Relación entre habilidades de empleados y el sector industrial de las empresas en Ecuador aplicando NPL

Juan Francisco Nebel Dunn   
ESPOLGuayaquil, Ecuador  
jfnebel@espol.edu.ec

Danny Fabricio Ucho Morocho   
ESPOLCuenca, Ecuador  
ducho@espol.edu.ec

Andres Xavier Vargas Vera  
ESPOLGuayaquil, Ecuador  
axvargas@espol.edu.ec

*Abstract*— Cada empresa según su actividad económica es parte de un sector industrial y a su vez el conjunto de empleados de estas empresas posee habilidades que les permiten una ventaja competitiva frente a otros empleados o postulantes. En muchos casos los conjuntos de habilidades de empleados de una empresa particular pueden no ser propios de su actividad económica por esto, mediante una técnica de ponderación(tf-idf) de habilidades determinamos que grupos de habilidades son transversales y que conjunto son característicos para cada grupo de industria en el ecuador.

Keywords—habilidades, tf-idf, industria

# Introduction (*Heading 1*)

La dinámica de contratación laboral en el Ecuador está en un cambio constante debido a las exigencias del mercado, por esta razón las empresas buscan mejorar sus estrategias de contratación a postulantes que mejor se adapten a sus necesidades y para esto se toman en cuenta varios factores entre estos las habilidades que tienen sus empleados.

En el mercado laboral ecuatoriano el conjunto de empresas según su actividad económica es diverso sin embargo muchas de las habilidades que reportan los empleados en plataformas como LinkedIn son habilidades que forman parte de un grupo transversal de habilidades y solo un grupo específico son las habilidades características propias del tipo de industria. En este informe se establece que conjuntos de habilidades son propias de un sector industrial mediante la adaptación de una técnica de ponderación tf-idf que es utilizada para el análisis de texto.

En nuestro análisis empezamos con determinar que empresas son las más relevantes en el mercado, así como el tipo de industria al que pertenecen, luego con datos de una red social de empleo capturamos los datos de habilidades de empleados para establecer la relación entre el tipo de industria y las habilidades[1].

# Dataset

## Dataset de Empresas

Este dataset este compuesto de registros de la Súper Intendencia de Compañías ecuatoriana (SUPERCIAS). El dataset cuenta con un total de 79043 registros. Cada registro representa una compañía almacenada y registrada en el organismo gubernamental.

### Estrategia de datos faltantes

El dataset cuenta con 3878 entradas con información incompleta. Un patrón detectado en las entradas incompletas es que les faltan casi todos los campos (no solo uno o dos campos). Esto genera un problema para trabajar con estos datos particularmente en columnas de tipo categóricas como “TAMAÑO” o “REGIÓN”.

Ya que preferimos no colocar un reemplazo como el promedio o moda para registros con datos incompletos continuos o categóricos, y ya que estos representan poco menos del 5% de los registros totales, elegimos descartar registros incompletos.

### Estructura del dataset original

1. Table Type Styles

| COLUMNA | DESCRIPCIÓN |
| --- | --- |
| 2020 | Posición en el ranking en el 2020. |
| 2019 | Posición en el ranking en el 2019. |
| EXPEDIENTE | Número de expediente. |
| TIPO DE COMPAÑÍA | Figura jurídica que ampara a la organización y asigna responsabilidades sociales, tributarias, etc. |
| ACTIVIDAD ECONÓMICA | Descripción de actividades comerciales. |
| REGION | Región ecuatoriana donde se encuentra la sede. |
| PROVINCIA | Provincia ecuatoriana donde se encuentra la sede. |
| CIUDAD | Ciudad ecuatoriana donde se encuentra la sede |
| TAMAÑO | Categoría brindada por la SUPERCIAS. |
| SECTOR | Corresponde a su apertura de acciones en el mercado ecuatoriano. |
| CANT. EMPLEADOS | Número de empleados trabajando para la compañía. |
| ACTIVOS2020 | Activo total 2020. |
| PATRIMONIO2020 | Patrimonio total 2020. |
| INGRESO POR VENTA2020 | Ingresos por venta 2020. |
| UTILIDAD ANTES DE 2020 | Utilidades previas al año del dataset. |
| UTILIDAD DEL EJERCICIO2020 | Utilidades producto de la labor realizada en el 2020. |
| UTIDLIDAD NETA2020 | Utilidad neta 2020. |
| IR CAUSADO | Impuesto a la renta causado 2020. |
| INGRESO TOTAL 2020 | Ingreso total 2020. |

1. Tabla descriptiva del dataset central.

### Overview del dataset empresas

Al explorar la información del dataset centra, se puede vislumbrar desde temprano ciertas características interesantes entre los registros.

