más avanzadas de procesamiento de lenguaje natural. Esto pudo haber limitado el potencial de extracción de características adicionales, afectando potencialmente los hallazgos de la investigación post-extracción.

En la sección de recomendaciones para trabajo futuro, se profundizará en el alcance y las limitaciones aquí establecidas, proponiendo sugerencias para proyectos futuros que puedan abordar aspectos fuera del alcance definido para este proyecto.

2. Marco Teórico y Estado del Arte

2.1. La Analítica de Datos en la Gestión del Talento Humano

Como se venía mencionando en la introducción, las empresas y áreas de reclutamiento vienen integrando herramientas analíticas para incrementar su efectividad. Al ser una subárea del sector de RRHH, es necesario entender cómo está impactando el área de datos dentro de este sector.

Para empezar, es posible entender que desde hace algunos años hasta la fecha se viene implementando la analítica en RRHH, conocida también como People Analytics. Así se puede ver en el artículo de Harris et al. (2011), que menciona la necesidad de adoptar un enfoque basado en datos en RRHH para continuar el desarrollo de las áreas de talento humano, con enfoques en diferentes áreas, incluyendo la generación de acciones predictivas para prevenir la fuga de talentos y optimizar las cadenas de suministro de talento por medio de información en tiempo real para las compañías y sus proveedores de talento.

Un estudio más reciente, el de Conte et al. (2023), analiza el uso de la analítica en recursos humanos frente a otras áreas organizacionales, encuestando a 90 líderes al respecto. Este revela que, en promedio, menos del 6.5% de las empresas participantes tenían implementadas herramientas analíticas en el área de RRHH, Aunque un estudio más reciente de LinkedIn (2025) sitúe la estadística en un 11% con respecto al año anterior, sugiriendo un crecimiento, estas cifras indican una adopción aún limitada. De hecho, el estudio de Conte et al. (2023) también señala que apenas el 3.6% de los esfuerzos analíticos se destinaban específicamente al área de reclutamiento. Por lo tanto, aunque la tendencia de adopción sea

creciente, existe un considerable margen de desarrollo en la aplicación de la analítica en RRHH, y particularmente en reclutamiento.

Por otro lado, un aporte importante del estudio de Walford-Wright et al. (2018) es la constatación de que empresas como Google utilizan herramientas analíticas en sus procesos de reclutamiento y selección para reducir el sesgo humano en sus filtros y seleccionar al mejor talento para sus posiciones, independientemente de su edad, educación o género. Este enfoque en la objetividad es relevante para contextualizar el presente estudio, que busca revisar principalmente los factores objetivos de los CVs.

En resumen, la literatura revisada indica que, si bien la analítica de datos ofrece un potencial significativo para transformar la gestión del talento humano y los procesos de reclutamiento, permitiendo optimizar decisiones y reducir sesgos, su adopción real, particularmente en selección, es aún muy baja en muchas organizaciones. Este contexto justifica la exploración de nuevas aplicaciones analíticas, como la que aborda este proyecto centrado en el análisis de características formales y de estructura de Hojas de Vida, con el fin de contribuir a la eficiencia y efectividad del reclutamiento en contextos específicos.

2.2. Técnicas y Desafíos en la Extracción de Información de Hojas de Vida

La extracción automática de información de las CVs es un paso importante para el desarrollo de este proyecto. De acuerdo con lo anterior, se tiene la oportunidad de ver las maneras y enfoques con los cuales se ha realizado dicho proceso en ocasiones anteriores.

Para este proceso, se han utilizado principalmente técnicas de Procesamiento del Lenguaje Natural (NLP), como puede verse en los estudios referenciados de Ali et al. (2022), Sowjanya et al. (2023), Saatçı et al. (2024) y Zu et al. (2019). El enfoque de NLP tiene total

sentido, ya que su objetivo es transformar el contenido de texto del CV en datos organizados con los que sea posible trabajar, lo cual es básicamente lo que se hará en el estudio actual también. En dichos estudios se encuentra un enfoque importante en la extracción del texto para el entendimiento posterior del contenido por medio de modelos de clasificación, como redes neuronales o KNN, que permitan organizar la información y posteriormente compararla frente a una descripción de cargo, realizando lo que se conoce comúnmente como un screening (Saatci, Kaya, & Ünlü, 2024).

De la misma manera, se hace énfasis en la extracción en los trabajos de Sowjanya et al. (2023) y Saatçı et al. (2024), usando la librería de NLP preentrenada llamada SpaCy para dicha extracción. Se explica en el trabajo de Satheesh et al. (2020) que SpaCy facilita un proceso de reconocimiento de entidades nombradas (NER, por sus siglas en inglés), el cual permite categorizar entidades específicas en un cuerpo de texto. Lo anterior puede ser bastante valioso para el proceso requerido en el trabajo actual.

Algo muy curioso que se percibió en los artículos estudiados es que en la gran mayoría de ellos el enfoque se encuentra en extraer el contenido del candidato, sin centrarse particularmente en el estudio de las características formales y de estructura; de hecho, en varios estudios se busca eliminar varias de estas características antes del procesamiento, como es el caso del de Zu et al. (2019). En este caso, se puede entender que este estudio puede llegar a sentar una base para investigaciones futuras que busquen enfocarse en este campo, particularmente en el estudio de las características formales de los CVs, considerando que la variedad de formatos es reconocida como una limitación al realizar este proceso de extracción (Ali, Mughal, Khan, Ahmed, & Mujtaba, 2022).

Se puede entender así que las técnicas de Procesamiento del Lenguaje Natural, como el reconocimiento de entidades nombradas (NER) facilitado por herramientas como SpaCy, son importantes para extraer el contenido de los CVs. Así mismo, se puede entender que la revisión anterior indica una tendencia a priorizar el contenido sobre la forma. Teniendo en cuenta que la diversidad en estructuras de los los CVs presenta un desafío, el presente estudio, al centrarse precisamente en estas características formales y estructurales, se posiciona como una exploración complementaria con potencial para aportar nuevos *insights* al campo de la analítica en el reclutamiento.

2.3. Madurez analítica, manejo y uso de Datos.

La base de cualquier iniciativa analítica exitosa es la calidad y el uso correcto de los datos. Desde aquí**,** es importante establecer procesos que permitan asegurar la calidad de los datos, como un paso fundamental antes de poner en marcha proyectos de analítica. Lo anterior no es solo un mero capricho; por el contrario, las empresas cuya madurez analítica es mayor cuentan con un mejor desempeño financiero frente a las menos maduras (Mazor, Goretsky, Moss, Cleary, & Fineman, 2018), lo cual se evidencia particularmente en su Retorno sobre Activos y sus márgenes de utilidad (Gonzales, 2025). Desde lo anterior, es importante entender lo que representa la madurez analítica en la organización y, de la misma manera, el manejo de la calidad de dichos datos para su implementación exitosa.

Relativo a la madurez, existen varios modelos que permiten evaluar dónde se encuentran las organizaciones en términos de analítica. Por ejemplo, está el modelo expuesto por la Transportation Research Board (*2024*) que divide la madurez en 5 etapas, donde la primera implica trabajar con datos pequeños y estáticos, para pasar por ser una organización

con analítica localizada, luego analítica en escalamiento, posteriormente a nivel organizacional, hasta llegar al quinto nivel, el cual consiste en ser un referente en analítica en el sector particular de la empresa. De la misma manera, otro modelo como el de Crunchr (*2025*) establece 4 cuadrantes, definidos por los ejes de uso (entendiendo si es para solo reportes o para analítica) y frecuencia de uso (siendo esta de forma oportunista o sistemática). Cuando una empresa usa los datos de forma oportunista y solo para reportes, es una empresa de baja madurez analítica, mientras que, si utiliza los datos de forma sistemática para la analítica, se puede definir como una empresa de alta madurez. Ambos modelos aportan perspectivas diferentes que permiten entender la situación actual de la organización, así como la oportunidad de ver dónde puede mejorar.

Para lo anterior, esté la organización en el nivel de madurez en el que se esté, es importante contar con calidad en los datos para obtener buenos resultados. Las empresas usualmente no miden el impacto financiero de una mala calidad de datos, lo cual posteriormente trae problemas en términos de costos y pérdida de posibles oportunidades de negocio (Moore, 2018). De esta forma, existen varias metodologías para entender la calidad de datos en las organizaciones y procesos, como lo es el del DAMA y sus 6 dimensiones de precisión, completitud, consistencia, puntualidad, unicidad y validez (IBM, 2025). De la misma manera, existe el marco de Gartner que estudia la precisión, amplitud, consistencia y profundidad de los datos (Luka & Karkunova, 2025). Ambas perspectivas permiten analizar de forma profunda la calidad de datos de la organización y seguramente serán aplicadas en este estudio para el entendimiento de los datos actuales de la empresa.

Así mismo, se entienden varias razones para que la calidad de los datos no sea la mejor, como lo pueden ser tecnologías desactualizadas o el ingreso manual de los datos (Luka

& Karkunova, 2025), la resistencia al cambio en la cultura organizacional, entre otras (Vorecol, 2024). Dado lo anterior, existen múltiples maneras en las cuales dichas dificultades pueden ser sobrepasadas, como eliminando barreras o intermediarios al momento de ingresar los datos (Vasquez, 2025). Por ejemplo, ingresándolos directamente en un sistema, en vez de pasarlos primero por un cuaderno o documento de texto adicional. También es útil centralizar la información y definir procesos estándar para la compañía (Luka & Karkunova, 2025), por medio de un manual de uso de datos o un documento que establezca buenas prácticas, por ejemplo.

En definitiva, es importante para implementar analítica en la empresa, entender el nivel de madurez analítica de la organización y asegurar la calidad de los datos de esta y sus procesos. Utilizar modelos de madurez y marcos de calidad de datos nos dará una guía estructurada para evaluar la situación actual y poder partir de ahí para planificar mejoras y garantizar el correcto funcionamiento del piloto analítico y posteriores proyectos analíticos puestos en marcha para la organización.

3. Metodología

3.1. Marco metodológico CRISP-DM

Para la creación de la herramienta analítica piloto y la posterior propuesta de mejoras en la gestión de datos de la empresa, se adoptó como referencia la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining), explicada por IBM (2021). Esta se divide en 6 fases, las cuales serán abordadas a profundidad en las siguientes secciones para mostrar el desarrollo del trabajo.

Las fases de CRISP-DM inician con la comprensión del negocio, enfocada en entender los objetivos organizacionales y cómo la analítica puede ayudarlos, para posteriormente realizar el entendimiento de los datos disponibles. Consecuentemente, se preparan los datos para construir un conjunto de datos con calidad que permita posteriormente el modelado. Así, se pasa a la fase de modelado para obtener un modelo óptimo que busque resolver las necesidades de negocio planteadas en la primera fase. Asimismo, se realiza la evaluación de este para validar que funciona correctamente y, finalmente, se despliega para su aplicación.

Aunque estas fases sugieren un modelo en secuencia, la metodología permite iterar varias veces con el objetivo de optimizar cada vez más los modelos resultantes en caso de que surjan nuevos requerimientos o hallazgos. Por lo tanto, esta metodología ofrece una oportunidad para sentar las bases con la herramienta piloto y permite que, posteriormente, se pueda iterar sobre la misma para su mejora y escalado.

3.2. Fase 1: Entendimiento del Negocio

Antes de iniciar cualquier tipo de proceso con datos, es necesario entender las expectativas de la organización frente al uso de estos. Esto es la fase de comprensión del negocio de acuerdo con la guía de IBM (2021), la cual consiste no solo en el entendimiento de dichas expectativas, sino también en la delimitación de objetivos de negocio que guíen todo nuestro proceso de análisis de datos.

De acuerdo con lo anterior, el proceso de entendimiento de negocio es lo que se viene realizando en el punto 1 del presente trabajo, estableciendo el contexto de la compañía, entendiendo su operación y su necesidad de implementar analítica en sus procesos.

Como un breve resumen, hay que recordar que la intención es generar una herramienta que permita incrementar la conversión en la fase 3 del embudo de reclutamiento. Esto, por medio de una herramienta que permita sugerir a los candidatos de las vacantes de desarrollo de software cambios en sus CVs, con el objetivo de que tengan más posibilidades de pasar a la fase de entrevista con el cliente. Esto se hará únicamente considerando los factores de estructura y forma del CV, ya que la idea es poder utilizar la herramienta en diferentes vacantes, por lo que la misma debe aplicar de forma general a CVs de diferentes grupos de habilidades dentro del desarrollo de software.

Asimismo, esta herramienta analítica piloto nos permite sentar las bases del escalado de la madurez analítica de la organización, de la mano de una mejora en la calidad y uso de los datos (esto a partir de un análisis y posteriores recomendaciones), con el objetivo de mejorar la posición competitiva de la organización en el mercado.

Con esto en mente, los criterios de éxito del piloto se definen, a nivel técnico, como el desarrollo efectivo de un modelo que nos permita definir las características de los CVs y entender cómo se agrupan frente a dichas características y respecto a la probabilidad que tienen o no de pasar a la siguiente fase. A nivel de negocio, el éxito viene a partir de la implementación de dicho modelo y su capacidad de permitir a los reclutadores sugerir a los candidatos cambios particulares en los CVs que, a mediano plazo, mejoren la tasa de conversión en la fase 3 del embudo de reclutamiento.

3.3. Fase 2: Entendimiento de los Datos

Una vez recordado el contexto de negocio y habiendo definido los criterios de éxito, es importante entender los datos disponibles para realizar el análisis.

3.3.1. Fuentes de Datos Utilizadas

3.3.1.1. El ATS

La compañía cuenta con un sistema de seguimiento de candidatos o ATS (por sus siglas en inglés). El ATS almacena la información de cada una de las vacantes, candidatos y clientes con los que se trabaja en la organización. El ATS contiene toda la información correspondiente al embudo de reclutamiento, o al menos debería tenerla, ya que, como se ve a continuación, hay bastantes cosas que corregir en términos de calidad de datos allí. Como reflexión, es importante para la creación de futuras herramientas analíticas entender la calidad de los datos dentro del ATS y generar las recomendaciones correspondientes para que esta se encuentre en la mejor forma posible.

En el caso particular del proyecto, la revisión de calidad de datos se centrará en la información de los candidatos, ya que en principio está más relacionada con el estudio a realizar, además de que es la más numerosa dentro del ATS, siendo que se procesan alrededor de 3000 candidatos por trimestre.

Para realizar la revisión propuesta, se descargó la información de alrededor de 3460 candidatos en formato CSV para su posterior importación al software que se utiliza para el análisis. La cifra corresponde a la cantidad de candidatos procesados en el último trimestre del año 2024. Esta base cuenta con 37 variables, que referencian principalmente datos demográficos del candidato; datos de experiencia general, como años de experiencia; datos relativos a información salarial, como compensación actual, beneficios y expectativa salarial; además de datos relacionados con el proceso mismo, como su fecha de creación, actualización o contratación.

3.3.1.2. CVs del ATS

En este caso y para este proyecto en particular, se separan las CVs dentro del ATS de la información del ATS en general por varias razones. En primer lugar, porque estas serán el enfoque del estudio de características particulares, donde la única información a obtener adicional desde el ATS es si el candidato que representa esa CV avanzó o no en el proceso (resaltando, igual, la importancia de la calidad de datos en el ATS). En segundo lugar, porque son datos no estructurados, por lo cual deberán ser objeto de una transformación a datos estructurados para su posterior descripción y modelado.

De esta manera, para esta ocasión en particular, se tomaron CVs procesadas en varias vacantes de desarrollo de software. Particularmente, el criterio para escoger las CVs a trabajar fueron las CVs a las cuales el autor del presente proyecto tuvo acceso directamente en el desarrollo de su trabajo, con el objetivo de no perturbar la operación de la empresa durante la realización del proyecto. En total, se pudieron obtener como muestra unas 635 CVs, las cuales se separaron en 2 carpetas para su posterior procesamiento: en una se encuentran las CVs que avanzaron y en otra las que no, de las cuales se eliminaron 21 que tenían valores nulos, dejándonos con 614 CVs como muestra final.

3.3.2. Análisis Exploratorio y de Calidad de Datos

3.3.2.1. Calidad en el ATS

Como se mencionó en el punto anterior, se realizará un entendimiento de la calidad de datos del ATS, particularmente de la información de los candidatos. Para dicho análisis de calidad, se toma como base el marco de 6 dimensiones de la calidad del DAMA (IBM,

2025), siendo estas las referidas a completitud, consistencia, precisión, validez, puntualidad y unicidad de los datos en una base de datos.

(El proceso completo de análisis se dejará como Anexo 1 si es de interés del lector ahondar en el mismo).

Frente a la dimensión de completitud, que como su mismo nombre lo indica, se refiere a que todos los valores de una columna se encuentren presentes (IBM, 2025), se encuentran hallazgos importantes. Por ejemplo, hay datos clave como el correo electrónico, con 265 registros faltantes, lo cual es crítico, ya que es el principal medio de contacto de la compañía con el candidato, además de un identificador clave para evitar duplicados. De la misma forma y referente a datos de contacto, hacen falta un total de 808 registros del número de teléfono. Esto es un poco más comprensible, ya que a muchos candidatos no les gusta revelar su número de teléfono por privacidad; aun así, es un medio de contacto importante en un proceso de selección y sería importante contar con más de ellos, ya que es el método por el cual se realizan procesos como agendamiento de entrevistas e incluso ofertas.

