Universidad Torcuato Di Tella
Escuela de Negocios

EXECUTIVE MBA 2017

APRENDIZAJE AUTOMATICO PARA LA GESTION Y RETENCION DEL TALENTO EN LAS ORGANIZACIONES

Área Temática: Gestión de Talento, Modelos de Aprendizaje Cognitivos Automatizados, Algoritmos Matemáticos y Planificación de Operaciones con Mano de Obra Calificada Intensiva.

Alumno: **Luis Mauricio Frieri**

Tutor: **Mariano Pérez**

Fecha: **Mayo de 2017**

Lugar: **Ciudad Autónoma de Buenos Aires**

Agradecimientos: Quiero agradecer a mi pareja y familia, quienes siempre estuvieron a mi lado y me ayudan a superarme todos los días. Especial agradecimiento a quien con su apoyo, me alentó a concluir esta tesis.

UNIVERSIDAD TORCUATO DI TELLA

APRENDIZAJE AUTOMATICO PARA LA GESTION Y RETENCION DEL TALENTO EN LAS ORGANIZACIONES

RESUMEN

Las renuncias en las organizaciones son una problemática abordada desde muchas perspectivas hoy en día. Desde el punto de vista empresarial, este tipo de situación trae aparejado dificultades para la planificación operacional, costos de inducciones altos y pérdidas de productividad en los meses previo a la salida.

En los últimos años, la minería de datos y el modelaje de algoritmos predictivos han evolucionado fuertemente. En la actualidad contamos con los recursos tecnológicos necesarios para el procesamiento de datos masivos y la construcción de modelos capaces de aportar información útil para la toma de decisiones. Dicha información podría ser utilizada por la compañía, no por el simple hecho de pronosticar con cierto grado de confianza cuando sucederá una renuncia y prepararse para su reemplazo, sino también para aprender a gestionar al talento formado durante tanto tiempo en pos de alcanzar un equilibrio mutuamente beneficioso entre empresa y profesional.

Esta tesis se basó en la investigación de la realidad del profesional para poder crear un algoritmo predictivo de renuncias con el objeto de optimizar los recursos financieros y humanos de la organización.

Para lograr este objetivo fue necesario recolectar datos observables relevantes sobre el historial de renuncias y entender sus potenciales causas a través de entrevistas con los responsables jerárquicos. Posteriormente se realizó el modelaje de algoritmos predictivos que conecten a todas las variables observables subyacentes al profesional y el análisis de los resultados parciales para comprender la eficiencia de nuestro modelo.

Con la eventualidad de que tales variables no fueran suficiente para el armado de un algoritmo preciso, nos planteamos cuales fueron las falencias de nuestro trabajo para recorrer un camino complementario. El mismo vendrá dado por la detección, estudio y medición de variables latentes que afectan el comportamiento de los profesionales en el momento de dejar la organización.

Palabras Claves: Renuncias, Algoritmos, Cadenas de Markov, Python, Variables Latentes.

INDICE

Tabla de Illustrationes

1 OBJETIVOS GENERALES

Históricamente el área de recursos humanos tuvo dentro de sus objetivos crear condiciones en el entorno laboral que permitan desarrollar a los empleados y posteriormente estos puedan contribuir a la organización en la búsqueda de sus metas financieras. Las herramientas más conocidas tiene como foco las conversaciones significativas entre supervisor y supervisado para la construcción de carreras profesionales. Es por tal motivo que actualmente las empresas invierten muchos recursos en la capacitación de sus líderes, dejando sobre estos la responsabilidad de administrar a su gente. Sin embargo, hay pocos antecedentes sobre análisis de datos concretos en las causas de las renuncias y sus consecuencias económicas.

La tesis tiene como objetivo general plantear un cambio de paradigma con respecto a la gestión del talento en las organizaciones donde los recursos humanos son los factores determinantes del éxito y forman las ventajas competitivas. Puntualmente perseguiremos la creación de un modelo cognitivo de aprendizaje para la retención del talento a través de la combinatoria de variables que afectan el tiempo de permanencia. En una primera instancia analizaremos las variables denominadas observables directamente y endógenas a la organización como puede ser la remuneración del empleado. Esto nos servirá como punto de partida para la construcción de un modelo multidimensional, donde la combinatoria de valores que tenga cada individuo en un momento especifico nos ayudará a predecir una posible renuncia. Es probable que nos encontremos con el obstáculo de las dimensiones, es decir, cuanto más variables agregamos en nuestro modelo más necesidad de datos tendremos para reducir el error de estimación. Con los resultados parciales de los algoritmos, llevaremos a cabo una matriz de confusión para determinar el grado de precisión del modelo y realizar las proyecciones para un periodo de tiempo determinado.

Por otro lado, intentaremos desarrollar un modelo cognitivo para entender como las variables latentes en los individuos asociadas a sus perfiles y personalidades pueden influir a la hora de la toma de decisiones. En otras palabras, conocer por qué un profesional que tienen las mismas condiciones laborales que otro puede optar por dejar la organización. Esta situación nos permite plantear como objetivo encontrar una forma de poder medir este tipo de variables a través de conductores “proxy”. Como parte de este análisis, utilizaremos los conceptos de validación cruzada para conjuntos de prueba y de entrenamiento. Además la información suministrada por los Gerentes y Lideres de equipo será filtrada junto con tabulada para diseñar matrices con los perfiles que predominan entre los integrantes de la organización.

Por ultimo nuestro objetivo general será combinar ambos modelos que representan las variables endógenas latentes y observables en un único algoritmo predictivo. Para ello utilizaremos el soporte de herramientas tecnológicas dedicadas al manejo de grandes datos (Big Data) como son R y Python.

Desde el punto de vista económico, nuestro modelo debería facilitar la optimización de los recursos atribuidos a cada variable. La idea es encontrar el punto de inflexión donde agregar más recursos conduce a rendimientos marginales decrecientes. Con esta información, los directores de la organización podrían extender al máximo posible el tiempo de permanencia de un profesional, con el menor costo operativo, hasta el límite en que continua siendo productivo.

2 OBJETIVOS ESPECÍFICOS

Examinar las causas de un alza en la tasa de renuncias, en los últimos años, y realizar entrevistas con gerentes para entender sus percepciones del asunto en cuestión. Cuantificar el costo directo de un reclutamiento y el costo implícito de una renuncia por pérdida de productividad. Interpretar los perfiles actuales y cuestionarlos, relevar datos a través de encuestas y entrevistas sobre la personalidad de los profesionales, conocer políticas de incorporación y retención por la organización, caracterizar al perfil óptimo en términos de permanencia. Relacionar productividad con permanencia. Relevar estadísticas descriptivas sobre la composición de la fuerza de trabajo para investigar configuraciones de perfiles que mejor se adapta al tipo de organización estudiada.

Desde el punto de vista del modelaje de algoritmos, definir qué tipo de distribución encierran las renuncias. Realizar cálculos para obtener el punto en que el sistema (negocio) entraría en colapso por renuncias masivas. Crear la matriz de estados con los valores que cada variable (observable como latente) tome durante un cierto lapso de tiempo. Determinar el error de estimación en el uso de un cuestionario proxy para inferir variables latentes.

Analizar algunas herramientas utilizadas en la actualidad para el estudio de redes, en especial, las relacionadas a las búsquedas laborales y estudiar la forma de integrar variables endógenas y exógenas a la organización.

3 BREVE RESEÑA HISTORICA Y ACTUAL

La historia del capitalismo, como sistema social y económico, desde el siglo XVI hasta nuestros días indica como distintos elementos han tomado el eje central en su matriz de valor, la fuerza de trabajo ha tenido un protagonismo único en los últimos 50 años, siendo un factor clave para las transformaciones que evitaron la ruptura definitiva del sistema capitalista profetizada por Marx en su obra **"El Capital"**[1].

Hace más de 200 años que Adam Smith escribía, en su obra capital **"La Riqueza de las Naciones"**[2], que la verdadera riqueza de los pueblos reside principalmente en sus personas y su creatividad. Y aunque las organizaciones han tardado en constatar esta realidad, hoy día se puede afirmar que los esfuerzos por retener a los mejores han pasado a ocupar un lugar clave en la estrategia empresarial. Buena muestra de ello es la incipiente preocupación de los altos directivos de empresas por la escasez de talento. No solo como motor de la economía moderna y la creación de valor, sino también por la dinámica del cambio, por eso el trabajador constituye la mayor fuente de innovación para las empresas.

Inicialmente la tierra era considerada como la fuente del valor principal, determinante de las ventajas comparativas entre los países. En el último siglo hemos notado como ese paradigma cambio, posicionando al factor humano como hacedor de las ventajas competitivas de las economías modernas. En efecto, los avances tecnológicos de las últimas décadas permitieron que la formación del capital humano fuese más económica y accesible, este hecho trajo aparejado mejoras en términos de la distribución del ingreso a nivel mundial. En 1970 la mayor parte de ciudadanos del mundo provenían de países pobres (China: 1.200 Millones; India: 1.000 Millones de habitantes). Sin embargo, desde esa fecha hacia adelante, los países pobres de Asia aumentaron el grado de educación de sus recursos humanos y la disponibilidad de capital físico en cada uno de ellos a un ritmo vertiginoso de modo que el ingreso de los ciudadanos de estos países fue convergiendo hacia el nivel de los países ricos (Europa y EE.UU.). De este modo, la distancia entre los países (ponderada por el número de habitantes) fue reduciéndose y esa es la causa principal por la cual las desigualdades en el mudo están disminuyendo y no creciendo. Basándonos en el **Índice de Theil**[3], entendemos como aquellos países que cuenten con mayor educación para sus habitantes lograrán mejoras en la productividad, lo que tiende a la disminución de la desigualdad. El proceso de ahorro e inversión de capital juega un rol clave al respecto.

Debemos el análisis más conocido del "**capital humano**" al economista norteamericano Gary Becker[4] Define el conjunto de las aptitudes y las habilidades acumuladas por el individuo y susceptibles de desempeñar un papel en el proceso de producción. Es la forma de capital cuya consideración es la más reciente. Desempeña un papel que crece en una sociedad cada vez más especializada y donde la investigación y las ciencias tienen un sitio crucial. Este capital es

sustancial al individuo y parece pues improbable que se le pueda desposeer. Existen no obstante unas excepciones notables. Los asalariados que dejan su empresa pueden estar sometidos por ejemplo a una cláusula de no-competencia, impidiéndoles entonces que una empresa competidora aproveche sus conocimientos por un cierto tiempo. Pero el capital humano pone verdaderos problemas: la "fuga de cerebros" por ejemplo (altos diplomados formados a expensas de un Estado y que otros aprovechan). De la misma forma, el riesgo de perder a sus asalariados desanima a las empresas de ofrecerles una formación onerosa. El capital humano representa una forma de capital de la que el capitalista todavía no puede apropiarse.

"El trabajador, hoy, no necesita más instrumentos de trabajo (es decir de capital fijo) que sean puestos a su disposición por el capital. El capital fijo más importante, el que determina los diferenciales de productividad, en lo sucesivo se encuentra en el cerebro de la gente que trabaja: es la máquina-herramienta que cada uno de nosotros lleva en sí mismo. Es esto la novedad absolutamente esencial de la vida productiva en nuestros tiempos."[5]

Teniendo en cuenta esta breve reseña, parece innegable que la clave para el éxito residirá en aquellas organizaciones que sepan optimizar a sus recursos humanos y retenerlos, una especie de bucle del desarrollo profesional pero también personal porque el individuo consciente de su valor buscara trabajar en aquellos lugares que satisfagan al máximo sus necesidades. En otras palabras, el capital y los recursos tecnológicos pueden ser conseguidos casi sin restricciones en el corto plazo, pero el desarrollo de las personas es una inversión a largo plazo, una política permanente, una visión del ente.

No es nada nuevo si decimos que los tiempos han cambiado. En el caso del ámbito laboral, uno de esos cambios es el impulso de las nuevas generaciones de trabajadores hacia una nueva forma de valorar las compensaciones que un trabajo les puede reportar. El dinero a final de mes ya no es tan importante si la empresa no ofrece otro tipo de incentivos no económicos por los cuales merezca la pena permanecer dentro de la compañía, como por ejemplo la forma en que las empresas desarrollan a su gente y contribuyen a la sociedad. Participar de las decisiones empresariales parece ser algo muy valorado por los profesionales porque estos están deseosos de demostrar lo que pueden aportar. Poco a poco la visión de las organizaciones como solo generadoras de dinero y rentabilidad está girando hacia una visión más holística del propósito de las empresas. Metas como fomentar la innovación, el uso de recursos renovables y tener un alto impacto en la sociedad están tomando el centro de la escena. Todo esto hace que los incentivos para retener a un profesional de alto rendimiento sean multidimensionales. Si los directivos no

[1] Ver **Karl Marx**. El Capital. TOMO I. "La Acumulación de Capital"
[2] Ver, **Adam Smith**. La Riqueza de las Naciones. (1776)
[3] Ver**, Índice de Theil, Claude E. Shannon**, Entropía. www.cepal.org/deype/mecovi/docs
[4] Ver **Gary Becker.** El Capital Humano. (1964)
[5] Ver **Gary Becker.** Economía de la Discriminación. (1957)

prestan atención y ponen en caja un modelo de integración que contemplen todas estas variables, correrán el riesgo de quedarse en el camino.

Algunos datos de la actualidad en Argentina no muestran una adaptación muy pronunciada ya que el 75% de los Millennials piensa que las organizaciones están demasiado enfocadas en sus agendas y restan importancia a ayudar a mejorar la sociedad. El liderazgo juega un papel clave en esta transformación, las nuevas generaciones tienen una visión muy diferente de cómo deben trabajar los equipos de líderes a como era hace 20 o 30 años atrás. ***"Los grandes líderes poseen una inteligencia social desbordante"***[6]

Los rasgos de personalidad de quienes son identificados como líderes inspiradores por parte de los Millennials incluyen:

1. Pensamiento Estratégico.
2. Ser fuente de Inspiración.
3. Fuertes habilidades interpersonales.
4. Visión.
5. Pasión y entusiasmo.

Las prioridades de los líderes y de las nuevas generaciones suelen ser foco de conflictos al momento de alinearse, mientras los primero tienen su objetivos orientados al corto plazo y a los resultados inmediatos, los segundo esperan señales de planificación estratégica para las organización junto con desarrollo de todos los integrantes. También encontramos diferencias en las expectativas, mientras que algunos aspira a convertirse en el líder su organización, el resto siguen teniendo incertidumbre acerca de su futuro laboral. Hay un marcado conflicto vocacional.

---

[6] Ver, **Generación del Milenio**. Deloitte. 2015

4 ESTRATEGIA METODOLOGICA

Es probable que el lector tenga cuestionamientos acerca de los modelos y la eficiencia de los mismos a medida que avance en el texto. Es lógico pensar que por más abarcativo que sean los algoritmos utilizados en cada etapa, siempre habrá un factor imponderable que determina la conducta de las personas ya que estas son según sus circunstancias. Por tal motivo debemos recordar que el objetivo general de esta tesis es investigar soluciones complementarias para la gestión de los recursos humanos en la organización a través del análisis de datos y de identificar patrones de comportamiento susceptibles de ser influenciados.