El dataset tiene un total de 79043 entradas cuando este no es sujeto a ninguna alteración. Luego de la eliminación de registros vacíos se obtiene un dataset con un total de 75165 entradas y 20 columnas.

Como factores importantes iniciales es necesario notar que los valores únicos para la columna “SECTOR” son: ‘MERCADO DE VALORES’, ‘SOCIETARO’.

Esto indica que todas las compañías del Ecuador tienen acciones abiertas a la bolsa de valores nacional y cerradas dentro de sociedades contractuales.

En cuanto a las categorías que la SUPERCIAS determina para las empresas, tenemos como valores únicos: “GRANDE”, “MEDIANA”, MICROEMPRESA”, “PEQUEÑA”. Respecto a las proporciones de estas tenemos que:

| TAMAÑO | VALOR |
| --- | --- |
| MICROEMPRESA | 47,787 (63.6%) |
| PEQUEÑA | 18,666 (24.8%) |
| MEDIANA | 6,124 (8.1%) |
| GRANDE | 2,588 (3.4%) |

1. Número de registros de empresas

Las microempresas superan al resto con un total de 47,787 (63.6%), seguido los registros de empresas pequeñas 18,666 (24.7%), luego vienen las empresas medianas con 6,124 (8.1%) de registros y finalmente las empresas grandes con un total 2,588 (3.4%).

Sería un error pensar que simplemente por cantidad de microempresas el impacto en la fuerza laboral de las empresas grandes no es significativo. Después de todo, las microempresas tienen facilidades de ser un nuevo entrante en el mercado. Por otro lado, las grandes empresas han logrado hasta cierto punto triunfar y crecer su emprendimiento; generalmente contratan a más personas que una microempresa por la diferencia de tamaño que existe entre los registros de ambas categorías.

Para corroborar esto, es propio visualizar la proporción de empleados que representan el top 1000 empresas del país (según el ranking oficial de SUPERCIAS 2020). Si comparamos el número de empleados que registran estas empresas con el número de empleados que registran el resto de las compañías ecuatorianas tenemos lo siguiente:

Se tiene que las top 1000 compañías (1.3%) representan desproporcionadamente el 23.9% de toda la fuerza laboral registrada en la SUPERCIAS.

Otra característica importante de mencionares, en el

dataset, 62,178 (82.72%) de todas las compañías

ecuatorianas tiene menos de 10 trabajadores. Esto

es congruente con que la vasta mayoría de empresas

son o pequeñas o microempresas (un total del

88.4% entre las dos categorías).

## Dataset LinkedIn

Siendo LinkedIn una de las redes sociales más relevantes para en el proceso de promoción de empleos, así como perfiles de postulantes se obtuvo el siguiente conjunto de datasets mediante Scraping.

* Habilidades de empleados.
* Lugar de estudio de empleados.
* Lugar donde viven los empleados.
* Estudios realizados por los empleados.
* Actividad que desempaña en la empresa.
* About de la compañía.

Es necesario aclarar que del total de compañías del dataset anterior únicamente se han tomado 438 empresas representativas para extraer la información.

# Dataset Habilidaes

Para este dataset cada fila corresponde a una compañía y cada columna según corresponda a una habilidad reportada por un empleado de la empresa.

## Estructura del Dataset

Este dataset tiene estructura de 438 filas que representan las empresas y 381 columnas que representan las habilidades, el valor de las celdas de este dataset es la frecuencia que se repite la habilidad en la empresa.