Por otro lado, se puede ver que hay una cantidad de datos importante, alrededor de un 30% faltante, en lo que respecta a las variables de salario actual y salario esperado. Esto representa una falta de información crítica, ya que los salarios son usualmente usados como uno de los principales factores para tomar decisiones frente a si se avanza o no con un candidato en las primeras etapas del proceso.

Se ve, de igual manera, una falta de datos considerable en diferentes variables, incluso en aquellas donde estas deberían ser automáticamente completadas por el sistema que alberga el ATS. Dado lo anterior, se debe revisar el papel del reclutador y los encargados de datos en garantizar la completitud.

En lo que respecta a validez, la cual se refiere a que los datos se ajustan a un formato específico (IBM, 2025), se estudiaron principalmente los datos cuyo formato debe ser coherente con cierto formato, el cual puede ser establecido por la empresa o por el sistema mismo, siendo estas variables los nombres, correos, números de teléfono y salarios. En lo que respecta a salarios y correo electrónico, el sistema les da formato automáticamente, por lo que se consideran datos con validez en todos sus registros en la base.

Por otro lado, en lo que respecta a los nombres, se busca que estos cumplan con no tener caracteres especiales para su fácil lectura, principalmente por parte de los clientes que se encuentran en países de habla inglesa. Aquí solo alrededor del 5% de los candidatos cuentan con nombres no válidos frente al formato buscado. Por otro lado, en lo relativo a números de teléfono, sí es posible encontrar bastantes disparidades dada la presencia de símbolos, espacios o la falta del código país. Al ser el teléfono importante para la comunicación, es importante que todos cuenten con el formato correspondiente para evitar confusiones, e incluyendo el código, al contar con candidatos de múltiples países.

En lo que respecta a unicidad, que busca que existan datos únicos donde corresponda (IBM, 2025), se encontraron 43 candidatos duplicados por medio del correo electrónico. Así mismo, en lo referente a consistencia, que refiere a que los valores en una columna cumplan con una regla específica (IBM, 2025), la columna más relevante frente a esta es que las columnas de “nombre” y “apellido” sean consistentes con lo mostrado en la columna “nombre completo”, lo cual no se cumple en el 40% de los casos, tema que debe ser revisado.

Por último, frente a precisión y puntualidad, los cuales respectivamente se refieren a que los valores se encuentren cercanos a valores reales y que la data represente la realidad desde un punto de tiempo requerido (IBM, 2025), no se encontraron temas particularmente

relevantes, dado que, en puntualidad, por proceso, todos los candidatos se suben al ATS tras la entrevista, por lo que no hay problemas de puntualidad; y frente a precisión, gracias a los campos predefinidos y validaciones automáticas del sistema CRM, se dificulta que existan valores atípicos. El único campo donde existieron valores atípicos fue en salarios, a lo cual se le deberá poner un especial cuidado.

El análisis anterior buscaba mostrar que existen bastantes retos y desafíos en lo respectivo a calidad y uso de datos en la organización, por lo cual se buscará brindar recomendaciones correspondientes a dicho análisis que, de la mano del piloto analítico, dejen unas bases fuertes para empezar a implementar nuevos proyectos analíticos en la organización en las diferentes fases del proceso de reclutamiento.

3.3.2.2. Calidad de las CVs del ATS

Teniendo en cuenta que son datos no estructurados, de los cuales se realizará la extracción como parte del proceso; el entendimiento de la calidad se realizará durante y posterior al proceso de extracción de información de la CV, el cual será explicado en la sección 3.4. del presente proyecto.

3.4. Fase 3: Preparación de los Datos

La fase de Preparación de Datos es bastante importante en CRISP-DM, ya que nos permite transformar los datos brutos o, en este caso, no estructurados, en un set de datos estructurado y apto para un proceso de modelado. De acuerdo con IBM (2021) se estima que esta etapa es de las más extensas en los proyectos de analítica, ya que suele llevar el 50%-70% de esfuerzo y tiempo del proyecto.

Para el caso particular, así fue este el caso, ya que hubo que enfocar esta sección del trabajo hacia el proceso de extracción de información de las CVs que nos permitiera posteriormente analizarla. Este proceso de extracción debía permitir extraer las características formales y estructurales de la CV, pasarlas a texto y este, posteriormente, transformarlo en un conjunto de datos estructurado. Lo anterior tomó 3 intentos diferentes hasta que se llegó a un script que permitió extraer la información de la mejor forma posible, el cual será descrito a continuación.

3.4.1. Extracción de información de las CVs

Para el proceso de extracción de las CVs, se contó con diferentes procesos, los cuales serán explicados durante el desarrollo de este numeral.

3.4.1.1. Uso de herramientas técnicas:

Como se había explicado anteriormente, se dividieron las CVs en dos carpetas, una con el grupo que pasó y otra con el que no pasó a la fase de entrevista con cliente. De acuerdo con esto, era necesario encontrar un recurso técnico que nos permitiera extraer la información detallada de las CVs, de forma relativamente rápida, ya que había que procesar más de 600 CVs y sobre las mismas realizar pruebas del funcionamiento del script y modelado.

Con esto en mente, se utilizó Python como lenguaje de programación, al contar con librerías para extracción de PDFs. En este caso, PyMuPDF, la cual es una librería cuyo proceso de extracción y lectura de PDFs es considerablemente más rápida que la otra opción disponible que es PyPDF (Gardiner, 2024). Con lo anterior en mente, es importante aclarar que para el presente estudio únicamente se utilizaron CVs en formato PDF, esto dado que es

el formato más común en el que se reciben CVs en la organización, por una amplia mayoría, además que facilita el proceso de extracción al tener que usar una única herramienta. PyMuPDF también cuenta con una herramienta para localizar colores, texto e imágenes dentro de los documentos.

Una vez definida la herramienta para extraer el texto, se definió para el entendimiento del mismo varias herramientas para su procesamiento. En esta ocasión y remitiéndonos los estudios realizados anteriormente para extracción y procesamiento de texto en CVs, se ve que se utilizaban procesos de NLP para el procesamiento del texto, lo cual se realizó desde la librería SpaCy (Sowjanya et al., 2023; Saatçı et al., 2024). La anterior es una librería de código abierto que permite realizar todo el proceso de tokenización, etiquetado, asignación de atributos, lematización y el posterior reconocimiento de entidades (Explosion, 2025). Así mismo, SpaCy cuenta con modelos pre-entrenados para detectar texto en diferentes idiomas, lo que facilitó el proceso de extracción y evito el entrenamiento particular de un modelo. Aunque, sería bueno tener esto a consideración, ya que podría aumentar la precisión de este tipo de estudios realizar el entrenamiento propio del modelo, aunque este se sale del alcance de la presente investigación, pero se abordará en la sección 7 al discutir el trabajo futuro.

Por otro lado, se utilizó la librería RapidFuzz, la cual no se podría decir que entiende lenguaje natural, pero es útil para el procesamiento de datos tipo String, los cuales son procesados más rápidamente por esta librería que por SpaCy. En este caso en particular, se utilizó para entender jerarquías dentro del texto.

Por último, se utilizó Pandas para organizar todo dentro de un dataframe y otras librerías, como lo son Os para acceder a los archivos, dateutil para las fechas y counter para

contar elementos. Así, habiendo entendido las herramientas usadas, se puede pasar al proceso mismo de extracción.

3.4.1.2. Proceso de extracción de información

Como se mencionó anteriormente, se realizaron varias iteraciones y 3 scripts de extracción antes de llegar al script final que nos permitiera extraer de manera completa y coherente la mayor cantidad de información de las CVs frente a forma y estructura. Con lo anterior, después de la revisión manual de las CVs y con ayuda de las herramientas, en la iteración final se llegaron a 4 grupos grandes de características a extraer, sobre las cuales se haría el análisis. Los grupos fueron variables correspondientes a longitud, ya fuera total o solo de una sección; variables correspondientes a formato; variables correspondientes a diseño; y variables correspondientes a completitud.

De la misma manera, antes de iniciar la extracción, hubo que definir tres diccionarios previos que le ayudaran al modelo de Spacy a interpretar cierto contenido. En primer lugar, se realizó un diccionario de secciones, que le permitió al modelo entender mejor los posibles títulos de las secciones más comunes y definir las mismas dentro del documento. Por otro lado, se realizó un diccionario de términos técnicos, para posteriormente ayudar al modelo a entender cuáles de estos son usados en la CV, además de no tomarlos como un error. Por último, se dio un diccionario que le brindara al modelo los formatos de fechas más comunes, con el objetivo de que pudiera identificar fechas más fácilmente y no confundirla con otro tipo de valores numéricos. Los diccionarios se pueden encontrar como Anexo 2.

Cuando se hablan de variables de longitud, se obtuvieron variables tales como el conteo total de palabras, de páginas, y la longitud de cada sección detectada. Particularmente,

estas variables nos ayudan a entender la longitud general del CV para posteriormente compararla con otras variables.

Con respecto a las variables de formato, se extrajo información correspondiente a los formatos en las fechas, identificando la manera en la que el candidato expresa la temporalidad de sus experiencias y educación, por ejemplo, viendo si solo pone el año, mes y año, o pone el día exacto también, además de la forma en la que lo hace. Así mismo, se extrajo el formato del texto, como la manera en la que está escrita la CV, ya sea con viñetas, párrafos o ambas, la tipografía y tamaño que este usa, además del uso de negritas y cursivas en el texto. De la misma manera, se definió la consistencia del mismo formato, entendiendo si este usaba diferentes tipos o tamaños de fuentes y la justificación del texto.

Con respecto al diseño, se obtuvo el uso de colores y gráficos en las CVs, se detectó si esta contaba o no con foto, o si tiene elementos gráficos como íconos, tablas o emojis. En particular para la detección de colores, se obtuvo el porcentaje de color usado en la CV y fue separado por su uso en el texto o en dibujos, ya que hay documentos que usan color de forma muy sutil, mientras que en otras es bastante protagónico en el documento.

Por último, la dimensión de completitud se enfocó en entender la presencia de las diferentes secciones definidas en el diccionario previamente mencionado. De la misma manera, se detectó el uso de links externos hacia LinkedIn o Github como páginas que refieren a perfil profesional y portafolio, o uso de otra página web dentro del CV, ya sea para referirse a un Website personal o a al website de algún proyecto en el que se trabajó. Así mismo, utilizando el diccionario de términos técnicos, se obtuvo una variable que define el porcentaje de uso de lenguaje técnico en la CVs.

Al finalizar el proceso de extracción, se obtuvo un dataframe con 55 variables, de valores numéricos, booleanos y categóricos. Más adelante se detallará la exploración y transformación de dicha información para el proceso de modelado. Así mismo, de acuerdo con la carpeta en la que se encontraba el CV, se le asignó en una nueva columna llamada 'Passed' un valor de 0 en caso de que se encontrara en el grupo de CVs que no pasaron, y de 1 en caso de que estuviera en el grupo de CVs que pasaron. En esta ocasión, contando con 352 CVs que no pasaron y 262 que, si pasaron, asumiendo una base balanceada frente a esta variable.

Una vez realizada la extracción automática, se realizó una revisión manual para entender que la misma tuviera sentido y fuera coherente con las CVs usadas. La anterior fue una de las formas de evaluación en cada iteración del proceso de extracción, que nos ayudó a llegar al script final que, finalmente, fue utilizado.

3.4.2. Ingeniería y Selección de Características

Una vez creado el dataframe original, se empezó a realizar la transformación de variables, se crearon nuevas características para el estudio y se empezó a entender variable por variable para escoger las más relevantes para la etapa de modelado.

De esta manera, iniciando por la transformación de características, se transformaron las variables booleanas de valores tipo String (true, false) a valores binarios (0,1), esto dado que los modelos que se probarán únicamente pueden procesar valores numéricos. Por otro lado, se agruparon variables como las de formato de fecha o tamaño de fuente con el objetivo de contar con grupos de variables mejores definidos al momento de modelar y obtener menor

ruido. Así como también se transformó la variable de lista de tamaño de fuente, que enlistaba los tamaños de fuente usada, en un promedio de estas.

Una vez realizada la agrupación, se utilizó la función get_dummies en Pandas para transformar las variables categóricas a variables binarias (pandas , 2024) y facilitar su procesamiento para el modelo.

Con respecto a la creación de nuevas variables, se creó una nueva variable que nos permite ver el número de secciones totales de la HV basado en las secciones extraídas anteriormente, complementando la dimensión de completitud. Por otro lado, para robustecer el análisis de longitud, se crearon variables que miden la densidad de la experiencia laboral frente al texto total, para descartar que la cantidad de experiencia laboral misma influya en algo. De la misma manera, se obtuvo la densidad del texto frente al número de secciones, con el objetivo de entender que tan densa es cada sección y si esta completitud afecta en la cantidad del texto también.

Una vez creadas y transformadas las variables, se obtuvo un dataframe únicamente con valores numéricos, compuesto de 80 variables, con el cual se procedería al modelado, el cual es visible en el Anexo 3. Aunque, antes de modelar, al contar con un dataframe nuevo, se realizó un entendimiento de los datos con este para entender su comportamiento a nivel estadístico y dirigir el modelado de mejor manera.

3.4.3. Descripción de la base de datos

Una vez lograda la extracción y la creación de nuevas características, se buscó realizar la descripción de las variables desde un punto de vista estadística para dejarlas listas para un eventual modelado.

En primer lugar, se obtuvo la distribución de cada una de las variables en gráficos de barras y circulares para su mejor entendimiento. Los anteriores fueron utilizados para comprender el comportamiento de las variables, particularmente, las categóricas con varias categorías y proceder al agrupamiento, explicado en el punto anterior.

Por otro lado, se tenía la intención de detectar outliers con dos intenciones particulares, la primera, identificar los mismos para eliminar ruido en el modelo y, en segundo lugar, entender la consistencia de la extracción, ya que, aunque sí deben ser diferentes, la gran mayoría de CVs son parecidas por su concepción misma, por lo que no debe haber muchos outliers. Siendo así, se utilizó la metodología z-score, la cual consiste en la estimación del valor Z de un elemento al normalizar los valores del conjunto de datos. Si este valor Z es mayor a 3, se considera este valor un outlier (R, Aakash, 2024). Dicha metodología elimina las propiedades de posición y escala de los datos, ayudando a asociar datasets que son disimiles, como lo es el presente (VenkataAnusha, Anuradha, Murty, & Kiran, 2019). De esta manera y entendiendo que las variables se distribuyen normalmente, se aplicó dicha metodología, encontrando no más de 20 outliers en todas las variables numéricas. Con esto en mente, es posible establecer que la extracción fue buena, y que en efecto existen CVs con valores extremos dadas varias particularidades, como lo pueden ser uso de fuentes muy grandes, una gran cantidad de imágenes o simplemente un exceso de texto. Los outliers serán eliminados para el posterior modelo; al no ser muchos, no se perderá mucha información.

Así mismo, se estudiaron las columnas cuya varianza fuera menor a un umbral definido para entender si las mismas se encontraban agrupadas cerca a la media, las cuales

no contribuyen mucho al modelado (C, Jahnavi, 2022). En este caso no se obtuvieron variables con varianza baja, por lo cual se mantenienen todas para el estudio.

Posteriormente, se realizó el estudio de correlaciones de las variables, tanto entre todas las variables entre sí, como entre todas las variables y la variable “Passed”, la cual es la variable para estudiar en este trabajo en particular.

Siendo así, se obtuvieron múltiples variables con correlaciones bastante altas entre sí, como por ejemplo la variable de ‘Cantidad de Palabras’ con la de ‘Longitud de texto extraído’, las cuales se correlacionan directamente con total lógica, ya que, entre mayor longitud de texto extraído, mayor cantidad de palabras debe haber. Otro ejemplo es la correlación entre las variables ‘Densidad de texto por sección’ y ‘Cantidad de palabras’, entendiendo que entre más palaras, más palabras deben existir por sección dentro de la CV. Así mismo, se dieron correlaciones perfectamente negativas para variables complementarias, como por ejemplo las de ‘Tiene foto’ y ‘No tiene foto’ o ‘Uso de colores’ y ‘No uso colores’. Por lo anterior, es necesario eliminar dichas variables que generan tanto ruido para un eventual modelo, lo que se hará después de entender la correlación de las variables existentes frente a la variable ‘Passed’.

De acuerdo con lo anterior, al evaluar la correlación de las variables del dataset con la variable ‘Passed’, se obtiene un resultado bastante curioso y es que no existe ninguna que siquiera supere siquiera un valor de 0,1 de correlación, lo cual es significativamente bajo y desde un análisis superficial nos puede llevar a decir que no existe ninguna variable que influya directamente en la variable ‘Passed’. Con lo anterior en mente, es importante continuar analizando las variables frente a la variable ‘Passed’ para entender cuáles pueden

ser las más útiles para el modelo, ya que del estudio de correlación no se obtuvo dicha información.