Todos los modelos e hipótesis planteados a lo largo del estudio están desarrollados en el campo teórico, solo los datos que sirven de base para los desarrollos de cada modelo fueron extraídos empíricamente de un contexto organización especifico. Es factible que nuevas investigaciones logren mejorar la precisión de las predicciones y que estas sirvan de plataforma para saltos cualitativos en este campo. Esto hace referencia al concepto de **"Machine Learning"**[1] que sin dudas debería aplicarse en este tipo de trabajos a la hora de llevarlo al campo práctico. La aplicación de decisiones soportadas con los resultados arrojados por los algoritmos no formará parte de este trabajo y tampoco serán medidas en términos de su eficiencia. No obstante, el presente trabajo incluye resultados obtenidos sobre pruebas para el testeo de los algoritmos parciales, pero el lector debe distinguir que estos resultados marcan el grado de precisión en las predicciones sobre el mismo conjunto de entrenamiento con que se construyó el algoritmo. Sobre otro contexto, los resultados podrían ser distintos.

Otro punto importante a destacar es el contexto necesario para llevar adelante este tipo de trabajo, como mostraremos durante el desarrollo del marco teórico es imperativo contar con una cantidad suficiente de sujetos que contribuyan con una cantidad de datos óptimo para su procesamiento. Este tema no debe pasar inadvertido y podría aclarar varias dudas del lector acerca de la factibilidad del estudio. El problema de las renuncias y pérdida de talento puede tener innumerable cantidad de aristas, situación que demanda un análisis previo al momento de aplicar herramientas para su control. Reiteramos la importancia de enfocar el estudio en el contexto adecuado como premisa principal antes de cualquier ejercicio al respecto. Esto sin perjuicio de un correcto aprovechamiento de los datos acumulados. La disponibilidad de esos datos es un importante activo para cualquier organización, en la medida en que puedan ser transformados en información de interés, utilizando técnicas y métodos de "**Data Mining**"[2]

Con la finalidad de simplificar la lectura de la tesis y definir claramente la metodología, se realizó una síntesis grafica de los pasos y modelos desarrollados para la construcción de un algoritmo predictivo de las renuncias en las organizaciones.

Identificación del problema en el entorno organizacional (Índice 5 - Problemática y Entorno)

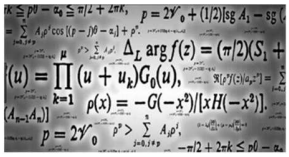
Cuantificar los costos asociados al problema y el impacto sobre la productividad

Explorar Ideas no convencionales y estudiar factibilidad. (Índice 6.1 – Variables Observables).

Recolectar datos de las fuentes información (Índice 6.1.1 - Descripción de Variables))

Adaptación de la Base de Datos. (Índice 6 – Descripción del Modelo)

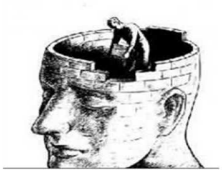
- Descripción de las Variables (Índice 6.1.1)
- Determinación de la Muestra (Índice 6.1.2)
- Tipo de Distribuciones (Índice 6.1.3)
- Selección de Variables Significativas (Índice 6.1.4)
- Regresares y Correlaciones (Índice 6.1.5)



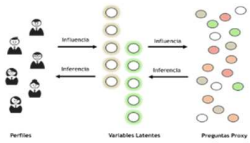
Algoritmo Predictivo de Renuncias (Índice 6.2)

- Cadenas de Markov (Índice 6.2.1)
- Matriz de Transición y Estados (Índice 6.2.2)
- Combinatoria de Variables en el Espacio-Tiempo (Índice 6.2.3)
- Matriz de Confusión (Índice 6.2.4).

Finalmente al introducir el concepto de Matriz de Confusión y el margen de error que tiene el algoritmo, se desprenderá naturalmente la conclusión de que hay razones para las renuncias que no pudieron ser cubiertas por el modelo predictivo creado. Por esto se propone complementar el modelo con el estudio de variables latentes según el siguiente flujo:



Explorar Nuevas Causas para las Renuncias (Índice 6.4 – Variables Latentes)



Creación de cuestionarios para revelar Variables Latentes (Índice 6.4.2)

Reorganización de datos y tabulación

Inferencia de Variables Latentes. Modelo Matemático (Índice 6.4.3)

Validación de Inferencia Proxy

Integración de Modelos Observable y Latente (Índice 7)

---

[1] Ver **Data Science for Business.** Foster Provost. Tom Fawcett. ED2013. O´Reilly Media.
[2] Ver **Minería de Datos Basada en Sistemas Inteligentes.** Britos P. y Hossian A. Sierra E. 2005.

5 PROBLEMATICA Y ENTORNO: DESARROLLO DEL CASO

"Por razones de confidencialidad no estamos autorizados a develar el nombre de la compañía de manera explícita y datos del personal clasificados como restringidos"

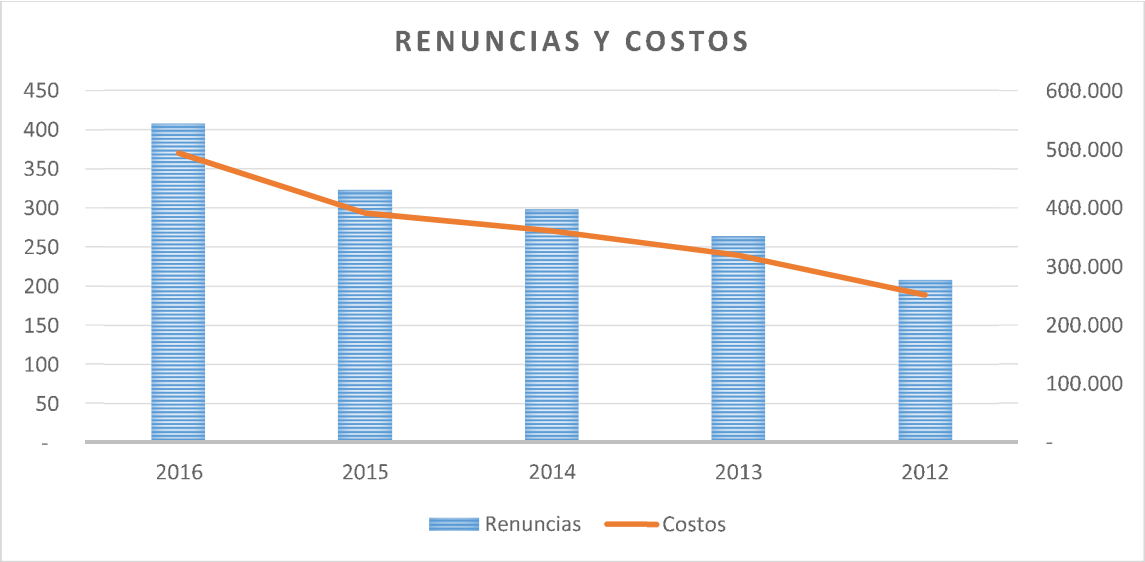
A modo de presentación podemos destacar que la compañía opera a nivel global como consultora de gestión, en servicios tecnológicos y procesos de negocios. Teniendo en su nómina a más de 370.000 empleados dando soporte a más 30 clientes diamantes en 120 países del mundo.

La fuerza de trabajo se divide en las siguientes cuatro funciones complementarias de acuerdo con las tareas que realizan:

- **Estrategia**
- **Consultoría**
- **Tercerización de procesos (servicios)**
- **Soluciones de tecnología**

La línea de negocios inicio sus actividades brindando servicios de administración financiera y control interno entre lo que se destacan los procesos de contabilidad, facturación, cobranzas, pagos a proveedores y administración en general. La empresa cuenta con una red de centros de servicios creados para atender las necesidades de los clientes (en su mayoría empresas multinacionales) en forma geográfica y por especialización de tareas. El presente trabajo tendrá como objeto de estudio a los profesionales del centro de servicios localizado en la Ciudad Autónoma de Buenos Aires en zona de microcentro, y cuenta con alrededor de 1800 profesionales. Como puede interpretarse, el contexto abarca a un entorno multicultural, donde los empleados tienen interacción con similares de otras regiones del mundo. Este tema no es menor, estudios recientes señalan la posibilidad de fronteras amplias como un elemento que atrae talento.

La propuesta de valor en los últimos años estuvo enfocada en operar estructuras de personal extensas de manera eficiente, lograr alta estandarización de procesos y abaratar costos operativos a los clientes, con el objeto de permitirle a estos enfocarse en sus verdaderos negocios y creación de riqueza. Siendo el capital humano el máximo exponente de la ventaja competitiva de la empresa, su alta rotaciones conlleva grandes problemas operativos y financieras para la empresa. Alrededor de esta problemática estará orientada nuestra tesis, con la intensión de aprovechar una oportunidad de negocios.



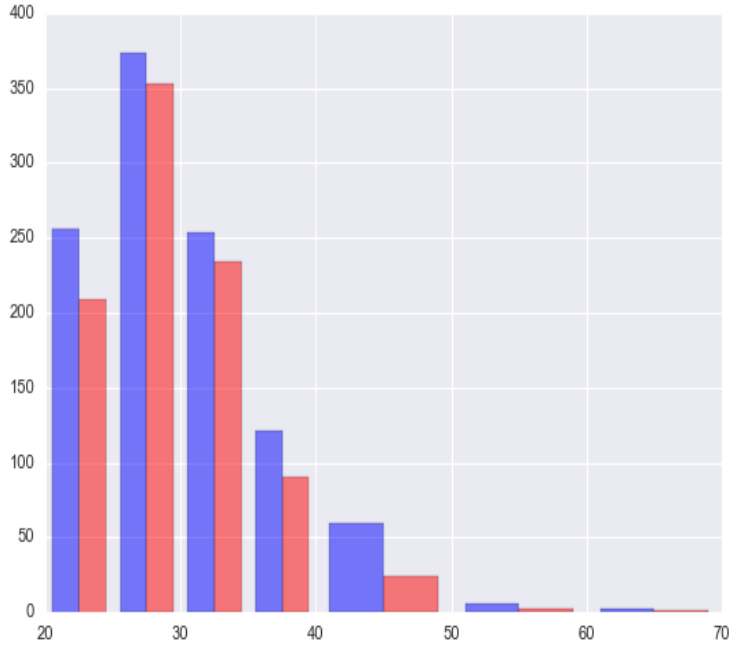
I. Renuncias y Costos – Fuente: Depto. de Finanzas.

El grafico muestra la tendencia de las renuncias en los últimos años, aquí vemos como no solo es una problemática para la operatoria y servicio al cliente sino también el impacto en costos que acarrea. Un modelo que permita extender el tiempo de permanencia se convertiría en una herramienta de planificación financiera.

La tesis intentara abordar el problema a través del estudio de las variables endógenas y exógenas que influyen en el tiempo de permanencia. La premisa consiste en que no es razonable intentar predecir el tiempo de permanencia exacto de cada empleado, pero es posible mejorar la toma de decisión a través del análisis asociativo de datos empíricos. El siguiente trabajo teórico tiene como objeto de estudio al comportamiento de los profesionales de una organización empresarial al momento de decidir su permanencia en esta o continuar su desarrollo en otro tipo de emprendimiento, más ambicioso aun, intentara mostrar que existe un tipo de perfil de persona que sea más adaptable al estilo de organización bajo estudio y por ende más proclive de retener con menores esfuerzos. El contexto contiene características que determinaran el estudio y justamente dichas características serán las variables claves para construir un modelo predictivo con el mayor grado de precisión posible. El modelo no solo buscara identificar variables influenciables sobre las cuales los directivos pueden actuar, sino también sobre variables predefinidas que están insertas en los perfiles y las cuales solo pueden identificarse al momento del reclutamiento.

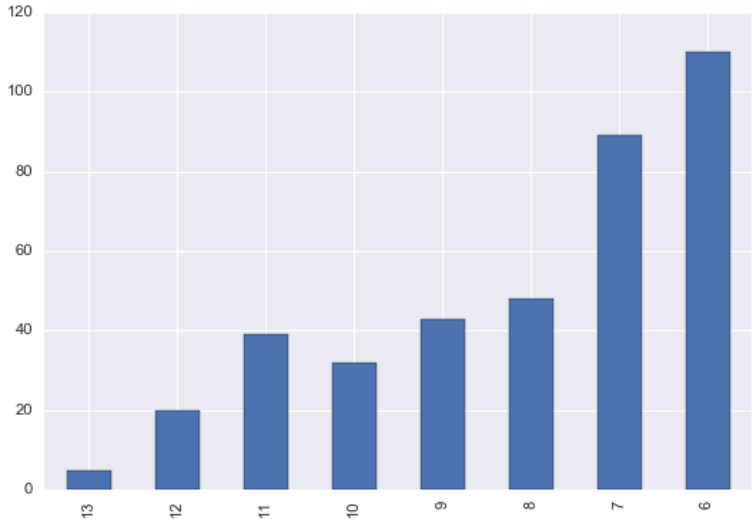
Los profesionales son de distintas disciplinas, desde el área de ciencias económicas e ingenierías en su preeminencia hasta derecho y otras ramas de índole social. Un evidente factor que caracteriza a la población bajo estudio es su edad, el 87% de los recursos son menores a 30 años. Por otro lado, solo el 48% ha completado sus estudios de grado, es decir, más de la mitad desarrolla su actividad laboral al mismo tiempo que completa su educación. Estos datos aportan mucho valor para armar los supuestos sobre el algoritmo de retención de talento. Si los analizamos de manera aislada es probable que las conclusiones sean inconducentes pero de manera integrada con otros datos los

resultados llevan a desmentir creencias generales. Más importante aún, los directivos encontraran en oportunidades que algunas decisiones no fueron tomadas con elementos lógicos y concretos para alterar comportamientos.



II. Distribución Población por Género. Fuente: Depto. de RR.HH.

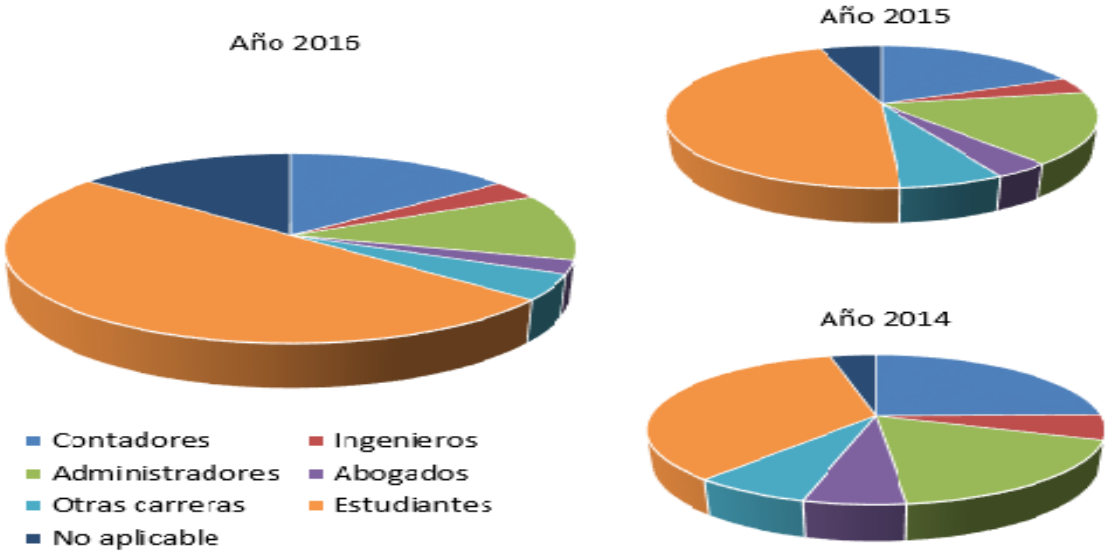
El grafico muestra la distribución de la población por su género (masculino=azul y femenino=rojo) junto con las edades de los profesionales. Como mencionamos, un rasgo distintivo es que analizaremos a integrantes de la generación Y en su mayoría. La totalidad del personal asciende a 1832 Desde el puesto de asistente hasta especialistas seniors.