# Dataset About compañía

Para este dataset cada fila corresponde a una compañía y cada columna corresponde a una característica de la empresa (tipo de industria, localidad, cantidad de empleados)

## Estructura del Dtaset

Este dataset está estructurado con 438 filas y 5 columnas, cada columna posee información de las empresas de este dataset el campo importante para nuestro análisis es el tipo de industria al cual pertenece la compañía por lo tanto los valores nulos obtenidos en este dataset son

# Metodología

## Empresas

Para el dataset de empresas se encontró que existen varios *outliers*, y no se desea eliminar a los *outliers* ya que cada empresa de este dataset puede ser representativa para buscar su información de empleados, por este motivo se opta por dividir a las empresas en cuartiles. Esto demuestra algo importante en lo que a la varianza de datos respecta:

| TAMAÑO | VALOR |
| --- | --- |
| MICROEMPRESA | 0.810 |
| PEQUEÑA | 0.476 |
| MEDIANA | 0.500 |
| GRANDE | 1063.884 |

1. Varianza de empleados por cuartil.

Se pude apreciar que casi toda la varianza se encuentra en el último cuartil. El resto de los cuartiles muestra valores particularmente uniformes. Se debe revisar qué significa esto en cuanto a cantidad de empleados.

Se analiza la cantidad de empleados que las compañías tienen de acuerdo con su región nacional en el primer cuartil. El primer cuartil muestra que el 88.02% (34,987) de los registros de este cuartil tienen entre 3 y 4 empleados. 4249 (10.69%) cuentan con 2 empleados y 513 (1.29%).

El segundo cuartil tiene de 5 a 6 empleados: 34.8% (4872) tienen 5, 65.2% (9129) tienen 6. Esto indiferentemente de la región nacional a la que pertenecen.

En el tercer cuartil ocurre lo mismo entre 7 u 8 empleados. Lo que significa que no se encuentra nada significativo ni en la distribución de valores y en la división por provincia.

En el último cuartil la gran mayoría de registros (13,200, 90.11%) tienen menos de 100 empleados, mientras que el tope del último cuartil se dispara en cuanto a la cantidad de empleados que tienen.

Este análisis nos permite determinar que si orientamos nuestro análisis a obtener la información de las empresas que están alrededor del top 1000 son representativas en el mercado laboral por la cantidad de empleados que estas poseen.

## Habilidades

En el análisis del dataset de habilidades encontramos que el total del conjunto de empresas el 7.53% poseen un único valor que hace referencia a una habilidad reportada generando un outliers en la distribución de habilidades por lo cual estos registros son descartados de nuestro análisis teniendo al final un total de 405 empresas.

Del conjunto de habilidades encontramos que un grupo marginales 20% reportan un único valor en la distribución por cada habilidad, mientras que existen un conjunto de habilidades 28% del total que reportan valores superiores a los 1000 y alcanzan valores hasta los 25000, por lo tanto un análisis de frecuencia no es ideal para determinar la relación entre las habilidades e industrias por lo tanto como método para evaluar la relación se utiliza tf-idf que da una ponderación a una palabra en un documento en nuestro caso será la ponderación de una habilidad en un grupo de industrias.

## Métodos

1. Procesamiento de Lenguaje Natural (NPL)

Utilizamos una técnica de NLP (Natural Language Processing) con el fin de identificar qué habilidades son características y qué habilidades son transversales de acuerdo con la industria en base a lo que hemos recopilado de LinkedIn. En este caso se trata de Term Frequency - Inverse Document Frecuency (TF-IDF).

TF-IDF consiste en ver qué tan frecuentemente ocurre un término en un documento (TF) y al mismo tiempo contraponerlo a si ha sido usado en otros documentos también (IDF).

Mientras más ocurra un término en un documento y menos se lo pueda encontrar en otros documentos mayor será su valor de TF-IDF. Importante recalcar, si un término se encuentra en todos los documentos, su TF-IDF será 0, ya que este término no es nada característico para un documento en específico (es completamente transversal).

La única diferencia en nuestro acercamiento a la técnica es que nosotros mapeamos términos a habilidades (skills) y documentos a industrias, no obstante, el problema es el mismo a resolver. Queremos saber qué habilidades son características de una industria y cuales son habilidades transversales (aparecen de manera nominal a través de todas las industrias).