Como fue mencionado anteriormente, para obtener un modelo efectivo, es necesario reducir variables que se encuentren correlacionadas entre ellas, diferentes a la variable a estudiar, las cuales fueron mencionadas anteriormente. Por lo que se utilizó la metodología del factor de inflación de la varianza (VIF por sus siglas en inglés), la cual estudia dentro de un modelo de regresión múltiple la manera en la que los errores estándares, por ende, la varianza, se expanden dada la existencia de multicolinealidad (The Pennsylvania State University, 2018). Con esto en mente, al realizar el análisis del VIF, de las 80 variables, se obtuvieron 53 variables con un VIF mayor a 10. Con esto en mente, es necesario escoger cuales se correlacionan y están generando dicho VIF para poder eliminarlas y dejar solo las adecuadas para el modelado.

Dado lo anterior, se tomaron 32 variables a eliminar, las cuales son las que tenían VIF infinito o se correlacionaban directa o indirectamente con otra variable en el dataframe. Así mismo, se tuvo cuidado de que, para esta eliminación de variables, se dejara por lo menos representación de alguno de los factores que mostraban dichas variables. Por ejemplo, lo mismo se mantuvieron variables para las categorías de formato como ‘Tipo de Fuente’, ‘uso de viñetas’, ‘consistencia de márgenes’, entre otras. En términos de completitud, se mantuvieron algunas variables que representan la completitud de secciones, como los ratios o longitud y la variable de ‘secciones completas’. Frente a diseño, se mantuvieron variables como las que muestran el uso de colores en el texto o en dibujos, además de la que muestra el uso de foto. Frente a longitud, se eliminó la variable ‘cantidad de palabras’, ya que es prácticamente la misma que la de ‘Longitud de texto extraído’. Así, entre otras eliminaciones

de variables, se obtuvo como resultado final un dataframe con 50 Variables, el cual se puede ver como ejemplo en el Anexo 4, o más resumidamente en la tabla a continuación:

Tipo de variable	**Nombre de Variable**
De longitud	- Numero de Paginas - Densidad Informacion (%) - Lineas_education - Lineas_work_experience - Lineas_skills - Lineas_certifications - Lineas_achievements - Lineas_professional_profile - Lineas_languages - Lineas_projects - Lineas_publications - Lineas_training_courses - Lineas_volunteer_work - texto_extraido_len
De formato	- Tamaño fuente probable - Variedad de fuentes - Variedad de tamaños - Uso de negritas (estimado %) - Uso de cursivas (estimado %) - Promedio tamaño fuente - Formato Texto (Lineas)_Párrafos - Formato Texto (Lineas)_Viñetas - Orden Temporal_Orden Temporal Detectado - Formato Fecha Más Común_MM-YYYY - Formato Fecha Más Común_MM/YYYY - Formato Fecha Más Común_Mon YYYY (EN) - Formato Fecha Más Común_YYYY - Fuente principal_ArialMT

	- Fuente principal_Calibri - Fuente principal_Otra - Fuente principal_Tahoma - Legibilidad general_Buena - Consistencia tamaños - Fuente_Consistente - Consistencia márgenes
De diseño	- Cantidad de imágenes - Tiene Elementos Graficos - Uso de colores (texto) - Uso de colores (dibujos) - Deteccion Foto Perfil
De completitud	- Porcentaje Lenguaje Técnico - Tiene LinkedIn - Tiene GitHub - Tiene Website/Otro - Seccion_languages - Seccion_publications - Seccion_training_courses - secciones_completas - Ratio_Lineas_Experiencia_len
Variable a estudiar	- Passed

Tabla 1: Dataframe de 50 variables

En ese dataframe, al volver a realizar el estudio del VIF, solo una variable cuenta con un VIF superior a 5, siendo esta 'Lineas work experience' que se correlaciona con la variable 'Ratio líneas de experiencia', pero en este caso, a pesar del VIF, ambas deben mantenerse, a una representar la proporción y la otra la longitud misma, factores que se deben entender para ver si son estos los que afectan o no la oportunidad de pasar a un siguiente filtro con el cliente o no.

Una vez entendidas las correlaciones con la variable 'Passed', era necesaria también entender qué variables se relacionan más con ella desde otros procesos, como lo son pruebas

estadísticas que nos permitan, mediante hipótesis, entender qué variables, ya sean numéricas o categóricas se relacionan más con la variable a estudiar. Por lo anterior, se realizó la prueba-t, la cual evalúa las medias de dos o más grupos mostrando si hay diferencias significativas entre ellas (JMP Statistical Discovery, 2025), para las variables numéricas, que se deben evaluar por sus medias. Por otro lado, se realizó la prueba chi-cuadrado, la cual consiste en comprobar si las frecuencias que se observan se ajustan con los valores esperados frente a otra variable (JMP Statistical Discovery, 2025), esto hecho para las variables categóricas (binarias en este caso), las cuales cuentan su frecuencia de ocurrencia para la realización de la prueba.

Es necesario recordar que el factor de éxito de las pruebas a realizar se encuentran el valor p, el cual representa la importancia del resultado de la prueba (Arias, 2017), en donde un valor p menor a 0,05 es un resultado significativo, un valor entre 0,05 y 1 es marginalmente significativo y un valor mayor a 1 es no significativo.

Siendo esto así, empezando por la prueba-t, se tomaron las variables numéricas para probarlas bajo la hipótesis de que la diferencia de medias entre las que su CV respectivo pasó o no a la siguiente fase, es diferente a cero. Teniendo en cuenta el factor de éxito establecido, únicamente 2 variables cuentan con un resultado significativo y dos con un resultado marginal. En este caso, las variables ‘Longitud del texto extraído’ y ‘Secciones completas (#)’ tuvieron un valor p menor a 0.05, que frente a la hipótesis nos dice que la diferencia de medias entre las CVs que pasaron o no pasaron para estas dos variables es significativa. Para las variables ‘Número de páginas’ y ‘Uso de negritas’ su diferencia de medias es marginalmente significativa entre las CVs que pasaron o no, frente a estas dos variables, teniendo un valor p entre 0,05 y 0,1.

En lo que respecta a la prueba chi-cuadrado, se probaron las variables categóricas, frente a la hipótesis de que la frecuencia de las variables estudiadas es diferente entre las CVs que pasaron y las que no pasaron. Al realizar la prueba, de forma similar a la prueba anterior, se obtuvieron 2 variables con un resultado significativo y 2 con un resultado marginal. Las variables 'Tiene link a un Website (otro)' y 'Tiene sección de cursos/entrenamiento' son significativamente diferentes entre las CVs que pasaron y las que no pasaron, con un valor p menor a 0,05. De la misma manera, las variables de 'Uso colores en el texto' y 'Formato de fecha más común: Mes/Año', las cuales tienen un valor p de entre 0,05 y 0,1, teniendo una diferencia marginal en la frecuencia observada en estas variables entre las CVs que pasaron y las que no.

Con lo anterior presente, y para no tener únicamente pruebas que representan relaciones principalmente lineales como las realizadas anteriormente, se tomó también en consideración realizar una prueba bajo un modelo de aprendizaje que permitiera entender la importancia de las variables en una posible predicción, particularmente de la variable 'Passed' la cual es la variable a estudiar. Siendo así, se usó el modelo de Random Forest, en el cual es fácil determinar la contribución de una variable en particular al modelo (IBM, s.f.). Al realizar este proceso, se obtuvo una nueva variable que mostró relevancia, la cual fue 'Porcentaje de Lenguaje técnico usado', que, aunque se muestra relevante en el modelo Random Forest, el modelo mismo muestra un score de precisión de apenas 0,53. Por lo cual, aunque será inicialmente considerado, ya que nos da una perspectiva nueva para evaluar en un eventual modelo. El hecho de que el modelo Random Forest tenga una precisión tan baja, nos empieza a dar señales de la dificultad de predecir la variable 'Passed', aunque esto será explorado más a detalle en la sección de Modelado.

Al tener en cuenta lo anterior, se obtuvieron sets de datos nuevos para considerar para el modelado: uno con 9 variables a parte de la variable 'Passed', que cuenta con las 8 variables con resultados significativos y marginales en las pruebas. El otro, con las 4 variables, más la variable 'Passed', únicamente aquellas con resultados significativos en las pruebas. Las variables se muestran en la siguiente tabla:

Tipo de variable	Dataset de 9 variables	Dataset de 4 variables
De longitud	- Numero de Paginas - texto_extraido_len	- texto_extraido_len
De formato	- Uso de negritas (estimado %) - Formato Fecha Más Común_MM/YYYY	
De diseño	- Uso de colores (texto)	
De completitud	- Porcentaje Lenguaje Técnico - Secciones_Completas - Tiene Website/Otro - Seccion_training_courses	- Tiene Website/Otro - Seccion_training_courses - Secciones_Completas

Tabla 2: Sets de datos de 9 y 4 variables

Una vez realizado este proceso, se puede proceder a la etapa de modelado, que nos permita construir un Modelo Analítico que nos permita contestar nuestra pregunta analítica, además de ser la base para construir la herramienta a utilizar y poder incorporarse en el proceso de negocio de la empresa.

3.5. Fase 4: Modelado

Una vez preparados y estructurados los datos extraídos de las CVs, la fase de modelado se centra en seleccionar y aplicar técnicas analíticas que nos permitan responder la pregunta planteada al inicio del trabajo.

3.5.1. Enfoque Analítico

Con lo anterior en mente, es relevante recordar la pregunta de investigación para contrastarla frente a los datos existentes y, desde allí partir hacia el modelado. Siendo esta la siguiente:

¿Qué tipos de CVs, definidos principalmente por sus características estructurales y formales, se asocian con una mayor probabilidad de superar la fase de revisión por parte del cliente, y qué características específicas definen a estos tipos?

Con esto en mente, se deben encontrar tipos de características y cómo estas se asocian con la probabilidad o no de pasar a una siguiente fase, es decir, de acuerdo con los datos existentes, que la variable 'Passed' en las mismas es 1, en caso de pasar o, 0 en caso de que no.

De la misma manera, al realizar el estudio descriptivo anterior, se encontraron hallazgos bastante particulares, como lo son la baja correlación lineal entre las características formales y de estructura frente a la variable 'Passed', además de un bajo rendimiento predictivo inicial sugerido por el modelo Random Forest para obtener la importancia de las características.

Sucede, así mismo, que, al tener una variable particular a estudiar, siendo 'Passed' en este caso, se asumió erróneamente que se podría utilizar un modelo supervisado de

aprendizaje con dicha variable como objetivo, realizando intentos con modelos como Regresión Lineal, Árboles de Decisión, xGBoost, entre otros. Dados los factores comentados anteriormente, además de una posible falta de datos para robustecer el modelo (esto se explorará mejor en la sección de trabajo futuro), ninguno de los modelos tuvo una precisión mayor a 0,53, teniendo sus respectivas curvas ROC un valor muy cercano al Azar. Todo esto, inclusive modificando hiperparámetros dentro de cada uno de los modelos probados. Así, con los indicadores de rendimiento en mente, estas pruebas fueron catalogadas como insuficientes (Martínez Pérez & Pérez Martin, 2023). Al tener en cuenta lo anterior, se determinó que un enfoque netamente predictivo supervisado no sería el más fructífero.

Dado lo anterior, con el objetivo, no solo de tener resultados más interpretables, sino de lograr contestar la pregunta de investigación, en donde se busca identificar si existen grupos o tipos de CVs que se asocien más con la probabilidad de pasar o no, desde sus características de forma y estructura; se tomó el camino exploratorio y descriptivo. Lo anterior se realizó utilizando una técnica de aprendizaje no supervisado, que nos permita agrupar dichos grupos e identificar las características que los asocian de forma más fuerte a la probabilidad de pasar a una siguiente etapa en el embudo de reclutamiento.

3.5.2. Selección del Modelo

Una vez entendiendo lo que se necesitaba para la realización del modelado, era importante escoger el modelo de agrupación adecuado. Con esto en mente, se seleccionó el algoritmo K-Means para segmentación, el cual es una técnica de aprendizaje no supervisado para dividir un conjunto de datos en un número de Segmentos (K) predefinido (Universidad de Oviedo, s.f.). Este se eligió sobre otros modelos de segmentación ya que permite agrupar

CVs con características similares sin asumir ninguna relación predictiva a priori, lo que nos brinda una mejor perspectiva para después comparar cada grupo y su probabilidad de pasar o no a la siguiente etapa.

3.5.3. Proceso de aplicación del Modelo

Una vez teniendo presente el modelo a utilizar, y después de las pruebas realizadas y la aclaración del objetivo, era momento de realizar su aplicación.

En primer lugar, se realizaron diferentes pruebas con los conjuntos de variables ya definidos anteriormente: El de 50 Variables que representaba cada uno de los datos extraído originalmente; el de 9 variables (con importancias significativas y marginales frente a la variable 'Passed' además de la variable más importante en la prueba de Random Forest); por último, el de 4 variables únicamente con las variables con alta significancia en las pruebas estadísticas realizadas.

Para aplicar este modelo, fue necesario estandarizar las características, ya que, al ser un algoritmo basado en distancias, deben todas contar con la misma escala (media 0, varianza 1) para evitar sesgos dentro de la misma.

Asimismo, se utilizó el método del codo, una técnica empleada para determinar el número de grupos óptimo, que consiste en identificar el punto en el cual se observa un cambio significativo en la tasa de disminución de la varianza entre clústeres (Rodríguez, 2023). Así, se obtienen diferentes números de segmentos óptimos para los diferentes sets de datos probados, siendo k=4 para el de 50 variables, k=4 para el de 9 variables y k=3 para el de 4 variables.

Utilizando este número de segmentos, se ejecutó el algoritmo en los diferentes sets de datos, obteniendo resultados mixtos. Para medir los resultados, se utilizaron 2 indicadores de éxito. El primero fue la prueba chi-cuadrado, la cual como se explicó anteriormente consiste en comprobar si las frecuencias que se observan se ajustan con los valores esperados frente a otra variable (JMP Statistical Discovery, 2025), en este caso bajo la hipótesis de que la frecuencia de aparición de la variable 'Passed' en cada segmento tiene relación con el grupo en el que se encuentra. La otra prueba utilizada, es el score de silueta, una métrica la cual mide la separación y estructura de los segmentos, en donde los valores cercanos o mayores a un score de 0,5 indican que el algoritmo produjo clústeres separados y válidos (SPSS Analysis, 2025).

Siendo así, para la primera base de datos (50 variables, k=4) se obtuvo un valor p de 0,83 en la prueba chi-cuadrado, mostrando casi nula significancia en la relación entre los segmentos y la variable 'Passed'. Para la base con 9 variables (k=4), se obtuvo un resultado en la prueba Chi-Cuadrado de 0.0156, lo que muestra significancia en la relación, pero su score de silueta fue de apenas 0,19, lo que nos lleva a concluir que tiene una estructura muy débil y los segmentos solapados entre ellos. Por último, al realizar el proceso con la base de 4 variables(k=3), la prueba Chi-Cuadrado arrojó un valor p de 0,0052, por lo que es posible decir que la frecuencia de aparición de la variable 'Passed' en cada segmento para este set de datos, tiene relación con el segmento en donde se encuentra. Asimismo, se obtuvo un coeficiente de silueta de 0,48, que, aunque podría ser más alto, dadas las limitaciones conocidas de la base, nos brinda igual un indicador importante de que, al ser un valor cercano a 0,5, la segmentación cuenta con una buena estructura y poco solapamiento con los otros clústeres.

Con esto en mente, este último (4 variables, k=3) será el algoritmo y set de datos finales con los que se realizará la creación de la herramienta analítica. Siendo así, es importante comprender los resultados que arrojó el modelo y comprenderlos a detalle.

3.5.4. Comprensión del modelo

Una vez realizado el modelado final, se obtuvo un modelo de segmentación por medio del algoritmo k-Means que representaba las 4 variables que más significancia tuvieron en las pruebas, siendo estas 'Longitud del texto extraído', 'Secciones Completas (#)', 'Tiene link a un Website (otro)' y 'Tiene sección de cursos/entrenamiento'. El modelo obtenido se ve representado en una gráfica bidimensional gracias a un proceso de reducción de dimensionalidad realizado por un análisis de componentes principales.

