SISTEMA DE RECOMENDACIÓN DE EMPLEO – EVIDENCIA PARA ALGUNOS PROGRAMAS DE PREGRADO DE LA UNIVERSIDAD DE LA SABANA

GABRIEL GERARDO AMAYA BECERRA

ECONOMISTA CON ÉNFASIS EN FINANZAS INTERNACIONALES

DOCUMENTO DE GRADO

FACULTAD DE INGENIERÍA

MAESTRÍA EN ANALÍTICA APLICADA

Universidad de
La Sabana

Universidad de La Sabana

PÁGINA DE ACEPTACIÓN

Tutor: Felix Vivian Mohr

_____

Jurado: Maria Fernanda Rodriguez Moreno

_____

Jurado: Dr. Cesar Enrique García Diaz

_____

Jurado: Dr. Juan Manuel Aranda López King

Universidad de La Sabana

TABLA DE CONTENIDO

Palabras Clave

Sistemas de recomendación, procesamiento en lenguaje natural, machine learning, modelos no supervisados, trabajo, empleo, perfilamiento, desempleo, habilidades, soluciones tecnológicas, analítica, recuperación de información.

Resumen

Uno de los aspectos preocupantes dentro de la dinámica social colombiana, es el alto nivel de desempleo que se viene presentando en los jóvenes, cifra que alcanzó el 23.5% entre febrero y abril de 2021 [1]. Y es que aunque ya existen plataformas tecnológicas como LinkedIn, El Empleo, CompuTrabajo, entre otras, las cuales contienen una gran cantidad de ofertas laborales, se encuentra una brecha entre los conocimientos que tiene el estudiante o el recién graduado que iniciará su inserción en el mundo laboral, frente a los requisitos de habilidades y competencias que presenta un mercado laboral cada vez más exigente, producto de la implementación de tecnologías propias de industrias 4.0 y la búsqueda de automatización de tareas repetitivas para optimizar tiempos y costos.

A partir de lo anterior, se ha propuesto desarrollar una aproximación de conexión laboral a través de un sistema de recomendación que permita identificar los perfiles que presentan un mayor grado de similitud con los requisitos expuestos en las ofertas laborales, teniendo como principal premisa el generar un empoderamiento de las personas respecto a los conocimientos que posee, el tipo de cargos que puede desempeñar, y al mismo tiempo abrir la posibilidad de identificar otras posibles habilidades que debería adquirir para mantenerse acorde a las tendencias del sector profesional en el que se desempeña o podría desempeñarse.

Introducción

Para el año 2019, en Colombia se graduaron 398.149 estudiantes de educación superior en los niveles técnico, tecnológico y profesional, de los cuales el 48.58% se encuentran ubicados en el departamento de Cundinamarca [2], en su mayoría jóvenes que han adquirido una serie de conocimientos de un programa de educación formal y que esperan salir a aplicarlos en el mercado laboral.

Desde el otro lado, las empresas vienen avanzando cada vez más en la búsqueda de personal con un conjunto de habilidades y conocimientos de mayor interdisciplinariedad, pero con un nivel de profundización de nicho[1], los cuales le permiten a la empresa generar eficiencias a través de la automatización de procesos y aprovechar la ventana de oportunidad que genera la adopción de las tecnologías de industrias 4.0, tales como la robótica avanzada, el análisis de datos, la inteligencia artificial, la simulación, los sistemas embebidos, la fabricación aditiva, la realidad virtual y la realidad aumentada [3], que se espera sean fuertemente fortalecidas con la implementación de las redes 5G en Colombia, debido a que mejoran la velocidad de la red, presentan una menor latencia, permite la conexión de múltiples dispositivos y presenta un mayor ancho de banda.

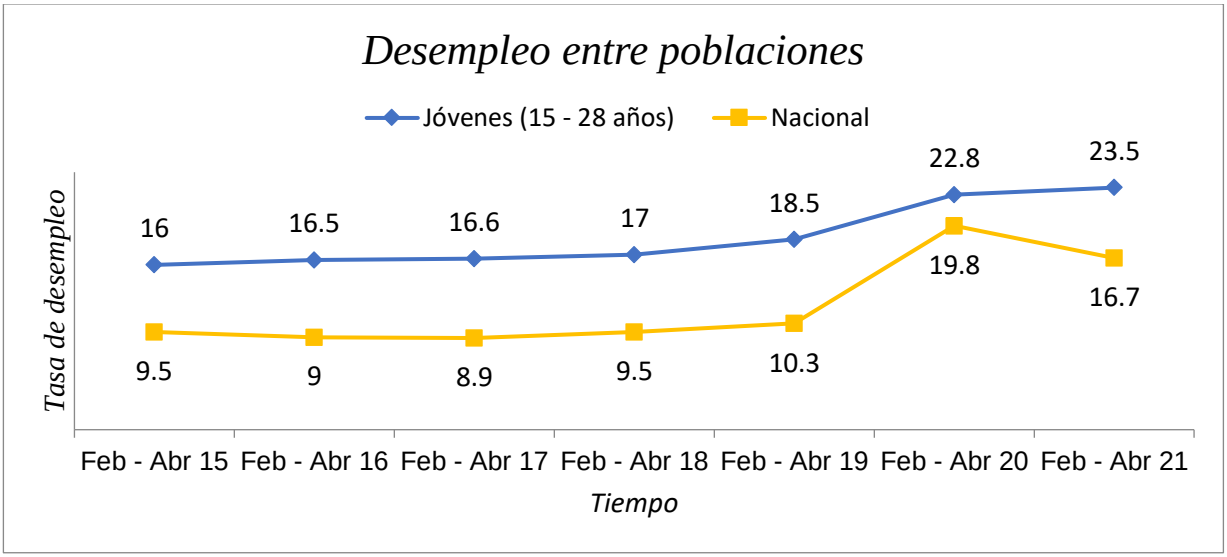
A partir de estas dos realidades, es interesante observar que, de acuerdo con el informe del DANE [1], en la población joven[2] desde el año 2015 se presentan niveles de desempleo de 16%, cifra que para el año 2019 había aumentado a niveles del 18.5% y que para la medición de cierre del trimestre febrero - abril del 2021 se ubicó en el 23.5%[3] (Ver figura 1), lo que sugiere que algo está pasando en Colombia

[1] El surgimiento por ejemplo de perfiles como el Growth Hacker o el Data Science que combina campos de conocimiento como la estadística, la programación, la inteligencia de negocios, la comunicación, entre otros, para el desarrollo de sus funciones.
[2] La población joven definida por el DANE se encuentra entre los 14 y los 28 años.
[3] Siendo importante reconocer, que una parte del aumento proviene también de la crisis generada por la pandemia, pero que la tendencia ha sido creciente en la población joven, mientras a nivel nacional el aumento se observa durante los meses de la pandemia.

dentro del mercado laboral en la población joven, lo que ha llevado a que la tasa de desempleo se encuentre en aumento constante.

Figura 1

Comparativo Desempleo entre población Joven y el nivel Nacional (Trimestre febrero – abril 2015 a 2021)



Fuente: Información DANE

Una de las posibles razones para explicar el comportamiento de los resultados obtenidos a nivel de desempleo puede estar asociado a los hallazgos de la investigación de Pal'Ova [4], en la que se identifica que existe un diferencial entre lo que requiere el mercado laboral y las habilidades con las que cuenta la persona o como lo define Singh [5] la diferencia entre la expectativa de la industria y los resultados que puede dar el graduado con los conocimientos que posee. Para ello, y con el objetivo de lograr alinear lo mejor posible los requerimientos empresariales frente a los de cada egresado, se ha propuesto el desarrollo de un sistema de recomendación de empleo, el cual además de permitir conocer las vacantes a las que puede aplicar cada persona conforme a su perfil profesional, identifica las habilidades o conocimientos que está requiriendo el mercado y que actualmente no posee o no se encuentra explicito en el perfil.

Para lograr estos dos objetivos, se hará uso de distintas herramientas analíticas. La primera será el Procesamiento en Lenguaje Natural - NLP, técnica que permitirá realizar el procesamiento de textos a fin de extraer los conocimientos y habilidades que se encuentran en cada uno de los perfiles y de las ofertas laborales. A partir de los resultados obtenidos se utilizará la segunda técnica que serán los modelos de recuperación de información, con los cuales se podrá realizar la conexión e identificación de los puestos laborales a los que una persona con sus habilidades actuales podrá presentarse, como también la detección de las habilidades más solicitadas que actualmente no se encuentran en el perfil.

El documento se encuentra compuesto por los siguientes elementos:, la primera parte corresponde a la revisión bibliográfica; el segundo fragmento se encuentra compuesto por la metodología de recolección de información y el proceso inicial de exploración de datos; la tercera sección contiene el proceso de limpieza de información, los 7 modelos de recuperación de información desarrollados y los resultados obtenidos de cada modelo; la cuarta parte presenta los resultados obtenidos del sistema de recomendación y la metodología utilizada para la evaluación del modelo; la quinta sección presenta las conclusiones; y finalmente en la sexta sección se presenta el trabajo futuro.

1 . Revisión bibliográfica

Las dinámicas del mercado laboral han sido objeto de estudio desde diferentes perspectivas. Teniendo presente la situación de coyuntura actual muy similar en sus características al comportamiento ocurrido durante la gran depresión, se hace pertinente traer los principales hallazgos de Tristan Potter entre los cuales se logró identificar que las personas desempleadas durante este tipo de episodios (eventos coyunturales), destinan al menos una hora diaria a encontrar empleo y en la medida que pasa el tiempo empiezan a tener un comportamiento diferente en esa dinámica de búsqueda, dónde uno de los factores que se destaca es la reducción del nivel de esfuerzo en la búsqueda. Potter también afirma que existe un alto optimismo entre

las personas que recién pierden el trabajo por encontrar uno nuevo y que conocer la expectativa salarial permitiría enfocar mejor los esfuerzos [6], obteniendo así que uno de los principales aspectos que espera aportar el sistema es una reducción en los tiempos de identificación de las vacantes más acordes al perfil y que la identificación de los salarios puede ser una pieza importante a tener en cuenta dentro del mismo.

De igual forma, en la actualidad existe una expectativa respecto al impacto que generará la inteligencia artificial y el machine learning en los trabajos futuros, expectativa que se enfoca principalmente en los desplazamientos laborales que se pueden generar, más que en la dinámica o beneficios que puede traer el trabajo en conjunto con las nuevas tecnologías [7]. Para el estudio, estos hallazgos sobre la dinámica laboral permiten dar una mayor relevancia a la identificación de las habilidades que está requiriendo el mercado y que puedan aportar a mejorar esa percepción de competencia con la tecnología, a una de trabajo en conjunto para obtener beneficios.

Frente a los actores de la dinámica laboral, las conclusiones presentadas por Di Pace [8] nos permiten observar cómo estas (las empresas) también tienen un proceso de aprendizaje frente a cuál es el mejor momento para publicar la vacante disponible, lo cual hace que se genere un factor cíclico del mercado laboral.

Adicionalmente, los cambios que se vienen presentando en las dinámicas empresariales, hacen que cobre importancia el marco de referencia aportado por Itsakov [9], pues el autor afirma que la aplicación de tecnologías de industrias 4.0 y otras alternas, generarán que consultores, analistas, diseñadores, ingenieros, traders, entre otros perfiles identificados, sean reemplazados producto de una mejora en temas asociados a costos, calidad, disponibilidad y producción, una vez se haya implementado la tecnología respectiva, de igual forma que todas estas capacidades digitales están enfocadas en el nuevo mercado laboral (población joven) más que en la adaptación de los trabajadores existentes, punto que también es reforzado por Nazareno [7] cuando sugiere que las personas con menores

habilidades serán las que tendrán un mayor nivel de riesgo a ser reemplazadas por la tecnología, aumentando por tanto la relevancia de identificar los conocimientos o habilidades que las personas deberían adquirir para mantenerse dentro de las dinámicas del mercado laboral, pero al mismo tiempo revelando que en la población joven que ingresa al mercado laboral se da por descontado el dominio de esas habilidades, mientras que en la población más adulta debe existir una adaptabilidad para la adquisición de estas.