1. Métricas
2. Tf

TF(Term Frequency) consiste en ver qué tan frecuentemente ocurre un término en un documento (TF). En nuestro análisis el valor de frecuencia ya representa un valor de tf sin embargo debido a la cantidad de habilidades que se reportan es necesario normalizar este valor por lo que la formula aplicada queda de la siguiente manera.

*tf= f/sum(f)*

1. IDF

IDF(Inverse Document Frecuency) permite determinar la inversa de la frecuencia de un término en un documento esto hace que a pesar de que un término aparezca en varios documentos no signifique que sea un valor que irrelevante en el conjunto de documentos en nuestro caso utilizamos la siguiente formula.

*idf= log(d/n)*

donde d es el número de industrias y n el número de industrias donde aparece esa habilidad.

# Resulados

## Resultados Habilidades Transversales

El resultado de aplicar tf-idf a nuestra matriz de frecuencia de habilidades versus industrias es una nueva matriz cuyo valor por cada celda representa la caracterización de una habilidad para la industria.

En primer lugar, a partir de la matriz idf cuyas dimensiones son 1xn donde n es el numero de habilidades podemos determinar que habilidades con transversales para nuestras industrias, se muestra un gráfico del resultado obtenido:

Gráfico, Gráfico radial

Descripción generada automáticamente

Ilustración 1 Habilidades Trasversales

|  |  |
| --- | --- |
| Habilidad | IDF value |
| Microsoft Excel | 0.033336 |
| Microsoft Office | 0.050431 |
| Strategic Planning | 0.067823 |
| Customer Service | 0.085522 |
| Microsoft Word | 0.103541 |
| Team Leadership | 0.103541 |
| Negotiation | 0.121890 |
| Project Management | 0.121890 |
| English | 0.179048 |
| Business Strategy | 0.198851 |
| Microsoft PowerPoint | 0.219054 |
| Management | 0.219054 |
| Leadership | 0.239673 |
| Teamwork | 0.260726 |
| Sales | 0.326684 |
| Business Management | 0.527355 |
| Research | 0.852777 |
| Marketing | 0.852777 |
| Analytical Skills | 0.852777 |
| Communication | 1.019831 |

Ilustración 2 Habilidades transversales factor IDF

En cuanto a las habilidades transversales hay un par de observaciones importantes. Si tomamos

las 20 habilidades más transversales usando nuestra matriz IDF tendremos lo siguiente.

Sorpresivamente no todas son habilidades suaves. Inclusive, las habilidades más transversales

son habilidades duras (Microsoft Excel, Microsoft Office).

Tabla 1Industria construcción

De manera general podemos ver que podemos vislumbrar que nuestras habilidades caen en: Microsoft tools, management (cosas como Team Leadership o Project Management), lenguajes, communication y soft skills (cosas como Communication o Teamwork). Esto tiene sentido pues son aspectos que

todo emprendimiento u organización necesita indiferentemente de lo que haga. Estas habilidades

se adaptarán a la industria en particular, pero son necesarias para casi todas las industrias

(al menos fueron mencionadas por casi todas las industrias).

Otro punto interesante es que ningún valor es 0, lo que quiere decir que ninguna habilidad fue

mencionada por todas las industrias. Esto no es concluyente, puede ser que una compañía fue

eliminada del análisis porque otro campo que presentada era inconsistente pero justo mencionaba

la habilidad en cuestión, no obstante, la jerarquía mostrada nos da una idea de la

transversalidad de estas habilidades entre sí.

## Resultado perfil tf-idf

Después de observar los perfiles preliminares que se obtienen a partir de la selección de las

habilidades con índice TF-IDF más alto por industria, tenemos ciertos comportamientos que se

repiten. Por este motivo, vamos a explicar a detalle un par de industrias que presentan

comportamientos interesantes que se pueden ver reflejados en otras y que nos parecen que tienen

insights valiosos. Comenzaremos por industrias que parecen valorar habilidades duras y técnicas

por encima de suaves. Gradualmente iremos transicionando a industrias que le ponen mucha más importancia a habilidades suaves (e incluso transversales).