El análisis de componentes principales o PCA (por sus siglas en inglés) es una técnica de reducción de dimensiones que extrae las características más importantes de grandes sets de datos, conservando a la vez la información relevante del conjunto de datos original (IBM, 2023.). Con esto en mente, se reducen a dos dimensiones los datos obtenidos del modelo para ser representados en la siguiente gráfica:

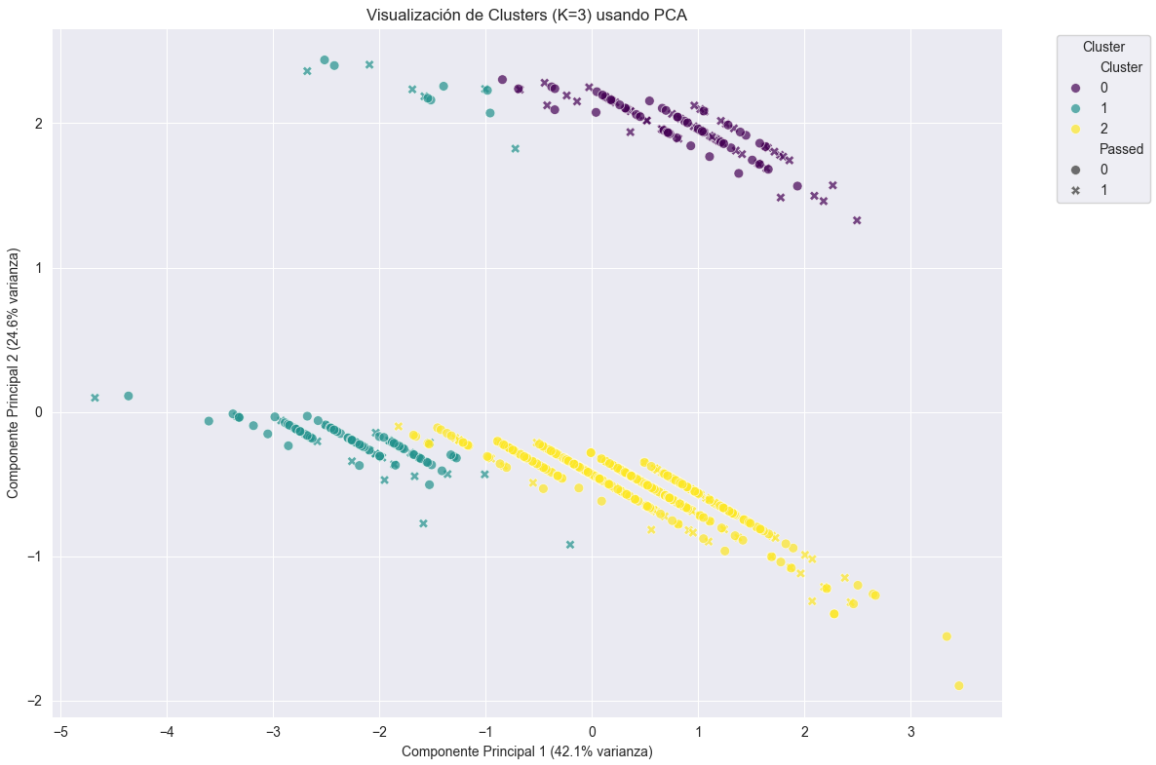


Ilustración 2: Representación del modelo (K=3, 4 Variables)

Para realizar la interpretación de la gráfica, es necesario primero entender que representa cada uno de los ejes, en este caso, los primeros dos componentes principales. Así, empezando por el Componente Principal 1, este representa principalmente si el texto cuenta o no con secciones completas y si tiene sección de cursos/entrenamiento; medianamente representa también la longitud del texto. El Componente Principal 2, representa prácticamente en su totalidad si el CV tiene o no un link a un website externo (Omitiendo LinkedIn y GitHub).

Con lo anterior en mente, es posible interpretar la gráfica, en donde se observan tres segmentos definidos, el 0, el 1 y el 2. Empezando por el Segmento 0, donde los CVs este siempre tiene link a un website externo y sección de cursos/entrenamiento, además de ser

ligeramente más largo que el promedio. El Segmento 1 agrupa CVs que casi nunca tiene link a un website ni sección de cursos/entrenamientos, además de ser más corto y con menos secciones que el promedio. Por último, el Segmento 2 es un segmento que es posiblle llamar promedio o estándar, ya que su longitud y secciones son promedio frente a la generalidad del set de datos; nunca tiene link a una web externa, pero siempre tiene sección de cursos/entrenamientos.

Al haber entendido los segmentos, es importante entender cómo se distribuye en si mismo los CVs que pasaron dentro de cada uno de estos, siendo que para el Segmento 0, el 54.4% de los CVs pasaron; para el Segmento 1, únicamente el 32.4%; y para el Segmento 2, el 42.5% (una medida cercana al promedio general).

Revisando estos hallazgos, se pudo entender que los perfiles definidos por estas 4 variables están asociados con diferentes tasas de éxito en el proceso. Particularmente, las CVs que tienen links a una web externa y una sección de cursos/entrenamiento (Segmento 0) presentan una mayor tasa de éxito. De la misma manera, los CVs más cortos y con menos secciones que el promedio (Segmento 1), se ven asociados a tener una tasa de éxito menor que la media.

Una vez comprendidos los hallazgos del modelado, se debe evaluar el mismo para un posterior despliegue dentro de la empresa, en el marco de ser un piloto analítico para la organización que permita escalar desde allí su madurez analítica.

3.6. Fase 5: Evaluación

Una vez construido el modelo, la fase de evaluación se enfoca en validar que el modelo es técnicamente correcto en función de sus criterios de rendimiento técnico, además

de evaluar si los resultados del mismo también cumplen con los factores de éxito establecidos al inicio del proyecto (IBM, 2021).

Con lo anterior en mente, es preciso recordar que los factores de éxito planteados al principio del proyecto. A nivel técnico, se propuso el desarrollo efectivo de un modelo que nos permita definir las características de los CVs y entender cómo se agrupan frente a dichas características y respecto a la probabilidad que tienen o no de pasar a la siguiente fase. Lo anterior fue evaluado por medio de las métricas internas de calidad del segmentación y la prueba de hipótesis que nos permita probar su relación con la variable 'Passed'.

Con respecto al negocio, el factor de éxito planteado viene a partir de la implementación de dicho modelo y su capacidad de permitir a los reclutadores sugerir a los candidatos cambios particulares en los CVs. Esto se evaluará mediante el análisis de la relevancia de este modelo frente al proceso actual de la empresa.

Iniciando con la evaluación técnica, se utilizaron varias métricas para comprobar la calidad técnica del modelo. Inicialmente, se utilizó el coeficiente de silueta, recordando que es una métrica que mide la separación y estructura de los clústeres, en donde los valores cercanos o mayores a un score de 0,5, indica que el algoritmo produjo clústeres separados y válidos (SPSS Analysis, 2025). El modelo final (4 variables, k=3) obtuvo un coeficiente de silueta de 0.48, que, si bien un valor ideal estaría más cerca de 1, al obtener un valor cercano a 0,5, indica que los clústeres tienen una estructura definida y existe una separación aceptable entre ellos. Desde una perspectiva técnica, el algoritmo logró encontrar agrupaciones con coherencia interna, logrando un modelo que nos permite definir las características de los CVs y entender cómo se agrupan frente a estas.

Continuando con lo técnico, es importante evaluar también la relación de dichos grupos con la probabilidad o no de pasar a una siguiente fase. Con esto en mente, se realizó una prueba Chi-Cuadrado, con la hipótesis de que la frecuencia de aparición de la variable 'Passed' en cada segmento tiene relación con el grupo en donde se encuentra. Siendo el resultado de esta prueba altamente significativo, con un valor p de 0,0052. En este caso, existe evidencia estadística fuerte para afirmar que la pertenencia a un determinado segmento está asociada con la probabilidad de que el CV pase o no el filtro del cliente. Con esta conclusión, es posible dar por hecho que se cumple el factor de éxito del proyecto a nivel técnico.

Por el lado del negocio, su evaluación, aunque será mejor explorada en el capítulo de implementación del presente trabajo, se puede concluir de la misma manera que es posible realizar la implementación del modelo analítico como herramienta que ayude a los reclutadores a sugerir cambios a los candidatos a sus CVs para incrementar su probabilidad de éxito frente a la revisión con el cliente.

3.7. Fase 6: Despliegue

La fase final de la metodología CRISP-DM consiste en el despliegue, que es básicamente la implementación de lo descubierto en la empresa (IBM, 2021). Con esto en mente, y dado el carácter de piloto analítico de este proyecto, esta fase se centró en la planificación y conceptualización de cómo dichos hallazgos serán utilizados en el corto y mediano plazo para cumplir los objetivos de negocio. Lo anterior, de la mano de la creación de un script a modo de herramienta analítica de uso diario para los reclutadores, como entregable final del ejercicio. Este será de fácil uso e integración dentro del proceso diario de

la operación, con el fin de que no interrumpiese (o interrumpa) el trabajo diario del reclutador de la empresa.

Con esto en mente, para abordar con mayor detalle este plan de implementación de la herramienta analítica, dicho proceso será abordado con total detalle en el capítulo 6 del presente trabajo.

4. Resultados

Aunque ya presentados inicialmente, este capítulo busca presentar a nivel general los resultados obtenidos durante el proceso expuesto en la metodología, de forma concreta para facilidad de lector.

4.1. Resultados del Modelo

Al realizar el proceso al final, inicialmente por medio del estudio de las variables y características extraídas, además de la realización de pruebas estadísticas sobre las mismas, se llegó a un modelo de segmentación K-Means, seleccionando 3 segmentos, basado en 4 características particulares, sobre una muestra de 614 CVs. Recordando que, al final, las 4 características más relevantes para el estudio, basado en las pruebas y análisis realizado, son: 'Longitud del texto extraído', 'Secciones completas (#)', 'Tiene link a un Website (otro)' y 'Tiene sección de cursos/entrenamiento'.

Teniendo en cuenta lo anterior, el modelo indicó 3 tipos de CVs, separados en la misma cantidad de clústeres. Empezando por el Segmento 0, este se refiere a un tipo de CV que cuenta con la presencia constante de un enlace a un website externo y la sección de cursos o entrenamiento, con una longitud de texto y número de secciones completas superior al promedio general. El Segmento 1, se refiere a un tipo de CV con ausencia casi total de la

sección de Cursos y Entrenamiento y de links a una web externa, además de que, en promedio, su longitud y número de secciones completas es menor. Por último, el Segmento 2, también entendido como el segmento estándar, al mostrar el comportamiento más cercano al promedio de la base, presenta CVs con sección de cursos y entrenamiento, pero ausencia total de un link externo a una website, así como su longitud y número de secciones son promedio.

Al lograr ubicar las CVs dentro de cada uno de estos clústeres, también lo se pueden asociar con su probabilidad de pasar a la siguiente ronda, siendo la del Segmento 0 la más alta, con 54.4% de probabilidad, la del Segmento 1 la más baja con 32.4% y la del Segmento 2 con un valor cercano al promedio general, con un 42.5%. Estas, aunque parecen no ser diferencias significativas, la diferencia de más de 20% que presenta el Segmento 1 con respecto al Segmento 0, o del 10% con el Segmento 2, puede representar una pérdida de candidatos en esta etapa que a mediano plazo signifique la pérdida de contrataciones. Lo mismo con la diferencia entre el Segmento 0 y el 2, teniendo la posibilidad de incrementar la probabilidad de pasar a los candidatos alrededor de un 10%, se facilitaría aumentar la conversión final en el embudo de reclutamiento.

Con esto en mente, al poder ubicar las nuevas CVs que se incorporen al proceso dentro de un segmento en particular, será posible brindar sugerencias a los candidatos sobre como modificar su CV para poder incrementar ligeramente su probabilidad de éxito con el cliente. Esto se explicará más a fondo en la sección siguiente, Aplicabilidad.

4.2. Aplicabilidad

Una vez entendidos los resultados, es mucho más factible lograr aplicarlos al día a día del negocio, particularmente, al proceso de reclutamiento que se viene abordando en el presente trabajo y la manera en la que este puede beneficiar al negocio.

En primer lugar, el resultado obtenido es un complemento a la evaluación del reclutador, en el sentido en el que, de igual manera, el reclutador sigue teniendo el criterio propio para entender si una CV es apropiada o no para enviar al cliente, pero, gracias a la posibilidad de tomar en cuenta estas características, el reclutador puede realizar sugerencias adicionales al candidato que antes escapaban su percepción particular, teniendo en consideración el segmento en el que la CV se encuentre.

Siendo así, lo positivo de este resultado, es en primer lugar, la identificación de una posible CV que puede ser riesgosa enviar de forma más sencilla; por ejemplo, si esta pertenece al Segmento 1, sabiendo que tiene menor probabilidad de pasar el filtro del cliente y, con esto, el reclutador puede poner especial cuidado al realizar las recomendaciones al candidato.

De la misma manera, se le puede brindar retroalimentación mucho más específica al candidato frente a su CV y qué cambiar. Usualmente, cuando se le pide a un candidato que cambie algo de su CV, no se sabe muy bien por dónde empezar. Con esta herramienta, es posible dar sugerencias particulares sobre las secciones que más afectan la probabilidad de pasar, además de cualquier sugerencia frente al contenido a criterio del reclutador. Como ejemplo, si una CV se encuentra en el Segmento 2, se le puede sugerir añadir enlace a un portafolio a los websites de sus proyectos, además de agregar nuevas secciones al CV en caso

de esto ser necesario (y posible). Las posibles sugerencias particulares para cada segmento se profundizarán más adelante.

Todo lo anterior, será realizado por medio de la implementación de una herramienta analítica dentro del flujo de trabajo de los reclutadores, la cual ubica las CVs en su segmento respectivo por medio de un script de Python. El funcionamiento y despliegue de esta herramienta será explicado en la sección de Implementación y Recomendaciones.

5. Discusión

Una vez entendidos los resultados del estudio, es necesario entender cómo estos se contrastan frente a los estudios existentes, además de las limitaciones de los mismos. Esto, habiendo entendido ya la forma en la que dichos resultados pueden afectar a la empresa, permite también entender como aportan a los estudios existentes en el campo.

En la presente revisión de literatura, se tuvieron en cuenta estudios que realizaban procesos de extracción de información de CVs utilizando, en su gran mayoría, procesamiento de lenguaje natural, como lo es el presente. La mayor diferencia, es que dichos estudios estuvieron enfocados en la extracción del contenido de la CV, relacionado principalmente con la experiencia, estudios, entre otros factores, como por ejemplo se puede ver en los trabajos de Zu, et al. (2019). En este caso, particular, el enfoque de la extracción de información únicamente para las características de estructura y forma se dio ya que funciona para el caso particular a estudiar en la empresa, pero, se puede complementar con estos trabajos al añadirlo al estudio de contenido.

De la misma manera, otra diferencia fundamental que existe al respecto se encuentra en el objetivo de dichos estudios, los cuales varios de los cuales buscaban ya sea comparar

la información de un CV frente a una descripción de cargo específica para clasificarlos, como en los estudios de, Sowjanya et al. (2023) y Saatçı et al. (2024) o para ranquearlos, como lo es el de Satheesh et al. (2020). Aquí, el objetivo al ser distinto y no depender de una descripción de cargo en particular, no da mucho campo a la comparación. Lo que si nos permite es sentar la base para un posible estudio en el cual se complementen ambos objetivos, como por ejemplo brindar sugerencias mucho más específicas para las CVs de los candidatos al comparar las características de forma y estructura (e inclusive incorporar contenido) frente a las que ya pasaron, para incrementar aún más sus probabilidades de pasar.

Por otro lado, también es de tener en cuenta que este estudio en particular tuvo varias limitaciones frente a lo técnico, dado que el acceso a modelos de lenguaje natural más avanzados que posiblemente hubieran permitido hacer una extracción más a profundidad son algo costosos. De la misma manera, existe la limitación de la cantidad de CVs a utilizar, lo que posiblemente nos esté limitando en este caso para lograr un mejor proceso de clasificación al contar con más información.

6. Plan de Implementación y Recomendaciones

El presente capítulo se centrará en entender la manera en la cual se podrán implementar los hallazgos del proyecto al flujo de trabajo de la empresa, con el fin de cumplir con los objetivos del proyecto frente a organización y calidad de datos, además de con base en estos, implementar el piloto analítico que a largo plazo nos permita mejorar la madurez analítica de la organización.

6.1. Recomendaciones de Gobernanza y Calidad de Datos

Uno de los objetivos principales del proyecto, se encuentra en el entendimiento de la calidad de los datos usados en el proceso de selección, ya que, al ser el proceso principal de la compañía, si se quiere implementar procesos analíticos en la organización, es necesario primero mejorar la calidad y uso de los mismos.

Con esto en mente, se realizó en la etapa de entendimiento de los datos una exploración general de los datos utilizados en la organización y su respectiva evaluación por medio de los criterios del DAMA (IBM, 2025), cuyas conclusiones fueron utilizadas para la creación de dos entregables.

El primer entregable es un documento titulado 'Recomendaciones de uso de datos en *la empresa* para el almacenamiento y procesamiento de la información de los candidatos.' El cual es un documento que, con base en los hallazgos de la exploración, brinda recomendaciones generales para los reclutadores de la empresa a la hora de recolectar, manipular y almacenar los datos de los candidatos durante el proceso de selección. Donde, en resumen, las principales recomendaciones se centran en garantizar la completitud de los datos al momento de recolectarlos y verificar la consistencia de estos al introducirlos en el ATS. Si se quiere profundizar en este tema, este documento se puede encontrar adjunto en el Anexo 5.