III. Tiempo Medio de Permanencia. Fuente: Depto. de RR.HH.

Aquí el tiempo medio de permanencia por niveles histórico (13 es el nivel más bajo y 6 el más alto). Como era esperado a medida que un profesional hace carrera en la empresa obteniendo promociones su tiempo de permanencia se extiende. Más adelante veremos que esta variable no es la única determinante de las salidas. Pero si podemos estudiarla y comprender como afecta la retención del talento.

Categorias de profesioanles según sus estudios:



IV. Perfiles de Profesionales. Fuente: Depto. de RR.HH.

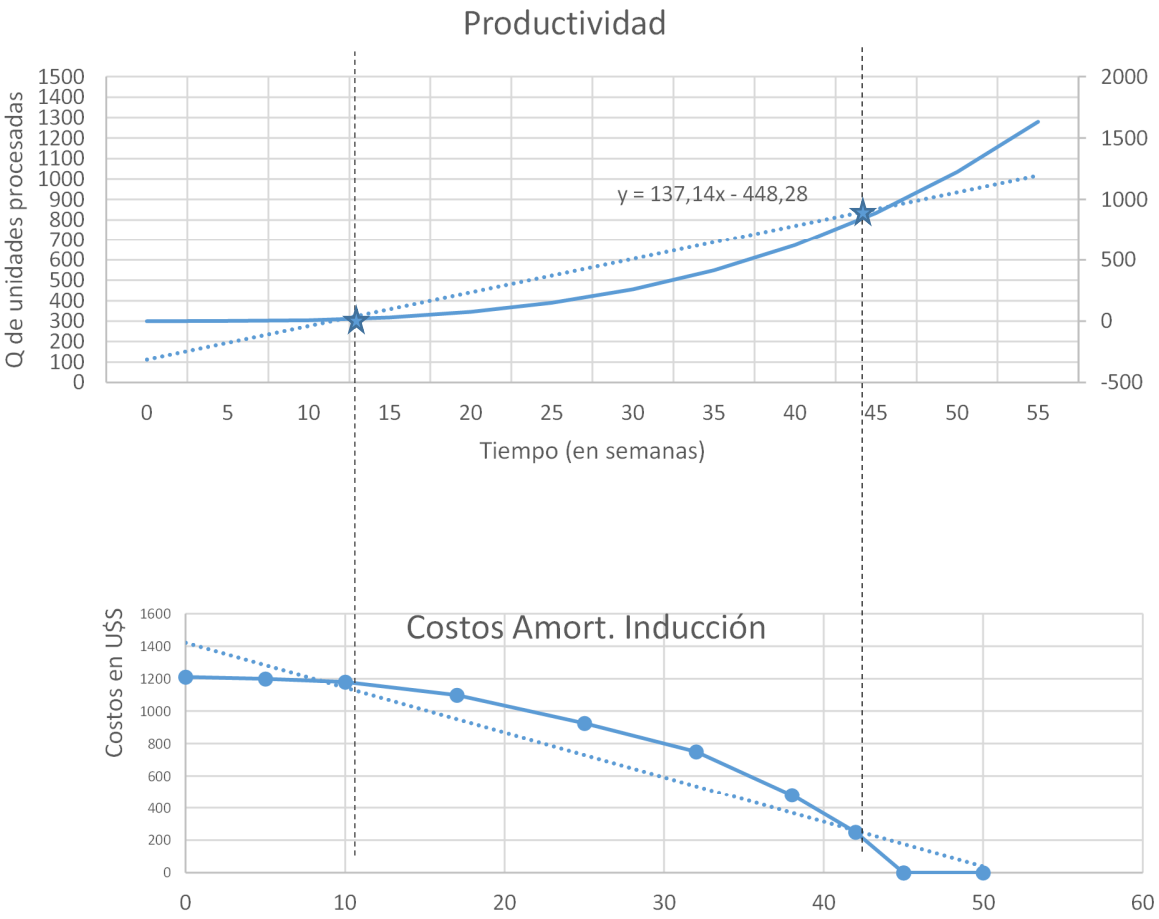
Como podemos notar la cantidad de profesionales egresados ha disminuido en tendencia, teniendo mayor predominación los estudiantes. Esto es un indicativo del tipo de organización en la actualidad, y como los candidatos ven a la empresa. Es por tal motivo también que la edad promedio haya bajado en los últimos años de 29 a 25 años.

Desde el punto de vista económico, es indudable que una alta tasa de renuncias genera distorsiones en la planificación financiera de la empresa. Cada evento de este tipo crea una contingencia para reemplazar al profesional de aproximadamente 1.210 dólares que incluyen costos de búsqueda y selección del candidato, proceso de entrevistas por expertos y otros gastos de ingresos como los exámenes previos. En el siguiente cuadro encontraremos el detalle de costos.

TABLA DE COSTOS

Actividad	U$S prom por candidato	Detalle
Aviso para reclutar	125 U$S	En páginas web y Consultoras de RRHH.
Costo Laboral entrevistadores	365 U$S	Incluye Sueldos, Cs Ss. y otros costos indirectos
Exámenes Pre-ingreso	432 U$S	Médico, Psicotécnico, Técnico.
Costo Laboral soporte técnico	288 U$S	Reaconcondicimiento equipos comput,

Sin embargo las contingencias no son solo del tipo monetarias. Cada baja representa una pérdida de conocimiento de difícil cuantificación en términos financieros pero que genera conflictos en las expectativas de calidad del cliente. La siguiente grafica muestra el tiempo de aprendizaje y adaptabilidad necesaria para que un nuevo recurso este operativo y su productividad sea comparable con el renunciante, en condiciones normales.



V. Productividad vs Costos Inducción. Fuente: Elaboración Propia.

Como podemos observar existe un primer periodo de tiempo donde el profesional adquiere información sobre sus funciones y aumenta su productividad con rendimientos marginales crecientes (Etapa 1). Sin embargo, a partir de determinado punto (primer intersección entre la curva y la recta) notamos una especie de estancamiento, y entonces comienzan a producirse rendimientos decrecientes. Esto parecería improbable si solo consideramos el aspecto determinístico del proceso, no olvidemos que tratamos con personas, pero por otro lado es el momento en que los costos de la

inducción comienzan a amortizarse por completo. Una renuncia en esta etapa 2 implicaría incurrir en más costos para ingresar a otro empleado y volver a entrenarlo. La grafica denota que en un cierto lapso de tiempo corto, todos disminuyen su productividad excepto por aquellos que sufrieron un cambio en los valores que toman las variables determinantes de la permanencia, lo que visualizamos a partir de la segunda intersección donde entramos en la etapa 3 con nuevamente rendimiento crecientes. Volveremos luego sobre estos temas con mayor detalle técnico, como conclusión preliminar podemos decir que trabajar sobre la retención del talento por el solo hecho de retenerlo carece de sentido y nos lleva a incurrir en costos ocultos de improductividad y baja calidad. El problema entonces es como medir tales costos a través del uso del algoritmo y más importante como transformarse al profesional para ayudarlo a transitar de la etapa 2 a la 3.

La función que desempeña los Gerentes en esta etapa es clave para lograr el objetivo, pero muchas veces sus acciones llegan a destiempo por no tener un sistema de información que alerte este tipo de problemática.

6 MARCO TEORICO

6.1 VARIABLES OBSERVABLES

Es imperativo explicar que el armado de un algoritmo predictivo del comportamiento humano requiere la búsqueda y obtención de una gran cantidad de datos. Como si esto fuese poco, en este caso, muchos de los datos no son cuantificables del modo tradicional y es necesario atribuirles series de valores para poder ingresarlos en el modelo. Pero antes de adentrarnos en el armado del conjunto de datos debemos conceptualizar el método aplicado. Una variable de resultado de un ensayo es toda característica medida en los sujetos de estudio que nos permita diferenciar el efecto encontrado en los grupos comparados y plantear el contraste con la hipótesis. En este trabajo, la tesis es la posibilidad de influenciar sobre las renuncias dentro de una organización pero para lograrlo debemos tener un modelo y ese modelo se alimenta de datos o mejor dicho de valores que toman la variables de la ecuación. A medida que agregamos más variables al modelo para lograr mayor precisión, nos encontraremos con un problema de dimensiones. En otras palabras, cada nueva variable agregada demandara buscar más y más datos debido a que el volumen del espacio aumenta exponencialmente. Cuando esto ocurre corremos el riesgo de perder significación estadística. Si esto sucede por la dispersión de datos, las conclusiones no serán sólidas y fiables. Para mitigar tales efectos es necesario realizar un proceso de selección de cuales variables serán tenidas en cuenta por el algoritmo y si los valores de dichas variables pueden ser conseguidos y cuantificados de manera fehaciente. En una primera etapa el proyecto buscara construir un algoritmo solo con el uso de variables observables. Para ejemplificar, una variable observable es aquella que puede estudiarse y medirse directamente. Cada cambio en una variable observable producirá un cambio en el resultado de la ecuación de manera directa y visible. Llegado a este punto y teniendo en cuenta el problema de las dimensiones, la forma de seleccionar la variables observables será con la experiencia de aquellos que influyen sobre ellas. Los gerentes y encargados de retener y potenciar al talento podrán darnos un tentativo conjunto de variables a incorporar en el algoritmo. Siendo los gerentes quienes reciben en forma constante feedback por parte de su supervisados y deben trabajar en el compromiso y motivación de estos para alcanzar el mayor rendimiento del negocio, es atinado considerar sus formas de manejar a los recursos. Cada gerente tiene a su cargo entre 60 y 80 profesionales, para administrar un equipo de ese tamaño utilizan aproximadamente 5 o 6 reportes directo.

A continuación detallamos cuales fueron las variables que los gerentes entienden como determinantes del modelo y sobre las cuales ellos trabajan para retener al talento.

Gerente 1	Gerente 2	Gerente 3	Gerente 4
Salario	Flexibilidad	Trabajo Remoto	Salario
Rotaciones Internas	Promociones	Rotaciones internas	Reconocimiento
Distancia al Trabajo	Salario	Flexibilidad	Salario

En un análisis preliminar podemos observar que todos coinciden con el mismo conjunto de variables pero no con el orden de importancia, esto puede deberse a las experiencias de cada gerente y el modo en que su estilo de liderazgo penetra en los integrantes de un equipo de trabajo. Este punto de partida nos ayudara a realizar las primeras pruebas para entender la significación de cada variable. Con el propósito de poner a prueba los algoritmos tentativos, necesitamos buscar los datos que completen las dimensiones que estas variables configuran en el espacio-tiempo (entiéndase como espacio-tiempo a la carrera profesional que el recurso realiza en su estadía en la organización).

6.1.1 Descripción de las Variables Observables.

Remuneración: El salario y otros tipos de remuneraciones pertenecen al grupo de variables directamente observable y puede ser medido sin mayores inconvenientes. Sin embargo, las remuneraciones presentan una restricción debido al tipo de organización donde desarrollamos el proyecto. Un aumento de salario significativo solo viene acompañado de una promoción jerárquica. Existen aumentos de menor cuantía sin cambio de jerarquía, pero no son percibidos por los empleados de igual modo. Es importante no confundir ambas variables, remuneraciones con promociones, si bien las promociones incluyen una mejor remuneración estas no se limitan solo a eso y engloban un cambio de status y desarrollo profesional. Evidentemente para el modelo no es intranscendente. Esta disgregación fue aclarada al momento de señalar las variables.

La forma de cuantificarlo ha sido por niveles desde analistas junior a especialistas sénior, cada categoría tiene una remuneración asociada que no muestra superposiciones. Es decir, una categoría mayor jamás tiene una remuneración menor a una categoría inferior.

Fuente de la información: Recursos Humanos – Área de compensaciones y beneficios.

Promociones y Reconocimientos: Como comentaba esta variable está estrechamente asociada con la anterior pero no tienen la misma significatividad. En ocasiones hay reconocimientos que son medibles en cantidad sin necesidad de que haya aumento de remuneración. Por otro lado, las promociones implican mucho más en la percepción del empleado y sus colegas que un incremento salarial porque el primero tiene connotaciones superiores. Por ejemplo, cada promoción conlleva

una mejora en las credenciales del recurso que podría aprovechar en otra organización además de la experiencia que la nueva jerarquía le otorgará. Sin dudas lo más importante de una promoción pasara por la autorrealización del profesional. Entonces es posible que esta variable de compleja cuantificación sea clave en los resultados.

La cuantificación es medida a través de la cantidad de veces que un recurso recibió reconocimientos, los cuales son enviados por un sistema y son parte de la calificación de performance del empleado. En cuanto a las promociones, estas serán consideradas en el armado de la combinatoria de variables que tienen cada empleado en el espacio-tiempo.

Fuente de la información: Recursos Humanos

Flexibilidad: Una de las condiciones de trabajo más valora por los profesionales de la nueva era es tener flexibilidad horaria y no verse atados a cumplir horarios. Especialmente cuando trabajan para otras regiones con diferentes huso horarios. En muchas ocasiones, son los propios recursos quienes consultan si la organización cuenta con este tipo de beneficios. Estudios recientes demuestran que la productividad es mayor en aquellos recursos que trabajan desde sus hogares o fuera de la oficina convencional. Al mismo tiempo hay un ahorro de costos por las empresas, en términos de infraestructura. La compañía tiene actualmente este tipo de práctica laboral implementada, siendo el ahorro generado en costos fijos de 267.340 U$S anuales sobre un total de 1.217.150 U$S presupuestados anuales. Sin embargo, el estudio nos enfoca hacia entender si también hay un ahorro implícito representado en la reducción de renuncias e incrementos sostenidos de productividad.

La manera de cuantificar la flexibilidad laboral está basada en la cantidad de días que cada recurso trabaja desde afuera de la oficina. Para el presente trabajo supondremos que no hay rendimientos crecientes de productividad por razones prácticas y de medición.

Fuente de la información - Recursos Humanos y sistemas internos de administración del personal.

Rotaciones internas: Este tipo de práctica representa una opción para motivar a un profesional con su aprendizaje de carrera y para que no ingrese en una zona de monotonía con sus tareas. Existen ciertas reglas para poder acceder a una rotación de proyecto, como por ejemplo tener al menos 1 año en el rol actual y haber entrenado a su backup. La empresa cuenta con un sistema para gestión de rotaciones donde gerentes hacen demandas de trabajo y los empleados pueden aplicar según su interés. La confirmación de la rotación depende del gerente actual y el nuevo.

Para medir este punto vamos a extraer información de las rotaciones a través del sistema usado por la organización y determinaremos cuantas rotaciones o cambios de rol tuvo un profesional en su carrera dentro de la compañía.

Fuente de la información – Recursos Humanos y sistemas de gestión internos.

Madurez del equipo: Como señalamos anteriormente, cada cliente representa un equipo de trabajo. Dependiendo del tiempo que se le ha prestado servicio a un cliente y del tipo de procesos que se realizan, algunos equipos están más estables que otros. Estables indica que todas las funciones y tareas están correctamente y el volumen de trabajo balanceado entre todos los empleados. Esporádicamente hay necesidades de horas extras de trabajo, pero no son regla. Desde el punto de vista de la retención de talento, esta condición de madurez y estabilidad no implica una ventaja para todos los perfiles. Es justamente uno de los objetivos del modelo encontrar que tipo de perfiles pueden adaptarse mejor a determinadas circunstancias y asignar recursos en esa línea.

La cuantificación de esta variable no es simple, ya que pueden existir muchos factores que afectan a la estabilidad de un grupo de trabajo. Por ende decidimos tomar como medida de la estabilidad la cantidad de años que un proyecto tiene en régimen. Por lo general, después del primer año podemos considerar que un proyecto esta estable si las demás condiciones permanecen constantes.

Fuente de la información: Directores y Gerentes de proyectos.

Distancia al Trabajo/Universidad: En este punto los gerentes intentaron remarcar un asunto que escuchan reiteradamente en las entrevistas. Consideran que los empleados no deberían viajar más de una hora para llegar a las oficinas. A su vez, la universidad a donde concurran debe ser cercanas a la oficina para evitar mayor trajín. Estas condiciones no son determinantes a la hora de incorporar un nuevo recurso pero si presentan indicios para medir los tiempos de permanencia, o al menos eso podrá ratificarlo nuestro modelo.

Para cuantificar estas variables, realizamos un análisis sobre las distancias recorridas por cada profesional desde su hogar a las oficinas y para aquellos que concurrir a la universidad incluyendo este punto también. Posteriormente estimamos el tiempo promedio que deben dedicar al viaje para arribar con un desvío de 15 minutos a cada lugar.

Fuente de la Información: Recursos Humanos – Investigación propia.

Título Universitario en Carrera afín: Otro punto que señalaron los gerentes hace referencia a si los empleados hayan o no completado sus carreras de grado. Según sus opiniones, existen mayores probabilidades de retención cuando el empleado obtuvo su título durante su estadía en la organización que aquellos que ya poseen su título al ingresar. Las estadísticas muestran este hecho pero no por ello podemos asociarlo como una causal directa que influya en la renuncia, probablemente funcione en combinatoria con otras variables.

La forma de medir esto depende de cuan actualizados estén los sistemas de información del personal dentro de la empresa. A riesgo de no contar con la información completa, realizamos pruebas aleatorias para corroborar la validez y precisión de los datos.

Fuente de la información: Recursos Humanos y búsquedas propias.

Edades y Género: Si bien esta podría ser la variable más evidente para incluir no por eso debe ser subestimada. Como mencionáramos, hay un 87% de la población total menor a 30 años. Esto nos podría indicar una tendencia hacia una organización de jóvenes profesionales. Perder esto de vista llevaría a incurrir en errores al momento de tomar decisiones.

La manera de cuantificar estas variables es simple, para el caso de las edades tomaremos el valor absoluto de cada recurso y para el género usaremos el sistema binario.

Fuente de la información: Recursos Humanos

6.1.2 Muestra Estadística

En todo proceso estadístico mientras mayores datos sean utilizados en el modelo para las mismas variables ganaremos en precisión. El ideal sería estudiar a toda la población en todas sus circunstancias, pero como definidos anteriormente agregar más variables al modelo demandaría un esfuerzo mayor en la búsqueda de los datos para completar tales dimensiones. A efectos de cuantificar la cantidad de datos necesarios para formar un modelo estable, seguimos la siguiente ecuación.

$$\mu_n^* = \sum_{i=1}^{n} X_i/n$$

Estimador del tiempo medio de permanencia.

$$SD(\mu_n^*) = SD(X)/\sqrt{n}$$

Desvío del estimador.

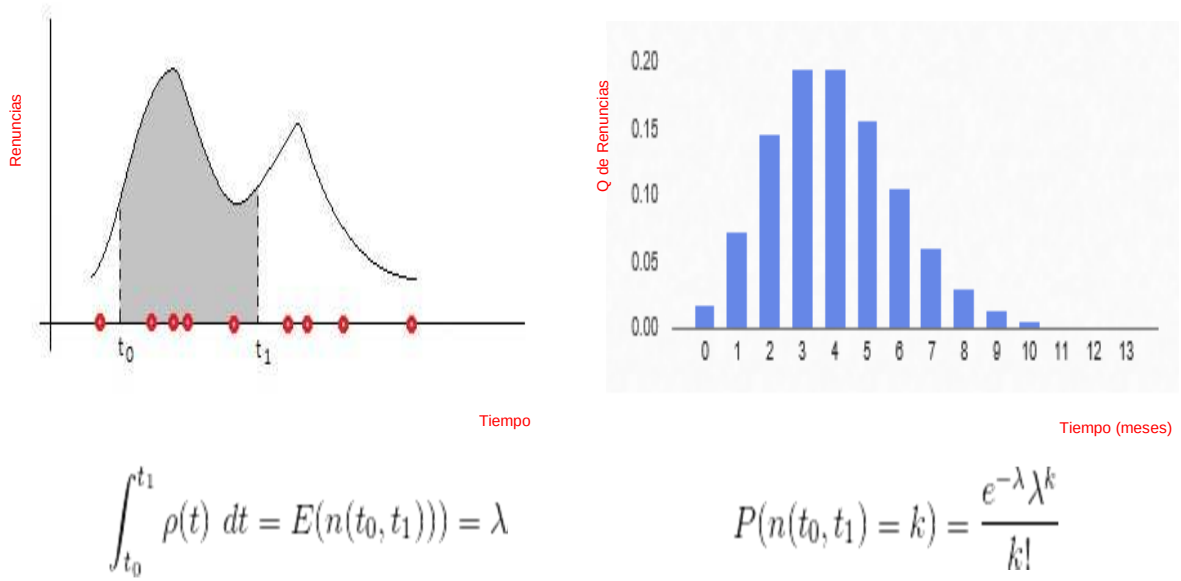
$$Var(\mu_n^*) = Var(X)/n$$

El error estándar decrece a medida que aumenta la muestra.

VII. Cálculo N muestral – Fuente: The Elements of Statistical Learning.

6.1.3 Modelo Poisson de distribución de probabilidades.

En teoría de probabilidad y estadística, la distribución de Poisson[1] es una distribución de probabilidad discreta que expresa, a partir de una frecuencia de ocurrencia media, la probabilidad de que ocurra un determinado número de eventos durante cierto período de tiempo. Justamente nuestra intención es poder predecir cuándo ocurrirá una renuncia, con determinada tiempo de anticipación, pero sería demasiado ambicioso conocer el momento justo, más bien un intervalo de tiempo en que podría suceder con probabilidad de x%. Dicho esto, entendemos que la distribución Poisson es la que mejor se ajuste a nuestra búsqueda por el siguiente análisis.



VIII. Distribución Poisson. Fuente: The Elements of Statistical Learning

Valor esperado de eventos dentro del rango (t_1 ; t_2) **Distribución de la cantidad de renuncias en (t_1 ; t_2)**

---

[1] Ver, **Distribución Poisson**, https://es.wikipedia.org/wiki/Distribución_de_Poisson

A modo de ensayo realizamos el siguiente ejercicio:

Según análisis de planificación operativos para atender a todos los clientes, es requerido tener un 87% de capacidad sobre el staff vigente de 1832 profesionales para llevar adelante el negocio. Es decir que si hubiese ~ 238 renuncias en un tiempo determinado de 3 meses (tomamos este periodo porque representa el tiempo necesario de reclutar y entrenar a un reemplazo), la operatoria sería inviable porque los reemplazos no cubrirían el nivel de actividad necesario. Utilizando la distribución de probabilidades de Poisson, podemos calcular con que probabilidad sucederían ese límite de renuncias.

Siendo, $P(n(t_0,t_1)=k)=\frac{e^{-\lambda}\lambda^k}{k!}$ …donde k es 238 y λ son las renuncias reales ocurridas en el lapso de tiempo estudiado.

Entonces:

$$P\,(n\,(t0,t1)=k)=\frac{e^{-87}*87^{238}}{238!}=0.1503\sim 15\%$$

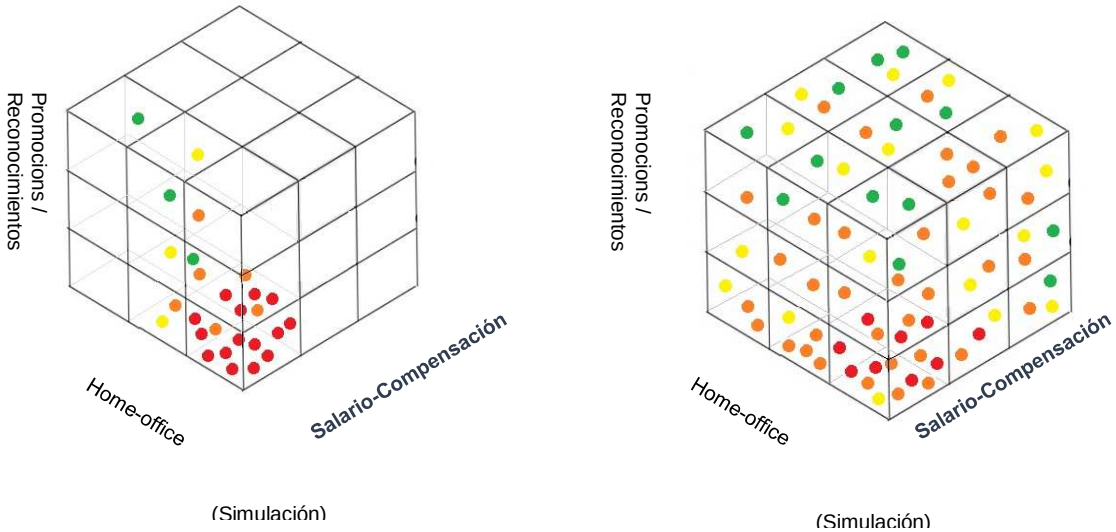
Si bien la probabilidad es relativamente baja, 15%, de ocurrir sería un caso extremo que involucraría un alto costo de resarcimiento a los clientes por no poder cumplir el contrato de servicios.

6.1.4 Selección de Variables Observables Significativas.

Habiendo descripto algunas de las variables identificadas por los Gerentes como críticas, y antes de detallar la forma en que el modelo cuantifica y analiza los datos, definiremos la herramienta a utilizar para determinar que variables son más significativas y como en conjunto influyen en la toma de decisiones de los empleados.

La herramienta elegida es la regresión lineal múltiple, donde la influencia de una variable explicativa (X) en los valores que toma otra variable denominada dependiente (Y). Como vamos utilizar más de una variable explicativa; esto nos va a ofrecer la ventaja de utilizar más información en la construcción del modelo y, consecuentemente, realizar estimaciones más precisas. El objetivo será responder cuales de las variables explicativas tienen mayor peso en el resultado. Con cada

variable explicativa tendremos una nueva dimensión. Aquí encontramos entonces una restricción de suma importancia. Cuando intentamos explicar un modelo podemos vernos tentados a incluir una gran cantidad de eventos que pensamos a priori condicionan dicho resultado. Sin embargo, al hacer esto nos encontramos con el problema de los datos. Al agregar variables es necesario juntar más datos para que el modelo tenga la suficiente información para hacer predicciones. La siguiente ilustración muestra la conocida "maldición de las dimensiones" en referencia al punto tratado.



VI. Dimensiones. Fuente: Data Science for Business

Las variables del modelo se clasifican en cualitativas y cuantitativas. Dentro del grupo cualitativo podemos decir que tendremos variables ordinales como por ejemplo la jerarquía del empleado (analista, especialista) siempre siguiendo un orden como de mayor a menor. Por otro lado tenemos a las cualitativas cardinales como ser el título de grado que poseen cada recurso, o su género. Este concepto representa a las variables **"dummies"**[1] también conocidas como dicotómicas o categóricas. Las cuantitativas pueden ser de tipo continuas o discretas. Del primer tipo encontramos al tiempo de permanencia o el salario. Mientras que de orden discretas tenemos a la cantidad de promociones o reconocimientos que recibió una persona.

---

[1] Ver **Dummy Variable**, Principles of Statistics M.G. Bulmer. ED1965. Dolver Publications.

Otras condiciones como que las variables explicativas no pueden estar altamente correlacionadas entre sí, las relaciones entre las causas y el resultado deben ser lineales, todas las variables deben seguir la distribución normal y deben tener varianzas iguales; se cumplen para el set que estamos analizando.

Antes de realizar la regresión, determinaremos cuantos datos deberíamos obtener para que la muestra sea significativamente estadística. Partiendo de que la población total que ingresará al algoritmo es de 1832 personas, podemos estimar cuanto seria la n muestral necesaria para reducir el desvío del estimador. La teoría que acompaña esta idea y los cálculos para nuestro caso son explicados a continuación.

6.1.5 Regresión Lineal Múltiple.

En estadística la regresión lineal o ajuste lineal es un modelo matemático usado para aproximar la relación de dependencia entre una variable dependiente Y, las variables independientes Xi y un término aleatorio ε. Este modelo puede ser expresado como:

$$Y_t = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

Donde:

- Variable dependiente, explicada o regresando. Y_t
- Variables explicativas, independientes o regresares. $X_1, X_2, \cdots, X_p$
- Parámetros, miden la influencia que las variables explicativas tienen sobre la dependiente.

  $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$

Siendo la ecuación para la regresión lineal múltiple, la siguiente identidad.

$$Y_i = \beta_0 + \sum \beta_i X_{ip} + \varepsilon_i$$

6.1.5.1 Supuestos del modelo de regresión lineal.

Para poder crear un modelo de regresión lineal es necesario que se cumpla con los siguientes supuestos:

- Linealidad: los valores de la variable dependiente están generados por el siguiente modelo lineal: $Y = X * B + U$
- Homocedasticidad: todas las perturbaciones tienen las misma varianza: $V(u_i) = \sigma^2$
- Independencia: las perturbaciones aleatorias son independientes $E(u_i \cdot u_j) = 0, \forall i \neq j$
- Las variables explicativas X_k se obtienen sin errores de medida.

Como nuestro objetivo es identificar cuáles son las variables que mayor relación tienen con el tiempo de permanencia de los empleados (aplicado al set de muestra determinado) de modo que podamos trabajar sobre ellas en términos de recolección de datos, medición e incorporación al algoritmo predictivo de renuncias; utilizaremos el método de la regresión para encontrar esta respuesta. El resultado fue el siguiente para el primer conjunto de variables analizadas:

Estadisticas Regresión	
Multiple R	0,916964207
R^2	0,840823357
R^2 Ajustado	0,790176243

	Coeficiente	*Error Standard*	*t Stat*	*P-valor*	*Bajo 95%*	*Alto 95%*
Interceptor	63,7285	40,2008	1,5853	0,1272	-19,6430	147,0999
X Salario 1	-42,4287	30,2479	-1,4027	0,1747	-105,1590	20,3016
X Home Office 2	10,2607	4,4084	2,3275	0,0295	1,1181	19,4032
X Flexibilidad 3	0,9425	2,5369	0,3715	0,7138	-4,3186	6,2037
X Promociones 4	19,6817	7,2502	2,7146	0,0127	4,6457	34,7178
X Sobrecarga 5	-0,5799	0,2873	-2,0188	0,0559	-1,1757	0,0158
X Madurez Eq 6	0,1492	2,5433	0,0587	0,9537	-5,1253	5,4238
X Distancia 7	1,6004	1,2970	1,2339	0,2303	-1,0894	4,2902

De estos cuadros, en primer término nos interesa el dato del R^2. Este valor explica cuánto las variables independientes explican la variable dependiente, indica el porcentaje de la varianza de la variable dependiente explicado por el conjunto de variables independientes. En nuestro caso un 82% es muy alto para este tipo de modelos, por eso es importante destacar los siguiente. "Correlación no implica causalidad", es decir, inferir que dos o más eventos están conectados causalmente porque ocurren juntos podría ser una falacia. Justamente este punto quedaría evidenciado en el siguiente título donde explicaremos el algoritmo predictivo de renuncia.