NOTA: Las habilidades que tienen una (T) son habilidades bastante transversales, pero

alcanzaron un puntaje más alto que otras habilidades en una determina industria. Esto no es

imposible. Puede ser que una industria esté mencione más ciertas habilidades por encima de que son transversales entre industrias.

## Resultados por Industrias Técnica

Tabla 2 Industria Construcción

|  |  |
| --- | --- |
| Industrias "Técnicas" | |
| Construcción | |
| Hardskills | Softskills |
| 1. Construction (0.0561) | 1. Contract Negotiation (0.01) |
| 2. AutoCAD (0.035) | 2. Administration (0.009) |
| 3. Engineering (0.0282) | 3. English (T) (0.008) |
| 4. Construction Management (0.0243) |  |
| 5. Civil Engineering (0.0178) |  |
| 6. Pre-construction (0.011) |  |

Como podemos apreciar, la industria de la construcción valora el uso de herramientas técnicas como AutoCAD y conocimiento en construcción (obviamente). Es de esperarse que este tipo de habilidades sean características de la industria puesto a que mencionan a menudo aquí pero no tanto en otras.

Fue un poco más complicado conseguir habilidades suaves con más alto TF-IDF, esto se lo puede apreciar por el salto en orden de magnitud entre administración (0.009) y construcción (0.056).

Podemos apreciar también que conforme bajamos más en cuanto a puntaje de TF-IDF nos topamos con una habilidad bastante transversal, inglés (0.008). Se debía bajar tanto en el puntaje TF-IDF

para encontrar habilidades suaves en la industria de construcción que empezamos a ver

habilidades transversales.

Tabla 3Industria Oil & Energy

|  |  |
| --- | --- |
| Oil & Energy | |
| Hard skills | Softskills |
| 1. Petroleum (0.1355) | 1. Spanish (0.0067) |
| 2. Energy (0.1004) | 2. English (T)(0.0065) |
| 3. Oil & Gas (0.0999) | 3. Negotiation (T) (0.0063) |
| 4. Gas (0.0923) |  |
| 5. Onshore Operations (0.0722) |  |

La industria de combustibles es otra bastante técnica. No solo eso, es fácil entender por qué las habilidades características más representativas de esta industria tienden a tener in TF-IDF inclusive más alto que las de la industria de la construcción (por ejemplo: Petroleum (0.13) vs. Construction (0.05)). Esto posiblemente se debe a lo fundamental que resultan estas habilidades y lo poco necesitadas que son en otras industrias; son habilidades particularmente específicas pero bastante solicitadas en esta industria.

Algo similar a lo que ocurría con las habilidades suaves de construcción ocurre en esta industria. Las habilidades suaves pasan a un segundo plano, tanto así que los valores de TF-IDF de las habilidades suaves difieren en 2 órdenes de magnitud con la habilidad más característica

(Petroleum (0.1355) vs. Spanish (0.0067)).

Aquí también podemos apreciar cómo las habilidades suaves, conforme bajan en TF-IDF empiezan a presentar habilidades transversales (no realmente características).

|  |  |
| --- | --- |
| It & Telecom Services | |
| Hard skills | Softskills |
| 1. Telecommunications (0.174) | 1. Teaching (0.0087) |
| 2. Business Intelligence (BI) (0.039) | 2. Time management (0.0087) |
| 3. PHP (0.0268) | 3. Leadership (0.007) |
| 4. Networking (0.02077) |  |
| 5. Internet Protocol (0.02044) |  |
| 6. HTML5 (0.2) |  |
| 7. SQL (0.018) |  |

|  |  |
| --- | --- |
| Software Development | |
| Hard skills | Softskills |
| 1. Software Development (0.1383) | 1. Accountability (0.015) |
| 2. JavaScript (0.1133) | 2. Communication (T) (0.012) |
| 3. Python (0.095) | 3. Research (T) (0.011) |
| 4. Open ERP (0.095) |  |
| 5. Java (0.082) |  |
| 6. Genexus (0.071) |  |
| 7. PHP (0.069) |  |

## Resultado por Industria Mixta

|  |  |
| --- | --- |
| Finance & Banking Sector | |
| Hard skills | Softskills |
| 1. Banking (0.1486) | 1. Teamwork (0.011) (T) |
| 2. Financial Analysis & General Finance (0.104) | 2. Customer Service (0.008) (T) |
| 3. Bookkeeping (0.011) | 3. Business Strategy (0.007) (T) |

En cuanto a las habilidades fuertes características de la banca, pensamos que son lógicas. Lo

interesante con esta industria es que marca el inicio de un patrón que observamos en sus

habilidades débiles.