Así mismo, de la mano de dichas recomendaciones, se creó otro documento que permita a los reclutadores aplicar estas recomendaciones en su día a día de forma más práctica, con esto en mente, se creó un segundo documento llamado 'Manual de uso del ATS para la entrada de datos en el sistema de información'. El cual aplica las recomendaciones brindadas en el documento anterior a un paso a paso con capturas de pantalla del ATS

actualmente usado en la empresa, para que sea visualmente más fácil para el reclutador almacenar y manipular los datos en el ATS existente. Este documento se puede encontrar adjunto en el Anexo 6.

Con los dos anteriores documentos entregados al liderazgo de operaciones de la organización, además del especial cuidado por su parte que todos los reclutadores estén cumpliendo con su parte para mantener los datos con la calidad esperada, se espera que se tenga una base sólida en términos de calidad y uso de datos en la organización. De esta manera, se sienta una base sólida para la implementación de futuros proyectos analíticos.

6.2. Propuesta de Implementación de la Herramienta Analítica

Una vez sentadas las bases en términos de uso y calidad de datos, para continuar este proceso de sentar una base importante de analítica en la organización, la implementación de la herramienta piloto creada es fundamental. Con esta herramienta implementada, es posible facilitar la identificación del tipo de CV por parte de los reclutadores y brindar sugerencias a los candidatos con base en esto. Al realizar esto, se busca que la conversión de candidatos en la fase de revisión por parte del cliente aumente. Pero más importante aún, que deje un precedente haciendo entender a la empresa que es posible incluir analítica en sus procesos del día a día, sin interrumpir la operación; logrando a largo plazo la madurez analítica esperada.

En este caso, como se mencionó en la fase de despliegue de la metodología CRISP-DM, la herramienta derivada del proceso realizado es un script de Python de uso general para todos los reclutadores que contiene el proceso de extracción de información de la CV así como la ubicación de la misma dentro de los clústeres ya definidos. Con base en el segmento

en el que se encuentre la CV, el script brindará una sugerencia al reclutador que le ayude a este a guiar a su candidato para modificar su CV y acercarla más al segmento más exitoso que se definió durante el estudio.

Los ejemplos de uso de esta herramienta se pueden encontrar en el Anexo 7. Aquí se podrán encontrar algunos ejemplos de su uso para entender su posterior implementación.

6.2.1. Integración en el Proceso de Selección

Habiendo entendido el funcionamiento básico de la herramienta, la implementación de la misma se hace sencilla dado su fácil uso. La idea de esta es que no interrumpa el día a día de los reclutadores ni genere reprocesos, al contrario, sea una medida que facilite este mismo flujo de trabajo.

Con esto en mente, recordando el proceso de reclutamiento y como el mismo se lleva a cabo, una vez es realizada la entrevista por parte del reclutador, antes de enviar la CV al cliente para su respectiva revisión, el reclutador le solicitará su CV al candidato como siempre sucede, solo que esta vez, en vez de revisarla manualmente, también la pasará por la herramienta. La herramienta dará un resultado específico de acuerdo al segmento en el cual ubique la hoja de vida del candidato y así mismo brindará una respectiva sugerencia.

El reclutador, al ver la sugerencia, la corroborará bajo su criterio frente a la CV del candidato, la posición y el cliente. Si el reclutador lo considera adecuado, brindará la sugerencia al candidato para modificar su CV de acuerdo a los criterios que la acerquen más al segmento más exitoso. Por ejemplo, si la CV pertenece al segmento con menor probabilidad de éxito, se le brindará la siguiente sugerencia:

"Sugerencia: Ayúdale al candidato a brindarle más contenido a su CV, si es posible agregar una sección de entrenamiento, cursos o certificaciones en caso de no tenerla. Sugiere añadir links a websites externos de proyectos anteriores o inclusive un portafolio si lo tiene. Si lo vez necesario, recomiéndale al candidato también incrementar el contenido de esta, tal vez agregando más texto en sus experiencias laborales o agregando secciones adicionales"

Con la cual el reclutador debe conversar con el candidato y ver si es posible agregar la información sugerida.

Una vez recibida la CV corregida por parte del cliente, el reclutador puede hacer una segunda revisión, así como pasarla de nuevo por la herramienta a ver si ahora si es posible ubicarla en el segmento más exitoso. En caso de ser necesario, se puede repetir el proceso, pero, en caso de estar satisfecho, el reclutador puede enviar la CV del candidato al cliente, esta vez con la tranquilidad de que ha aumentado la probabilidad de que esta CV avance a la fase de entrevista con cliente.

6.2.2. Guía de uso para reclutadores.

Para facilitar el uso diario de la herramienta, es fundamental contar con una guía que les permita a los reclutadores usar la herramienta de forma sencilla.

Con esto en mente, se incluyó un archivo README dentro de la carpeta en la cual será distribuida la herramienta. Un archivo README es un estándar en herramientas tecnológicas que permiten al usuario entender lo que hace el software y cómo funciona el mismo (Github, 2025)

Siendo así, el archivo README que funciona como guía de uso para los reclutadores de la organización, cuenta con la descripción del funcionamiento de la herramienta, explicando las características que extrae de una CV y los segmentos en la que las ubica respectivamente. Así mismo, muestra los requisitos previos que debe cumplir un equipo para

ejecutar el script, en términos de software y archivos. Posteriormente, pasa a explicar cómo instalar las librerías que necesita el programa para el análisis. Por último, muestra cómo seleccionar la CV a analizar y obtener el resultado. El archivo README se puede encontrar en el Anexo 8. Por otro lado, en el Anexo 10 se puede encontrar un enlace que muestra un ejemplo del funcionamiento de la herramienta.

Así mismo, se brindarán las instrucciones de los pasos a seguir una vez se tenga la sugerencia brindada con el programa, siendo estas las explicadas en la sección anterior de este documento. Cabe aclarar que el uso de la herramienta queda bajo el criterio del reclutador teniendo en cuenta el tipo de perfil, vacante y cliente para el cual se hace el proceso.

6.2.3. Consideraciones para su escalamiento

Una vez implementada la herramienta, es importante entender cómo es posible a futuro continuar escalando la implementación de analítica dentro de la organización para alcanzar una mayor madurez analítica y potencialmente mejorar la eficiencia e ingresos de la empresa.

Con esto en mente, se tiene en cuenta que se ha sentado una base sólida de uso y calidad de datos, de la mano de una primera implementación frente a el proceso de selección. Se busca que con esta primera implementación aumente la conversión y a corto plazo, aumenten así mismo las contrataciones.

Si, el caso es positivo, puede ser posible iniciar con la creación de nuevas herramientas analíticas para los otros pasos dentro del embudo de reclutamiento, aumentando las conversiones en cada una de ellas. Un ejemplo expuesto por el gerente de operaciones de la organización, puede ser la automatización de la creación de reportes de los candidatos, con

el objetivo de ser más eficientes a la hora de que los reclutadores entrevisten a los candidatos y contar con reportes más completos para presentar frente a los clientes y que les permita evaluar de forma más sencilla a los candidatos que se les envía.

Esta, además de otras ideas sobre la mesa y otras que pueden ir surgiendo en el camino; pueden ser posibles siguiendo un paso a paso parecido al que se realizó en el presente estudio para el desarrollo de la herramienta analítica.

7. Conclusiones y Trabajo Futuro

El trabajo se centró en analizar el uso de los datos dentro del proceso de reclutamiento de la organización, así como en identificar la oportunidad de diseñar e implementar una herramienta analítica que, a largo plazo, siente las bases para incorporar nuevas soluciones y procesos basados en analítica. El objetivo final es que la compañía mejore sus procesos de selección mediante el uso estratégico de datos, fortaleciendo así su posición competitiva en la industria del reclutamiento.

Con este propósito, se inició una exploración del proceso organizacional, identificando que el proceso de selección es el componente más crítico. A partir de esto, se estableció como objetivo principal entender cómo se utilizan los datos dentro de dicho proceso. Una vez logrado este entendimiento, se entregaron al área de operaciones un manual para el uso de datos y una guía para el ingreso de información en el sistema interno de la compañía. Estas acciones buscan asegurar la mejor calidad posible de los datos, con miras a facilitar futuros análisis dentro del proceso de selección.

Asimismo, al analizar uno de los puntos clave del proceso: la revisión de las hojas de vida por parte del cliente. Se identificó que esta fase representa una oportunidad estratégica para implementar un piloto analítico. Con esto en mente, se realizó un análisis de múltiples

CVs, lo que permitió identificar los factores de forma y estructura que aumentan la probabilidad de que un candidato avance a la etapa de entrevista con el cliente. A partir de este análisis, se desarrolló una herramienta que permite a los reclutadores brindar sugerencias sobre ajustes específicos en los CVs, incrementando así la tasa de conversión en esta fase del embudo de reclutamiento.

La implementación de estas recomendaciones, junto con el uso de la herramienta desarrollada, sienta las bases para la creación progresiva de procesos analíticos dentro de la organización. A largo plazo, esto permitirá mejorar la eficiencia del proceso de reclutamiento, incrementar la conversión de candidatos y, en consecuencia, aumentar los ingresos de la empresa.

7.1. Respuesta a los Objetivos del Proyecto

El proyecto tenía el objetivo de diagnosticar el proceso de reclutamiento de la organización y con base en este, desarrollar herramientas que permitieran esta incrementar su madurez analítica. Como se estableció anteriormente, este objetivo fue alcanzado, al menos en una etapa inicial, sentando las bases necesarias para la creación e implementación de dichas herramientas.

En relación con la pregunta de investigación, se obtuvo una respuesta satisfactoria. Se concluyó que los CVs que presentan ciertas características estructurales y formales; como incluir todas las secciones completas, contar con una sección de cursos y entrenamientos, tener un enlace a una página web personal o de proyectos, y ser ligeramente más extensos que el promedio, están asociados a una mayor probabilidad de superar la fase de revisión por parte del cliente.

Así mismo, frente a los objetivos específicos planteados, se logró entender el estado actual de la gestión y calidad de los datos asociados al proceso de selección en la empresa, además de que se diseñó un conjunto de recomendaciones para estandarizar y mejorar la recolección, calidad y uso de la información. De la misma manera se dio el desarrollo e implementación de una herramienta analítica a modo de piloto con el fin de identificar distintos tipos o perfiles de CVs basados principalmente en aspectos formales y estructurales.

7.2. Recomendaciones para Trabajo Futuro

En consideración al trabajo futuro, este trabajo deja múltiples brazos por donde puede continuar una investigación relacionada al análisis de hojas de vida, así como a la implementación de modelos analíticos en empresas tipo start-up o dentro del área de reclutamiento en general.

En primer lugar, respecto a la oportunidad de profundizar en la extracción y análisis de información de las hojas de vida, sería interesante continuar iterando sobre el ejercicio propuesto en este estudio desde dos frentes. Por un lado, en el frente de extracción de información, los futuros trabajos podrían incorporar modelos más avanzados de procesamiento de lenguaje natural. Esto permitiría extraer un volumen mayor de información sobre los factores de forma y estructura, facilitando una clasificación más precisa de los CVs dentro de los clústeres previamente identificados, o incluso permitiendo descubrir nuevos clústeres o características relevantes.

Por otro lado, sería altamente valioso repetir el ejercicio con una muestra significativamente mayor de hojas de vida. Esto permitiría evaluar si, además de

clasificaciones, es posible comenzar a realizar predicciones sobre la probabilidad de éxito de los candidatos, con base en sus características de forma y estructura.

Adicionalmente, se sugiere una nueva línea de investigación enfocada en combinar lo abordado en este trabajo con hallazgos de estudios previos, con el fin de desarrollar herramientas de screening o ranking de CVs que no se basen únicamente en el contenido, sino que también incorporen estos factores de forma y estructura como variables relevantes para la toma de decisiones.

En cuanto a la aplicación práctica en entornos empresariales, este trabajo sienta una base importante para entender cómo se pueden implementar nuevas soluciones analíticas en otras etapas del proceso de reclutamiento. En este sentido, la metodología CRISP-DM podría aplicarse en diferentes procesos dentro de la misma organización, o también a modo de consultoría para otros clientes de la empresa.

Todos los archivos utilizados para la realización del proyecto se pueden encontrar en el repositorio de Github expuesto en el Anexo 9.

8. Referencias

Toxigon. (23 de Marzo de 2025). *LinkedIn Statistics 2024: A Deep Dive*. Obtenido de Toxigon: https://toxigon.com/linkedin-statistics-2024

Amplitude Marketing. (15 de Febrero de 2024). *51 LinkedIn Statistics You Need to Know in 2024*. Obtenido de Amplitude Marketing: https://amplitudemktg.com/social-media/51-linkedin-statistics-you-need-to-know-in-2024/

appcast. (2025). *Recruitment marketing benchmark report.* Obtenido de appcast: https://1859609.fs1.hubspotusercontent-na1.net/hubfs/1859609/FINAL%20CONTENT%20PDFS/Whitepapers/%5BWhitepaper%5D%20Appcast%20Recruitment%20Benchmark%20Report%202025.pdf

American Staffing Association. (s.f.). *Staffing Industry Statistics*. Obtenido de American Staffing Association: https://americanstaffing.net/research/fact-sheets-analysis-staffing-industry-trends/staffing-industry-statistics/

LinkedIn. (2025). *The Future of Recruiting 2025: How AI redefines recruiting excellence*. Obtenido de LinkedIn: https://business.linkedin.com/talent-solutions/resources/future-of-recruiting

Microsoft and LinkedIn. (8 de Mayo de 2024). *2024 Work Trend Index Annual Report*. Obtenido de Microsoft: https://www.microsoft.com/en-us/worklab/work-trend-index/ai-at-work-is-here-now-comes-the-hard-part

Mazor, A., Goretsky, C., Moss, K., Cleary, B., & Fineman, D. (2018). *Talent acquisition analytics: Driving smarter sourcing and hiring decisions with data.* Obtenido de Deloitte: https://www2.deloitte.com/content/dam/Deloitte/us/Documents/human-capital/us-talent-acquisition-analytics.pdf

AWS. (2024). *¿Qué es el ciclo de vida del desarrollo de software (SDLC)?* Obtenido de AWS: https://aws.amazon.com/what-is/sdlc/

IBM. (17 de Agosto de 2021). *Conceptos básicos de ayuda de CRISP-DM.* Obtenido de IBM SPSS Modeler: https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview

Harris, J. G., Craig, E., & Light, D. A. (2011). Talent and analytics: new approaches, higher ROI. *Journal of Business Strategy, 33*(6), 4-13. doi:10.1108/02756661111180087

Conte, F., & Siano, A. (2023). Data-driven human resource and data-driven talent management in internal and recruitment communication strategies: an empirical survey on Italian firms and insights for European context. *Corporate Communications: An International Journal, 28*(4), 618-637. doi:10.1108/CCIJ-02-2022-0012

Gavin Walford-Wright, W. S.-J. (2018). Talent Rising; people analytics and technology driving talent acquisition strategy. *Strategic HR Review, 17*(5), 226-233. doi:10.1108/SHR-08-2018-0071

Ali, I., Mughal, N., Khan, Z. H., Ahmed, J., & Mujtaba, G. (2022). Resume Classification System using Natural Language Processing and Machine Learning Techniques. *Mehran University Research Journal of Engineering and Technology, 41*(1), 65-79. doi:10.22581/muet1982.2201.07

Sowjanya, Y., Keerthana, M., & Suneeksha, P. (Marzo de 2023). Smart Resume Analyser. *International Journal of Research in Engineering and Science (IJRES), 11*(3), 409-418.

Saatci, M., Kaya, R., & Ünlü, R. (2024). Resume Screening With Natural Language Processing (NLP). *Alphanumeric Journal, 12*(2), 121-140. doi:10.17093/alphanumeric.1536577

Zu, S., & Wang, X. (Octubre de 2019). RESUME INFORMATION EXTRACTION WITH A NOVEL TEXT BLOCK SEGMENTATION ALGORITHM. *International Journal on Natural Language Computing (IJNLC), 8*(5), 29-48. doi:10.5121/ijnlc.2019.8503

Satheesh, K., Jahnavi, A., Iswarya, L., Ayesha, K., Bhanusekhar, G., & Hanisha, K. (Mayo de 2020). Resume Ranking based on Job Description using SpaCy NER model. *International Research Journal of Engineering and Technology (IRJET), 7*(5), 74-77.