Justamente desde esta perspectiva de conocimientos y habilidades, Pal'Ova [4] da a conocer en los resultados de su investigación enfocada en las áreas financieras y de contabilidad, que se identificó que dentro de los principales problemas para la vinculación laboral se encuentra la falta de preparación de los graduados para resolver problemas prácticos en un 76%, y el cual desde la perspectiva del estudio podría evaluarse como la experiencia profesional que tiene la persona; la falta de trabajo independiente se presenta en un 72%; la falta de personal calificado para los requerimientos laborales 64% (para el cual se ha propuesto el diseño del sistema de recomendación), seguido de las debilidades en lenguas extranjeras 58% y altas expectativas salariales 57%. Estos últimos dos factores son más complejos de validar, el primero debido a que el alcance del proyecto no llega a evaluar el nivel de competencia o dominio de la lengua (o cualquier otra habilidad o competencia) y el segundo debido a que esta información no siempre se encuentra disponible o que la expectativa salarial puede variar conforme al cargo, las responsabilidades a desempeñar o el salario de reserva que requiere la persona para cumplir con sus responsabilidades.

De igual forma Singh [5] nos plantea que desde la parte de habilidades tenemos distintos componentes como lo son las habilidades core, las habilidades transferibles, las habilidades genéricas, las habilidades funcionales y las habilidades emprendedoras; las cuales pueden variar en su nivel de importancia dependiendo del tipo de cargo que se desarrollará o la empresa en que se desempeñará, siendo por tanto un marco de referencia respecto a las categorías que deberían ser tenidas

en cuenta en el momento de construir y desarrollar el sistema propuesto e incluso generando una nueva posibilidad de evaluar las habilidades o competencias con grados de importancia, acorde a la necesidad de cada vacante o el nivel de dominio de cada persona.

Desde la perspectiva técnica se destaca el marco teórico realizado por Kumar [10], el cual permite tener en contexto los diferentes tipos de sistemas de recomendación, sus ventajas y limitaciones, se resalta que para el presente estudio el sistema de recomendación más apropiado será el CBRS (Content-based recommender system) debido a que tiene en cuenta las descripciones entre el usuario y el ítem (para este caso la vacante). Sin embargo, es importante reconocer que estos modelos presentan una alta dependencia de desempeño a la información que se presente del usuario, este problema es conocido como limitación de análisis de contenido, otro aspecto importante está asociado a la sinonimia (mismos o similares ítems con diferentes nombres) que fácilmente son asociados por las personas pero que a nivel computacional son diferentes, al igual que las abreviaciones las cuales pueden generar confusiones o no ser identificadas dentro del modelo.

Se resaltan de igual forma los hallazgos de Gorb [11] enfocados en las principales características que poseen los modelos de recuperación de información desarrollados, entre los que se destaca que una técnica indispensable es la tokenización debido a que permite dividir el texto para hacerlo analizable, la segunda técnica común es retirar los stopwords ya que permite quitar aquellas palabras que no tienen importancia semántica y finalmente el uso de TF-IDF el cual permite evaluar la importancia de una palabra y realizar recomendaciones.

Respecto al desarrollo de modelos de machine learning, se destacan las aproximaciones realizadas por Alksasbeh [12] quién plantea el uso de ANNIE POS Tagger y Fuzzy Inference como una alternativa de solución al problema y obtiene resultados por encima del 80% en las métricas de precisión, recall y F1 Score, sin embargo, este método presenta requerimientos específicos para su funcionamiento, dentro de los que se encuentran el idioma en el que debe estar la vacante (Inglés)

y el tamaño que se puede incluir en la descripción, rango que oscila entre 300 a máximo 1.000 palabras de longitud, por su parte Giabelli [13] propone un modelo a partir de coocurrencias de las habilidades de las personas alcanzando un nivel de precisión superior al 70%, este modelo está compuesto de tres partes, la primera es un set de ocupaciones (vacantes), la segunda un set de habilidades y la tercera una relación asociativa entre la ocupación y las habilidades.

La propuesta más interesante en términos de estructura global[4] para abordar el problema es la planteada por Bañeres [14], quién hace uso de técnicas de procesamiento de lenguaje natural para la estandarización de información, respecto a la identificación de vacantes que no son fáciles de cubrir utiliza un modelo de recomendación basado únicamente en términos lingüísticos que son iguales entre la vacante y el perfil, y frente a la recomendación de los cursos dónde se puedan desarrollar las habilidades faltantes, se conectan a partir de los conocimientos no evidenciados en el perfil pero que requieren las diferentes vacantes.

De igual forma el autor reconoce que estos modelos son altamente sensibles a la conexión con palabras claves, la estructura o forma en que se redacta, e incluso para conectar con el life-long learning es altamente dependiente de la información que se encuentra sobre el programa académico respectivo. A partir de ello, es que el foco de la investigación que el autor desarrolló estuvo orientada a las áreas de tecnología y de evidencia para algunos estudiantes en Barcelona - España, lo que en palabras de Singh [10] sería definido como una sobre especialización del modelo, sin por ello desconocer el alto potencial que tienen los resultados obtenidos allí.

Finalmente, a partir de la revisión de literatura se puede determinar que, desde la perspectiva técnica, una buena práctica para extraer la información de habilidades y competencias tanto de las vacantes como de los perfiles a través del procesamiento en lenguaje natural [14], adherido a un sistema de recolección de información que, tal como se describía previamente, presentará una alta

[4] Estructura global entiéndase como la combinación entre el conocimiento técnico y el conocimiento de negocio que permite llegar a una solución aproximada.

dependencia de las palabras claves que sean definidas en la bolsa de palabras y la cual debería tener en cuenta los diferentes campos de habilidades propuestos por Singh [5].

Ahora bien, desde la perspectiva de conocimiento de negocio, es importante tener en mente los hallazgos realizados por Itsakov [9] y Nazareno [7] en temas de los perfiles altamente susceptibles a ser reemplazados y la relación con Pal´Ova [4] respecto a los principales motivos por los que una persona presenta dificultades para vincularse al mercado laboral, pues estos serán los factores motivantes y principales para determinar la calidad del modelo final obtenido, de igual forma es importante mencionar que tal como lo plantea Allal-Chérif [15] el sistema de recomendación propuesto es uno de los métodos que se encuentra enfocado en generar una mayor eficiencia en las áreas de contratación y selección, más que en reemplazar este tipo de cargos.

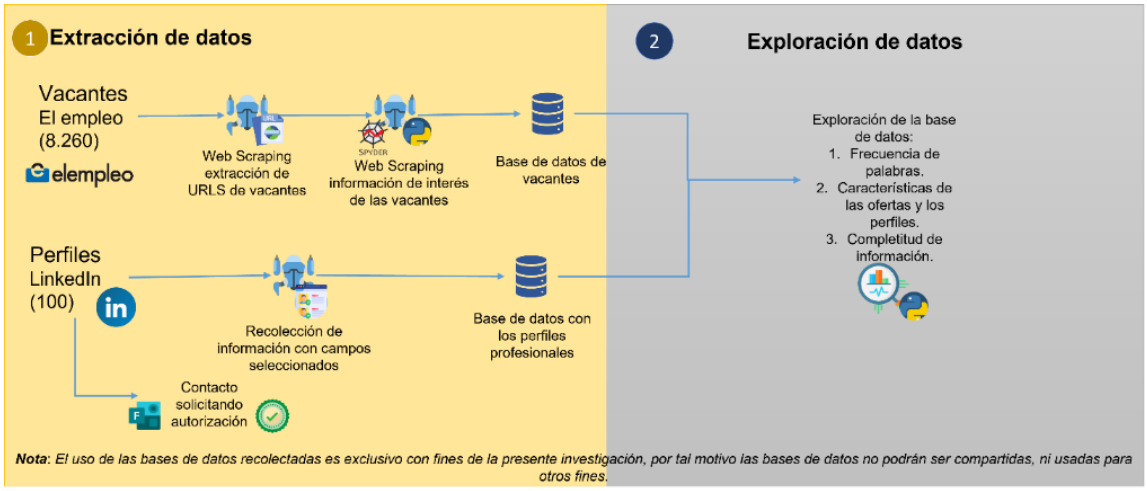
2. Metodología de recolección y exploración de datos

El proceso que se realizó para la recolección y exploración de información será descrito a lo largo de este segmento, se recomienda revisar el diagrama 1 que resume la información presentada (ver diagrama 1).

Uno de los primeros retos para el desarrollo del modelo se encontraba en la recolección de información actualizada, disponible y verídica. Durante el proceso de selección del portal del que sería descargada la información de las vacantes se tuvo en cuenta variables como: la cantidad de ofertas, la variedad de ofertas, la disponibilidad de información como la ubicación dónde se desarrollará el empleo, la posibilidad de obtener alguna otra información adicional asociada a la vacante como el nombre de la empresa, el salario y la posibilidad del manejo de información acorde a las políticas de protección de datos de las plataformas, a partir de estos criterios se consideró como mejor opción el portal de El Empleo.

Diagrama 1

Proceso desarrollado para la extracción y exploración de datos



Fuente: Realización propia

La extracción de información se realizó a través de dos procesos automáticos de web scraping[5], el primer proceso permite extraer la url de las vacantes y el segundo proceso permite extraer la información de interés. Una vez ejecutados los procesos se obtuvo un total de 8.260 vacantes conformadas por dos tipos de búsquedas, la primera en cargos de analistas, coordinadores, ingenieros y profesionales de nivel junior. La segunda asociada a conocimientos o habilidades generales que fuesen requeridas en las ofertas tales como el análisis de datos, calidad, comunicación, desarrollador, marketing, procesos y psicología; lo anterior teniendo presente los tipos de perfiles profesionales que se incluyeron dentro del marco del estudio y que serán presentados más adelante.

La base de datos descargada del portal seleccionado cuenta con un total de 7 campos presentados en la tabla 1 (ver tabla 1).

[5] La información de las vacantes extraídas fue utilizada únicamente con fines del presente estudio, no será utilizadas para fines comerciales, ni podrá ser compartida, respetando y acogiendo así la línea ética de manejo de información determinada por CEET – Casa Editorial El Tiempo.

Tabla 1

Campos descargados del portal El Empleo

Nombre de la variable	Descripción
Codigo_oferta	Contiene el código de la oferta que le asigna el portal del El Empleo.
Descripción	Contiene la descripción de la oferta laboral.
Empresa	Contiene el nombre de la empresa que postula la oferta laboral.
Rango salarial	Contiene el rango salarial ofertado por la empresa.
Tipo contrato	Contiene el tipo de contrato que oferta la empresa.
Título	Contiene el título de la oferta laboral (acotado por los dos tipos de búsqueda mencionados previamente).
Ubicación	Contiene la ubicación en la que se desarrollará el cargo.

Fuente: Realización propia

Las vacantes obtenidas se encontraban distribuidas en 178 ubicaciones diferentes[6], sin embargo, teniendo presente que el estudio se enfocó en los programas de pregrado de la Universidad de La Sabana y que estos se desarrollan de forma presencial en el campus ubicado en el municipio de Chía – Cundinamarca, se decidió filtrar las vacantes a partir de la zona de influencia de la universidad, correspondiente a los municipios de Sabana Centro, Sabana de Occidente y Bogotá D.C., pasando a una base de datos con un total de 24 ubicaciones y 5.763 vacantes.