Los soft skills de la banca tienen sentido, pero, tienen algo peculiar. Todos pueden ser considerados de los mas transversales del corpus. Si bien siguen separados del skill más característico (Banking (0.1486)) por todo 1 orden de magnitud, ahora ocupan más espacios en el

top de skills con mayor puntaje TF-IDF. En otras palabras, no están tan separados del ámbito

técnico por encima de si se tratan de habilidades transversales.

Este fenómeno lo vemos replicado en varias industrias consideradas tradicionalmente como menos técnicas. Se desarrollará esta idea al final del análisis. Veamos otro ejemplo.

|  |  |
| --- | --- |
| Pharmaceuticals | |
| Hard skills | Softskills |
| 1. Pharmaceutics (0.097) | 1. Teamwork (T) (0.0136) |
| 2. Good Manufacturing Practice   (GMP) (0.034) | 2. Personal Development (0.013) |
| 3. Pharmacy (0.017) | 3. Lecturing (0.012) |
| 4. Quality Control (0.0148) | 4. Leadership (T) (0.009) |

ara la industria farmacéutica podemos apreciar que la habilidad suave con mayor TF-IDF

(Teamwork (0.0136)) y el hard skill con mayor TF-IDF (Pharmaceutics (0.097)) comparten orden de magnitud. Lo cual refuerza aquello que hemos discutido previamente sobre disminuir la brechaentre ambos tipos de habilidades.

# discusión

Al tomar industrias tradicionalmente consideradas como técnicas, podemos apreciar que se le da

una importancia visible a habilidades fuertes. No solo eso, especulamos que las habilidades

técnicas tienden también a ser más específicas y relacionadas a tareas puntuales. Es por eso

que tenemos industrias como la de combustibles que tiene una habilidad muy característica y

técnica (conocimiento en petróleo (0.1355)) que no vamos a ver comúnmente en otras industrias.

Como resultado de esto, el algoritmo de Term Frequency - Inverse Document Frequency (TF-IDF) va

a asignarle un valor alto.

Otro punto importante, es que las industrias técnicas tienden a tener una brecha en cuanto al

puntaje TF-IDF entre las habilidades duras y las habilidades suaves. En otras palabras, las

industrias técnicas tienden a no ser caracterizadas por las habilidades suaves. Es muy

importante recalcar que esto no significa que no sean importantes, solo que no son

características de la industria.

Conforme vamos a industrias que se orientan más al trato de personas a los negocios y a la

gestión empezamos a vislumbra un patrón contrario a lo que vemos en industrias técnicas. Se

empieza a ver más habilidades suaves como características de la industria, como por ejemplo la

industria de gestión de educación con habilidad de oratoria (0.274). Aparte de que porsupuesto

vemos habilidades técnicas orientadas a las personas como lo son los recursos humanos (0.012).

Mientras menos técnica la industria menos brecha existía entre las habilidades técnicas y las

habilidades suaves. Esto puede ser porque las habilidades técnicas están mas relacionadas a

temas de trato de personas y por ende son necesarias en varios tipos de organizaciones a través

de las industrias.

Finalmente es de observar que conforme la brecha entre habilidades se cerraba, las industrias

tenían, de manera general, un valor de TF-IDF menor en sus habilidades; en otras palabras.

Las habilidades de estas industrias eran menos características de la industria en sí y podían

ser más holísticas, transferibles y transversales. Por ejemplo, la gerencia o el trabajo en

equipo, que, si bien se benefician de tener 'know how' de una industria en particular, son

habilidades mucho menos específicas que saber sobre el lenguaje de programación PHP.

[1]Bjarke Mønsted, Anders Mollgaard, Joachim Mathiesen,” Métrica basada en el teléfono como predictor de rasgos básicos de personalidad”, 2018,.