Gonzales, S. (2025). *People Analytics Maturity Linked to Better Financial Performance, Research Shows*. Obtenido de VISIER: https://www.visier.com/blog/people-analytics-maturity-linked-to-better-financial-performance-research-shows/

Transportation Research Board. (s.f.). *Maturity Model and Evaluation Matrix*. Obtenido de TRB's COOPERATIVE RESEARCH PROGRAMS: https://crp.trb.org/acrpwebresource18/maturity-model/

crunchr. (2025). *Understanding Crunchr's HR Analytics Maturity Model*. Obtenido de crunchr: https://www.crunchr.com/resources/blog/crunchrs-hr-analytics-maturity-model/

Moore, S. (18 de Enero de 2018). *How to Stop Data Quality Undermining Your Business*. Obtenido de Gartner: https://www.gartner.com/smarterwithgartner/how-to-stop-data-quality-undermining-your-business

IBM. (30 de Abril de 2025). *Data quality dimensions*. Obtenido de IBM Cloud Pak for Data: https://www.ibm.com/docs/en/cloud-paks/cp-data/5.1.x?topic=quality-data-dimensions

Luka, S., & Karkunova, D. (2025). *HR data quality: Its definition, importance, and 5 actionable steps to improve it*. Obtenido de NAKISA: https://nakisa.com/blog/the-power-of-accurate-hr-data-introducing-nakisa-hanelly-hr-data-quality/

Vorecol. (28 de Agosto de 2024). *¿Cuáles son los principales desafíos en la implementación de analítica predictiva en recursos humanos?* Obtenido de Vorecol: https://vorecol.com/es/articulos/articulo-cuales-son-los-principales-desafios-en-la-implementacion-de-analitica-predictiva-en-recursos-humanos-111253

Vasquez, G. (2025). *Five steps to efficiently manage your HR data quality*. Recuperado el 12 de Mayo de 2025, de greenhouse: https://www.greenhouse.com/blog/five-steps-to-efficiently-manage-your-hr-data-quality

Gardiner, T. (2024). *NLP-PyPDF vs PyMUPDF Speed Test*. Obtenido de kaggle: https://www.kaggle.com/code/toddgardiner/nlp-pypdf-vs-pymupdf-speed-test

Explosion. (2025). *Trained Models & Pipelines*. Obtenido de spacy: https://spacy.io/models

pandas . (2024). *pandas.get_dummies*. Obtenido de pandas: https://pandas.pydata.org/docs/reference/api/pandas.get_dummies.html

R, A. (29 de Octubre de 2024). *Outlier Detection and Treatment: Z-score, IQR, and Robust Methods*. Obtenido de medium: https://medium.com/@aakash013/outlier-detection-treatment-z-score-iqr-and-robust-methods-398c99450ff3

VenkataAnusha, P., Anuradha, C., Murty, P. S., & Kiran, C. S. (10 de Noviembre de 2019). Detecting Outliers in High Dimensional Data Sets Using Z-Score Methodology. *International Journal of Innovative Technology and Exploring Engineering (IJITEE), 9*(1), 48-53. doi:10.35940/ijitee.A3910.119119

C., J. (9 de Marzo de 2022). *Data Cleaning - Variance*. Obtenido de mage: https://pro.mage.ai/blog/data-cleaning-variance

The Pennsylvania State University. (2018). *Detecting Multicollinearity Using Variance Inflation Factors*. Obtenido de STAT 462 Applied Regression Analysis: https://online.stat.psu.edu/stat462/node/180/

JMP Statistical Discovery. (2025). *La prueba t*. Obtenido de Portal de formación estadística: https://www.jmp.com/es/statistics-knowledge-portal/t-test

JMP Statistical Discovery. (2025). *La prueba de ji cuadrado*. Obtenido de Portal de formación estadística: https://www.jmp.com/es/statistics-knowledge-portal/chi-square-test

Arias, M. M. (Diciembre de 2017). ¿Qué significa realmente el valor de p? *Pediatría Atención Primaria, 19*(76), 377-381.

IBM. (s.f.). *What is random forest?* Obtenido de IBM: https://www.ibm.com/think/topics/random-forest

Martínez Pérez, J., & Pérez Martin, P. (Febrero de 2023). La curva ROC. *Medicina de Familia. SEMERGEN, 49*(1), 1-3. doi:10.1016/j.semerg.2022.101821

Universidad de Oviedo. (s.f.). *El algoritmo k-means aplicado a clasificación y procesamiento de imágenes*. Obtenido de unioviedo: https://www.unioviedo.es/compnum/laboratorios_py/kmeans/kmeans.html

Rodríguez, D. (9 de Junio de 2023). *Método del codo (Elbow method) para seleccionar el número óptimo de clústeres en K-means*. Obtenido de Analytics Lane: https://www.analyticslane.com/2023/06/09/metodo-del-codo-elbow-method-para-seleccionar-el-numero-optimo-de-clusteres-en-k-means/

SPSS Analysis. (2025). *Silhouette Cluster Analysis*. Obtenido de SPSS Analysis: https://spssanalysis.com/silhouette-cluster-analysis-in-spss/

IBM. (8 de Diciembre de 2023). *¿Qué es el análisis de componentes principales (PCA)?* Obtenido de IBM: https://www.ibm.com/es-es/think/topics/principal-component-analysis

OpenAI. (2025). *Precios de API*. Obtenido de OpenAI: https://openai.com/es-419/api/pricing/

Github. (2025). *About READMEs*. Obtenido de GitHub Docs: https://docs.github.com/en/repositories/managing-your-repositorys-settings-and-features/customizing-your-repository/about-readmes

9. Anexos Técnicos

9.1. Anexo 1: Análisis de calidad de datos del proceso de selección

Análisis de calidad

Base de datos de candidatos Q3 2024

Proyecto empresarial 2 - Manuel Vecino

Universidad del Rosario

Maestría en Busines analytics

1ra dimensión: Completitud:

Completitud se define como el grado en que los atributos requeridos están presentes.

Se mide como el porcentaje de registros cuyos campos requeridos están informados / registros totales.

	Datos Faltantes	% Datos Faltantes
Candidate Name	0.000000	0
Candidate Updated Date	0.000000	0
Candidate Created Date	0.000000	0
Source	0.000000	0

	% Datos Faltantes	Datos Faltantes
Candidate Last Name	0.115607	4
Candidate First Name	0.115607	4
Candidate Creator	5.086705	176
Current Company	5.578035	193
Current Position	6.011561	208
Candidate Email Address	7.658960	265
Candidate Owner	11.560694	400
Resume Added Date	15.867052	549
Candidate Phone Number	23.352601	808
Candidate Location	25.751445	891
Expected Salary	27.803468	962

	Datos Faltantes	
% Datos Faltantes		
Current Salary	29.710983	1028
Notice Period	38.843931	1344
Years of Experience	40.462428	1400
Region	40.924855	1416
Current Benefits	67.369942	2331
Candidate industry	71.242775	2465
salary_min	91.300578	3159
salary_max	92.080925	3186
Hired Date	95.520231	3305
Employment Status	95.520231	3305
Employee Job	95.635838	3309

	% Datos Faltantes	Datos Faltantes
('Placement Client',)	95.635838	3309

Conclusiones de Completitud

- En primer lugar, entendiendo la naturaleza de los datos, y como son recolectados, iniciamos el análisis de completitud. En primer lugar, podemos ver que el grupo de datos "Source", "Candidate Updated Date", "Candidate Created Date" son datos automáticamente generados por el sistema, por lo que en la creación de todos los registros van a ser creados consecuentemente, por lo que no existe en si mismo un análisis de completitud, ya que siempre van a encontrarse. Dentro de este grupo de variables automáticas, una que genere curiosidad que esté vacía es el "Candidate Creator" ya que por regla del sistema todos los candidatos deberían ser creados por algún reclutador y así mismo, tener un "Candidate Owner", variable que por alguna razón también cuenta con varios registros vacíos y valdría la pena entender los casos particulares. Así mismo, vemos que hay candidatos sin "Resume added date", aunque es entendible la razón por la cual esto puede estar pasando, es incongruente con el proceso con el que trabajamos que los candidatos registrados en el CRM no cuenten con un resume.
- Por otro lado, el dato "Candidate Name" existe al ser el identificador principal del candidato en el sistema por lo que es un dato obligatorio y no es posible crear un nuevo candidato sin esta variable. Así mismo, sus variables complementarias "Candidate First Name" y "Candidate Last Name" aunque deberían ser obligatorias a la par, vemos que aquí se empiezan a generar inconsistencias, ya que existen 4 registros donde no estan, cosa que no debería ser ya que todos los candidatos deberían contar con un nombre y un apellido.

- Avanzando con el siguiente grupo de datos, podemos ver que faltan datos en las variables "Current Company" y "Current Position" las cuales aunque son generadas automáticamente por el sistema al momento de realizar el proceso de scrapping del LinkedIn o la CV del candidato, este proceso puede fallar.
- De la misma manera, entendiendo los métodos de contacto, contamos con el "Candidate Email Addres" y el "Candidate Phone Number" donde contamos con registros faltantes en el caso del primero en 265 ocasiones, por otro lado, en el teléfono es bastante más, alrededor de 808. Al ser el Email, además del medio de comunicación más relevante durante un proceso de selección, la manera más sencilla de identificar a un candidato y evitar que se duplique en la base de datos, es crítico tenerlo, por lo cual es crítico que este se incluya en absolutamente todos los candidatos. Por otro lado, el teléfono es un poco más entendible que existan datos faltantes, ya que múltiples candidatos no lo incluyen en su CV o datos de contacto ya que prefieren no ser contactados por este medio.
- Antes de avanzar con el siguiente grupo, dos de las variables donde perdemos bastante información son las variables "región" y "Candidate Industry". Variables las cuales Aunque no brindan información relevante en si misma para el proceso, si brindan información importante para la operación y el entendimiento del mercado, los cuales pueden brindar insights significativos a futuro sobre el proceso.
- El siguiente grupo, y uno de los grupos de variables más importantes para el análisis del proceso y fit en la posición se trata de las variables de salario, las cuales como podemos ver, cuentan con bastantes datos faltantes. Empezando por el salario esperado, no tiene sentido que haya tantos datos faltantes, ya que es uno de los datos más importantes a preguntar en una entrevista, inclusive si el candidato no avanza en el proceso. De la misma manera, podemos ver que el current salary cuenta con una cantidad de datos faltantes aun mayor, lo cual es entendible ya que no todos los candidatos se sienten cómodos compartiendo su compensación actual, aun así, la

gran cantidad de datos faltantes denota que junto con la variable "current salary" sea un descuido del reclutador no agregarlo en el CRM.

- Por último, datos relevantes para el proceso son "Years of Experience" y "Notice Period" donde vemos que los datos faltantes se encuentran cercanos al 40%. Esto probablemente por la falla del reclutador al subir los datos, ya que los años de experiencia son un dato con el cual se cuenta incluso antes de tener una entrevista. De la misma manera, el Notice period se debe discutir en una entrevista para entender la disponibilidad de un candidato, aunque, es entendible que no se tenga si el candidato no avanzó.
- Recomendaciones:
- Las variables de Candidate Creator y Candidate Owner en dado caso de no ser llenadas por el sistema automáticamente y quedar vacías, deben de alguna forma poder llenarse manualmente, particularmente el candidate owner, ya que es la demostración de que el candidato está siendo procesado (o no) y la persona quién lo procesa. Es decir **No puede haber candidato sin candidate owner**
- Los candidatos registrados deben contar con un Resume en el ATS, así sea el generado automáticamente en LinkedIn.
- En caso de que el Scrapping de LinkedIn no vincule ningún candidato con un Empleo y Puesto actual (o inmediatamente anterior), debe ser escrito manualmente por el reclutador. Así mismo, en un caso excepcional que el candidato no cuente con trabajo actual o anterior, se debe escribir un "N/A", mas no dejar la columna vacía.
- El email debe ser un campo obligatorio para crear un candidato y no es posible avanzar con el resto de pasos de creación de un candidato sin contar con el correo.
- Tanto el candidate location como la región deben ser un campo obligatorio. Inclusive, si es posible dentro del sistema, generar la posibilidad que desde la selección de la ubicación se seleccione automáticamente la región.

- La industria del candidato debe ser un campo obligatorio, aunque posiblemente y para facilitar el workflow de los recruiters, podría ser recomendable aumentar la lista de industrias en el sistema, tal vez copiando la misma lista de categorías que maneja LinkedIn para que sea incluso familiar para el equipo de recruiting.
- El salario actual y anterior deben ser obligatorios a agregar, así como los beneficios. En dado caso que el candidato no haya querido discutir al respecto del salario actual, es recomendable poner un "0" con tal de no dejar el campo vacío, en caso de realizar un análisis posterior, para que estos ceros no dañen las mediciones, serán imputados anteriormente. Así mismo, si el candidato no cuenta con beneficios actuales o no se discutieron, se pondrá un "N/A" o un "Not discussed" dependiendo del caso."
- En dado caso de no contar con el notice period de un candidato, con tal de no dejar el campo vacío, se puede poner el estándar de "2 semanas" como dato por defecto.

2da dimensión: Validez:

Validez se define como el grado de conformidad con formato, tipado y rango.

Se mide como el porcentaje de registros cuyos campos requeridos cumplen con los requerimientos de validez / registros totales

validación de nombres/apellidos en Candidate Name:

Candidate Name_Valid

True 3309

False 151

Name: count, dtype: int64

Validación de nombres/apellidos en Candidate First Name:

Candidate First Name_Valid

True 3451

False 9

Name: count, dtype: int64

Validación de nombres/apellidos en Candidate Last Name:

Candidate Last Name_Valid

True 3444

False 16

Name: count, dtype: int64

Conclusiones de validez de nombres

- Podemos ver que con respecto a los nombres la validez no es un problema tan grande, aunque encontramos que principalmente en el nombre completo, encontramos más candidatos que no se adhieren a la regla general de no contener caracteres especiales.

Recomendaciones:

- La principal recomendación es, al ser una empresa que tiene candidatos en todo el mundo pero cuyo lenguaje principal es el inglés, nos debemos a tener a tener nombres sin caracteres especiales para encontrar fácilmente a los candidatos además de poder mantener su registro de manera uniforme.

Validación de correos electrónicos:

Candidate Email Valid

True 3458

False 2

Name: count, dtype: int64

Ejemplos de correos electrónicos inválidos:

2550 '+5491163609923-matiassandoval980@gmail.com

2793 nursedaturkcan123@gmail.com

Name: Candidate Email Address, dtype: object

Conclusiones de validez de los correos:

- Con respecto a la validez, los correos son bastante ordenados gracias que el CRM tiene su propio sistema de auto validación. A parte de un par de errores como uno que incluye un número de teléfono, no es un campo del que nos debemos preocupar mucho.

Recomendaciones

- Para evitar el error en este campo importante para la verificación de duplicados, es importante que el reclutador revise dos veces que este sea correcto.

Validación de números de teléfono:

Candidate Phone Valid

True 2433

False 1027

Name: count, dtype: int64

Ejemplos de números de teléfono inválidos:

1 '+19784516966

6 '+5519-99111.1562

8 '+52 7715673710

15 '+529848775633

16 '+541122669666

...

3448 '+57 300 279 2226

3451 '+573016477114

3456 (+90)536-792-8328

3457 (+36) 70 319 0639

3459 '+573218823946

Name: Candidate Phone Number, Length: 1027, dtype: object

Revisión Validez Teléfonos

Aqui uno de los campos a entender que debemos contar con formatos válidos para evitar confusiones posteriores. Podemos ver que no se tiene in formato uniforme para el número de teléfono, el cual al mismo tiempo debe ser el más simple posible para que los reclutadores no pierdan mucho tiempo revisandolo, pero que cuente con la información completa, en este caso código pais, código ciudad (si aplica) y el número como tal.

Recomendaciones:

- De esta manera, es importante mantener un formato uniforme pero simple, el propuesto es el siguiente +## ########## siendo los dos primeros (o tres primeros en ciertos casos) el código país y el resto, el número de teléfono como tal, sin nigún simbolo adicional de por medio.

3ra dimension: Unicidad

Unicidad se define como el porcentaje de registros únicos que existen. Se mide como el porcentaje de registros únicos / registros totales.

Utilizaremos los correos y los teléfonos que son valores que no deberían ser iguales para los candidatos.

Cantidad de correos duplicados: 43

Cantidad de teléfonos duplicados: 35

- Conclusiones de unicidad

El análisis aquí se realizó principalmente sobre las variables de correo y teléfono, los cuales son únicos entre los candidatos. En este caso, vemos que el registro cuenta con 43 candidatos duplicados y 35 teléfonos duplicados.

Recomendación:

- El reclutador debe verificar manualmente si el correo ya se encuentra creado.
- Así mismo, si el sistema detecta que el correo ya está en otro registro; el registro no podrá ser creado.

4ta dimensión: Consistencia

Consistencia mide el nivel en que se dispone de la misma información independiente de en qué fuente se consulte. Se mide como el porcentaje de registros consistentes / registros totales.

Una de las fuentes de consistencia más importantes es validar que exista consistencia entre el nombre y apellido del candidato y el nombre completo del mismo.