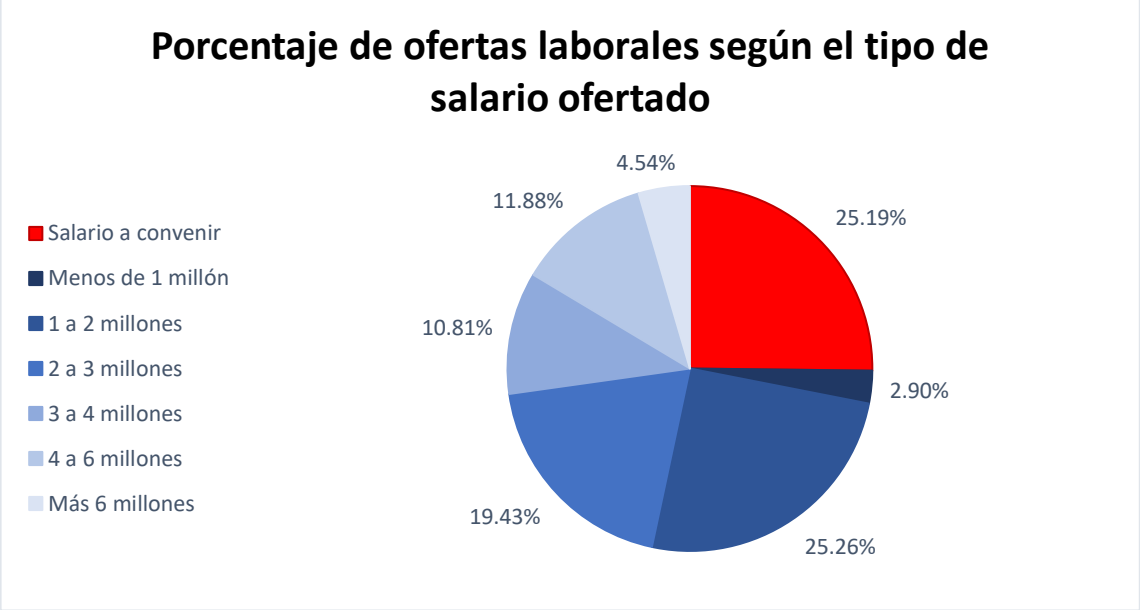
Durante el proceso de análisis exploratorio se evidenció que las 50 palabras más frecuentes dentro de las vacantes correspondían a temáticas como la experiencia, el tipo de trabajo, habilidades, horarios, entre otros, dónde se destacan las habilidades o conocimientos en los siguientes campos: sistemas, procesos, datos, excel, servicio, gestión, entre otros.

[6] Principalmente municipios, también se encuentran otras descripciones como “Bogotá – Alrededores” o “Toda Colombia”.

Teniendo presente los hallazgos encontrados durante la revisión de literatura de Pal´Ova [4] respecto a las expectativas salariales, se validó la disponibilidad de la información del salario en las vacantes, en la cual se encontró que el 75% del total de las ofertas laborales presentan un rango salarial base y el restante presenta un salario a convenir (ver figura 2).

Figura 2

Porcentaje de ofertas laborales según el tipo de salario ofertado.



Fuente: Realización propia.

Respecto al siguiente punto, asociado a la experiencia laboral, se evidenció que un 10% del total de las ofertas recolectadas presentan el nivel de experiencia requerido para desempeñar el cargo y que puede ser identificado utilizando combinaciones de hasta 5 ngramas, por lo cual esta variable no podrá ser tenida en cuenta dentro del modelo a desarrollar, aunque había sido identificada como una variable relevante desde la revisión bibliográfica.

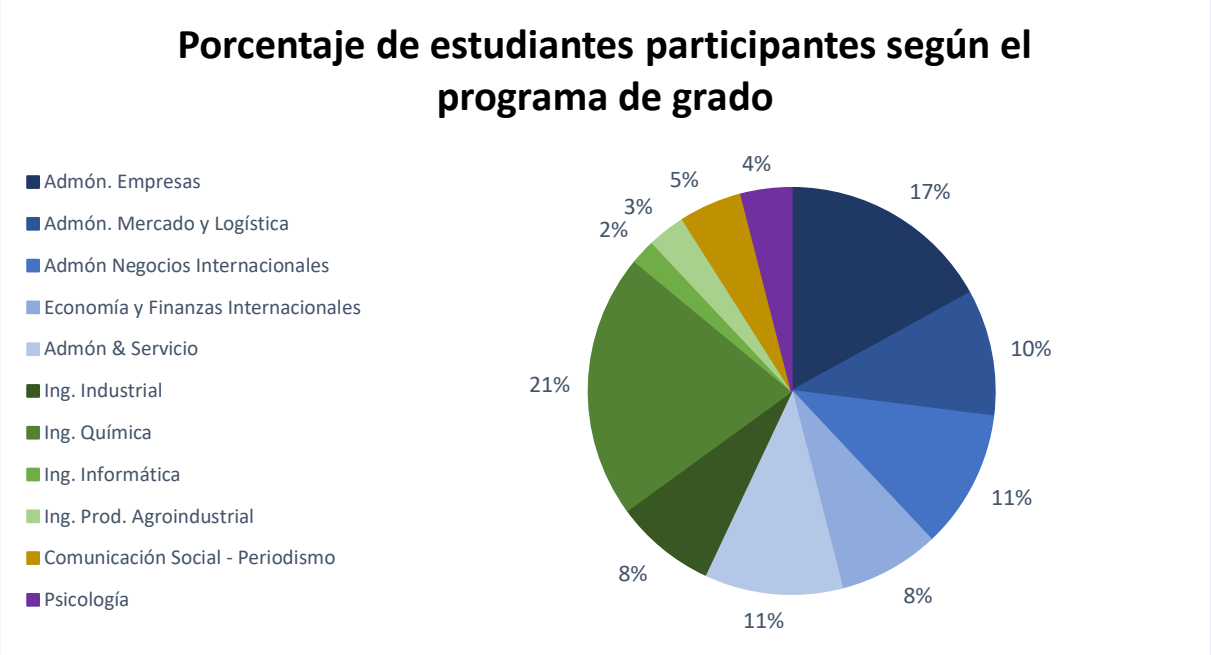
Desde la demanda laboral, se recolectaron 100 perfiles de graduados de la Universidad de La Sabana de las cohortes 2018 a 2021 de programas académicos de ingeniería, administración, economía, comunicación y psicología. La fuente de

recolección disponible, estándar y vigente seleccionada para este estudio fue el perfil profesional disponible en LinkedIn[7], teniendo en cuenta que en las otras plataformas de empleo no es posible acceder a los perfiles de cada persona, el uso de la información recolectada se hará bajo el marco de la política de privacidad de manejo de información de LinkedIn, por lo cual no será entregada o divulgada para ningún otro fin fuera del presente estudio.

La figura 3 permite observar la diversidad de los perfiles de los graduados según el programa académico (ver figura 3).

Figura 3

Porcentaje de estudiantes participantes según el tipo de programa de grado



Fuente: Realización propia.

La tabla 2 permite observar los 14 campos de información recolectada para cada uno de los perfiles. (ver tabla 2).

[7] Se envió a cada uno de los participantes la respectiva autorización para el manejo de información personal y la posibilidad de participar dentro del estudio en la fase de evaluación de resultados.

Tabla 2

Campos recolectados para cada perfil de LinkedIn

Nombre de la variable	Descripción	Fuente
ID	Código que identifica al estudiante.	Automático
Nombres y Apellidos	Nombres y apellidos del participante.	LinkedIn
Profesión	Programa de pregrado del que se graduó el participante.	LinkedIn
Contactar	Identificador Si o No dependiendo de si desea participar en la fase de evaluación de los resultados del modelo.	Formulario de encuesta enviado.
Año de grado	Identificador del año de grado	LinkedIn
Link LinkedIn	Enlace del perfil profesional del participante	LinkedIn
Cargo actual	Descripción del cargo actual que desempeña (si aplica).	LinkedIn
Empresa	Nombre de la empresa dónde desarrolla el participante su cargo actual (si aplica).	LinkedIn
Acerca de	Descripción general que cada participante da a su perfil (si aplica).	LinkedIn
Experiencia total	Tiempo de experiencia total que tiene el participante.	LinkedIn
Empleo	Cargos que ha desempeñado el participante con sus respectivas descripciones (si aplica)	LinkedIn
Educación	Educación formal que ha incluido el participante (si aplica)	LinkedIn
Cursos	Educación no formal que ha realizado el participante (si aplica).	LinkedIn
Aptitudes	Aptitudes que ha identificado el participante posee (si aplica)	LinkedIn

Fuente: Realización propia

Se destaca que dentro de estos perfiles las 50 palabras más frecuentes se encuentran asociadas a los diferentes empleos desempeñados, fechas y conocimientos, siendo de las habilidades o conocimientos más comunes los asociados a Microsoft, marketing, análisis, management, entre otros.

Respecto a la experiencia laboral, se evidenció que el 76% de la muestra recolectada presenta una experiencia igual o inferior a 3 años, pero teniendo en cuenta que esta variable fue descartada en las ofertas laborales, no podrá incluirse la información referente a la experiencia laboral. Frente a la expectativa salarial de cada uno de los participantes, no se encontró la información respectiva en la fuente de datos seleccionada, por lo cual esta variable no podrá ser tenida en cuenta dentro del desarrollo del modelo.

Durante el proceso de análisis exploratorio de la base de datos se evidenció que no todas las personas presentaban información completa del cargo actual, descripción, cursos realizados y aptitudes (ver tabla 3), los cuales son componentes que aportan valor durante el proceso de recomendación dado el enfoque desarrollado en habilidades y conocimientos. Una vez validado con la fuente de información se confirmó que efectivamente se encuentra vacíos y que debido a que no existe forma de imputación de datos, se realizará todo el proceso de análisis sobre las variables que poseen información en cada participante.