Aclarado este tema, tenemos otros resultados interesantes para analizar en nuestra regresión. Por ejemplo veamos que sucede con las variables de manera individualmente. Empecemos por el salario, aparentemente los resultados marcan que esta variable está en relación inversa al tiempo de permanencia y por otro lado tiene un P-valor mayor a 0,05, lo que indicaría que no es influyente. A priori cualquiera podría pensar que el salario es la principal variable para extender la permanencia, y seguramente tenga razón. El motivo por el cual la regresión pareciera mostrar lo contrario son al menos estos: 1. Por lo explicado antes, correlación no implica causalidad y su opuesto es válido también para este caso. 2. Los datos están sesgados del contexto laboral, cuando un profesional decide irse por temas salariales generalmente esta comparado contra el mercado. Esta regresión no tiene datos del mercado y solo considera los relación entre el salario que ofrece la organización contra la permanencia. El error standard es el mayor de todas las variables.

Como variables a destacar podemos mencionar a las promociones/reconocimientos y al trabajo remoto. En el primer caso, el P-valor es mucho menor al 0.05 lo que muestra alta relación y el coeficiente es positivo, es decir, dicha relación es directa. Por otro lado tenemos al trabajo remoto con indicadores similares. Como podemos suponer que estas variable tiene restricciones físicas y de negocios para ser aumentada indiscriminadamente.

Por ultimo algunos análisis generales, la variable sobrecarga de trabajo es entendible tenga coeficiente negativo porque obviamente a medida que aumente, la permanencia del empleado se verá amenazada. A su vez la distancia al trabajo y la madurez de equipo de trabajo muestran relación directa. En todos estos casos el P-valor es mayor a 0.05 pero no por mucho margen. Esto es correcto porque la escala de valores usada para la cuantificación fue inversa. Por ejemplo, el mayor tiempo de recorrido para llegar tenía valor 1 y el menor 5. Para el caso de la flexibilidad horaria parece no ser una variable significativa y hasta podría estar solapada por la de trabajo remoto.

A efectos de convalidar la apreciación acerca de la influencia del salario en el tiempo de permanencia decidimos excluir esa variable de la regresión y analizar los nuevos resultados, siendo estos:

Estadisticas Regresión	
Multiple R	0,909168535
R^2	0,826587425
R^2 Ajustado	0,781349362

	Coeficiente	*Error Standard*	*P-valor*	*Bajo 95%*	*Alto 95%*
Intercept	7,901	5,780	0,185	-4,055	19,857
X Home Office 2	10,480	4,497	0,029	1,176	19,784
X Flexibilidad 3	0,566	2,575	0,828	-4,761	5,893
X Promociones 4	10,693	3,462	0,005	3,531	17,856
X Sobrecarga 5	-0,421	0,269	0,132	-0,978	0,136
X Madurez Eq 6	1,133	2,496	0,654	-4,029	6,296
X Distancia 7	1,226	1,296	0,354	-1,454	3,906

Lo importante aquí es entender como vario el R^2 ajustado, ya que R^2 siempre aumentara al incluir una nueva variable o disminuiría al quitarla. Como observamos el R ajustado nos muestra que si bien los salarios no eran significativos a nivel individual, ayudaban a comprender el modelo en su integridad. Por ende tenemos un descenso de este indicador al eliminar la variable (79% vs 78%). Como conclusiones podemos decir que el salario debe ser parte del algoritmo.

6.2 ALGORITMO PREDICTIVO DE RENUNCIAS

6.2.1 Cadenas de Markov.

La teoría que aplicaremos para el desarrollo de nuestro proyecto es conocida como Cadenas de Márkov. Un modelo de Márkov es un tipo especial de proceso estocástico discreto en el que la probabilidad de que ocurra un evento depende *en gran parte* del evento inmediatamente anterior. No debemos confundir como juega la variable *tiempo* en el modelo, la cual analizaremos a continuación. En este caso el índice temporal t asume un rango continuo (usualmente en los números reales). Un ejemplo simple de un proceso estocástico es una sucesión de ensayos de Bernoulli, por ejemplo, una sucesión de lanzamientos de una moneda. En este caso, el resultado en cualquier etapa es independiente de todos los resultados previos (esta condición de independencia es parte de la definición de los ensayos de Bernoulli). Estas cadenas tienen memoria, recuerdan el último evento y eso condiciona las posibilidades de los eventos futuros con mayor fuerza que

cualquier otro evento anterior. Esto justamente las distingue de una serie de eventos independientes como el hecho de tirar una moneda al aire.

El tiempo no es continuo en nuestro modelo, es necesario particionarlo en la unidad de medida más pequeña posible. Para la cual se tienen registrados datos de cambios ocurridos.

Debido a la recopilación de datos realizada sobre cada evento sucedido y conociendo la historia del sistema hasta el momento actual, intentaremos describir en probabilidad su estado futuro, o sea si sucederá una renuncia.

$$P(X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_2 = x_2, X_1 = x_1) = P(X_{n+1} = x_{n+1} | X_n = x_n).$$

La entidad mostrada es conocida como la propiedad de Márkov

Una cadena de Márkov es de carácter homogénea, cuando la probabilidad de pasar del estado i (permanencia) al estado j (no permanencia o renuncia) en un paso no depende del tiempo en el que se encuentra la cadena. Evidentemente esta afirmación se aplica para nuestro estudio debido a que el tiempo es una dimensión del modelo y el momento 0 no es igual al momento 1 por el deterioro de las condiciones de cada recurso.

- La probabilidad de pasar del estado i al estado j, en x unidades de tiempo es:

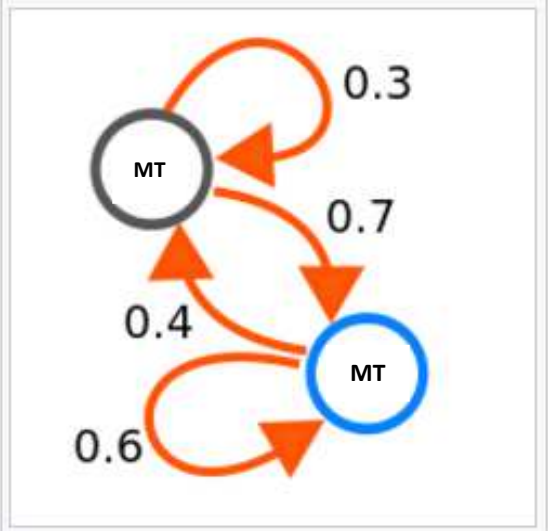
$$P_{ij}^{(n)} = \Pr(X_n = j \mid X_0 = i)$$

6.2.2 Matriz de Transición y Estados.

En la realidad organizacional existen solo dos estados posibles para cada profesional en nuestro estudio. Uno estado es la permanencia, el otro su opuesto o sea la renuncia. El estado inicial seria la permanencia, para que un individuo decida cambiar su estado, dependerá de cómo fluctúe su matriz de transición. Llamamos matriz de transición a la combinatoria de las variables identificadas como significativas en la decisión. Los elementos de dicha matriz determinan las probabilidades de que en el próximo valor el estado del sistema cambie.

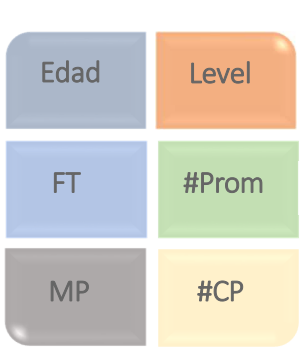
Los círculos A y B se denominan nodos y representan los estados del proceso, las flechas que van de un nodo a sí mismo o al otro son los arcos y representan la probabilidad de cambiar de un estado al otro.

Definición: Consideremos un proceso de Márkov en que el sistema posee varios estados posibles, dados por los símbolos MT (matriz de Transición). Denotemos IJ(p) a la probabilidad de que el sistema pase al estado j después de cualquier cambio en donde su estado era i antes del mismo. Los valores IJ (p) se denominan probabilidades de transición y la matriz P (ij) = p se conoce como matriz de transición del sistema.



IX. Matriz de Transición. Fuente: ***Markov Chains – Statistical Laboratory.*** *James R. Norris. ED1997*

Ejemplo de la matriz de transición:



Level: es la jerarquía de cada recurso desde analista junior a especialista sénior.

FT: es el tiempo flexible medido en cantidad de días de trabajo remoto.

MP: representa la madurez del equipo donde trabaja el profesional.

#CP: son la cantidad de reconocimientos que recibió el profesional.

#Prom: cada ascenso de categoría que tuvo el empleado.

A esta altura podemos inferir que nuestro modelo representara un proceso estocástico. Intuitivamente podemos decir que es estocástico por 2 razones fundamentales.

1. Que un recurso haya decidido abandonar la organización cuando el conjunto de sus variables marcaba determinado valor combinatorio en un punto específico del espacio tiempo no implica que otro recurso tome la misma decisión en el mismo lapso de tiempo bajo la misma matriz. Sin embargo podemos apoyarnos en la hipótesis de que en algún momento renunciara. El objetivo del algoritmo es estimar ese momento y extenderlo hasta el punto de lo económicamente rentable. Por ejemplo, si un recurso percibiera un salario

de X y según el pasado estuviera en la zona de riesgo, entonces podríamos aumentar su ingreso por 10 veces y muy probablemente logremos extender su permanencia mucho más allá de la media pero no sería rentable para la organización.

2. El subsiguiente estado del sistema está determinado tanto por las acciones predecibles del proceso como por elementos aleatorios. En otras palabras si las variables elegidas para insertar en el modelo son las indicadas y explican el comportamiento de los profesionales en un 100% de los casos, entonces diremos que es determinístico. Aunque para el enriquecimiento del ensayo esperemos no lo sean…

La teoría de la ciencia social estocástica es similar a la "teoría de sistemas" en que los eventos son interacciones de los sistemas, aunque con un marcado énfasis sobre los procesos inconscientes. El evento crea sus propias condiciones de posibilidad, haciéndolo impredecible para las variables que participan de él. La teoría de la ciencia social estocástica puede verse como una elaboración de un tipo de "tercer eje" en el que puede situarse el comportamiento humano en la línea de la oposición o naturaleza irracional, es decir, que algunos actos no tienen comprensión racional a priori.

Un proceso estocástico puede entenderse como una familia no-paramétrica de variables aleatorias indexadas mediante el tiempo t. Para nuestro caso podemos conceptualizar de la siguiente manera.

6.2.3 Combinatoria de Variables en el Espacio-Tiempo.

La combinatoria de valores que pueden tomar las variables definidas en la sección anterior con significatividad estadística conforma una matriz única en el espacio tiempo de cada profesional. A medida que transcurre el tiempo alguna, todas o ninguna pueden variar desembocando en consecuencias siendo la buscada por nuestro modelo la renuncia. Matemáticamente el algoritmo debe trabajar con variables discretas y continuas. Sin embargo la matriz solo podrá tomar valores discretos dentro de un mínimo conjunto numerable, es decir, no acepta cualquier valor solo los que están dentro del conjunto como valores observados sucesivos.

En todo problema combinatorio hay varios conceptos claves que debemos distinguir:

1. Población:

Se llama así al conjunto de los elementos que estamos estudiando. Designaremos con una m al número de elementos del conjunto.

2. Muestra:

Se trata de un subconjunto de la población. Se denominará con la letra n al número de elementos que forman la muestra.

Los tipos de la muestra vienen determinados por dos aspectos:

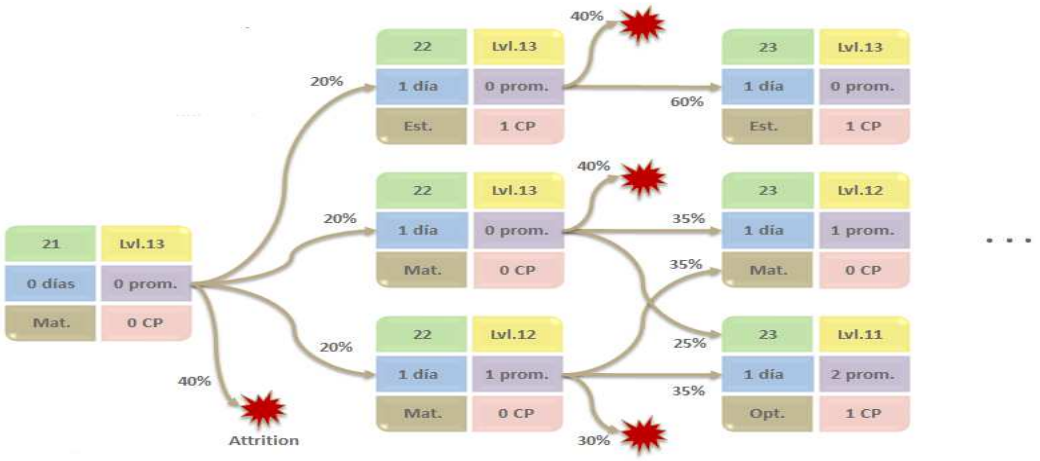
- Orden

Determina si es importante o no que los elementos de la muestra aparezcan ordenados.

- Repetición

La posibilidad de repetición o no de los elementos.

A modo de representación gráfica, el siguiente esquema ejemplifica la forma que estamos estudiando.



t_0 t_1 t_2

Momentos en la carrera del profesional

X. Cadenas de Markov – Algoritmo. Fuente: Sistema Operativo Python

El espacio tiempo del esquema esta discretizado pero cada matriz o combinatoria de valores puede adoptar un solo estado, dicho estado es la sumatoria del valor que cada variable del sistema toma en un determinado momento. A medida que el tiempo avanza, si no hay cambios en las matrices la probabilidad de que un recurso abandone la organización comienza a crecer. Básicamente el modelo contiene toda la información del pasado, donde otros recursos se han encontrado en la misma condición y decidieron dejar la empresa. Con base en esa información, la intención del algoritmo es establecer un punto de comparación con el pasado y ser capaz de estimar el tiempo de permanencia restante si ninguna variable sufriese cambios. Cada mutación de estado debe sumar el 100% de probabilidades, es decir, hay una probabilidad de x% a que el individuo pase a otra matriz y de (1-x) % a que renuncie en un lapso de tiempo (t_0; t_1).

6.3 MATRIZ DE CONFUSIÓN

Esta herramienta sirve para analizar los resultados arrojados por el algoritmo que se emplea con aprendizaje supervisado. Representa una tabla de doble entrada donde veremos en las filas los 2 estados que puede tomar cada individuo, siendo estos de carácter absoluto. No hay estados intermedios. En las columnas tenemos las predicciones hechas por nuestro algoritmo clasificadas en verdaderos y falsos o sea cuando el algoritmo dijo que sucedería una renuncia o una permanencia para un individuo y acertó, eso constituye su porcentaje. Los casos que no acertaron la renuncia o permanencia, entonces son los falsos de nuestra matriz. Evidentemente solo las filas deben tener una sumatoria del 100%, porque indican los resultados para la totalidad de los casos testeados.

Predicciones	Verdadero	Falso
Renuncia	30%	70%
Permanencia	86%	14%

Contar con una matriz de confusión implica la necesidad de tener un set de prueba donde podamos evaluar la precisión de nuestro modelo. Nuestro set de prueba fue la totalidad de la población para un periodo de tiempo de 2 meses. El algoritmo mostrará la probabilidad de renuncia de determinadas personas, configurado con un intervalo de confianza del 90%, en ese lapso de tiempo. Para el armado de la matriz no nos serviría trabajar con probabilidades porque dijimos que los estados eran absolutos. Por ende, asumimos que aquellos con probabilidad mayor al 80% serían renuncias y el resto serian permanencias para la comprobación de los verdaderos y falsos.

En conclusión tendremos 2 resultados que analizar, el primero es con que predicción nuestro modelo predice una renuncia. Según lo observado en la matriz de confusión, hablamos de un 30%. En segundo término tenemos que la precisión para predecir si el candidato va a permanecer es de 86%. A simple vista vemos que el modelo tiene mucha mayor precisión para las situaciones de permanencia, mientras que para las renuncias su eficacia baja significativamente. El periodo de tiempo analizado para determinar los valores de la matriz fue de 3 meses.

6.4 VARIABLES LATENTES

En el titulo anterior estudiamos como ciertas variables del contexto directamente observables y medibles guardan determinados grados de correlación tanto positiva como negativa con el fenómeno que intentamos predecir e influenciar, es decir, las renuncias. A su vez descubrimos que las variables tienen un límite de saturación y que no podemos darles valores más allá de cierto límite por razones económicas o de otras índoles, lo que nos obliga a encontrar el punto de mejor optimización de recursos disponibles. La construcción de un modelo con métodos matemáticos nos arrojó resultados interesantes que pueden ser utilizados por los tomadores de decisión a la hora de administrar a sus equipos de trabajo. No obstante, la cantidad de falsos positivos que muestra la matriz de confusión sigue siendo considerablemente alto, siendo el nivel de precisión hasta aquí del algoritmo de 47%, el cual combina los resultados de predicciones tanto de renuncia como de permanencias para un determinado lapso de tiempo.

Como el modelo desarrollado no explica con exactitud la totalidad de casos, esto nos lleva a pensar que probablemente haya otras razones que influyen en las personas cuando deciden abandonar la organización. Más complejo aún, los comportamientos humanos podría estar predefinidos por reglas que actúan como limitantes de la toma de decisiones y en cuyo caso trabajar sobre las variables observables endógenas en la empresa podría no ser suficiente para retener al profesional durante un tiempo adicional. Nuestra teoría en esta línea intentaría explicar, por qué si la matriz de estados ofrece valores iguales para dos individuos diferentes, sucede que uno renuncia y otro permanece en la organización durante un mismo periodo de tiempo analizado. El objetivo de este segundo capítulo será justamente reconciliar las variables observables endógenas con un conjunto de variables desconocidas en ese instante pero que a priori intuimos que están ocultas y que llamaremos ***latentes***.

6.4.1 Definición de Variables Latentes

Una variable latente por definición es toda aquella que no se observan directamente sino que son inferidas, a veces pertenecen a aspectos de la realidad física que son medidos indirectamente pero otras veces corresponden a conceptos abstractos como categorías, estados de comportamientos o estructuras de datos. Buscaremos que estas variables latentes nos hablen sobre perfiles de personalidad y comportamientos que entregarían información útil para conformar una red alrededor de la matriz de estados. En otras palabras, estaremos buscando los límites al espacio tiempo de nuestro modelo de Márkov que nos mostrará cuando las mejoras en variables observables dejan de funcionar porque entran en conflicto con rasgos de la personalidad de los profesionales.

Como primer paso debemos definir qué características distintivas de la personalidad pueden adaptarse mejor a las condiciones del ambiente y tipo de trabajo, siendo estos últimos factores flexibles pero conformados por el colectivo y no alterados individualmente. Para el punto de partida en la búsqueda de la información necesaria, decidimos realizar entrevistas con gerentes y líderes de equipos para que estos expliquen qué características observan en los integrantes de sus equipos. De más está aclarar que puede existir más de un perfil apto para el trabajo y a los cuales algunas variables observables lo afecten más que otras. La competencia laboral es un comportamiento, o un conjunto de comportamientos necesarios para llevar a cabo la tarea de trabajo específica o meta, el nivel más básico de la habilidad que se requiere para obtener un rendimiento de trabajo exitoso. Distinguir las características que tienen los candidatos será el primer paso para construir del algoritmo y la base para clasificar las variables latentes. El siguiente cuadro muestra lo que cada entrevistado relevo de sus profesionales, las reuniones fueron individuales.

	Entrevistado 1	Entrevistado 2	Entrevistado 3	Entrevistado 4	Entrevistado 5
1	Responsabilidad	Orden	Comunicación	Inteligencia	Pro actividad
2	Compañerismo	Inteligencia	Puntualidad	Enfoque	Diligencia
3	Integridad	Pro actividad	Entusiasmo	Orden	Colaboración
4	Predisposición	Responsabilidad	Prudencia	Carisma	Inteligencia

A simple vista vemos que hay cierto consenso en cuáles son las características que describen a la población bajo estudio, y si bien el orden de prelación puede variar para cada entrevistado, la tendencia es evidente.

6.4.2 Introducción al Concepto de "Proxy"

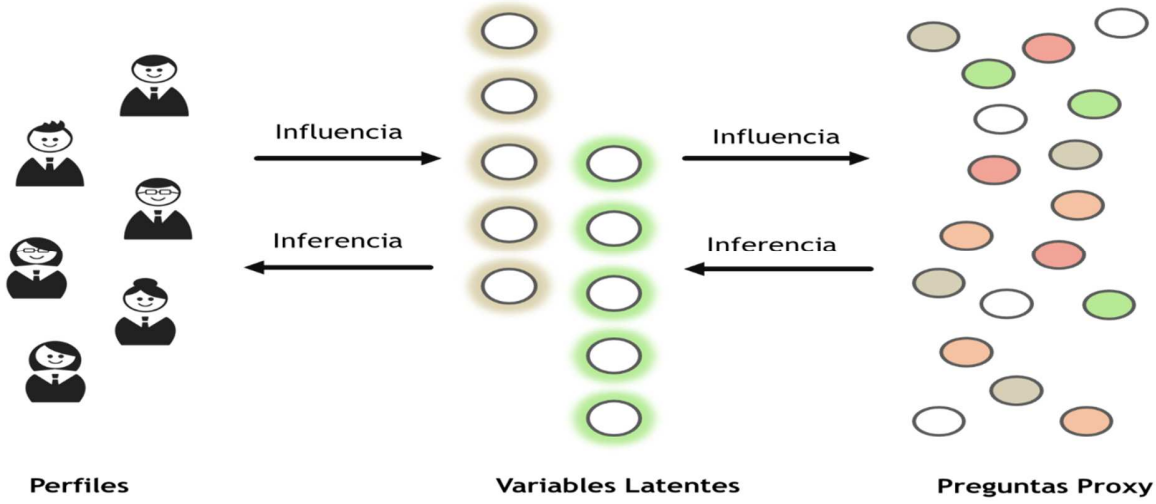
Ya tenemos idea de que buscar pero como mencionamos antes este tipo de variables son de carácter latentes, es decir, no podemos observarlas directamente. Por ende es imperativo encontrar un método que nos permita medir estas variables y posteriormente incluirlas al modelo que desarrollamos en la Sección I. El universo de características de la personalidad no tiene definido sus límites, sin embargo nuestra modelo intentará buscar a través de preguntas proxys cuáles son las características claves que mayor influencia tiene en el ámbito de trabajo.

El conjunto de preguntas será aplicado en primera instancia a un grupo de entrenamiento del modelo. Este es un punto fundamental de nuestro análisis, porque justamente ese grupo estará

conformado por aquellos profesionales que mayor contacto tienen con sus gerentes y por ende de quienes mayor conocimiento se tiene en cuanto a los rasgos de sus personalidad. Por otro lado es importante también determinar la cantidad de participante, es decir, para que el grupo satisfaga las restricciones de la ***n*** muestral en estadística.

Para nuestro trabajo utilizaremos el concepto de agente Proxy. En estadística una variable proxy es algo que de por sí no tiene gran interés, pero de la cual se pueden obtener otras de mucho interés. Para que esto sea posible, la variable proxy debe poseer una fuerte correlación, pero no necesariamente lineal o positiva, con el valor inferido. No tiene ningún valor si los datos no se ajustan a alguna relación (los datos se representan en una nube de certidumbre). A través de las variables proxy intentaremos acceder a la información que buscamos sobre los rasgos de la personalidad que podrían incidir en la renuncia. Este método de infiltración es necesario porque obviamente no podríamos preguntar a un profesional si tiene pensado renunciar y en qué momento. En caso de recibir respuesta seria inconducente por varios motivos.

Representación gráfica del modelo:



XI. Modelo Proxy. Fuente: Elaboración Propia.

Adentrándonos en los detalles vemos viable la configuración de una lista de preguntas que llamaremos proxy para entender a que nos referimos según lo explicado en el párrafo anterior. El conjunto de preguntas serán preparadas en relación con determinadas características buscadas y las mismas serán entregadas para responder a la misma muestra de personas, intercalando las preguntas a responder. Lo importante es que muchos individuos contesten las mismas preguntas para poder parametrizar y comparar resultados. Desde ya que las preguntas no podrán ser de libre respuestas, sino que debe haber 2 opciones para responder que indiquen un solo estado en su valor positivo o negativo según las respuesta elegida. Con esta restricción nos será posible tabular los resultados e ingresarlos al sistema operativo (Python) en el lenguaje adecuado. Esta restricción está sustentada en el modelo de prueba de Fisher.

La prueba de Fisher[1] es útil para los datos categóricos que resultan de clasificar los objetos en dos formas diferentes (en nuestro ejemplo, responsable o no responsable), se utiliza para examinar la significación de la asociación (de contingencia) entre los dos tipos de clasificación. Así en el ejemplo de modelo, uno de los criterios de clasificación podría ser si la persona es responsable, y el otro podría ser si el gerente piensa que es responsable o no. Queremos saber si estas dos clasificaciones están asociados —es decir, si el gerente puede realmente decir si el profesional es responsable o no. La mayoría de los usos de la prueba de Fisher implican, como en este ejemplo, una tabla de 2×2 de contingencia. El valor de p de la prueba se calcula como si los márgenes de la tabla son fijos, es decir, como si en el ejemplo, el gerente sabe el número de profesionales con cada característica y por lo tanto proporcionará conjeturas con el número correcto en cada categoría. Como se ha señalado por Fisher, esto conduce bajo una hipótesis nula de independencia a una distribución hipergeométrica de los números en las celdas de la tabla.

A continuación listamos a modo de ejemplo una serie de preguntas[2] que fueron preparadas para los profesionales seleccionados. Es importante aclarar que para realizar esta actividad, solo se comunicó que contesten las preguntas pero nunca se mencionó el objetivo del proyecto.

Pregunta	Indicativo de…
Si tuvieses que levantar un punto, y supieses que personas importantes podrían oponerse, qué harías? **A. Levantar el punto** **B. No hacerlo para evitar confrontación**	Responsabilidad
En ocasiones todos pueden tomar ventaja sobre el empleador en determinado aspecto, sin conocimiento de este. Como actuarias? **A. Aprovechar la situación.** **B. No aprovecharla y comunicar.**	Integridad
Desarrollar una buena relación con el cliente es importante y a veces clave para el negocio. Si un cliente no prestara atención a tus solicitudes o comunicaciones. Como reaccionarias? **A. Persistirías en la fomentar la relación.** **B. Pedirías a un compañero que trabaje con este cliente.**	Colaboración
Si tuvieses que resolver un problema poco definido y no documentado, que opción tomarías? **A. Escalar el asunto a un superior.** **B. Ejecutar una acción diseñada por ti.**	Proactividad

Si para realizar un proceso contases con información detallada de como ejecutarlo, luego de reiteradas ejecuciones… **A. Seguirías el proceso de igual modo.** **B. Cambiarias la forma de operar.**	Flexibilidad
Si un compañero de trabajo no cumpliese rigorosamente con los horarios de las reuniones. **A. Estarías molesto** **B. No te afectaría**	Puntualidad
En caso de haber tareas no realizadas pero tampoco asignadas a un responsable. **A. Realizarías las tareas que puedas.** **B. Alertarías que no se están realizando.**	Predisposición
En que circunstancia es más efectiva la comunicación escrita que la verbal? **A. Para comunicar sobre problemas.** **B. Siempre.**	Comunicación
Si tuvieses un entregable con mucho detalle técnico pero con poco tiempo para realizarlo. **A. Pedirías más tiempo para realizarlo** **B. Cumplirías con el plazo, atendiendo a los puntos más importantes.**	Enfocado en detalles
Si detectases un error de calidad que podría ocasionar una perdida para la compañía, que harías? **A. Intentarías resolverla sin llamar la atención.** **B. Alertarías del problema.**	Orientación en calidad
Cuando el supervisor directo está equivocado y tú tienes conocimiento de su error… **A. Lo corregirías en público/privado.** **B. No le comentarías su error y lo corregirías por tu lado.**	Toma de riesgos

---

[1] Ver Prueba de Fisher, **"The Elements of Statistical Learning".** Trevor Hastie, Robert Tibshirani, Jerome H. Friedman. ED2001

[2] **Preguntas del cuestionario,** elaboración del propio grupo observado y corregidas por el autor. Los cuestionarios intercalaban las preguntas de manera aleatoria antes de su entrega para ser contestados.

El siguiente paso de nuestro proyecto de variables latentes será confeccionar una tabla de valores para cada dimensión de personalidad, dicha tabla debe ser ponderada por los Gerentes según las cualidades que observen en los candidatos basándose en el conjunto de variables latentes definidas. Por ultimo debemos determinar cuantas preguntas debe contestar cada persona. Esta cantidad es importante porque nos permitirá trabajar para disminuir el error de la estimación hasta un margen aceptable, que podemos llamar α. A continuación debemos proveer a todos los integrantes del grupo de entrenamiento las preguntas realizadas para que sean contestadas. Es importante entender cómo funciona el proceso de entrenamiento o supervisado. Como conocemos las variables latentes de estos candidatos a través de los gerentes podremos determinar los aportes que cada pregunta proxy haga. Es también probable que haya variables latentes que no puedan ser encontradas a través de ninguna pregunta. Una vez que este subconjunto haya contestado intentaremos calcular la vinculación entre las respuestas a las preguntas proxy y su correspondiente variable latente.

6.4.3 Inferencia de Variables Latentes. Modelo Matemático.

Para continuar con la construcción del modelo debemos buscar una función para calcular como una pregunta proxy se vincula con una variable latente en términos de probabilidad. Supongamos que una pregunta pensada para inferir la característica “responsabilidad” (variable latente) es contestada en su opción A y B con una distribución del 50% en cada una y entre el grupo encuestado hay personas definidas como responsables y no, si las respuestas fuese intercaladas entre estas dos condiciones (que contestaron y quienes contestaron) entonces podríamos decir que esta pregunta no nos serviría para vincularla a dicha variable. La razón es porque evidentemente no podemos asociar a los responsables con la respuesta A, por ejemplo, porque contestaron indistintamente entre A y B.

En términos matemáticos, el proceso es el siguiente. Se calcula la probabilidad de ocurrencia de un estado latente por medio de la regla de Bayes:

$$P\left(v = v' \middle| r_1 = r'_1, \ldots, r_n = r'_n\right) = \frac{P\left(v = v'\right) P\left(r_1 = r'_1, \ldots, r_n = r'_n \middle| v = v'\right)}{P\left(r_1 = r'_1, \ldots, r_n = r'_n\right)}$$

Esta fórmula indica la probabilidad de que una variable latente tenga cierto estado, sabiendo cuales respuestas dio a las preguntas el sujeto bajo estudio, siendo, r = respuestas y v = variables. El lado izquierdo de la igualdad representa una probabilidad conjunta que no se puede medir de forma directa por falta de datos suficientes que aporten significatividad estadísticas.

Entonces, la formula se factoriza a través de ciertos supuestos de independencia de las variables lo cual permite realizar el cálculo con los datos disponibles.

Se realiza la siguiente aproximación, que asume que las respuestas a las preguntas son entre sí independientes.

$$\cong \frac{\sum \{ v = v' \} \pi_i P(r_i = r_i' \mid v = v_i')}{\pi_i P (r_i = r_i')}$$

Este supuesto es obviamente falso, porque no siempre son independientes aunque haya muchos casos en que sí. La aproximación tiene como secuela que se subestime la probabilidad de ocurrencia de varias respuestas inusuales en la misma persona y se sobreestime la probabilidad de que todas las respuestas pertenezcan al grupo de las más comunes. No obstante, la aproximación es necesaria para poder realizar la estimación del clasificador con suficiente significatividad estadística.

Se emplean los siguientes estimadores para calcular las probabilidades de la fórmula:

$$\mathrm{P}\ (r_i = r_i') = \frac{\sum k\ I\ \{ r_i = r_i' \}\ (R_k)}{k}$$

Reemplazando k, en la primera fórmula mostrada entonces:

$$\mathrm{P}\ (r_i = r_i' \mid v = v') = \frac{\sum k\ I\ \{ r_i = r_i' \mid v = v' \}\ (R_k)}{\sum k\ I\ \{ v = v' \}\ (R_k)}$$

La varianza de este tipo de estimadores se puede acotar por:

$$\mathrm{V}\ (\mathrm{P}) = \frac{1}{k}\ p(1-p)\ \leq\ \frac{1}{4k}$$

Esta fórmula es utilizada para saber cuántas personas deberían contestar las mismas preguntas y de esta manera entender si dicha pregunta tiene significatividad estadística, es decir, infiere una variable latente. Siendo k, la cantidad de personas que contestaron y p la siguiente igualdad:

$$p = P\ (r_i = r_i')$$

Luego determinamos el desvío standard[1]:

$$\text{SD (P)} \leq \frac{1}{2\sqrt{k}}$$

Donde k es la cantidad de respuestas consideradas. Entonces imponemos la condición:

$$\text{SD (P)} \leq \frac{1}{2\sqrt{k}}$$

Garantiza que,

$$\text{SD (P)} \leq \ \alpha$$

Siendo α el error maximo que podría tolerar el modelo, el cual puede controlar el error de estimación.

---

[1] Desvio Standard, Ver **"Principles of Statistics".** M.G. Bulmer. ED1965. Dolver Publications.

Por otro lado, es necesario poder definir (dada una cantidad prefijada de preguntas y la significatividad alpha) cuantas preguntas deben ser contestadas por cada persona de la población para que se obtenga una significatividad aceptable. Esta cantidad resulta ser, para el caso de una única variable latente:

$$[\,\mathrm{R}\,] = \frac{1}{4\alpha^2}\,\frac{|\,P\,|}{\sum k\; I\,\{\,v = v'\,\}\,(Rk)}$$

Pero para el caso de que existan muchas variables latentes que el estudio aborda simultáneamente, entonces la formula se generaliza a:

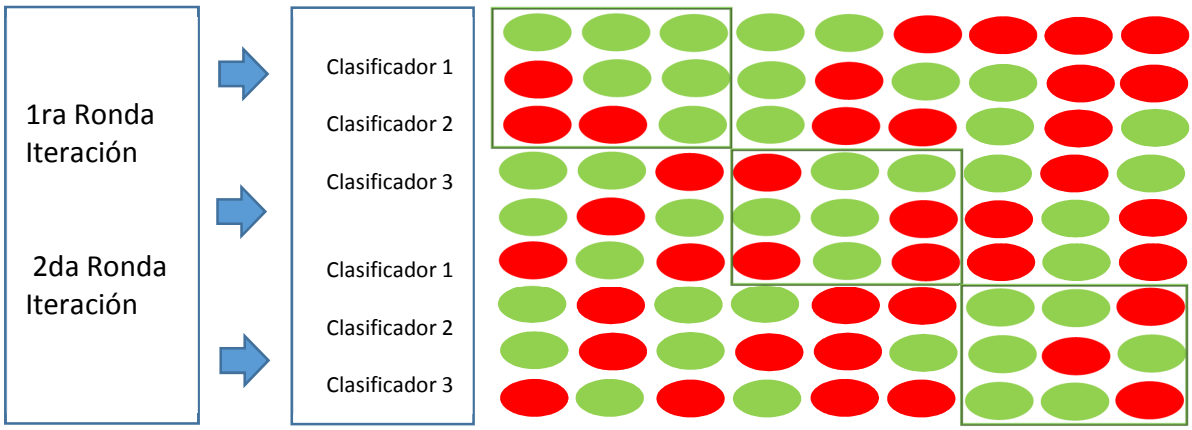
$$[\,\mathrm{R}\,] = \frac{1}{4\alpha^2}\,\frac{|\,P\,|}{\min i \sum k\; I\,\{\,v = v'\,\}\,(Rk)}$$

Una vez determinada la cantidad de preguntas que nos asegure la significatividad estadística, reduciendo el error de estimación y por otro lado, la cantidad de candidatos que deberán ser el conjunto de entrenamiento podemos tomar las respuestas y tabularlas según la tabla de valores que determinaron los gerentes y encontrar la vinculación de una variable latente con cada pregunta proxy. De este proceso podremos determinar que preguntas serán eficientes desde el punto de vista del modelo y descartar las preguntas inconducentes.

La validación cruzada es una técnica utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba. Para nuestro proyecto el conjunto denominada de entrenamiento no puede ser elegido de manera aleatoria porque recordemos la necesidad de crear la preguntas proxys trabajando sobre el conjunto de personas conocidas por los gerentes. Dicho esto, la forma de estimar el error del modelo deberá hacerse con una validación cruzada dejando afuera el grupo de entrenamiento y para la prueba al resto de la población. Debemos realizar tantas iteraciones como muestras (N) tenga el conjunto de datos, de forma que para cada una de las N iteraciones se realiza un cálculo de error. El resultado final lo obtenemos realizando la media aritmética de los N valores de errores obtenidos, según la fórmula:

$$\mathrm{E} = \frac{1}{N}\sum_{i=x}^{n} \mathrm{E}_i$$

El siguiente grafico representa el mecanismo de validación cruzada:



XII. Validación Cruzada. Fuente: Data Science for Business-

Para ejecutar la validación cruzada, tomamos un conjunto de persona quienes deberán responder las preguntas "proxy", a través de ajuste de la estimación aplicado con las ecuaciones demostradas somos capaces de entender si tales preguntas infieren variables latentes en base a las respuestas. Los círculos verdes representan los casos en donde la inferencia marcaba la presencia de la variable latente en sentido afirmativo y esto era constatado con el dato aportado por los gerentes, mientas que los rojos son los casos de no coincidencia. A continuación tomamos un grupo (llamado de prueba) y comprábamos la precisión de nuestro modelo al extrapolar el ejercicio al resto de los integrantes. Este proceso deber ser iterado n veces según la ecuación.

Con todas las respuestas recolectadas, es momento de tabularlas y buscar los patrones de comportamiento para incluirlos en la base de datos que alimenta nuestro algoritmo junto con las variables observables. Para tal propósito, debemos construir nuevas matrices de transición para cada persona. La información encontrada sobre las personalidades funcionará como cota para los valores que la matriz pueda tomar siempre de manera eficiente. Por ejemplo, aquella persona que tenga como rasgo distintivo la proactividad y curiosidad por el cambio en su matriz de transición la variable de rotación deberá tener valores altos para extender el tiempo de permanencia. Por otro lado, para quienes la ambición y crecimiento sea trascendental entonces el nivel de jerarquía y las promociones serán sus variables observables más significativas. Estos son solo ejemplos de cómo funcionaría esta parte complementaria del algoritmo. La clave es poder transformar la información para expresarla en valores que puedan interpretarse en acciones eficientes para aumentar la retención del talento.

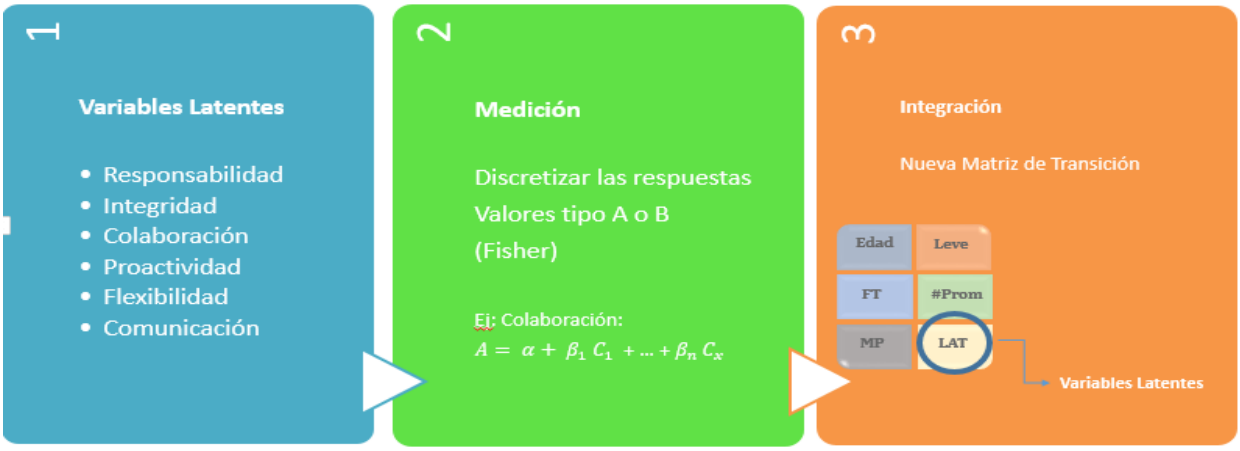
7 INTEGRACION DE LAS VARIABLES LATENTES AL MODELO

7.1 METODOLOGIA

Hasta el momento hemos trabajado en la identificación de las variables que afectan la retención del talento, sean estas observables directamente o latentes. Los diversos métodos desarrollados en la secciones I y II nos permiten aproximarnos a su medición y entender que rol juegan cuando una persona decide abandonar la organización. Sin embargo no hemos analizado aun como sería el algoritmo que integre a ambas variables dentro del mismo modelo predictivo y si sus resultados serían más precisos que de manera individual. El propósito de este título es estudiar la forma de conectar ambo conjunto de variables haciendo uso de los modelos descriptos anteriormente.

En primera instancia debemos trabajar sobre las variables latentes, siendo necesario ordenarlas, categorizarlas y atribuirles algún valor dentro de una escala. Esto es necesario porque partimos de la idea de agregar un nuevo estado a la matriz analizada en el punto **3.4. Matriz de Transición**. Como estudiáramos, las variables observables fueron discretizadas para que cada una de ellas tomara un único valor en el espacio tiempo del profesional en la organización. Por ende, debemos realizar el mismo procedimiento para las variables latentes.

Al respecto mostramos los pasos a seguir…



XIII. Integración de los Algoritmos

En el paso 3, el recuadro de la matriz de transición seleccionado representa el conjunto de variables latentes que determinada persona tiene en un momento dado. Para sintetizar esa información es necesario hacer clusters donde agrupemos todas las variables latentes y podamos asignarle un valor de ingreso al sistema. Con este último paso, estamos en condiciones de volver al punto de las cadenas de Markov para ejecutar el algoritmo pero esta vez con las variables latentes incluidas.

7.2 VARIABLES EXOGENAS

7.2.1 Redes Sociales y Recolección de Grandes Datos.

Hemos mostrado hasta el momento que una amplia gama de atributos personales aplicables al ámbito laboral pueden ser automáticamente inferidos usando cuestionarios con variables proxy. Esta información es incluida en nuestro modelo a efectos de predecir potenciales renuncias. Sin embargo, es posible encontrar otra fuente de información relevante para el algoritmo predictivo y la planificación de operaciones en esta época, que en cierta forma actúa como un sustituto más abarcativo. Estamos hablando de los dispositivos digitales en conjunto con el sistema de redes sociales, los cuales son capaces de contribuir muchos datos acerca del comportamiento individual. Existen en la actualidad varios sitios de Internet, como por ejemplo LinkedIn, Facebook o foros en general, donde se realizan gran cantidad de búsquedas y los profesionales comparten sus opiniones. Los comentarios y recomendaciones laborales pueden contabilizarse y ayudar a mejorar los modelos a través de añadir otras dimensiones a la matriz de estados. Los registros digitales sobre comportamientos podrían proveer bases confiables para medir potenciales cambios buscados por los usuarios. Más aun, las inferencias basadas en observaciones de tendencias en las redes abrirían nuevas puertas para la investigación de la psiquis humana en cuanto a su ámbito de desarrollo.

Una creciente proporción de actividades humanas, tales como interacciones sociales, entretenimiento y opiniones son registradas por los nuevos dispositivos electrónicos. Es necesario en principio distinguir entre la información fácilmente colectable y aquella que debe estadísticamente ser predicha. El diseño del modelo estaría sustentado en la identificación de atributos y comportamientos recolectados en base al tipo de búsquedas que realizan en línea.

7.2.2 Breve descripción del modelo basado en redes sociales.

El estudio debe basarse en una muestra N de usuarios (quienes deben brindar su consentimiento para suministrar el historial de navegación) que utilicen los sitios web para estar en contacto con el mundo profesional y de oportunidades laborales. Dichos usuarios son incluidos en una matriz de doble entrada, cuyos valores serán establecidos con 1 si se interesaron en otros trabajos y 0 en caso contrario, mientras estuvieron navegando. Las dimensiones de la matriz serán reducidas usando el modelo de Descomposición en Valores Singulares (SDV). Este concepto pertenece al algebra lineal y tiene como objeto la factorización de una matriz real o compleja.

Dada una matriz real $A = R^{n*m}$, los autovalores de la matriz cuadrada, simétrica y semidefinida positiva $A^tA = R^{n*m}$ son siempre reales y mayores o iguales a cero. Teniendo en cuenta el producto interno canónico vemos que:

$(A^t\ A)^t = A^t(A^t)^T\ =\ A^tA$ o sea que es simétrica y

$(A_x, A_x) = x^tA^tAx = ||Ax||^2 > 0$ es decir que A^TA es semidefinida positiva y todos sus valores son mayores a 0.

Si α^i es el i-ésimo autovalor asociado al i-ésimo autovector, entonces $\alpha^i \in R$. Esto es una propiedad de las matrices simétricas. Por definición:

Sean $\alpha_1 > \alpha_2 > \ldots > \alpha_n > 0$ los auto valores de la matriz A^tA ordenados de mayor a menor. Entonces $\sigma_i = \sqrt{\alpha_i}$ es i-ésimo Valor Singular de la Matriz A.

En paralelo y teniendo en cuenta que las características o variables observables y latentes ya fueron estimadas con los otros modelos (Regresiones, Preguntas Proxys y Cadenas de Markov), entonces este modelo utilizando la descomposición de valores singulares intentara estimar las mismas variables pero usando otros datos. Puede simbolizarse y abreviarse con esta representación:

N busquedas

1812 Usuarios

	Responsabilidad	Estabilidad Emoc.	Inteligencia	Genero	(...)
Usuario 1	1	1	0	0	...
Usuario 2	0	1	1	1	...
Usuario 3	1	0	1	0	...
(...)	...	...	...	...	...
Uusario n	1	1	0	0	(...)

Usuarios - Matriz de Busquedas

N Componentes

1812 Usuarios

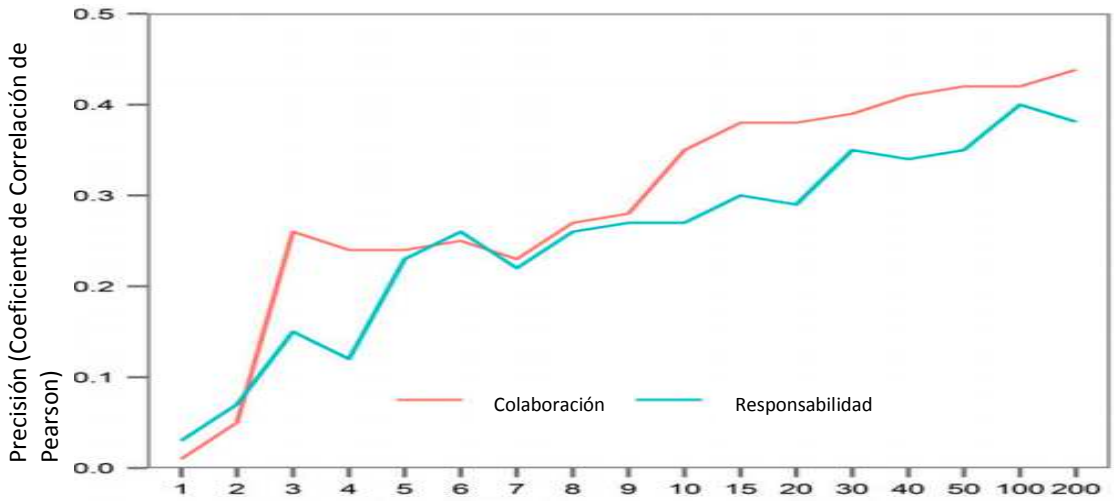
	Comp 1	Comp 2	Comp 3	(...)	Comp n
Usuario 1	1,5	0,7	0,5	0	1
Usuario 2	0,3	0,4	0,2	1	0,8
Usuario 3	0,6	0,1	0,9	0	1,1
(...)	...	...	...	...	...
Uusario n	1,2	0,85	1,05	(...)	1,15

Usuarios -Matriz de Componentes

XIV. Matriz Descomposición Valores Singulares

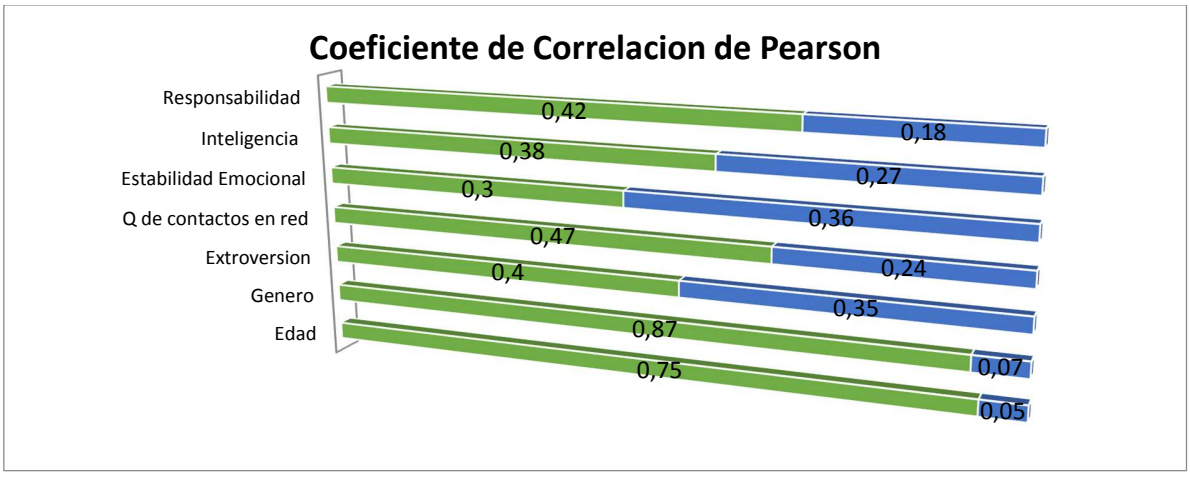
Los componentes están representados por las aplicaciones, comentarios o cualquier otro tipo de rastro que deje la persona al navegar en relación directa o indirecta con puestos de trabajo. (Por ejemplo, cantidad de contactos en redes sociales de profesionales o seguir a compañías en sus posteos). Para determinar el número óptimo de los componentes de la matriz, debemos examinar las predicciones con la validación cruzada como una función del número de componentes.

Curiosamente, al incluir algunos componentes aumenta abruptamente la precisión de la predicción de ciertos rasgos. Por ejemplo, incluyendo más de 3 componentes en el modelo aumenta la exactitud de las estimaciones de apertura de r = 0,1 a r = 0,4. De forma similar, más de 5 componentes aumentan drásticamente la precisión obtenida en la predicción de responsabilidad. Esto sugiere que los componentes particulares están específicamente relacionados con un atributo dado en la matriz similar al usuario. La figura debajo muestra que la exactitud de la predicción aumenta abruptamente en los principios, pero se aplana relativamente temprano (tener en cuenta que la horizontal no es lineal).



XV. Precisión de la Matriz DVS. Fuente:

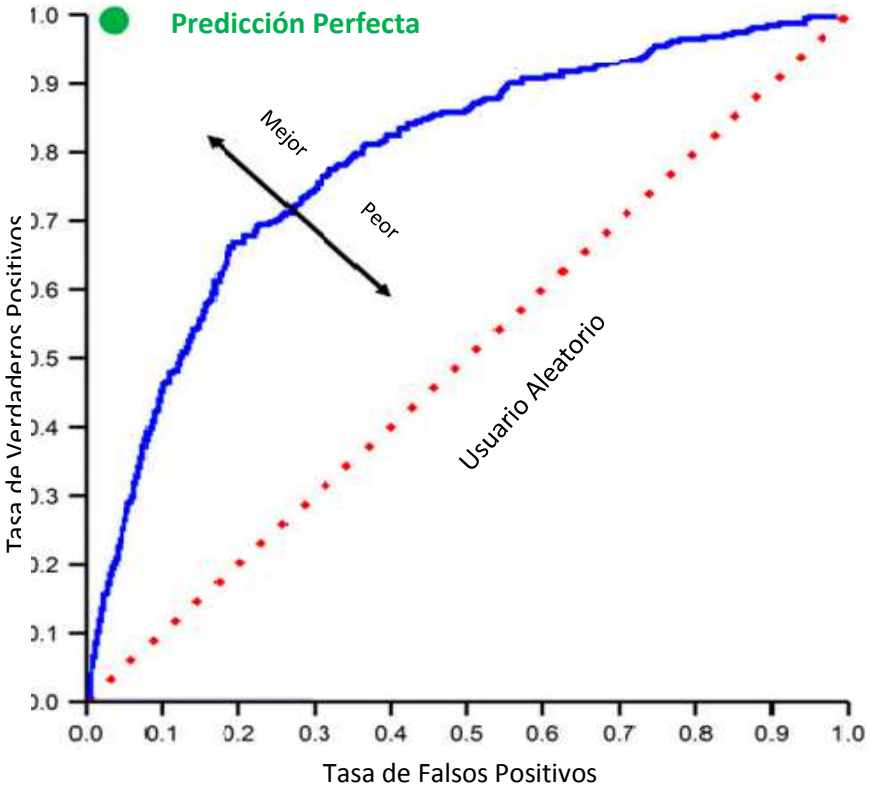
Las precisión de las predicciones de variables dicotómicas es equivalente a la precisión de correctamente clasificar 2 usuarios seleccionados aleatoriamente entre cual está interesado en un nuevo trabajo y cual no. La regresión para atributos numéricos y rasgos de la personalidad medidos en términos de la correlación Pearson entre la predicción y los valores actuales de tales características pude ser inferido a través de la matriz de los usuarios y la descomposición de valores singulares.



XVI. Coeficiente de Correlación de Pearson

Todas las correlación son significativas a niveles de P < 0.1. Las barras transparentes indican la precisión del cuestionario base, utilizado en la sección II – Preguntas Proxy de este trabajo. Evidentemente la mayor correlación corresponde a los atributos de género y edad por ser dicotómicos. Los atributos psicológicos (responsabilidad o extroversión) solo pueden ser aproximados a través de evaluar las respuestas del cuestionario. Las correlaciones con baja precisión en su predicción podrían ser atribuibles a la dificultad para separar la constancia en la variable de determinados saltos en el comportamiento como ser el caso de estabilidad emocional. En definitiva podemos asegurar que el modelo basado en las redes social para determinar comportamientos es en cierta forma la otra cara de nuestro modelo para medir variables latentes no observables directamente. Variables numéricas, como la edad o la inteligencia, se pueden calcular utilizando un modelo de regresión lineal basado en los usuarios k = 100 componentes como covariables. Variables dicotómicas como el género o estabilidad emocional, son modeladas usando la logística de regresión basada en los mismos componentes de DVS. En ambos casos, puede utilizar una validación cruzada de 10 veces para evaluar la predicción fuera de la muestra. Tamaño de los usuarios, y las predicciones para cada subconjunto se calculan basados en parámetros determinados para los usuarios restantes.

La precisión de la predicción se mide de dos maneras. Para el numérico variables, como la edad en años, se reporta el producto de Pearson- Coeficiente de correlación de momento entre el valor real y el Valores entre usuarios. Para las variables dicotómicas como el género, informamos del área bajo la característica de operación del receptor (COR), que puede interpretarse como el coeficiente de probabilidad de clasificar correctamente dos objetos seleccionados aleatoriamente, uno de cada clase.



XVII. Curva de Validación COR. ***Identifying Participants in the Personal Genome Project by Name****.* Latanya Sweeney, Akua Abu, Julia Winn. Harvard College Cambridge, Massachusetts

Representa al conjunto de verdaderos-positivos (o sensibilidad) frente a la tasa de falsos positivos (o 1 especificidad) para detección o clasificación. Los casos positivos son los clasificados por el modelo para pertenecer a una clase objetivo (por ejemplo, "proactivo" o "responsable"). Así, los casos verdaderamente positivos son los casos que fueron correctamente clasificados por el modelo como pertenecientes a una clase objetivo, mientras que falsos positivos fueron clasificados incorrectamente como pertenecientes a una clase objetivo. La tasa de verdadero-positivo es la proporción del número de verdaderos positivos con respecto de todos los casos en la clase objetivo, Mientras que la tasa de falsos positivos es la relación entre el número de falsos positivos al respecto de todos los casos en la clase. Los modelos de regresión logística utilizados en este estudio para predecir la dicotomía resultan de asignar una probabilidad de pertenecer a un objetivo clase para cada uno de los usuarios. Para evitar tener que seleccionar una para asignar usuarios a una categoría de destino determinada, una curva ROC (Receiver-Operator Curve) se puede utilizar para analizar todo el espectro de posibles umbrales En general, las curvas ROC para modelos aleatorios (o nulos) deben ser cerca de la diagonal, porque la probabilidad de ver un verdadero positivo no es mayor que la probabilidad de ver un falso positivo.

Lo más que una curva ROC se inclina hacia la parte superior izquierda, es la precisión del modelo, porque las tasas verdaderamente positivas más altas son logrado para un número dado de falsos positivos. El área por debajo de la curva es igual a la probabilidad que un clasificador clasificará una instancia positiva elegida al azar mayor que una negativa elegida aleatoriamente.

Para el aprendizaje a gran escala, los algoritmos en línea para modelos lineales (por ejemplo, regresión logística) tienen ventajas. Aunque el vector de características x podría tener millones de dimensiones, típicamente cada instancia sólo tendrá cientos de valores distintos de cero. Esto permite un entrenamiento eficiente en grandes conjuntos de datos mediante ejemplos de flujo de disco, ya que cada ejemplo de formación sólo necesita ser considerado una vez.

Es importante destacar que las predicciones de comportamientos humanos soportadas por registros digitales podrían traer implicancias negativas, porque estos son aplicados a un gran número de individuos sin su consentimiento o conocimiento. Es factible imaginar situaciones donde las predicciones lleven a tomar ciertas decisiones que amenazan su libertad de acción sin que las personas sepan sobre ello. Por otro lado, este tipo de connotaciones negativas puede derivar en el abandono de las redes por parte de las personas, para evitar esto es imprescindible que haya transparencia en el uso de los datos recolectados.

8 CONCLUSIONES

Las renuncias en las organizaciones tienen múltiples efectos negativos. El más visible para los directivos suele ser el costo de reincorporación de la nueva mano de obra, pero como hemos presentado en este trabajo no es el único.

El presente trabajo recorrió el análisis de diversos indicadores tales como la productividad, el ausentismo, la impuntualidad, etc., y su valor para estimar las potenciales renuncias. Entendiendo estas últimas como los desenlaces de procesos iniciados tiempo atrás. Cada vez que un empleado decide dejar la organización genera incertidumbre entre sus colegas generando en muchos casos un efecto "dominó". En la mayoría de las ocasiones este tipo de problemas son detectados tardíamente por lo que es difícil realizar acciones preventivas. La identificación temprana de tales acciones es el principal objetivo buscado de esta tesis. Es importante recordar que cuanto más grande sea la organización mayor utilidad tendrá la gestión de datos a través de algoritmos predictivos. Por este motivo es que mediante un abordaje determinista se buscaron las tendencias que desencadenan las renuncias intentando crear un modelo predictivo y probabilístico capaz de concluir dicho comportamiento. También se mencionan los sucesos o variables aleatorias que no pueden ser medidas pero que influencian dichas decisiones y sobretodo complementan el modelo presentado.

El análisis de datos históricos nos aporta información significativa para inferir conductas y decisiones. Sin pretender remplazar el aporte de los datos aleatorios que pueden abordarse más detenidamente desde el área de Recursos Humanos, proponemos juntar ambos modelos para lograr datos más exactos y con esto poder tomar acciones a tiempo para revertir dichas tendencias. Esta tesis buscó complementar y no remplazar la labor de quienes gestionan el capital humano y a la vez ofrecer herramientas sobre todo para organizaciones con gran número de empleados donde se dificulta más conservar y clasificar la información importante contenida en el sistema. Como mostramos en el título "Problemática y Entorno", cada renuncia conlleva un costo de inducción por el reemplazo de aproximadamente 1.200 dólares. Si esta metodología permitiera extender el tiempo de permanencia, la amortización de dicho monto sería más prolongada, representando un ahorro significativo para la organización. El beneficio en mantener una productividad alta y no entrar en los rendimientos marginales decrecientes, es mucho más compleja de medir y solo pareciera ser observable al largo plazo. No debemos olvidar que el nivel de productividad marca el primer indicio antes de una renuncia, aunque los datos muestren que pueden existir distintos perfiles de profesionales que manejen de maneras diferentes este proceso.

A futuro resulta imperativo trabajar sobre el desarrollo de un sistema computarizado que permita procesar todos los datos de los profesionales de manera conjunta en tiempo real y arroje recomendaciones individuales para accionar en cada caso. Es decir, una **"Machine Learning"**

9 REFERENCIAS

[1] **The Elements of Statistical Learning.** Trevor Hastie, Robert Tibshirani, Jerome H. Friedman. ED2001

[2] **Markov Chains – Statistical Laboratory.** James R. Norris. ED1997. University of Cambridge.

[3] **Data Science for Business.** Foster Provost. Tom Fawcett. ED2013. O´Reilly Media.

[4] **Principles of Statistics.** M.G. Bulmer. ED1965. Dolver Publications.

[5] **Probability Theory and Stochastic Processes with Applications (Second Edition).** Oliver Knill

[6] ***Minería de Datos Basada en Sistemas Inteligentes***, Britos P. y Hossian A. Sierra E. 2005.

[7] **Learning Python.** 5th Edition Mike Lutz.

[8] **Identifying Participants in the Personal Genome Project by Name.** Latanya Sweeney, Akua Abu, Julia Winn. Harvard College Cambridge, Massachusetts