Validación de consistencia entre Candidate Name y la combinación de First y Last Name:

Name Consistent

False 2096

True 1364

Name: count, dtype: int64

Filas con inconsistencias en los nombres:

	Candidate Name	Candidate First Name	Candidate Last Name	\
0	★ Jonathan Orozco Ruiz ★	Jonathan	Ruiz	
2	aalok kumar bhunjiya	Aalok	Bhunjiya	
5	Abigail Rijo Morales	Abigail	Morales	
7	Abraham Hernández Venegas	Abraham	Venegas	
8	Abraham López	Abraham	Pez	
...	...	...	...	
3455	Zi Jie (Jay) Ang	Zi	Ang	
3456	Zia Soroush.h	Zia	h	
3457	Zsolt Barkó	Zsolt	Bark	
3458	🌐Alejandro🌐 D.	Alejandro	D	
3459	💻 Mauricio Guzmán-Salazar	Mauricio	N-salazar	

	Full Name Combined
0	Jonathan Ruiz
2	Aalok Bhunjiya

5 Abigail Morales

7 Abraham Venegas

8 Abraham Pez

... ...

3455 Zi Ang

3456 Zia h

3457 Zsolt Bark

3458 Alejandro D

3459 Mauricio N-salazar

[2096 rows x 4 columns]

Conclusiones de consistencia

El análisis particular de consistencia va hacia los nombres dado que es donde más cambios pueden haber entre una misma fuente de datos en esta base. De esta manera, podemos ver que en efecto en la gran mayoría de casos, el nombre completo NO es la concatenación del Nombre y Apellido del candidato.

Recomendaciones:

- Es necesario que el reclutador verifique que el campo de "Candidate First Name" tenga el nombre(s) del candidato y el campo "Candidate Last Name" apellido(s) con el objetivo de que

la información sea consistente y sea más sencillo encontrar a un candidato en particular en la base.

5ta dimensión: Precisión

Precisión describe el grado en que los datos representan la realidad. Se mide como el porcentaje de registros precisos / registros totales.

Valores fuera de rango en 'Years of Experience':

Empty DataFrame

Columns: [Candidate Name, Candidate First Name, Candidate Last Name, Current Company, Current Position, Candidate Location, Candidate Email Address, Candidate Phone Number, Region, Employment Status, Source, Hired Date, Employee Job, ('Placement Client',), Candidate Created Date, Candidate Updated Date, Resume Added Date, Candidate Creator, Candidate Owner, Years of Experience, Current Salary, Current Benefits, Notice Period, Expected Salary, Candidate industry, salary_min, salary_max, Candidate Name_Valid, Candidate First Name_Valid, Candidate Last Name_Valid, Candidate Email Valid, Candidate Phone Valid, Full Name Combined, Name Consistent]

Index: []

[0 rows x 34 columns]

Conclusiones precisión en años de experiencia

No existe ningún dato irregular o impreciso para esta variable.

6ta dimension: Puntualidad

Puntualidad se define como el grado de disponibilidad de la información cuando se necesita. Se mide como el porcentaje de registros cuyos campos estuvieron disponibles antes del momento requerido / registros totales

Al no contar particularmente con una variable que nos permita validar la puntualidad requerida de la información en este caso, por el momento este será omitido en este análisis.

9.2. Anexo 2: Diccionarios para guía del procesamiento de lenguaje natural.

TECH_TERMS = {

 # Lenguajes de programación

"Java", "Spring Boot", "Hibernate", "JPA", "Java EE",

"JavaScript", "TypeScript", "React", "Angular", "Vue",

"Node.js", "Express.js", "NestJS", "HTML", "CSS", "SASS",

"Python", "Django", "Flask", "FastAPI", "C#", ".NET", "ASP.NET",

"Ruby", "Rails", "PHP", "Laravel", "Go", "Rust", "Kotlin", "Swift",

 # Bases de datos

"SQL", "PostgreSQL", "MySQL", "MariaDB", "MongoDB", "NoSQL",

"Redis", "Elasticsearch", "Firebase", "DynamoDB", "GraphQL",

 # DevOps y Cloud

"Docker", "Kubernetes", "AWS", "Azure", "GCP", "Terraform",

"Jenkins", "CI/CD", "GitHub Actions", "Ansible", "Linux",

Arquitectura y Metodologías

"Microservices", "API", "REST", "SOAP", "GraphQL",

"Agile", "Scrum", "Kanban", "DDD", "TDD", "SOLID", "Clean Architecture",

 # Seguridad y Testing

"OWASP", "JWT", "OAuth", "SAML", "Penetration Testing",

"Selenium", "JUnit", "Mockito", "Jest", "Cypress",

 # Otros conceptos importantes

"Multithreading", "Concurrency", "Asynchronous Programming",

"WebSockets", "Event-Driven Architecture", "Kafka", "RabbitMQ",

```
    "gRPC", "WebAssembly"
}

SECCIONES_CV_DICT = {
    "education": ["education", "academic background", "studies", "university studies",
"formación académica"],
    "work_experience": ["experience", "work experience", "employment history", "career
history", "experiencia laboral", "professional experience"],
    "skills": ["skills", "technical skills", "competencies", "habilidades", "conocimientos"],
    "certifications": ["certifications", "licenses", "accreditations", "certificaciones"],
    "achievements": ["achievements", "accomplishments", "milestones", "logros"],
    "professional_profile": ["profile", "summary", "about me", "professional summary",
"objective", "perfil profesional", "resumen"],
    "languages": ["languages", "linguistic skills", "spoken languages", "idiomas"],
    "projects": ["projects", "case studies", "portfolio", "proyectos"],
    "publications": ["publications", "research papers", "articles", "books", "publicaciones"],
    "training_courses": ["training", "courses", "workshops", "seminars", "courses and
seminars", "Other Studies", "cursos", "formación complementaria"],
    "volunteer_work": ["volunteer work", "volunteering", "community service", "social
impact", "non-profit", "voluntariado"],
}
PATRONES_FECHAS_FORMATOS = [
    # Formatos con Mes (texto) y Año: Jan 2020, January 2020, Enero 2020, etc.
```

```
    # Separar nombre completo e abreviado puede dar más detalle
    {'name': 'Mon YYYY (EN)', 'pattern':
r"\b(?:Jan|Feb|Mar|Apr|May|Jun|Jul|Aug|Sep|Oct|Nov|Dec)\.?\s+\d{4}\b"},
    {'name': 'Month YYYY (EN)', 'pattern':
r"\b(?:January|February|March|April|May|June|July|August|September|October|November|
December)\s+\d{4}\b"},
    {'name': 'Mes YYYY (ES Abr)', 'pattern':
r"\b(?:Ene|Feb|Mar|Abr|May|Jun|Jul|Ago|Sep|Oct|Nov|Dic)\.?\s+\d{4}\b"},
    {'name': 'Mes YYYY (ES Comp)', 'pattern':
r"\b(?:Enero|Febrero|Marzo|Abril|Mayo|Junio|Julio|Agosto|Septiembre|Octubre|Noviembre|
Diciembre)\s+\d{4}\b"},

    # Formatos numéricos: MM/YYYY, MM-YYYY
    {'name': 'MM/YYYY', 'pattern': r"\b(0?[1-9]|1[0-2])/\d{4}\b"}, # Asegurar mes válido
    {'name': 'MM-YYYY', 'pattern': r"\b(0?[1-9]|1[0-2])-\d{4}\b"}, # Asegurar mes válido

    # Formatos con día: DD/MM/YYYY, DD-MM-YYYY (o MM/DD/YYYY - ambiguo
sin contexto)
    # Ser más específico si es posible, o usar un nombre genérico
    {'name': 'DD/MM/YYYY', 'pattern': r"\b(0?[1-9]|[12]\d|3[01])/(0?[1-9]|1[0-
2])/\d{4}\b"},
    {'name': 'DD-MM-YYYY', 'pattern': r"\b(0?[1-9]|[12]\d|3[01])-(0?[1-9]|1[0-2])-
\d{4}\b"},
```

```
    # Podríamos añadir variaciones con año de 2 dígitos si son comunes: \d{1,2}[/-]\d{1,2}[/-]\d{2}\b

    # Rangos de años: 2018-2020, 2018 – 2020
    {'name': 'Rango YYYY-YYYY', 'pattern': r"\b(19[89]\d|20[0-3]\d)\s*[-–]\s*(19[89]\d|20[0-3]\d)\b"},

    # Año hasta presente: 2019 - Present, 2020 - Actualidad
    {'name': 'Rango YYYY-Presente', 'pattern': r"\b(19[89]\d|20[0-3]\d)\s*[-–]\s*(?:Present|Actual|Actualidad|Today|Now)\b"},

    # Años sueltos (menos prioritario, poner al final)
    # Usar (?<!\d) y (?!\d) para asegurar que no es parte de un número más grande
    {'name': 'YYYY', 'pattern': r"(?<!\d|\.|/|-)(19[89]\d|20[0-3]\d)(?!:|\d|\.|/|-)\b"} # Año aislado 1980-2039
]
```

9.3. Anexo 3: Dataframe de 80 variables

<class 'pandas.core.frame.DataFrame'>

Index: 614 entries, 0 to 634

Data columns (total 80 columns):

#	Column	Non-Null Count	Dtype
0	Passed	614 non-null	int64
1	Numero de Paginas	614 non-null	int64
2	Cantidad de Palabras	614 non-null	int64
3	Densidad Informacion (%)	614 non-null	float64
4	Fechas Detectadas (Count)	614 non-null	int64
5	Tamaño cuerpo probable	614 non-null	int64
6	Variedad de fuentes	614 non-null	int64
7	Variedad de tamaños	614 non-null	int64
8	Uso de negritas (estimado %)	614 non-null	float64
9	Uso de cursivas (estimado %)	614 non-null	float64
10	Porcentaje Lenguaje Técnico	614 non-null	float64
11	Porcentaje Lenguaje Genérico	614 non-null	float64
12	LinkedIn	614 non-null	int64
13	GitHub	614 non-null	int64
14	Website/Otro	614 non-null	int64
15	Cantidad de imágenes	614 non-null	int64
16	Tiene Elementos Graficos	614 non-null	int64

17	Lineas_education	614 non-null	int64
18	Lineas_work_experience	614 non-null	int64
19	Lineas_skills	614 non-null	int64
20	Lineas_certifications	614 non-null	int64
21	Lineas_achievements	614 non-null	int64
22	Lineas_professional_profile	614 non-null	int64
23	Lineas_languages	614 non-null	int64
24	Lineas_projects	614 non-null	int64
25	Lineas_publications	614 non-null	int64
26	Lineas_training_courses	614 non-null	int64
27	Lineas_volunteer_work	614 non-null	int64
28	Seccion_education	614 non-null	int64
29	Seccion_work_experience	614 non-null	int64
30	Seccion_skills	614 non-null	int64
31	Seccion_certifications	614 non-null	int64
32	Seccion_achievements	614 non-null	int64
33	Seccion_professional_profile	614 non-null	int64
34	Seccion_languages	614 non-null	int64
35	Seccion_projects	614 non-null	int64
36	Seccion_publications	614 non-null	int64
37	Seccion_training_courses	614 non-null	int64
38	Seccion_volunteer_work	614 non-null	int64
39	texto_extraido_len	614 non-null	int64

40	Tamaño de fuente más usado	614 non-null	int64
41	Promedio tamaño fuente	614 non-null	float64
42	Formato Texto (Lineas)_Mixto	614 non-null	int64
43	Formato Texto (Lineas)_Párrafos	614 non-null	int64
44	Formato Texto (Lineas)_Viñetas	614 non-null	int64
45	Orden Temporal_Orden Temporal Detectado	614 non-null	int64
46	Orden Temporal_Pocas Fechas	614 non-null	int64
47	Formato Fecha Más Común_DD/MM/YYYY	614 non-null	int64
48	Formato Fecha Más Común_MM-YYYY	614 non-null	int64
49	Formato Fecha Más Común_MM/YYYY	614 non-null	int64
50	Formato Fecha Más Común_Mes YYYY (ES Comp)	614 non-null	int64
51	Formato Fecha Más Común_Mon YYYY (EN)	614 non-null	int64
52	Formato Fecha Más Común_Month YYYY (EN)	614 non-null	int64
53	Formato Fecha Más Común_Rango YYYY-YYYY	614 non-null	int64
54	Formato Fecha Más Común_SpaCy DATE (Sin clasificar)	614 non-null	int64
55	Formato Fecha Más Común_YYYY	614 non-null	int64
56	Formato Fecha Más Común_YYYY (SpaCy)	614 non-null	int64
57	Fuente principal_ArialMT	614 non-null	int64
58	Fuente principal_Calibri	614 non-null	int64
59	Fuente principal_Lato-Regular	614 non-null	int64
60	Fuente principal_OpenSans-Regular	614 non-null	int64
61	Fuente principal_Otra	614 non-null	int64
62	Fuente principal_Roboto-Regular	614 non-null	int64

```
 63  Fuente principal_Tahoma                                  614 non-null    int64
 64  Legibilidad general_Buena                                614 non-null    int64
 65  Legibilidad general_Potencialmente Deficiente            614 non-null    int64
 66  Consistencia tamaños fuente_Consistente                  614 non-null    int64
 67  Consistencia tamaños fuente_Inconsistente                614 non-null    int64
 68  Consistencia márgenes (aprox)_Consistente                614 non-null    int64
 69  Consistencia márgenes (aprox)_Inconsistente              614 non-null    int64
 70  Uso de colores (texto)_No                                614 non-null    int64
 71  Uso de colores (texto)_Sí                                614 non-null    int64
 72  Uso de colores (dibujos)_No                              614 non-null    int64
 73  Uso de colores (dibujos)_Sí                              614 non-null    int64
 74  Deteccion Foto Perfil_No se detectaron imágenes candidatas  614 non-null    int64
 75  Deteccion Foto Perfil_Posible Foto Detectada             614 non-null    int64
 76  secciones_completas                                      614 non-null    int64
 77  Ratio_Lineas_Experiencia_len                             614 non-null    float64
 78  Ratio_Lineas_Experiencia_words                           614 non-null    float64
 79  Densidad_Texto_Por_Seccion                               614 non-null    float64
dtypes: float64(9), int64(71)
memory usage: 388.5 KB
```

9.4. Anexo 4: Dataframe de 50 variables

<class 'pandas.core.frame.DataFrame'>

Index: 614 entries, 0 to 634

Data columns (total 50 columns):

#	Column	Non-Null Count	Dtype
0	Numero de Paginas	614 non-null	int64
1	Densidad Informacion (%)	614 non-null	float64
2	Fechas Detectadas (Count)	614 non-null	int64
3	Tamaño cuerpo probable	614 non-null	int64
4	Variedad de fuentes	614 non-null	int64
5	Variedad de tamaños	614 non-null	int64
6	Uso de negritas (estimado %)	614 non-null	float64
7	Uso de cursivas (estimado %)	614 non-null	float64
8	Porcentaje Lenguaje Técnico	614 non-null	float64
9	LinkedIn	614 non-null	int64
10	GitHub	614 non-null	int64
11	Website/Otro	614 non-null	int64
12	Cantidad de imágenes	614 non-null	int64
13	Tiene Elementos Graficos	614 non-null	int64
14	Lineas_education	614 non-null	int64
15	Lineas_work_experience	614 non-null	int64
16	Lineas_skills	614 non-null	int64

```
17  Lineas_certifications                    614 non-null    int64
18  Lineas_achievements                      614 non-null    int64
19  Lineas_professional_profile              614 non-null    int64
20  Lineas_languages                         614 non-null    int64
21  Lineas_projects                          614 non-null    int64
22  Lineas_publications                      614 non-null    int64
23  Lineas_training_courses                  614 non-null    int64
24  Lineas_volunteer_work                    614 non-null    int64
25  Seccion_languages                        614 non-null    int64
26  Seccion_publications                     614 non-null    int64
27  Seccion_training_courses                 614 non-null    int64
28  texto_extraido_len                       614 non-null    int64
29  Promedio tamaño fuente                   614 non-null    float64
30  Formato Texto (Lineas)_Párrafos          614 non-null    int64
31  Formato Texto (Lineas)_Viñetas           614 non-null    int64
32  Orden Temporal_Orden Temporal Detectado  614 non-null    int64
33  Formato Fecha Más Común_MM-YYYY          614 non-null    int64
34  Formato Fecha Más Común_MM/YYYY          614 non-null    int64
35  Formato Fecha Más Común_Mon YYYY (EN)    614 non-null    int64
36  Formato Fecha Más Común_YYYY             614 non-null    int64
37  Fuente principal_ArialMT                 614 non-null    int64
38  Fuente principal_Calibri                 614 non-null    int64
39  Fuente principal_Otra                    614 non-null    int64
```

40 Fuente principal_Tahoma 614 non-null int64

41 Legibilidad general_Buena 614 non-null int64

42 Consistencia tamaños fuente_Consistente 614 non-null int64

43 Consistencia márgenes (aprox)_Consistente 614 non-null int64

44 Uso de colores (texto)_Sí 614 non-null int64

45 Uso de colores (dibujos)_Sí 614 non-null int64

46 Deteccion Foto Perfil_Posible Foto Detectada 614 non-null int64

47 secciones_completas 614 non-null int64

48 Ratio_Lineas_Experiencia_len 614 non-null float64

49 Passed 614 non-null int64

9.5. Anexo 5: Recomendaciones de uso de datos en la empresa para el almacenamiento y procesamiento de la información de los candidatos.