Tabla 3

Validación de completitud de información para las variables de interés

Disponible	Cargo actual	Empresa	Acerca de	Experiencia total	Empleo	Educación	Cursos	Aptitudes
Si	82	82	79	99	100	94	24	91
No	18	18	21	1	0	6	76	9
Total	100	100	100	100	100	100	100	100

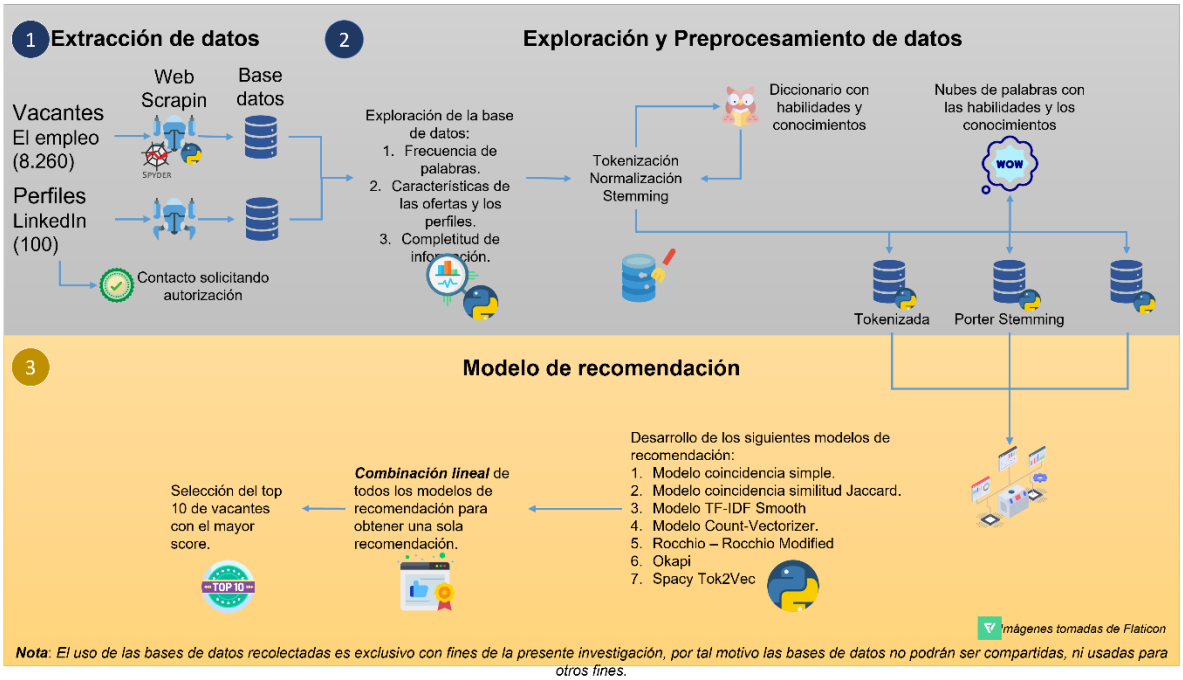
Fuente: Realización propia

3. Preprocesamiento de información y modelos de recuperación de información

El diagrama 2 recoge los procesos descritos y realizados durante los segmentos 1 y 2, adicionalmente se presentan los procesos de preprocesamiento de información, los modelos de recuperación de información que se utilizaron para el desarrollo del sistema y la metodología bajo la cual se seleccionaron las vacantes más relevantes (ver diagrama 2).

Diagrama 2

Proceso desarrollado para el preprocesamiento y desarrollo del sistema.



Fuente: Realización propia.

3.1. Preprocesamiento de información

El objetivo en este caso es extraer la información del contenido de la vacante enfocado a las habilidades y conocimientos requeridos. Para lograr el objetivo se procedió a realizar la limpieza de datos en dos etapas. La primera se encuentra compuesta por la tokenización y estandarización de la base de datos, en la cual se

retiran las palabras que dentro del contexto trabajado no agregan valor, se retiran las acentuaciones de las palabras y se dejan todas las palabras en minúsculas.

A partir de los resultados obtenidos en el paso descrito previamente, se realizó el análisis de los unigramas para generar un diccionario con la información de las habilidades y conocimientos que se estaban requiriendo alrededor de todas las vacantes recolectadas y para los cuales se tuvo en cuenta las habilidades core, las habilidades transferibles, las habilidades genéricas, las habilidades funcionales y las habilidades emprendedoras acorde al marco de referencia dado por Singh [5].

Este diccionario fue utilizado para la segunda etapa de limpieza, en la cual se dejaron solamente los monogramas que componen el diccionario desarrollado en cada una de las vacantes y se procedió a generar 3 bases de datos, la primera compuesta por las palabras o tokens en su forma original conforme a la propuesta realizada por Bañeres [14]; las otras dos bases se generaron a partir de Stemming, este es un proceso lingüístico que permite eliminar prefijos y sufijos hasta reducir las palabras a su origen raíz a partir de heurísticas, lo que permitiría aproximar mejor los resultados de cada una de las vacantes procesadas con cada uno de los perfiles procesados, ya que por ejemplo la palabra Ciclista y Ciclismo se esperaría queden reducidas al stem “cicli”, basados en esta premisa se generó la segunda base utilizando el Porter Stemming Algorithm y la tercera utilizando el Snowball Stemming.

La tabla 4 es un ejemplo del cambio que presenta la base de datos para una de las 5.763 vacantes una vez se ha aplicado el primer proceso, correspondiente a la tokenización, estandarización y el uso del diccionario generado, el cambio que presenta la base cuando se aplica el segundo proceso usando Porter Stemming Algorithm y finalmente el cambio que presenta la base cuando se aplica el tercer proceso usando Snowball Stemming (ver tabla 4).

Tabla 4

Diferencia entre los resultados obtenidos del preprocesamiento para una vacante.

Elemento	Descripción
Código de la oferta	1884831558
Descripción original de la oferta	Administrar Y Gestionar La Ejecución De Los Procesos Operativos Relacionados Con Control Y Aplicación De Las Operaciones Sobre Cuentas De Ahorro, Cats, Bolsillos, Embargos Y Demás Procesos Alineados A La Operación Del Pasivo, Con El Fin De Asegurar Su Correcto Cumplimiento Y Oportuna Ejecución. Requisitos: , Experiencia En Área Operativa (Control Contable, Cuadre Contable De Cdt´S, Cuentas De Ahorro, Cuentas Corrientes, Entre otros) Dentro Del Sector Financiero Superior A 1 Año. Grado En Administración, Economía, Contaduría, Ingeniería Industrial O Carrera Afines. Manejo Avanzado De Excel. Preferible Nivel Avanzado Access.
Descripción de la oferta Tokenizada	[‘access’, ‘administrar’, ‘ahorro’, ‘alineados’, ‘asegurar’, ‘carrera’, ‘cats’, ‘contable’, ‘control’, ‘corrientes’, ‘cuentas’, ‘embargos’, ‘excel’, ‘financiero’, ‘gestionar’, ‘industrial’, ‘operaciones’, ‘procesos’]
Descripción de la oferta Porter Stemming	[‘access’, ‘administrar’, ‘ahorro’, ‘alineado’, ‘aplicación’, ‘asegurar’, ‘carrera’, ‘cat’, ‘contabl’, ‘contaduria’, ‘control’, ‘corrient’, ‘cuenta’, ‘economia’, ‘ejecucion’, ‘embargo’, ‘excel’, ‘financiero’, ‘gestionar’, ‘industri’, ‘ingenieria’, ‘aplicación’, ‘operativo’, ‘pasivo’, ‘prefer’, ‘proceso’]
Descripción de la oferta con SnowBall Stemming	[‘access’, ‘administr’, ‘ahorr’, ‘alin’, ‘aplicación’, ‘asegur’, ‘carrer’, ‘cats’, ‘contabl’, ‘contaduri’, ‘control’, ‘correct’, ‘corrient’, ‘cuadr’, ‘cuent’, ‘economi’, ‘ejecucion’, ‘embarg’, ‘excel’, ‘financier’, ‘gestion’, ‘industrial’, ‘ingenieri’, ‘oper’, ‘pasiv’, ‘proces’, ‘relacion’]

Fuente: Realización propia

El resultado general obtenido del proceso de limpieza de tokenización se presenta en la figura 4, el principal motivo de utilizar dos tipos de Stemming diferentes se debe a que el algoritmo de Porter trabaja con heurísticas en inglés, mientras Snowball lo hace con heurísticas en español, los resultados de cada procesamiento se presentan en la figura 5 y figura 6 respectivamente (ver figuras 4, 5 y 6).

Uno de los primeros modelos que se desarrollará, buscará evidenciar si existe alguna mejora en la clasificación de las vacantes con el cambio de

preprocesamiento de la base de datos, lo anterior aportaría respecto al marco de referencia desarrollado por Bañeres [14], en caso tal que el preprocesamiento usando Stemming presente un mejor desempeño de emparejamiento que el obtenido por la tokenización.

Figura 4, Figura 5 y Figura 6

Figura 4: Nube de palabras tokenizadas para las ofertas laborales

Fuente: Realización propia.

Figura 5: Nube de palabras con Porter Stemming en las ofertas laborales

Fuente: Realización propia

Figura 6: Nube de palabras con Snowball Stemming en las ofertas laborales

Fuente: Realización propia.

Respecto a los perfiles, la etapa de limpieza de información se realizó de igual forma en dos partes, siendo la primera compuesta por la tokenización y estandarización de cada una de las variables recolectadas en la base de datos, retirando las palabras que dentro del contexto trabajado no agregan valor, se eliminan las acentuaciones de las palabras, se dejan todas las palabras en minúsculas,

adicionalmente se realiza la consolidación de todas las variables en una sola información agregada en la cual se identifican las palabras que actualmente no estaban en el diccionario desarrollado y bajo la propuesta de habilidades descrita por Singh [5] durante la revisión bibliográfica.

A partir de la revisión generada sobre cada uno de los unigramas, se realizó la realimentación del diccionario agregando los conocimientos y habilidades que no se encontraban. Una vez finalizado este proceso, se procedió a la segunda etapa de limpieza que correspondió a dejar únicamente las palabras que se encontraban disponibles en el diccionario construido y se procedió a generar 3 bases de datos: la primera compuesta por las palabras tokenizadas en su origen, la segunda utilizando el Porter Stemming Algorithm y la tercera utilizando el Snowball Stemming.

El resultado general obtenido de cada una de las tres bases de datos descritas previamente se presenta en las figuras 7, 8 y 9 respectivamente (ver figuras 7, 8 y 9).

Figura 7, Figura 8 y Figura 9

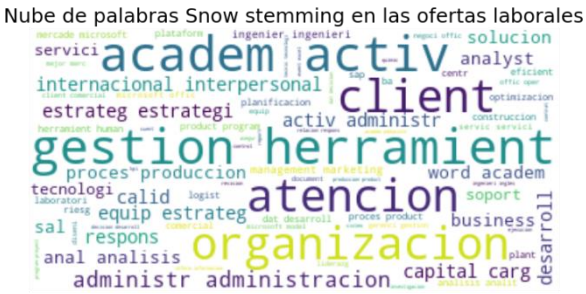
Figura 7: Nube de palabras tokenizadas para los perfiles

Fuente: Realización propia.

Figura 8: Nube de palabras con Porter Stemming en los perfiles

Fuente: Realización propia

Figura 9: Nube de palabras con Snowball Stemming en las ofertas laborales



Fuente: Realización propia.

La tabla 5 permite observar el cambio que presenta uno de los 100 perfiles tras el procesamiento descrito anteriormente (ver tabla 5).

Tabla 5

Diferencia entre los preprocesamientos realizados para uno de los perfiles

Elemento	Descripción
ID	5567
Fragmento perfil recogido (Acerca de - LinkedIn)	Profesional en Economía y Finanzas Internacionales de la Universidad de La Sabana, Con experiencia en desarrollo de analíticas estratégicas como elemento clave para la toma de decisiones, acercamiento a modelos de machine learning y el uso de herramientas de visualización como Power Bi, Tableau, herramientas de procesos ETL como Excel, SQL Server, R Studio, Python, y otras herramientas como DWH, AWS, Bloomberg y Reuters.
Fragmento Perfil profesional Tokenizado	['access', 'administrativas', 'administrativos', 'agile', 'analysis', 'analytics', 'aplicaciones', 'asertiva', 'auditorias', 'aws',…, 'control', 'controles', 'data', 'datos', …, 'descriptivas', 'descriptivos', 'dwh', …, 'estrategias', 'etl', 'excel', 'financial', 'finanzas', 'gastos', 'indicadores', 'informes', 'intelligence', 'interpersonales', 'learning', 'liderazgo', 'machine', 'management', 'mejora', 'microsoft', …, 'predictiva', 'presupuesto', 'presupuestos', 'proceso', 'procesos', 'python', 'query', 'r',..]
Perfil profesional Porter Stemming	['academica', 'access', 'acompaniamiento', 'administracion', 'administrativa', 'administrativo', 'agil', 'analisi', 'analitica', 'analysi', 'analyt', 'aplicacion', 'aprobado', 'asertiva', 'atencion', 'auditoria', 'automatizacion', 'aw', …, 'control', 'data', 'dato',

	…, 'descriptiva', 'descriptivo', 'disenio', 'dwh', …, 'estrategia', 'estrategica', 'estructuracion', 'etl', 'excel', 'financi', 'finanza', 'formacion', 'gasto', 'gestion', 'herramienta', 'indicador', 'inform', 'informacion', 'informatica', 'ingenieria', 'innovadora', 'institucion', 'insumo', 'intellig', 'interpersonal', 'learn', 'liderazgo', 'machin', 'macroeconomia', 'manag', 'mejora', 'microsoft', …, 'predictiva', 'presupuesto', 'proceso', 'publico', 'python', 'queri', 'r',…]
Perfil profesional SnowBall Stemming	['academ', 'access', 'acompani', 'activ', 'administr', 'administracion', 'agil', 'analisis', 'analit', 'analiz', 'analysis', 'analytics', 'aplic', 'aprob', 'asegur', 'asert', 'atencion', 'auditori', 'automatizacion', 'aws',…,'descript', 'diseni', 'dwh', …, 'estrateg', 'estrategi', 'estructuracion', 'etl', 'excel', 'financial', 'finanz', 'formacion', 'gast', 'gestion', 'herramient', 'indic', 'inform', 'informacion', 'informat', 'ingenieri', 'innov', 'insum', 'integr', 'intelligenc', 'internacional', 'interpersonal', 'learning', 'liderazg', 'machin', 'macroeconomi', 'management', 'manten', 'mejor', 'microsoft', … , 'predict', 'presupuest', 'proces', 'program', 'public', 'python', 'query', 'r',…]

Fuente: Realización propia.

3.2. Modelos de recuperación de información

Una vez realizado el preprocesamiento de información, el siguiente objetivo es el diseño de los modelos de recuperación de información con los que se busca establecer el rango de coincidencia existente entre cada una de las 5.763 vacantes y cada uno de los 100 perfiles profesionales, dónde será a través del sistema de recomendación que se establecerá el top 10 de las vacantes con mayor puntuación y bajo estas se realizará la evaluación de calidad del modelo.

Para evitar el sesgo dentro del desarrollo del sistema, se realizaron un total de 7 aproximaciones de modelos de recuperación de información algunos con variaciones en el preprocesamiento de información, otros en las métricas de evaluación de coincidencia y otros en la estructura del modelo, para los cuales se evalúo de forma individual las 576.300 combinaciones posibles entre la cantidad de vacantes y perfiles, agrupando el resultado obtenido en rangos de coincidencia de 10% y rangos de coincidencia de 20% para poder analizar la dispersión generada por cada aproximación. Los modelos, sus características, principales hallazgos y

nivel de incidencia sobre el sistema de recomendación final se presentan a continuación:

A. *Modelo base de coincidencia*

El primer modelo desarrollado tiene en cuenta el nivel de similitud existente entre cada una de las palabras que se identificaron en cada perfil y cada vacante. Para lograr obtener ese nivel se estableció que la métrica de similitud utilizada sería el número de habilidades y conocimientos que tienen en común el perfil con la vacante, dividido entre el total de requerimientos técnicos solicitados por la vacante, de esta forma cuanto mayor es el nivel de similitud más alto será el resultado obtenido.

$$Índice\ de\ coincidencia = \frac{Intersección\ de\ palabras\ entre\ perfil\ y\ vacante}{Total\ de\ requisitos\ que\ presenta\ la\ vacante}$$

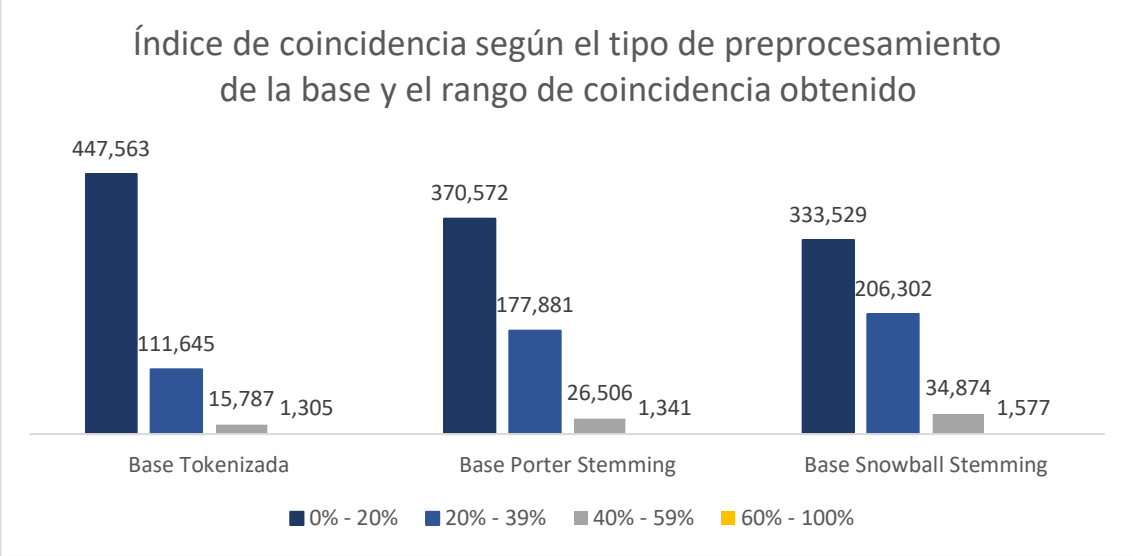
Teniendo en cuenta que se desarrollaron tres tipos de bases en el preprocesamiento, el diseño del modelo fue aplicado para cada una de ellas. Los resultados del índice de coincidencia agrupados por rangos de 20% se presentan en la figura 10 (ver figura 10).

Los resultados del modelo nos permiten observar que el cambio en el preprocesamiento de la base de datos si genera diferencias de entrada en la forma en cómo se distribuyen los índices de coincidencia, su principal explicación se encuentra en que la base Tokenizada contiene las palabras en su forma original, mientras que las bases con los estimizadores retiran las flexiones de las palabras, lo que aporta a obtener una mejor coincidencia de palabras de contextos similares pero en formas escritas distintas.

Teniendo en cuenta estos resultados, se observó que la base con Snowball Stemming presentó una mejor distribución en los distintos rangos evaluados en comparación con las otras dos bases, por lo cual los siguientes modelos serán desarrollados utilizando esta base de datos.

Figura 10

Índice de coincidencia simple entre bases procesadas



Fuente: Realización propia.

A partir de los resultados obtenidos en cuanto a la distribución de los índices de coincidencia entre vacantes y perfiles, las ventajas y desventajas presentadas para cada uno de los modelos frente al tipo de preprocesamiento, se determina que:

- El modelo de base tokenizada tendrá un nivel de relevancia **bajo** en el indicador global de coincidencia.
- El modelo con preprocesamiento de Porter Stemming tendrá un nivel de relevancia **bajo** en el indicador global de coincidencia.
- El modelo con preprocesamiento de Snowball Stemming tendrá un nivel de relevancia **medio** en el indicador global de coincidencia.

Estos niveles de relevancia serán de importancia para el sistema de recomendación que se presentará más adelante (ver sistema de recomendación)

B. *Modelo de coincidencia con similitud de Jaccard*

El segundo modelo desarrollado continúa con la idea de evaluar el nivel de coincidencia entre los perfiles y las vacantes a partir de las palabras que tienen en común ambas bases de datos. Para este caso se utiliza la métrica de similitud de Jaccard la cual toma como referencia la intersección de palabras entre cada perfil y vacante, sin embargo, el denominador en este caso corresponde a la unión entre los conocimientos del perfil y la vacante.

$$Índice\ coindiencia = \frac{Intersección\ de\ palabras\ perfil\ y\ vacante}{Total\ de\ palabras\ entre\ perfil\ y\ vacante}$$

Una de las desventajas absolutas del modelo es que, al realizar la unión entre los conocimientos del perfil y la vacante, el denominador se hace muy grande, por lo cual se obtiene que el 99% de los resultados se concentran en el rango de coincidencia del 0% - 20% y generando que el modelo no presente un buen nivel de dispersión.

A partir de la distribución obtenida, las ventajas y desventajas del modelo se estableció que el nivel de relevancia del modelo dentro del sistema es **muy bajo**.

C. *Modelo TF-IDF Smooth*

El modelo TF-IDF es un método que permite transformar las palabras en valores numéricos facilitando así el trabajo computacional para la identificación de elementos comunes entre los textos, para este caso se realiza el cálculo de la matriz TF (Term Frequency) y la matriz IDF (Inverse Document Frecuency), uno de los aspectos más importantes es que este modelo asigna pesos a las palabras, teniendo un peso mayor aquellas palabras que presenten una menor frecuencia entre todos los textos, la representación matemática del cálculo de similitud es:

$$Índice\ de\ coincidencia\ (d_j, q) = \frac{d_j * q}{|\ d_j| * |q|} = \frac{\sum_{t=1}^{|V|} w_{tf}\ x\ w_{tq}}{\sqrt{\sum_{t=1}^{|V|} W_{tj}^2}\ x\ \sqrt{\sum_{t=1}^{|V|} W_{tq}^2}}$$

Dónde el perfil que se está buscando se introduce en la variable q, y la validación se realiza sobre cada uno de los documentos que para este caso corresponde a las variables dj, para más información sobre el funcionamiento del modelo se recomienda revisar la documentación disponible en scikit learn [16].

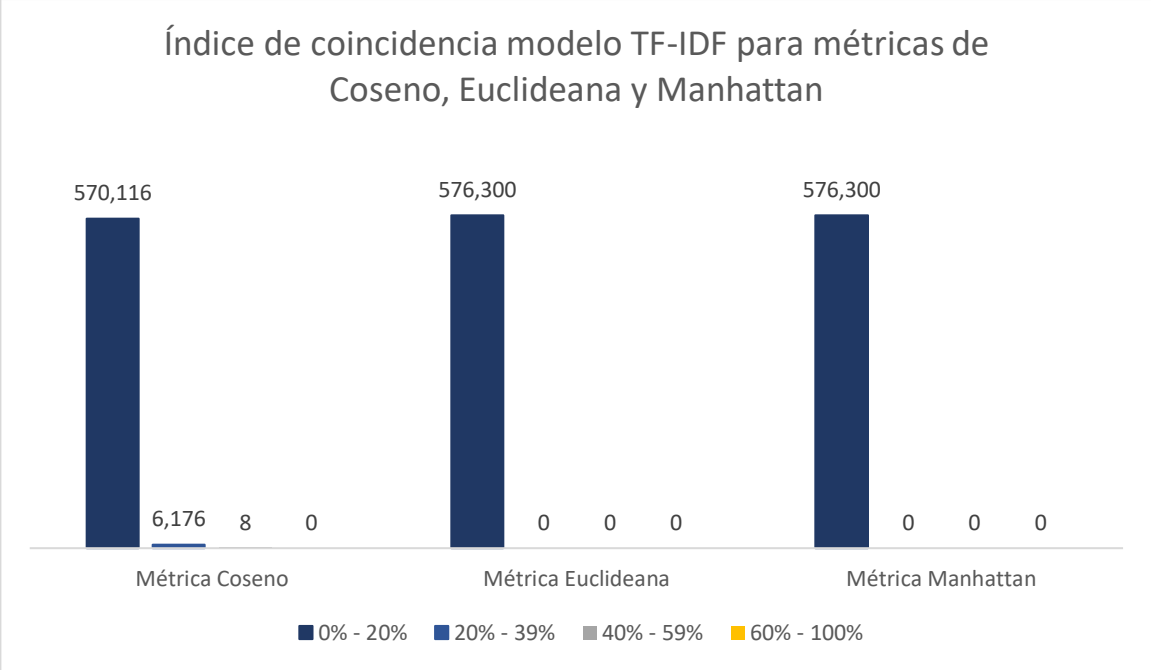
En este punto es importante destacar que dada la estructura de la base de datos, se utilizó el método smooth para mantener la importancia sobre aquellas palabras que aparecen por lo menos una sola vez en todos los documentos, lo anterior teniendo en cuenta que durante el preprocesamiento se retiraron las palabras duplicadas debido a que se buscada establecer los conocimientos y habilidades únicos que tenía la persona o que requería la vacante, de esta forma se busca que la estructura de la matriz tenga la posibilidad de tener el efecto diferenciador deseado en el momento de establecer el índice de coincidencia.