El presente reporte se basa en el análisis de calidad de datos realizado basado en las dimensiones de completitud, validez, unicidad, consistencia y precisión.

Las recomendaciones serán categorizadas de acuerdo con los deberes del reclutador y a lo que debe ser solicitado al candidato en la entrevista para que la data sea lo más precisa posible.

Recomendaciones para los reclutadores

- Los candidatos registrados deben contar con un Resume en el ATS en toda situación, así sea el generado automáticamente en LinkedIn con el objetivo de entender la persona que reclutó y entrevistó al candidato registrado, así como para evitar que el candidato quede sin ‘Owner’ (Caso visto en 400 registros).
- En caso de que el scrapping de LinkedIn no vincule ningún candidato con un Empleo y Puesto actual (o inmediatamente anterior), debe ser escrito manualmente por el reclutador. Así mismo, en un caso excepcional que el candidato no cuente con trabajo actual o anterior, se debe escribir un "N/A", mas no dejar la columna vacía (Cosa que pasó en 208 registros).
- El email debe ser un campo obligatorio para crear un candidato en el cual el reclutador debe poner particular atención, siendo que es un identificador principal para el candidato, así como para evitar duplicados del mismo. Este no puede estar vacío (Visto en 265 registros). El email en la gran mayoría de casos debe tenerse

antes de la entrevista, al ser necesario para agendar la misma, pero en un caso excepcional que no sea así, se le debe solicitar al mismo junto con su CV.

- El reclutador debe verificar manualmente si el correo ya se encuentra creado.
- Tanto el candidate location como la región deben ser un campo obligatorio. Dado que en múltiples ocasiones los candidatos no ponen su ubicación real en LinkedIn, debe ser un dato a preguntar en la entrevista.
- La industria del candidato debe ser un campo obligatorio, en caso de que el reclutador no conozca la empresa actual o anterior del candidato o no sea claro del todo la industria de la misma, puede referirse al LinkedIn de la compañía y copiar la industria del encabezado de esta en LinkedIn, con tal de no dejar el campo vacío (2465 registros).
- El salario actual y anterior deben ser obligatorios a agregar. En dado caso que el candidato no haya querido discutir al respecto del salario actual, es recomendable poner un "0" con tal de no dejar el campo vacío (Visto en 1028 registros).
- Si el candidato no cuenta con beneficios actuales o no se discutieron, se pondrá un "N/A" o un "Not discussed" dependiendo del caso, más no dejarlo como un campo vacío (Visto en 2331 casos).
- En dado caso de no contar con el notice period de un candidato, con tal de no dejar el campo vacío, se puede poner el estándar de "2 semanas" como dato por defecto.
- Los nombres deben registrarse sin caracteres especiales para encontrar fácilmente a los candidatos además de poder mantener su registro de manera uniforme.
- Es necesario que el reclutador verifique que el campo de "Candidate First Name" tenga el nombre(s) del candidato y el campo "Candidate Last Name" apellido(s) con

el objetivo de que la información sea consistente y sea más sencillo encontrar a un candidato en particular en la base.

- Es importante mantener un formato uniforme pero simple, el propuesto es el siguiente "+## ##########" siendo los dos primeros (o tres primeros en ciertos casos) el código país y el resto, el número de teléfono como tal, sin ningún símbolo adicional o espacio de por medio.
- Debe ser obligatorio en todos los casos llenar los 4 campos de información para el salario tanto actual como esperado. Como se mencionó anteriormente, en caso de no contar con la información, los valores predeterminados serán 0 (monto), USD (moneda), monthly (frecuencia), permanent (contrato).
-

Recomendaciones para la entrevista (información a solicitar)

- Aunque no se avance con el candidato en el proceso, el salario esperado es un campo el cual debe ser solicitado siempre al candidato con el fin de contar con información para posibles análisis salariales y tener la posibilidad de contar con dicho candidato para eventuales vacantes futuras.
- El reclutador debe verificar si las expectativas del candidato para el tipo de contrato ofrecido. Así mismo, si el candidato accede a compartir su compensación actual, también debe ser indagado el tipo de contrato. Dicha información debe ser registrada en el ATS.
- En un caso excepcional que no se cuente con el email del candidato de antemano, se debe solicitar durante la entrevista para su posterior registro en el ATS.
- Verificar en todos los casos con el candidato su ubicación real.

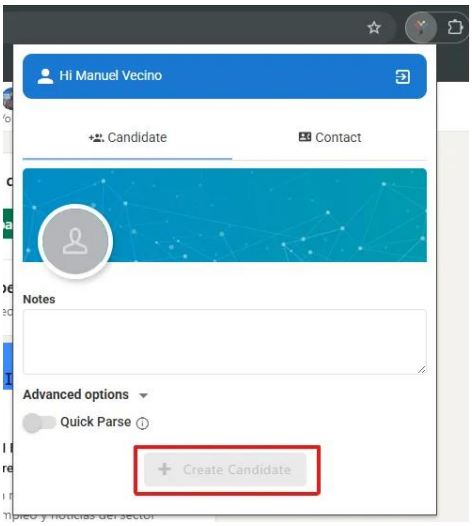
- En dado caso que el CV del candidato no cuente con el número de teléfono, se debe solicitar el número de teléfono a los candidatos en caso de avanzar con el objetivo de garantizar una comunicación más fluida con los mismos durante el proceso.

9.6. Anexo 6: Manual de uso de Manatal para la entrada de datos en el sistema de información

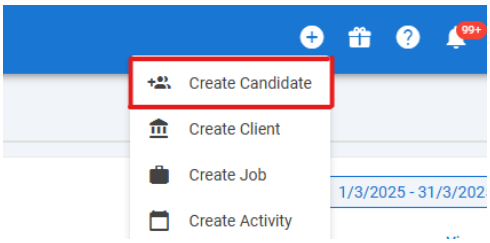
Registro de candidatos

Al momento de registrar un candidato hay dos opciones:

Usando la extensión de people match:

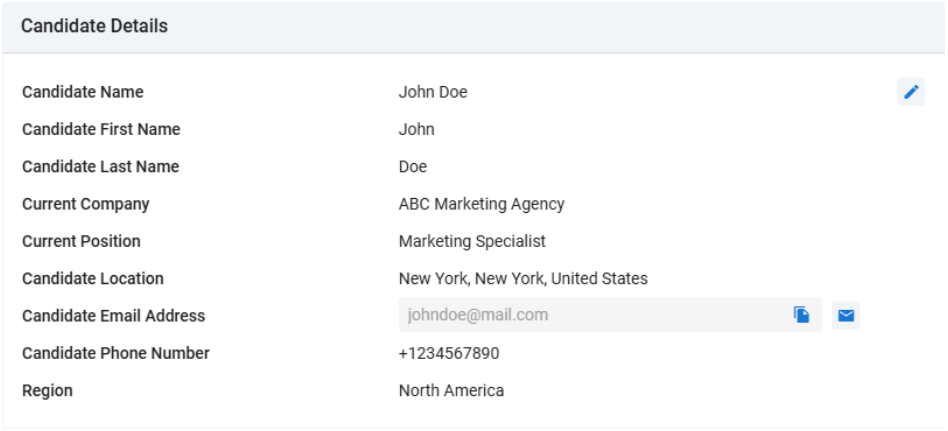


Subiendo la CV del candidato directamente a Manatal:



Ingreso de información

Una vez se encuentre creado el candidato, en primer lugar, se debe verificar que la información del campo “Candidate Details” se encuentre completa y sea consistente.



- En este caso, podemos ver como los campos “Candidate First Name” y “Candidate Last Name” son consistentes con lo visto en “Candidate Name”.
- El campo “Candidate name” **no debe tener** ningún tipo de carácter especial, emoji, tilde o simbolo. En caso de que el nombre del candidato cuente con alguno de estos, debe ser modificado.
- Así mismo, los campos “Current company” y “Curren position” deben coincidir con lo descrito por el candidato en la entrevista.
- De la misma manera, la ubicación debe corresponder con la ubicación real del candidato, verificada en la entrevista. Esta debe corresponder con la región seleccionada por el reclutador.

- El email debe ser verificado y consistente con el que el candidato proporcionó para el agendamiento de la HV. Así mismo debe ser verificado en el buscador para evitar algún tipo de duplicado.
- El teléfono cuenta con un formato numérico, sin símbolos o espacios diferentes del símbolo "+" el cual denota el código de país del candidato.

Al haber registrado la información básica, se continúa con la sección "Additional information"

Additional Information

Years of Experience	7+ years
Current Salary	1000 USD Monthly (permanent)
Current Benefits	Health Insurance
Notice Period	10 Days
Expected Salary	1500 USD Monthly (contract)
Candidate industry	Marketing
salary_min	1400
salary_max	1600

- Se deben registrar manualmente los años de experiencia del candidato.
- El salario actual idealmente debe obtenerse en la entrevista, junto con la frecuencia de remuneración y el tipo de contrato del mismo. En caso de que el candidato sea reacio a compartir dicha información, se debe dejar un "0" en la casilla.
- Los beneficios son escritos manualmente, de forma breve. En caso de no contar con beneficios se escribe "None" o "N/A" si no se llegaron a discutir durante la llamada.
- El notice period del candidato se registra de acuerdo a la lista dispuesta en Manatal. En caso de no tener notice period, se usa "10 Days" por defecto.
- La expectativa salarial debe incluir no solo el número y moneda en la cual el candidato espera ser remunerado, sino también el tipo de contrato y periodicidad del

mismo, ya que indica el monto que un candidato quisiera dependiendo del contrato ofrecido.

- La industria se escoge de la lista dispuesta en la app, o en caso de que no se liste, se puede escribir manualmente al pulsar la opción “other”. En caso de que no se tenga conocimiento sobre la industria de la empresa, se puede ir a la página en LinkedIn de la misma y copiar la descrita por esta.

Consideraciones adicionales:

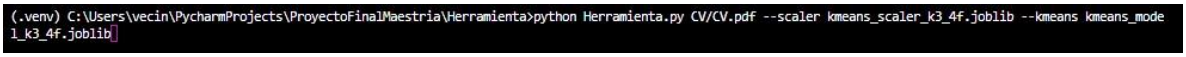
- Todos los candidatos deben tener un CV en la plataforma antes de ser agregados a un rol.
- Los candidatos después de ser entrevistados deben tener las respectivas notas para el rol agregadas en la pestaña de “Notes”, en la sub-pestaña del rol para el cual está siendo considerado.
- Se pueden agregar notas generales del candidato en la sub-pestaña “General” en la pestaña de “Notes”
- Es ideal que cada vez que un candidato pase de étapa, pueda ser registrada la consideración de este paso en una nota.
- Al rechazar un candidato, debe siempre ser explicita la razón de rechazo, adicional a la razón dada por defecto por Manatal

9.7. Anexo 7: Ejemplos de uso de herramienta analítica

Comando ejemplo 1:

Output ejemplo 1:

Comando ejemplo 2:



Output ejemplo 2:

```
(.venv) C:\Users\vecin\PycharmProjects\ProyectoFinalMaestria\Herramienta>python Herramienta.py CV/CV.pdf --scaler kmeans_scaler_k3_4f.joblib --kmeans kmeans_mode
l_k3_4f.joblib
2025-05-11 16:23:17,986 - INFO - Scaler y modelo K-Means cargados exitosamente.
2025-05-11 16:23:18,002 - INFO - Procesando PDF: CV.pdf
2025-05-11 16:23:18,020 - INFO - Features extraídas: {'texto_extraido_len': 1803, 'secciones_completas': 7, 'Seccion_training_courses': 0, 'Website/Otro': 0}
2025-05-11 16:23:18,137 - INFO - Cluster asignado: 1

--- Resultados del Análisis ---
Features Extraídas:
  - texto_extraido_len: 1803
  - secciones_completas: 7
  - Seccion_training_courses: 0
  - Website/Otro: 0

Cluster Asignado (K=3): 1
Cluster 1:  Sin Training (y Mayormente sin Website) - Corto

Sugerencia: Ayúdale al candidato a brindarle más contenido a su CV, si es posible agregar una sección de entrenamiento,
 cursos o certificaciones en caso de no tenerla.
Sugiere añadir links a websites externos de proyectos anteriores o inclusive un portafolio si lo tiene
Si lo vez necesario, recomiéndale al candidato también incrementar el contenido de esta, tal vez agregando más texto en sus experiencias laborales o agregando se
cciones adicionales
```

9.8. Anexo 8: Archivo README Incluido en la herramienta

==========

==========

Herramienta de Análisis de CVs y Asignación de Perfil (Cluster)

==========

==========

Fecha de Creación: 6/5/2025

Autor: Manuel Vecino

Versión del Script: 1.0

Descripción

Este script en Python analiza un archivo de Curriculum Vitae (CV) en formato PDF para extraer 4 características clave:

1. Longitud total del texto extraído (`texto_extraido_len`).
2. Número de secciones principales detectadas (`secciones_completas`).
3. Presencia de un enlace a un sitio web personal o portafolio (`Website/Otro`).
4. Presencia de una sección dedicada a cursos o formación (`Seccion_training_courses`).

Utilizando estas 4 características, el script asigna el CV a uno de los 3 perfiles (clusters) predefinidos, los cuales han mostrado tener diferentes tasas promedio de éxito ('Passed') en análisis previos. Los perfiles identificados son:

* **Cluster 0:** "Con Website y Sección Training" (Tasa de 'Passed' más alta)
* **Cluster 1:** "Sin Training (y Mayormente sin Website) - Corto" (Tasa de 'Passed' más baja)
* **Cluster 2:** "Con Sección Training, Sin Website (Grupo Estándar)"

Este análisis proporciona una visión descriptiva del perfil del CV.

Requisitos Previos

Antes de ejecutar el script, asegúrate de tener instalado lo siguiente:

1. **Python:** Versión 3.8 o superior recomendada. Puedes descargarlo desde https://www.python.org/downloads/.
   * Durante la instalación en Windows, asegúrate de marcar la casilla "Add Python to PATH".

2. **Archivos de Modelo:** Debes tener los siguientes archivos en la misma carpeta que el script `Herramienta.py` (o ajustar las rutas en el comando de ejecución):
   * `kmeans_scaler_k3_4f.joblib`: El objeto StandardScaler entrenado.
   * `kmeans_model_k3_4f.joblib`: El modelo KMeans (K=3) entrenado.

Instalación de Dependencias

Este script requiere varias librerías de Python. La forma más sencilla de instalarlas es usando `pip` dentro de un entorno virtual.

1. **Abrir una Terminal o Símbolo del Sistema:**

* **Windows:** Busca "cmd" o "PowerShell".

* **macOS/Linux:** Abre la aplicación "Terminal".

2. **Navegar al Directorio del Proyecto:**

Usa el comando `cd` para moverte a la carpeta donde descargaste este script y los archivos de modelo. Ejemplo:

cd ruta/a/la/carpeta_del_script

3. **(Recomendado) Crear y Activar un Entorno Virtual:**

* Crea el entorno (solo una vez):

```bash
python3 -m venv mi_entorno_cv o
python -m venv mi_entorno_cv
```

* Activa el entorno:

  * Windows:

```bash
mi_entorno_cv\Scripts\activate
```

  * macOS/Linux:

```bash
source mi_entorno_cv/bin/activate
```

Deberías ver `(mi_entorno_cv)` al inicio del prompt de tu terminal.

4. **Instalar las Librerías Requeridas:**

Copia el archivo `requirements.txt` (que debe estar en la misma carpeta que este README)

y ejecuta:

```bash
python3 -m pip install --upgrade pip o
python -m pip install --upgrade pip

pip install -r requirements.txt
```

Esto instalará: PyMuPDF, rapidfuzz, numpy, pandas, joblib, scikit-learn, y sus dependencias.

Uso del Script

Una vez instaladas las dependencias y con el entorno virtual activado (si creaste uno), puedes ejecutar el script desde la terminal.

Comando Básico:

```bash
python3 Herramienta.py CV/[AQUI MODIFICA EL NOMBRE POR EL DEL ARCHIVO DE CV A ANALIZAR].pdf --scaler kmeans_scaler_k3_4f.joblib --kmeans kmeans_model_k3_4f.joblib o
python Herramienta.py CV/[AQUI MODIFICA EL NOMBRE POR EL DEL ARCHIVO DE CV A ANALIZAR].pdf --scaler kmeans_scaler_k3_4f.joblib --kmeans kmeans_model_k3_4f.joblib

python3 Herramienta.py CV/sample2.pdf --scaler kmeans_scaler_k3_4f.joblib --kmeans kmeans_model_k3_4f.joblib
o python Herramienta.py CV/sample2.pdf --scaler kmeans_scaler_k3_4f.joblib --kmeans kmeans_model_k3_4f.joblib
```

Copia y pega esto en la terminal.

9.9. Anexo 9: Link del repositorio de GitHub del proyecto

https://github.com/mastervecino/ProyectoFinalMaestria

9.10. Anexo 10: Video muestra del uso de la herramienta analítica

https://youtu.be/H-JXFMUOBYE

Muchas gracias por leer hasta aquí :)