Para la evaluación del índice de coincidencia del modelo TF-IDF se decidió revisar tres métricas a fin de observar si se presentaban diferencias en la distribución de los resultados obtenidos. Las métricas seleccionadas fueron la similitud de coseno (la más recomendada desde la perspectiva teórica al medir la distancia de ángulos entre los vectores), la distancia euclideana (calcula la distancia total entre dos puntos) y la distancia de manhattan (calcula distancia absoluta entre dos puntos), los resultados obtenidos se presentan en la figura 11 (ver figura 11).

A partir de los resultados obtenidos, se destaca que la mejor métrica en aspectos de distribución, cuando el nivel de recomendación se agrupa en rangos de 20% es la similitud de coseno, explicado principalmente en que mide la distancia entre los ángulos de los vectores.

Figura 11

Índice de coincidencia modelo TF-IDF para distintas métricas



Fuente: Realización propia.

Uno de los principales inconvenientes asociados al modelo TF-IDF es la penalización de las palabras comunes entre las diferentes vacantes, lo cual puede generar que el resultado sea más orientado por los conocimientos o habilidades más exclusivas que se requieren, en vez de por un alto grado de coincidencia entre el perfil y la vacante.

De igual forma a nivel técnico es importante mencionar que en caso de ingresar una nueva vacante para el análisis, se debe realizar el cálculo completo de la matriz, ya que los pesos de la matriz podrían presentar variaciones al ingresar nuevos datos.

A partir de los resultados obtenidos en cuanto a la distribución de los índices de coincidencia entre vacantes y perfiles, las ventajas y desventajas presentadas para cada uno de los modelos frente al tipo de métrica utilizada para establecer el índice de coincidencia se determina que:

- El modelo TF-IDF con métrica de coseno tendrá un nivel de relevancia **bajo** en el indicador global de coincidencia.
- El modelo TF-IDF con métrica de euclideana tendrá un nivel de relevancia **muy bajo** en el indicador global de coincidencia.
- El modelo TF-IDF con métrica de manhattan tendrá un nivel de relevancia **muy bajo** en el indicador global de coincidencia.

D. *Modelo Count-Vectorizer*

El modelo Count-Vectorizer genera una matriz de vectores con dimensionalidad del diccionario que se establece, se asigna el valor a la matriz con base al total de apariciones que tiene la palabra evaluada del diccionario en el documento, y un valor de 0 en caso de que la palabra no se encuentre en el documento evaluado, para este caso el índice de coincidencia se establece a partir de la distancia que existe entre el vector generado por el listado de documentos y el vector del query que se busca.

Para más información sobre el funcionamiento y las características del modelo count-vectorizer se pueden consultar las siguientes fuentes de información [17].

Para nuestro caso y teniendo en cuenta que dentro del proceso de limpieza de datos se retiraron las palabras duplicadas dentro del diccionario, las vacantes y los perfiles dado que el objetivo principal es tener los conocimientos y habilidades, el desarrollo de la matriz Count-Vectorizer tiende a parecerse más a una matriz del tipo one-hot encoding, dónde el valor máximo que puede tener una palabra del diccionario es 1 cuando aparece en la vacante evaluada y 0 cuando la palabra no aparece.

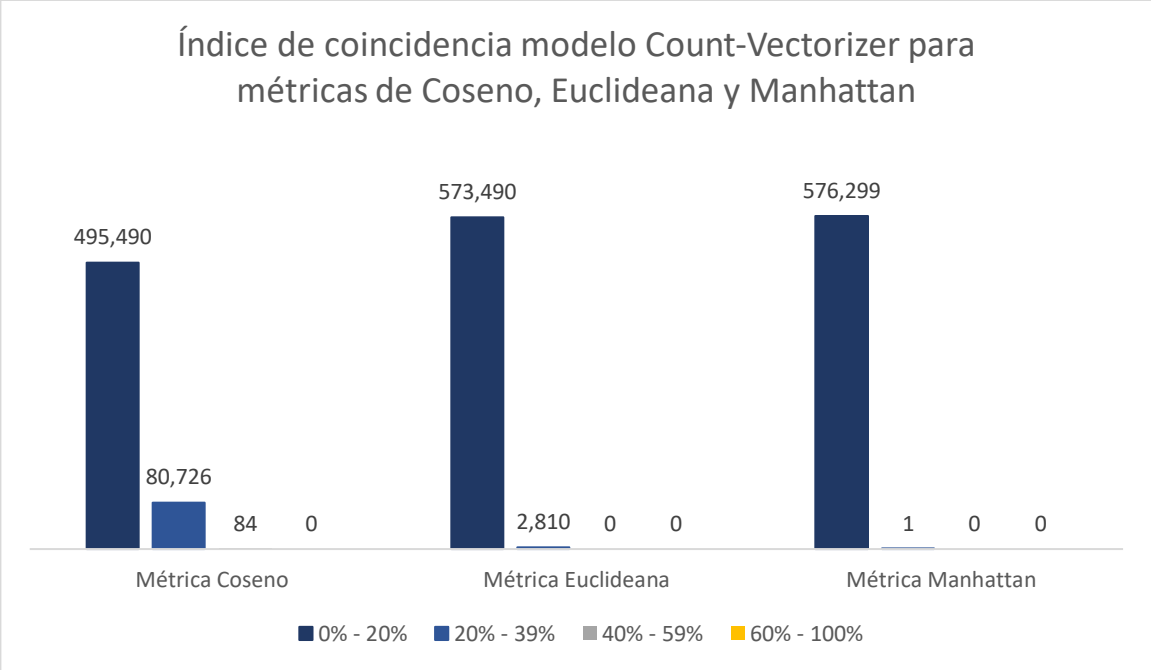
Es importante mencionar que una de las razones por la que se incluyó este modelo es buscando minimizar el impacto que se evidenció en la matriz TF-IDF, la cual penaliza las palabras con una mayor frecuencia de repetición a lo largo

de las vacantes, mientras que para este caso el modelo Count-Vectorizer asigna el mismo valor a cada una de las habilidades y conocimientos.

Al igual que el modelo anterior, se decidió evaluar su desempeño a partir de las tres métricas, lo anterior debido a que en términos de distribución no se observó una diferencia importante en el modelo anterior (modelo c), los resultados obtenidos de la evaluación del modelo se presentan en la figura 12 (ver figura 12).

Figura 12

Índice de coincidencia Count-Vectorizer para distintas métricas



Fuente: Realización propia.

En este caso se destaca que el nivel de dispersión respecto a la matriz TF-IDF mejoró al ubicar dentro del índice de coincidencia una mayor cantidad de combinaciones entre perfiles y vacantes con niveles de entre 40% - 59%, de igual forma que prevalece la métrica de coseno como la de mejor desempeño

respecto a las otras dos métricas y en este caso si se observó un cambio en la forma de distribuir los índices de coincidencia.

A partir de los resultados obtenidos en cuanto a la distribución de los índices de coincidencia entre vacantes y perfiles, las ventajas y desventajas presentadas para cada uno de los modelos frente al tipo de métrica utilizada para establecer el índice de coincidencia se determina que:

- El modelo Count-Vectorizer con métrica de coseno tendrá un nivel de relevancia **medio** en el indicador global de coincidencia.
- El modelo Count-Vectorizer con métrica de euclideana tendrá un nivel de relevancia **muy bajo** en el indicador global de coincidencia.
- El modelo Count-Vectorizer con métrica de manhattan tendrá un nivel de relevancia **muy bajo** en el indicador global de coincidencia.

E. *Modelo Rocchio – Rocchio Modified*

El algoritmo de Rocchio es un sistema de realimentación para el cual se puede determinar la cantidad de documentos relevantes y no relevantes, y a partir de estos parámetros se aproxima el vector Query (perfil) a los documentos más relevantes (vacantes). Este modelo toma como base la matriz TF-IDF y como métrica la similitud de coseno.

Para más información del funcionamiento y las características del modelo se recomienda consultar a Mandala [18].

Teniendo en cuenta los inconvenientes previamente discutidos de la matriz TF-IDF para este estudio en particular, se decidió realizar la misma estructura teórica del modelo respecto al sistema de realimentación, pero tomando como base la matriz del Count-Vectorizer a fin de determinar si se obtiene un mejor nivel de distribución del modelo con este ajuste, esta modificación a la estructura se presenta como Rocchio Modified.

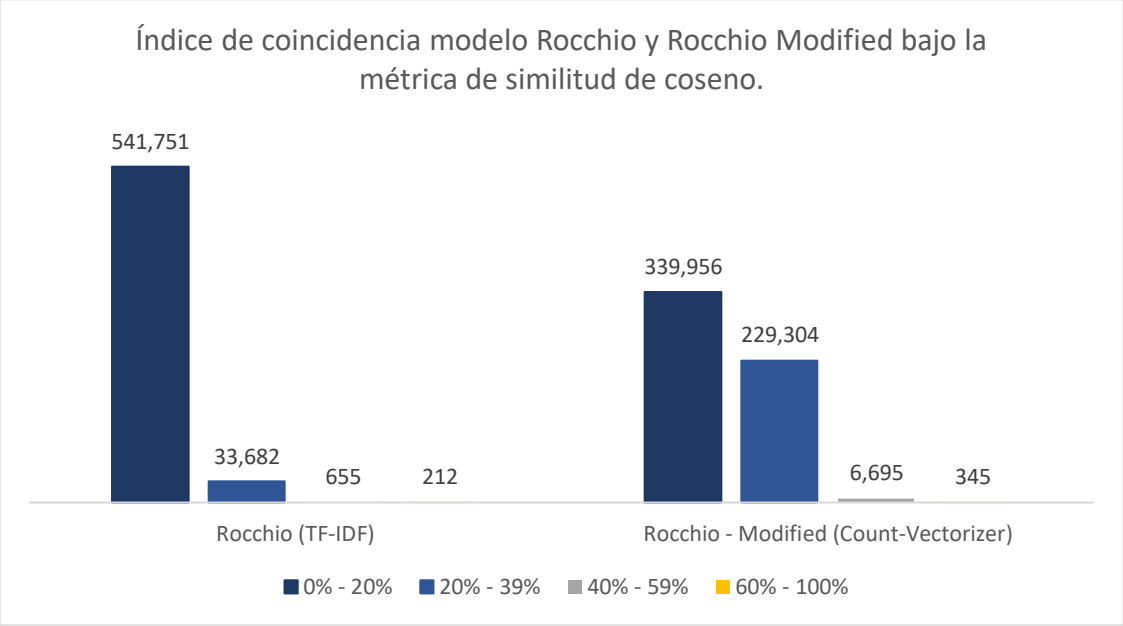
Se evaluaron ambos modelos con la métrica de coseno la cual durante los modelos previos fue la que presentó el mejor desempeño. Los resultados obtenidos se muestran en la figura 13 (ver figura 13).

Uno de los aspectos importantes a resaltar de este modelo, es que es el segundo después del índice de coincidencia base (modelo a) en generar una dispersión de resultados que alcanzan un rango de coincidencia de entre el 80% y 100%.

También se puede apreciar como el cambio de matriz base de TF-IDF a Count-Vectorizer para este caso genera una distribución diferente de los rangos de coincidencia cuando se mantienen el resto de los parámetros iguales, por lo cual se puede concluir que si existe un efecto generado por la penalización de la matriz TF-IDF a las palabras más comunes para el ejercicio que actualmente se está desarrollando.

Figura 13

Índice de coincidencia Rocchio – Rocchio Modified



Fuente: Realización propia.

A partir de los resultados obtenidos en cuanto a la distribución de los índices de coincidencia entre vacantes y perfiles, las ventajas y desventajas presentadas para cada uno de los modelos frente al tipo de métrica utilizada para establecer el índice de coincidencia se determina que:

- El modelo Rocchio (TF-IDF) tendrá un nivel de relevancia **alto** en el indicador global de coincidencia.
- El modelo Rocchio Modified (Count-vectorizer) tendrá un nivel de relevancia **muy alto** en el indicador global de coincidencia.

F. *Modelo Okapi*

El modelo Okapi toma como base el diccionario de palabras relevantes que se había creado previamente y a partir de la frecuencia de aparición de las palabras, el tamaño de las palabras y el tamaño del documento realiza el ranking de cada uno de los perfiles con las vacantes. En este punto es importante mencionar que este modelo toma como base la matriz TF-IDF.

$$Índce\ de\ coincidencia\ (D, Q) = \sum_{i=1}^{n} IDF\ (q_i) \frac{f(q_i, D\ ) * (k_1 + 1)}{f(q_i, D) + k1 * (1 - b + b * \frac{|D|}{avgdl})}$$

Dónde IDF es la matriz IDF, f(qi, D) representa la frecuencia de las palabras en el documento D, |D| representa el tamaño de palabras que contiene el documento, avgdl representa el promedio de tamaño de los documentos y k1 junto con b son parámetros de ajuste libre.

Para más información sobre el funcionamiento y las características del modelo okapi se recomienda consultar a Mohammad [19].

Dentro de los resultados más relevantes obtenidos del modelo es que ubica en el rango de coincidencia 60% - 79% un total de 476 combinaciones entre perfiles

y vacantes, de igual forma que en el rango anterior 40% - 59% ubica un total de 11.492 combinaciones, presentando por tanto niveles similares de distribución a los del modelo de índice de coincidencia base (modelo a).

A partir de la distribución obtenida, las ventajas y desventajas del modelo se estableció que el nivel de relevancia del modelo dentro del sistema es **muy alto**.

G. *Spacy Tok2Vec*

El último modelo desarrollado fue el Spacy Tok2Vec, pues tiene como una de sus mayores ventajas la extracción de información y una robusta arquitectura que le permite calcular similitudes, identificación de secuencias de tokens y serializaciones, adicional a que realiza procesos de encoding con valores hash lo que mejora el desempeño del modelo.

Para más información respecto al funcionamiento técnico del modelo se sugiere revisar la documentación oficial de spaCy [20].

Sin embargo, es importante reconocer que este modelo presenta una alta dependencia del orden en que se ingresa el vector, puesto que, aunque dos vectores sean iguales en intención comunicativa, pueden tener un bajo nivel de similitud producto de estar escritos en formas diferentes. Teniendo en cuenta lo anterior, se resalta que el modelo logró ubicar en el rango de 40% - 59% un total de 737 combinaciones entre perfiles y vacantes y en el rango anterior (20% - 39%) ubicó 531.564 coincidencias.

A partir de la distribución obtenida, las ventajas y desventajas del modelo se estableció que el nivel de relevancia del modelo dentro del sistema es **muy alto**.

Sistema de recomendación

Teniendo en cuenta que el objetivo principal es el diseño de un sistema de recomendación que permita obtener las vacantes que más se ajustan al perfil, se utilizará como insumo los resultados de los 7 modelos descritos previamente. Sin embargo, debido a que no es viable que cada persona evalué los 5.763 resultados en cada uno de los 14 posibles resultados[8], se realizará una combinación lineal de todos los resultados para cada perfil en cada vacante.

El nivel de influencia que tendrá cada modelo sobre el resultado final será determinado a partir de las ventajas y las desventajas que se han revisado a lo largo del documento, en conjunto con la distribución que ha generado en cada uno de los rangos, la tabla 6 presenta el consolidado del nivel de relevancia de cada modelo (ver tabla 6).

El indicador global de coincidencia será calculado para cada perfil en cada vacante de acuerdo con lo presentado en la fórmula 1 (ver fórmula 1), dónde el aporte de cada modelo al indicador global ha sido determinado conforme al nivel de relevancia que obtuvo a partir del análisis realizado (ver tabla 6). La conversión del nivel de relevancia al aporte dentro del indicador se presenta en la tabla 7 (ver tabla 7).

La tabla 8, presenta un ejemplo de cómo se visualizan los datos de recomendación obtenidos para cada perfil en cada vacante y cuál sería el valor del indicador global de coincidencia una vez se aplica la fórmula 1 (ver tabla 8).

[8] Son 7 modelos desarrollados pero algunos modelos presentan hasta 3 recomendaciones dependiendo del preprocesamiento de la base, las métricas utilizadas o las modificaciones realizadas a la estructura del modelo.

Tabla 6

Consolidado del nivel de relevancia de cada modelo.

Modelo	Muy bajo	Bajo	Medio	Alto	Muy Alto	Detalle
A Base Tokenizada		X				Ver modelo A
A Base Porter Stemming		X				Ver modelo A
A Base Snowball Stemming			X			Ver modelo A
B Jaccard	X					Ver modelo B
C TFIDF métrica coseno		X				Ver modelo C
C TFIDF métrica euclideana	X					Ver modelo C
C TFIDF distancia manhattan	X					Ver modelo C
D Count-Vectorizer métrica coseno			X			Ver modelo D
D Count-Vectorizer métrica euclideana	X					Ver modelo D
D Count-Vectorizer distancia Manhattan	X					Ver modelo D
E Rocchio				**X**		Ver modelo E
E Rocchio Modified					**X**	Ver modelo E
F Okapi					**X**	Ver modelo F
G Spacy Tok2Vec					**X**	Ver modelo G

Fuente: Realización propia.

Tabla 7

Equivalencia del nivel de relevancia y aporte al indicador global

Relevancia	Aporte al indicador global
Muy bajo	1.79%
Bajo	5.28%
Medio	7.14%
Alto	10.71%
Muy alto	14.29%

Fuente: Realización propia.

Fórmula 1

Índice de recomendación global del sistema de recomendación

$$Recomendación = M1_{B1} * 5{,}28\% + M1_{B2} * 5{,}28\% + M1_{B3} * 14{,}29\% + M2_{B3} * 1{,}79\% + M3_{B3Me1} * 5{,}28\% + M3_{B3Me2} * 1{,}79\% + M3_{B3Me3} * 1{,}79\% + M4_{B3Me1} * 7{,}14\% + M4_{B3Me2} * 1{,}79\% + M4_{B3Me3} * 1{,}79\% + M5_{B3T1} * 10{,}71\% + M5_{B3T2} * 14{,}29\% + M6_{B3} * 14{,}29\% + M7_{B3} * 14{,}29\%$$

Dónde:

$M_x = M$ *indica el Modelo y la* x *va de* 1 *a* 7 *conforme a los modelos descritos.*

$B_i = B$ *es la Base y la* i *señala el tipo de preprocesamiento que se realizó que va de* 1 *a* 3.

$Me_t = Me$ *indica la métrica y la* t *indica el tipo de métrica utilizada que va de* 1 *a* 3.

$T_m = T$ *indica el tipo de modelo y* m *indica el tipo de transformación realizada que va de 1 a 2.*

Tabla 8

Resultados obtenidos para una vacante y diferentes perfiles con el cálculo de indicador global de coincidencia (IG).

ID	Vacante	Modelos desarrollados														IG
		M1 B1	M1 B2	M1 B3	M2 B3	M3 B3Me1	M3 B3Me2	M3 B3Me3	M4 B3Me1	M4 B3Me2	M4 B3Me3	M5 B3T1	M5 B3T2	M6 B3	M7 B3	
5567	1884811242	11	15	21	2	2	2	58	5	8	12	25	5	12	20	14
136291	1884811242	33	23	21	4	9	3	27	12	10	8	11	10	11	19	15
2814	1884811242	33	31	28	5	12	2	39	14	10	8	14	7	16	22	18

Fuente: Realización propia.

Una vez aplicada la fórmula 1 a cada perfil y cada vacante, se obtendrán un total 5.763 puntuaciones para cada perfil, similar a la que se presentó en la tabla anterior (ver tabla 8), sin embargo, debido a que no es viable, ni óptimo que cada persona valide cada una de las vacantes evaluadas a través del sistema, se establecerá el top 10 de las vacantes que obtengan el mayor puntaje global en cada perfil y estos son los que se le recomendarán a cada persona.

Recomendación de habilidades o conocimientos por desarrollar

Teniendo en cuenta que para cada una de las vacantes y perfiles se hizo la extracción de las habilidades y conocimientos, a su vez que a través de la combinación lineal (ver fórmula 1, ver tabla 8) se obtuvo un índice de coincidencia para cada perfil en cada vacante, se decidió realizar el filtro de las 100 vacantes que obtuvieron el mayor puntaje de índice global de coincidencia.

A partir de las 100 vacantes se identificaron las habilidades o conocimientos únicos con su respectiva frecuencia, y a partir del cruce de información con el perfil se identificaron aquellos conocimientos requeridos por el mercado (comunes para las 100 vacantes) pero que no se encontraban de forma explícita en el perfil

recolectado, de esta forma se recomendaba desarrollar estas habilidades conforme al nivel de frecuencia observado en estas 100 vacantes.

4. Resultados

Previo a la presentación de los resultados y la evaluación del desempeño del sistema en las 10 vacantes recomendadas, se desarrolló el diagrama 3 que permite resumir el flujo que se realizó y que ha sido descrito a lo largo del texto, el cual agrega la forma en la que se evaluará el desempeño del sistema de recomendación desarrollado (ver diagrama 3).

Diagrama 3

Proceso desarrollado para el sistema de recomendación

Fuente: Realización Propia.

A partir de la combinación lineal, se realizó la evaluación de desempeño del modelo tomando el top 10 de las vacantes con mayor puntaje global para cada uno de los

perfiles. El proceso de evaluación se dividió en dos partes, las cuales se presentan a continuación:

a. Evaluación por parte de los participantes en la investigación

El proceso de evaluación por parte de los participantes en la investigación estuvo compuesto por tres secciones, la primera recoge la información respecto a los factores o elementos que tiene en cuenta en el momento de recibir o postularse a una oferta laboral, el segundo segmento contiene el top 10 de las vacantes que obtuvieron el mayor puntaje dentro del sistema y para los cuales el participante debía indicar si aplicaría o no a la vacante junto con los principales motivos que lo llevaban a tomar esa decisión, si aplica a la vacante recomendada se toma como un punto positivo por parte del modelo y la tercera parte presentaba una serie de habilidades y conocimientos que se habían detectado requería el mercado laboral pero que no se había evidenciado en el perfil.

Para el proceso de evaluación de los resultados del modelo se indagó a cada uno de los 100 participantes si deseaban participar de este proceso, recibiendo 17 respuestas positivas. A cada uno de estos participantes se envió el proceso de evaluación recibiendo solo 9 respuestas, los resultados obtenidos a cada una de las fases se presentan a continuación:

1. Fase - Elementos o factores que tiene en cuenta para recibir o postularse:
   El factor determinante en el momento de evaluar una propuesta laboral por parte de los participantes es el salario. Este aspecto es muy importante ya que no fue posible incluirlo dentro del modelo, pues en la fuente de datos no se tenía la expectativa salarial de los participantes. El segundo aspecto es el horario laboral, y el tercero es la forma en que se desarrollará el empleo (presencial, remoto o híbrido).

Todos estos factores podrían ser incluidos dentro del modelo para realizar un filtrado adicional sobre las vacantes que se muestran a cada persona, de forma tal que se encuentren alineadas al perfil profesional y a los gustos o intereses particulares de cada persona.

2. Fase – Evaluación Top 10 vacantes recomendadas:
   El resultado del modelo respecto a la aplicación directa en las 10 vacantes que más se ajustaban al perfil recolectado fue del 30%, es decir que en promedio una persona aplicaría a 3 de las 10 vacantes que recomienda el modelo de forma directa.

   El 70% restante junto con el análisis de las razones por las que no aplicaría a la vacante propuesta se presenta a continuación:

   a. 30% explicado en que las personas no cuentan con el nivel de experiencia que se solicita. Es importante recordar que este factor no fue incluido debido a que la mayoría de las vacantes recolectadas no contaban con esta información al realizar una combinación de hasta 5 ngramas para su extracción.

   b. 15% explicado en que la oferta propuesta no se encuentra en un campo de interés de la persona, aunque reconociendo que el perfil se ajusta para desempeñar este trabajo. En este punto es importante resaltar que en contacto posterior con las personas que se identificaron en este grupo, se concluyó que modificarían la hoja de vida, de forma tal que, la información allí presentada se ajuste mejor a sus intereses o gustos.

   c. 5% explicado en que el rango salarial ofrecido no concuerda con el nivel que está requiriendo el empleo, con la expectativa salarial o con

el salario de reserva de la persona para desempeñar esas funciones, un hallazgo completamente alineado a la fase de evaluación previa, dónde el salario es uno de los factores claves en el momento de evaluar una vacante de empleo.

d. 2% indicó que la oferta laboral no corresponde con los conocimientos que tiene la persona y por tanto se puede asumir como error del modelo.

e. El 18% restante identificó que dentro de los factores para no aplicar se encontraban alguno de los siguientes: aunque cuenta con algunos de los conocimientos requeridos en la vacante, les hace falta trabajar en más conocimientos que consideran muy importantes para poder desempeñarse en el cargo; que en algunas vacantes no existe la suficiente información para tomar una decisión; o que el horario o la modalidad de trabajo no se ajusta a lo esperado o deseado.

3. Fase – Recomendación de habilidades a desarrollar:
   Se encontró que del total de habilidades recomendadas a desarrollar o incluir en los perfiles, el 57% de las habilidades, las personas ya cuentan con ellas o actualmente se encuentran en proceso de adquirirlas ya que las habían identificado como necesarias para el campo en que se desempeñan.
   Por otro lado, el 21% de las habilidades restantes están en campos de conocimientos que no son de interés de las personas y el restante 22% son habilidades que actualmente las personas no tenían en mente que fueran requeridas.

b. Evaluación por parte de los expertos en áreas de selección.

La evaluación parte de los expertos en áreas de selección y contratación estuvo enfocada en conocer si basado en las 10 ofertas laborales recomendadas llamarían al perfil respectivo para iniciar un proceso de selección por la vacante, se asume como punto positivo cuando el evaluador responde que sí lo llamaría a entrevista, para ello se contactaron a 4 expertos en áreas de selección y contratación, de los cuales respondieron 2, los resultados obtenidos fueron:

1. El modelo tiene un nivel de acierto promedio del 47%, sin embargo, durante la revisión de la distribución sobre los perfiles que se encontraban por debajo del promedio, se encontró que el modelo presentaba un mayor índice de confusión en los que perfiles que tenían una menor cantidad de información, validando por tanto uno de los hallazgos encontrados durante la revisión bibliográfica, en el que se destaca que este tipo de modelos presenta una alta dependencia a la calidad de la información obtenida para cada perfil [10].

2. Respecto a los principales motivos de error presentados por el modelo, se encontró que el 60% estaban asociados a que el perfil no se ajusta con **todos** los requerimientos de la oferta laboral, el 21% que la persona no cuenta con la experiencia solicitada y el restante 19% que la persona no cumple con el nivel de formación requerido o la persona se encuentra sobre calificada para los aspectos que solicita la vacante.

5. Conclusiones

El sistema de recomendación de empleo propuesto permite la evaluación de múltiples vacantes y entrega el top 10 de vacantes que acorde a la metodología utilizada obtienen el mayor puntaje global de similitud más alto entre la vacante y el perfil, donde es muy importante reconocer el alto nivel de dependencia con el listado de palabras en términos de conocimientos y habilidades que fue construido.

De igual forma se destaca que el modelo alcanzó un 30% de predicción acertada en la evaluación por parte de los participantes, pero que su desempeño podría alcanzar un 65% o más si se incluyen las variables de nivel de experiencia requerida, las variables salariales y los componentes de gustos personales, variables que en su mayoría ya habían sido detectadas como importantes desde la parte teórica y fueron reforzadas durante el proceso de evaluación del modelo, pero no fue posible incluirlas debido a que no se contaba con la información.

Se identificó que las habilidades recomendadas a desarrollar se encuentran bastante alineadas al sector dónde se desempeñan las personas, sin embargo, también llama la atención que existen una serie de habilidades que, aunque están alineadas al perfil de la persona, no son de interés de esta, caso muy similar a lo encontrado en las vacantes recomendadas.

Finalmente, desde la evaluación de expertos se analizó si desde su perspectiva, basados en la descripción de la oferta laboral y en el perfil respectivo, llamaría a la persona para iniciar un proceso de selección, obteniendo que el modelo tiene un nivel de acierto del 47%.

Dentro de los factores que afectaron el desempeño del modelo se encontró que el nivel de experiencia especifica requerida en la vacante no era igual al que tenía la persona, también que el nivel de formación mínimo requerido no era igual al máximo nivel de formación de la persona, e incluso se evidenció que en algunos casos se presentaba una sobre calificación del perfil respecto a la vacante propuesta.

6. Trabajo futuro

A futuro se puede trabajar en un sistema de recomendación contextual, basado en el marco presentado por Singh, el cual puede incluir un componente de realimentación respecto a los intereses de cada persona en términos de habilidades, conocimientos, horarios, rangos salariales e incluso sectores o empresas de interés [10], variables que durante el proceso de evaluación se identificaron como importantes y que podrían aportar a mejorar el desempeño.

Respecto a la perspectiva del nivel de expertos se evidenció nuevamente la importancia de validar la experiencia profesional específica, y de agregar los componentes asociados al nivel de formación de la persona respecto a lo solicitado en la vacante para mejorar el desempeño del modelo, de igual forma basado en los hallazgos de Singh se puede considerar el desarrollar un sistema que tenga en cuenta un peso diferente en cada una de las habilidades dando prioridad a aquellas que son de primera necesidad tanto para la empresa como para la función que desarrollará la persona [5].

Finalmente se podría evaluar si con estas modificaciones el modelo podría ser escalable a perfiles con una mayor trayectoria profesional, e incluso se podría considerar realizar la misma evaluación sobre estudiantes en proceso formativo para validar los posibles cargos a los que podrían aplicar con los conocimientos adquiridos hasta la fecha.

7. Bibliografía y recursos

7.1. Bibliografía

1. DANE. (2020). Boletín Técnico Mercado laboral de la Juventud Trimestre móvil junio - agosto 2020 Boletín Técnico. 1–14. https://www.dane.gov.co/files/investigaciones/boletines/ech/juventud/Bol_eje_juventud_jun20_ago20.pdf
2. Sistema Nacional de Información de la Educación Superior - SNIES. (2019). SNIES. Obtenido de Ministerio de Educación Nacional - MEN: https://hecaa.mineducacion.gov.co/consultaspublicas/content/poblacional/index.jsf
3. Moraes, M. A., & de Mesquita, M. A. (2022). Industry 4.0: a tertiary literature review . Technological Forecasting & Social Change, 11.
4. Pal'Ova, D., & Vejacka, M. (2013). On-line E-learning platform supporting education and practice of accountants in EU space. 2013 36th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2013 - Proceedings, January 2013, 641–646.
5. Antra Singh, S. S. (2021). Do Employability Skills Matter in Placement: An Exploratory Study of Private Engineering Institutions and IT Firms in Delhi NCR. The Indian Journal of Labour Economics, 21.
6. Potter, T. (2020). Learning and job search dynamics during the great recession. Journal of Monetary Economics, 117, 706–722. https://doi.org/10.1016/j.jmoneco.2020.04.006
7. Nazareno Luis, S. D. (2021). The impact of automation and artificial intelligence on worker well-being. Elsevier, 24.
8. Di Pace, F., Mitra, K., & Zhang, S. (2021). Adaptive Learning and Labor Market Dynamics. Journal of Money, Credit and Banking, 00(00). https://doi.org/10.1111/jmcb.12764
9. Itsakov, E., Kazantsev, N., Yangutova, S., Torshin, D., & Alchykava, M. (2019). Digital Economy: Unemployment Risks and New Opportunities. Communications in Computer and Information Science, 1038 CCIS(January 2020), 292–299. https://doi.org/10.1007/978-3-030-37858-5_24
10. Pradeep Kumar Singh*, P. K. (2021). Recommender systems: an overview, research. 39.

11. Gorb, A. (2021). Text-based Recommendation Systems for Software. Journal of Physics: Conference Series , 9.
12. Malek Alksasbeh, T. A.-k. (2021). Smart job searching system based on information retrieval techniques and similarity of fuzzy parameterized sets. International Journal of Electrical and Computer Engineering (IJECE), 10.
13. Anna Giabelli, L. M. (2021). Skills2Job: A recommender system that encodes job offer embeddings. ELSEVIER, 10.
14. Bañeres, D., & Conesa, J. (2017). A Life-long learning recommender system to Promote Employability. International Journal of Emerging Technologies in Learning, 12(6), 77–93. https://doi.org/10.3991/ijet.v12i06.7166
15. Oihab Allal-Chérif, A. Y. (2021). Intelligent recruitment: How to identify, select, and retain talents from. ELSEVIER, 11.
16. Scikit, L. (2022). Scikit Learn TFIDF. Obtenido de Scikit Learn TFIDF: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html
17. Scikit, L. (2022). Scikit Learn Count Vectorizer. Obtenido de Scikit Learn: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html
18. Elvina, & Mandala, R. (2020). Improving Effectiveness Information Retrieval System Using Pseudo Irrelevance Feedback. IEEE International Conference on Sustainable Engineering and Creative Computing (ICSECC), 463-468.
19. Mohammad Ubaidullah Bokhari, M. K. (2020). Evaluation of News Search Engines Based On Information. Springer, 22.
20. spaCy. (s.f.). spaCy v3.4. Obtenido de spaCy v3.4: https://spacy.io/models, https://spacy.io/usage/spacy-101
21. Vanbaelen, R., Harrison, J., & Van Dongen, G. (2015). Lifelong learning in a Fourth World setting. IEEE International Professional Communication Conference, 2015-Janua. https://doi.org/10.1109/IPCC.2014.7020347

7.2. Diagramas

1. Proceso desarrollado para la extracción y exploración de datos.
2. Proceso desarrollado para el preprocesamiento y desarrollo del sistema.
3. Proceso desarrollado para el sistema de recomendación.

7.3. Figuras

7.4. Tablas:

7.5. Fórmulas: