REVISTA

TIA

Tecnología, Investigación y Academia -Red Avanzada – RITA

Publicación Facultad de Ingeniería y Red de Investigaciones de Tecnología

Modelos Predictivos para la Rotación del Talento Humano

Citar este documento:

José Luis Barbosa Fontecha, (2021), Modelos Predictivos para la Rotación del Talento Humano. Academia TIA, ISSN: 2344-8288, 8 (1), pp. 54-71- Bogotá-Colombia

Artículo de investigación
Fecha de recepción: 2019-11-30 Fecha de aceptación: 2020-03-30
ISSN: 2344-8288 Vol. 8 No. 1 2020

Modelos Predictivos para la Rotación del Talento Humano

Predictive Models for Human Talent Turnover

José Luis Barbosa Fontecha[1]

Resumen: En este artículo se presenta una investigación de los diferentes modelos predictivos que se utilizan en la rotación de personal en las empresas. Se consultaron artículos científicos que explican los diferentes métodos que se pueden implementar en este fenómeno que tiene gran impacto en las organizaciones, de la misma manera, se investigaron las aplicaciones de estos algoritmos predictivos en las empresas, con la información recopilada se presentaron resultados generales de lo que esperan las organizaciones en sus procesos, costos e imagen, al implementar estos modelos predictivos. Se tomarán los datos del proceso de rotación de personal de la empresa "Colsubsidio" (Caja de compensación social de Bogotá).

Palabras clave: Analítica, Modelo Predictivo, Rotación, Talento Humano.

Abstract: In this article we will present an investigation of the different predictive models that are used in the rotation of personnel in companies. We consulted scientific articles that explained the different methods that can be implemented in this phenomenon that has great impact in organizations, in the same way, we investigated the applications of these predictive algorithms in companies, with the collected information we presented general results of what organizations expect in their processes, costs and image, when implementing these predictive models. The data of the personnel rotation process of the company "Colsubsidio".

Key Word : Analytics, Human Resources, Predictive Model, Turnover.

[1] Universidad Distrital Francisco José de Caldas, jlbarbosaf@correo.udistrital.edu.co

1.Introducción

Las cajas de compensación Familiar son entidades de derecho privado, sin ánimo de lucro, que se hallan sometidas a la vigilancia del Estado a través de la Superintendencia de Subsidio Familiar, entidad adscrita al Ministerio de Trabajo y Seguridad Social. Entre las cuales se encuentra la caja de compensación familiar de Colsubsidio, siendo una de las que presenta un mayor número de afiliados en Colombia, aproximadamente son 1.200.000.

A pesar de que Colsubsidio es una de las empresas más grandes del país, en cuestiones de calidad de la información presenta fallas esto se evidencia en la dificultad de consecución de información para las personas que tuvieron un retiro antes del año 2017. Esta problemática se repite en muchas organizaciones y se puede explicar porque en su momento, cuando los sistemas de información fueron creados, tener una gran cantidad de información por persona no era necesario, además que las capacidades tecnológicas no podían soportar el manejo de bases de datos tan grandes, ni el análisis de las mismas, como en el caso de Colsubsidio. Con el paso del tiempo y gracias a la demanda por información cada vez más detallada y de calidad, empresas como esta se ven en la necesidad de reconstruir sus sistemas de información o completar la información histórica para poder generar modelos predictivos que permitan tomar decisiones acertadas.

● Para la revisión del estado del arte se consultarán métodos de análisis predictivo que pueden ser utilizados en el ámbito organizacional, más detalladamente, analítica predictiva para la rotación del talento humano. Con esta investigación se pretende obtener un amplio conocimiento sobre el tema, las herramientas que se utilizan comúnmente y las ventajas que puede traer la implementación de estos modelos a las empresas.

2.Contenido

2.1 Caja de compensación familiar colsubsidio

Colsubsidio se encuentra entre las 60 mejores empresas del país, según el último listado proporcionado por la revista “DINERO” en 2018. Actualmente esta empresa tiene un porcentaje de rotación del 2 % promedio mensual [1], que equivale aproximadamente a 340 personas. La rotación del talento humano en Colsubsidio se ha identificado como un proceso clave, ya que este afecta directamente a los procesos y la operación de la organización.

Esta situación se vive diariamente en las empresas, por lo cual, con la ayuda de modelos analíticos y predictivos se han creado algoritmos que permiten predecir el número de retiros que tendrá en un determinado periodo de tiempo una organización.

2.2 Rotación De Personal

La Rotación de personal es algo propio del ámbito empresarial y se puede presentar tanto por causas internas como por factores externos. Las causas internas son producto de las condiciones de la compañía, y se representan en renuncias o despidos, por el contrario, las causas externas se presentan por factores ajenos a la empresa, problemas personales o familiares, mejor oferta laboral, viajes, entre otros [2].

Por esta razón, existen dos tipos de rotación de personal: la voluntaria y la involuntaria. La primera se da cuando los empleados renuncian a sus cargos por algún motivo que no les permite continuar en la organización, mientras que la involuntaria se genera sin que el empleado tenga el deseo de salir de la organización, lo que puede ser propiciado por una falta disciplinaria grave, un bajo desempeño, recorte de personal o una reestructuración [3].

Los efectos que se producen por la rotación de personal pueden ser problemas entre los colaboradores. Puede dificultar la integración y la coordinación del equipo de trabajo, retrasando la adaptación de estos y, por ende, la productividad del área. También se generan altos costos a causa de la desvinculación, contratación y capacitación de los nuevos colaboradores que ingresen a la compañía, afecta la imagen de la organización, dejándola con perspectivas negativas como empresa empleadora. Produce barreras que pueden afectar la relación con los clientes y proveedores, porque no se logra establecer un vínculo con un representante de la compañía que les brinde confianza y que asuma cada uno de los procesos que manejan entre sí [2].

Pero no todo es negativo, también se pueden encontrar efectos positivos como los son la renovación del personal. Esto puede generar nuevas ideas y proyectos, un nuevo aire al interior de la organización, una visión diferente y contemporánea del negocio. Puede tener efectos positivos tanto interna como externamente, siempre y cuando ese índice de rotación no esté por encima de los parámetros y estándares normales. De ahí la importancia de medir el Índice de Rotación de Personal, con la finalidad de identificar problemas de insatisfacción laboral entre los empleados o deficiencias en los procesos de selección y contratación [4].

Existen diferentes métodos de cálculo, uno de los más utilizados es el siguiente:

$$R = (V + D)/(2\,P)$$

Ecuación 1. Fórmula Rotación Mercado

Donde:

- R = Índice de rotación.
- V = Vinculados.
- D = Desvinculados.
- P = Promedio de trabajadores

Para la investigación, la fórmula a utilizar, elegida por Colsubsidio, es la siguiente:

$$R = (Dx100\%)/P$$

Ecuación 2. Fórmula Rotación Colsubsidio

La rotación de personal es uno de los procesos que más plantea problemas al área de recursos humanos dado que requiere de un esfuerzo importante desde el área de incorporación para el cubrimiento de estas vacantes.

En Colsubsidio la rotación se divide en rotación evitable e inevitable. La evitable hace referencia a las renuncias y los abandonos de cargo, por lo que la inevitable hace referencia a los vencimientos de contratos. Para la empresa es importante descubrir la verdadera razón del abandono y renuncia por parte del personal en la misma, esto posibilita predecir y actuar directamente, iniciando un proceso de contratación para cubrir las vacantes [3].

2.3 Minería De Datos

Una parte primordial en la implementación de un modelo de análisis predictivo es la calidad de la información. La utilización de software para la gestión de información inició en los años ochenta, pero dada la cantidad que se manejaba en ese tiempo, esto no era esencial para la mayoría de las empresas. En los noventa las compañías comenzaron a utilizar algunos programas o herramientas como Visual Basic, Visual FoxPro o Microsoft Office para hacer la gestión de su información.

Con el rápido desarrollo de la tecnología, se han creado herramientas de análisis como la minería de datos, que es ampliamente utilizada para la gestión de los recursos humanos, actualmente la utilización de ésta puede ayudar a las empresas a tomar mejores decisiones, gestionar nuevos empleados y también analizar el volumen de negocios de los empleados antiguos.

En su investigación Xiao-Li Qu [5], usó el algoritmo del árbol de decisiones para analizar las razones principales de los empleados que se retiran, por el método de minería de datos. Analizó los datos por medio de los árboles de decisión ID3, C4.5 y CART utilizando Excel y R. Todos los datos se recopilan de kaggle.com (que es un software, usado para realizar minería de datos). Este conjunto de datos tiene más de diez mil registros y considera muchas características de los empleados, como el nivel de satisfacción, la puntuación de las últimas evaluaciones de desempeño, números de proyectos en gestión, promedio de horas de trabajo por mes y la cantidad de accidentes de trabajo.

Entre las posibles causas analizadas se encuentra: algunas personas pueden preocuparse por su futura oportunidad de carrera, no están satisfechas con su salario, algunas pueden pensar que en las horas de trabajo son extensas y malas condiciones de trabajo [6].

Si una empresa pierde un buen empleado, tendrá costo de dinero 1.5 veces mayor que reclutar y capacitar a un nuevo empleado, adicionalmente, también puede disminuir la satisfacción del cliente y la retención; esto influirá en el tiempo, la dinámica del equipo, la productividad y la continuidad y etc.

2.4 Preparación De Información

La eficacia de un modelo predictivo repercute en cómo se preparan los datos para que el algoritmo analice la información. Existen diferentes métodos de preparación de información, para la creación de modelos predictivos es usual determinar, a partir de la información, una variable objetivo "dicotómica" [7]. Las variables dicotómicas expresan dos modalidades, sí o no= o falso y verdadero, o 1 y 0, [8].

2.5 Sobreajuste y Subajuste

Uno de los problemas que se pueden presentar al intentar modelar son el sobreajuste y el subajuste (overfitting and underfitting) [9], los cuales son producidos al intentar encajar los datos de salida con los datos de entrada. Esto genera que el modelo no reconozca las verdaderas razones por la que un trabajador se retira (underfitting) o que variables no tienen influencia en que un trabajador se retire, mostrándose como prioritarias.

El overfitting se produce cuando las variables que se introducen solo permiten al modelo predecir con estas variables, pero será incapaz de aprender de nuevos datos [10]. A continuación, se muestra el equilibrio de aprendizaje [11].

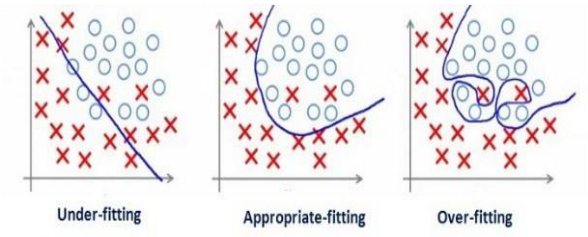

Ilustración 1. Tipos de ajuste

Para verificar la solución de este problema se debe dividir los datos en dos conjuntos [10] , esta validación se debe hacer con la regla del 80% para entrenar y el 20% para validar. Como se muestra a continuación:

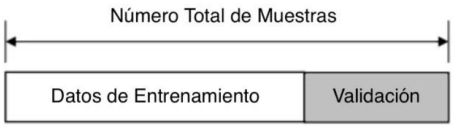

Ilustración 2. Participación Muestras

También se deben tener en cuenta las siguientes consideraciones:

- Las clases deben estar equilibradas en cantidad: Para este caso se debe tener la misma cantidad de información para los diferentes tipos de trabajadores, esto hace que no se sobreajuste los datos.

- Se debe elegir un conjunto de personas que permita validar los datos donde se encuentren todos los perfiles de los trabajadores.

- Se deben ajustar los parámetros, en este caso se modifica el número de interacciones.

- ●Reducir la cantidad de dimensiones: se deben eliminar las características. Es usual emplear PCA (Análisis de Componentes Principales) [10].

Si el modelo entrenado con el conjunto validador ofrece un 90 % de aciertos se considera un problema de overfitting [9].

Según Nevala [12], se deben responder a las siguientes preguntas para corroborar que el modelo creado tenga sentido dentro del contexto:

- ¿Qué estamos tratando de predecir?
- ¿Son predictivas las correlaciones resultantes? ¿Causal? ¿Hay sesgos inherentes?
- ¿Están los resultados en línea con las expectativas? ¿Hay excepciones al ¿Ser abordado?
- ¿Cuál es el valor predictivo y este puede generalizarse?
- ¿Se pueden aplicar el modelo y los resultados en la vida real?
- ¿Cuál es la respuesta adecuada?

Para la correcta implementación se requiere [12]:

- Plantear un problema que necesita resolverse.
- Establecer una mentalidad experimental, los modelos son iterativos y se requiere realizar pruebas (prueba- aprendizaje).
- Alistar un equipo con conocimiento estadístico elevado, un equipo conocedor del negocio que proporcione el contexto, y un equipo de TI que implemente y conozca el sistema técnico.
- Desarrollar una estrategia de extracción de datos robusta. Se debe crear una gobernanza de datos para asegurar la calidad y la oportunidad en la entrega de los datos.
- Establecer el grado de aceptación de la empresa a los riesgos, mientas el modelo se ajusta.
- Adaptar o modificar los procesos que el modelo pueda afectar.
- Adoptar las nuevas prácticas y requerimiento para el equipo de TI.

2.6 Análisis de Supervivencia

Los análisis de supervivencia se utilizan para evaluar la probabilidad que tiene un trabajador de retirarse (muerte) en un tiempo determinado. Desde el punto de vista médico, la supervivencia es una medida de tiempo a una respuesta, fallo, muerte, recaída o desarrollo de una determinada enfermedad o evento. El término supervivencia se debe a que en las primeras aplicaciones de este método de análisis se utilizaba como evento la muerte de un paciente [13].

La observación inicia en t=0 y continua hasta la muerte (retiro) o hasta que el tiempo de seguimiento se interrumpe. Cuando el tiempo de seguimiento termina antes de producirse la muerte o antes de completar el período de observación se habla de paciente "censurado".

Los datos requeridos deber tener:

1. Definir el origen y el inicio de seguimiento.
2. Definir la escala de tiempo.
3. Definir el evento.

La metodología estadística se realiza mediante pruebas paramétricas y no paramétricas.

Entre las paramétricas, los estudios más frecuentes se llevan mediante el análisis de distribución exponencial, Weibull y distribución normal. Y dentro de las no paramétricas se encuentran:

- Kaplan- Meier.
- Logrank.
- Regresión de Cox.

Estos son los métodos más conocidos y los más utilizados. Las curvas de supervivencia se producen mediante los dos métodos:

- Análisis actuarial: divide el tiempo en intervalos y calcula la supervivencia en cada intervalo, este método da aproximaciones.
- Método de límite producto de Kaplan Meier: calcula la supervivencia cada vez que un paciente muere, da proporciones exactas de supervivencia debido a que utiliza tiempos de supervivencia precisos.

La comparación de las curvas de supervivencia se realiza a través de pruebas no paramétrica de suma de rangos de Wilcoxon [13] o a través de la prueba de logaritmo del rango ("Logrank"), esta prueba compara en esencia el número de eventos (muertes, fracasos) en cada grupo con el número de fracasos que podría esperarse de las pérdidas en los grupos combinados.

3. Técnicas Comunes De Machine Learnning

3.1 Aprendizaje Supervisado

En el aprendizaje supervisado la máquina aprende mediante ejemplos, para esto es necesario proporcionar las entradas y salidas deseadas. La "máquina" (también conocido como el algoritmo) utiliza esta entrada para determinar las correlaciones y la lógica que se puede usar para predecir la respuesta, esto es como dar a los alumnos una clave de respuesta y pedirles que "muestren su trabajo". En el aprendizaje supervisado, se proporcionan ejemplos de preguntas y respuestas, a la máquina se le explica cómo pasar de A a B. Una vez que se identifica el patrón lógico, se puede aplicar para que se puedan resolver problemas similares [12].

Algunos de los algoritmos más utilizados son:

- Bayesian Statistics
- Decision Trees
- Forecasting
- Neural Networks
- Random Forests
- Regression Analysis
- Support Vector Machines [SVM]Aplicaciones.
- Detección de fraude.
- Interacciones personalizadas: recomendados dependiendo de los gustos y consumos de las personas.
- Reconocimiento de imagen, habla y texto: leer texto, correos. Crea comandos con las reglas expresadas mediante el habla.
- Evaluación de riesgos: Determinar las probabilidades de que un paciente enferme, así como evaluar la probabilidad de que una persona sufra un accidente.

3.1 Aprendizaje Semi Supervisado

El aprendizaje semi-supervisado se utiliza para abordar problemas similares a los de aprendizaje supervisado, sin embargo, en el aprendizaje semi-supervisado a la máquina se le proporcionan algunos datos con la respuesta definida (también conocida como etiquetada) junto con datos adicionales que no están etiquetados con la respuesta. En otras palabras, algunos de los datos de entrada son etiquetados con la salida deseada (respuesta) mientras que el resto está sin etiquetar.

El aprendizaje semi-supervisado se utiliza en los casos en que hay demasiada información o variaciones sutiles en los datos para poder proporcionar un conjunto completo de ejemplos, en este caso, las entradas y salidas proporcionadas determinan el patrón de la máquina que puede extrapolar y aplicarse a los datos restantes. [12].

Algoritmos más Utilizados

- See Supervised Learning.
- Aplicaciones.
- Reconocimiento de voz o del habla.
- Reconocimiento de imágenes
- Clasificación de páginas Web.

3.3 Aprendizaje No Supervisado

En el aprendizaje no supervisado, la máquina estudia los datos para identificar patrones. En este caso, no hay clave de respuesta, la máquina determina correlaciones y relaciones mediante el análisis de los datos disponibles. El aprendizaje no supervisado se basa en cómo los humanos observamos naturalmente el mundo. A medida que crece nuestra experiencia (o en este caso la máquina y la cantidad de datos a los que está expuesta) [12].

Algoritmos más conocidos:

- Affinity Analysis.
- Clustering.
- Clustering: K-Means.
- Nearest-Neighbor Mapping.
- Self-Organizing Maps.
- Singular Value Decomposition.

Aplicaciones mas utilizadas:

- Análisis de comportamientos de compra.
- Análisis de comportamientos inadecuados o anómalos para detectar fraudes.
- Agrupamiento de clientes dependiendo su consumo identificando sus gustos probables dependiendo de sus intereses.

3.4 Aprendizaje Por Refuerzo

En el aprendizaje por refuerzo, a la máquina se le proporciona un conjunto de acciones permitidas, reglas y estados finales potenciales. En las reacciones resultantes, la máquina aprende a explotar las reglas para crear una salida deseada, determinando así qué serie de acciones y en qué circunstancias, conduce a un resultado óptimo. El aprendizaje por refuerzo es el equivalente a enseñar a alguien a jugar. Las reglas y objetivos están claramente definidos. Sin embargo, el resultado depende del juicio del jugador que debe ajustar su enfoque en respuesta al entorno y la habilidad y las acciones de un oponente dado. [12].

Algoritmos o técnicas comunes:

- Artificial Neural Networks (ANN)● Learning Automata.
- Markov Decision Process (MDP).
- Q-Learning.

Aplicaciones:

- Video juegos, la maquina ejecuta la mejor estrategia posible dentro del juego.
- Robótica.
- Navegación: Identifica que ruta es óptima, con todas las variables que influencian en el tráfico.

4. Métodos Comunes En Machine Learning

4.1 Deep Learning

Técnica moderna y avanzada de aprendizaje automático que hace uso de redes neuronales extremadamente sofisticadas. Se denomina aprendizaje profundo porque los modelos generados son significativamente más complejos o profundos que las redes neuronales tradicionales. Los modelos de aprendizaje profundo también ingieren cantidades de datos mucho más grandes que sus predecesores. [12].

4.2 Cognitive Computing

Sistemas que buscan entender y emular el comportamiento humano. Además de proporcionar una interfaz más natural e intuitiva entre el hombre y la máquina. Por lo general, esto implica implementar sistemas que interactúen con las personas en su "lengua nativa". En otras palabras, sin requerir que un usuario escriba o comprenda el código. Las plataformas de computación cognitiva logran esto utilizando una gran variedad de técnicas que incluyen el procesamiento del lenguaje natural, los algoritmos avanzados de aprendizaje automático (incluido el aprendizaje profundo) y la generación del lenguaje natural [12].

4.3 Modelo Knn

El algoritmo de los k vecinos más cercanos (k-NN Nearest Neighbor) es un sistema de clasificación supervisado basado en criterios de vecindad. En particular, k-NN se basa en la idea de que los nuevos ejemplos serán clasificados a la clase a la cual pertenezca la mayor cantidad de vecinos más cercanos del conjunto de entrenamiento más cercano a él. [11]. El algoritmo del vecino más cercano (aquel que asigna a una nueva muestra la clasificación de las muestras de ejemplo más cercana) explora todo el conocimiento almacenado en el conjunto de entrenamiento para determinar cuál será la clase a la que pertenece una nueva muestra, pero únicamente tiene en cuenta el vecino más próximo a ella, por lo que

es lógico pensar que es posible que no se esté aprovechando de forma eficiente toda la información que se podría extraer del conjunto de entrenamiento [14].

En problemas prácticos donde se aplica esta regla de clasificación se acostumbra a tomar un número k de vecinos impar para evitar posibles empates (aunque esta decisión solo resuelve el problema en clasificaciones binarias). En otras ocasiones, en caso de empate, se selecciona la clase que verifique que sus representantes tengan la menor distancia media al nuevo ejemplo que se está clasificando. En última instancia, si se produce un empate, siempre se puede decidir aleatoriamente entre las clases con mayor representación.

Una posible variante de este algoritmo consiste en ponderar la contribución de cada vecino de acuerdo con la distancia entre él y el ejemplar a ser clasificado, dando mayor peso a los vecinos más cercanos frente a los que puedan estar más alejados. Por ejemplo, se puede ponderar el voto de cada vecino de acuerdo con el cuadrado inverso de sus distancias:

Si x es el ejemplo que se quiere clasificar, V son las posibles clases de clasificación, y {xi} es el conjunto de los k ejemplos de entrenamiento más cercanos, se define:

$$w_i = \frac{1}{d(x, x_i)^2}$$

Ecuación 3. Modelo KNN.

Entonces la clase asignada a x es aquella que verifique que la suma de los pesos de sus representantes sea máxima:

$$argmax_{v \in V} \sum_{i=1...k,\, x_i \in v} w_i$$

Ecuación 4. Asignación de clase.

Esta mejora es muy efectiva en muchos problemas prácticos. Es robusto ante el ruido de los datos y suficientemente efectivo en conjuntos de datos grandes [11].

A continuación, se denota un seudocódigo para un clasificado KNN básico [15]

COMIENZO

Entrada: D = {(x1, c1), . . ., (xN, cN)} x = (x1, . . . , xn) nuevo caso a clasificar

PARA todo objeto ya clasificado (xi, ci)

Calcular di = d (xi, x)

Ordenar di (i = 1, . . . , N) en orden ascendente

Quedar con los K casos DKx ya clasificados más cercanos a x Asignar a x la clase más frecuente en DK x

FIN

Donde D indica un fichero de N casos cada uno caracterizado por n variables predictoras, X1,….., Xn y una variable a predecir C.

Se tienen, además, algunas variantes sobre el algoritmo básico:

- K-NN con rechazo.
- K-NN con distancia media.
- K-NN con distancia mínima.
- K-NN con pesado de casos seleccionados.
- K-NN con pesado de variables.

4.4. Algoritmos Genéticos

El algoritmo genético es un algoritmo de búsqueda introducido por Holand [16]. Este método utiliza algoritmos de optimización para encontrar la mejor respuesta dentro de un campo de posibles soluciones, esta respuesta es denominada máximo global [17]. Los algoritmos genéticos exploran todo un conjunto de valores y los compara entre sí, determinando estocásticamente cuál es el mejor de ellos, esperando que esta respuesta sea cercana o igual al máximo global.

En la ingeniería industrial y de producción se utiliza para encontrar la solución que optimice la línea de producción, como complemento del Job-shop, que optimiza el tiempo de producción pero que a su vez contiene muchas soluciones probables [18].

En la ingeniería ambiental se utiliza para predecir los comportamientos de los sistemas de dispersión de gases. Entre los trabajos más destacados se encuentra el de Dang [19], el cual encontró la combinación óptima de sensores que permiten modelar el clima del sistema del Rio Columbia en Estados Unidos.

En la ingeniería de sistemas los algoritmos genéticos pueden ser implementados en la detección de fallos en los sistemas de seguridad de la información [20]. Este trabajo identifica patrones de tráfico, lo cuales son insertados en un sistema de inteligencia que establece ámbitos de seguridad informática, detectando así a intrusos a través de anomalías en el tráfico de la red. Castro [21], hace una comparación de las redes neuronales artificiales con SNMs:

Tabla1. Comparación de las redes neuronales artificiales con SNMs

Redes neuronales artificiales	SVMs
Capas ocultas transforman a espacios de cualquier dimensión.	Kernels transforman a espacios de dimensión muy superior.
El espacio de búsqueda tiene múltiples mínimos locales.	El espacio de búsqueda tiene sólo un mínimo global.
El entrenamiento es costoso.	El entrenamiento es muy eficiente.
La clasificación es muy eficiente.	La clasificación es muy eficiente.
Se diseña el número de capas ocultas y nodos.	Se diseña la función Kernel y el parámetro de coste C.
Muy buen funcionamiento en problemas típicos.	Muy buen funcionamiento en problemas típicos.
	Extremadamente Robusto para generalización, menos necesidad de heurísticos para entrenamiento

Son un conjunto de algoritmos de aprendizaje supervisado desarrollados por Vladimir Vapnik, que pueden desarrollar problemas de clasificación [22]. Este algoritmo tiene aplicaciones comunes como calificadores de crédito y series de tiempo para la predicción.

4.5. Modelo Random Forest

Es un método de aprendizaje conjunto que construye árboles de decisión múltiples. Las muestras de este algoritmo forman al azar datos con reemplazo en la construcción de cada árbol de decisiones que se denomina empaquetamiento. Cada árbol de decisión devuelve una clase y luego el empaquetado los combina para alcanzar una decisión única [14].

5. Medidas De Desempeño

AUC: AUC o AUROC es un parámetro estadístico de orden de jerarquía [23]. La interpretación del AUC es fácil: cuanto más alto es el AUC, mejor, con 0.50 que indica un rendimiento aleatorio y 1.00 que denota un rendimiento perfecto. [24]

ROC: La curva ROC representa gráficamente la relación entre la sensibilidad y 1 menos la especificidad [23]. Ahora según López [25], la sensibilidad es la probabilidad de clasificar correctamente a un individuo cuyo estado real sea el definido como positivo respecto a la condición que estudia la prueba, razón por la que también es denominada fracción de verdaderos positivos (FVP). La especificidad es la probabilidad de clasificar correctamente a un individuo cuyo estado real sea el definido como negativo. Es igual al resultado de restar a uno la fracción de falsos positivos (FFP). En la siguiente imagen se muestran los diferentes tipos de curva ROC [26]:

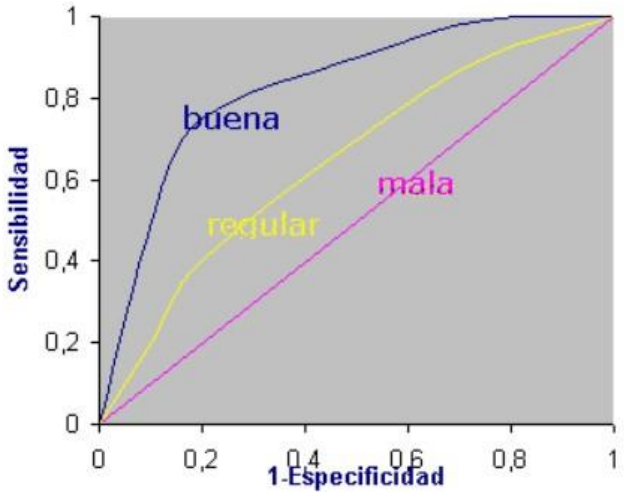

Ilustración 3. Tipos de Curva ROC

La curva ROC describe el rendimiento de un modelo en todo el rango de umbrales de clasificación. [27]

Conclusiones

Para realizar un modelo predictivo en la rotación del talento humano se debe tener especial cuidado en la selección de las variables que puedan influir en la decisión de retiro de un trabajador, pero más aún, en la aplicación de un contexto organizacional que relacione la combinación adecuada de características en un tiempo dado, para así evitar el overfitting y el underfitting, siendo estos uno de los principales problemas en el entrenamiento de un modelo.

Son muchas las herramientas para el entrenamiento y modelamiento del análisis predictivo, para la rotación de personal se debe realizar el ejercicio para todos los modelos posibles y después por lógica de negocio o por medio de las medidas de desempeño de los modelos, descartar los modelos, llevarlos a la realidad y aplicar aquel que tenga mayor ajuste.

Referencias Bibliograficas

[1] Colsubsidio, «Informe de Gestión y Sostenibilidad,» Bogotá, 2018.

[2] H. Zhang, «Big data research on driving behavior model and auto insurance pricing factors based on UBI,» in Proc. ICSINC, pp. 1-8, 2017.

[3] Colsubsidio, «Hoja de vida indicadores,» Bogotá, 2015

[4] Ishaan Ballal1, Shlok Kavathekar2, Shubham Janwe3, Pratik Shete4, Prof. Nivedita Bhirud, «Analysis and Prediction of Employee Turnover Characteristics based on Machine Learning.,» The 18th International Symposium on Communications and Information Technologies, 2018.

[5] L. Qu, «A decision tree applied to the grass-roots staffs' turnover problem —take C-R Group as an example,» IEEE International Conference on Grey Systems and Intelligent Services (GSIS), pp. 378382, 2015.

[6] D. ,. &. O. D. Carey, «The human side of M & A: how CEOs leverage the most important asset in deal making.,» Oxford University Press, 2004.

[7] T. Eichenberg, Supervised Weight of Evidence Binning of Numeric Variables and Factors, R documentation, 2018.

[8] S. d. l. Fuente, Tablas de contingencia, Madrid: Universidad Automoma, 2011.

[9] C. Chien y L. Chen, «Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry.,» Expert Systems with applications, vol. 34, pp. 280-290, 2008.

[10] J. I. Bagnato, «Qué es overfitting y underfitting y cómo solucionarlo,» 12 12 2017. [En línea]. Available:http://www.aprendemachinelearning.com/que-es-overfitting-y-underfitting-y-comosolucionarlo/.

[11] F. S. Caparrini, «Clasificación Supervisada y No Supervisada,» 26 12 2018. [En línea]. Available: http://www.cs.us.es/~fsancho/?e=77.

[12] K. Nevala, The machine learning, Cary: SAS, 2017.

[13] S. Pita Fernández, «Análisis de supervivencia,» Cad Aten Primaria, pp. 130-135, 2001.

[14] A. M. E. Sikaroudi, «A data mining approach to employee turnover prediction n (case study: Arak automotive parts manufacturing),» Journal of Industrial and Systems , pp. 106-121, 2015.

[15] A. Moujahid, «Tema 5,» de Clasificadores K-NN, Universidad del Pa´ıs Vasco–Euskal Herriko Unibertsitatea, 2011, pp. 1-8.

[16] J. Holand, Adaptation in natural and artificial systems",, USA: university of, 1975.

[17] E. Veslin, Aplicación de algoritmos genéticos en problemas de Ingeniería, Rio de janeiro: Federal University of Rio de Janeiro, 2014.
[18] M. C. &. M. J. A. Vélez, «Metaheurísticos: una alternativa para la solución de problemas,» Revista EIA, pp. 99-115, 2007.
[19] T. F. S. B. N. F. W. &. B. Dang, «A Near Optimal Sensor Selection in The Columbia RIvEr (CORIE) Observation Network for Data Assimilation Using Genetic Algorithms,» Distributed Computing in Sensor Systems Lecture Notes in Computer Science, pp. 253-266, 2007.
[20] C. G. G. Carlos Catania, «Reconocimiento de Patrones en el Tr´afico de Red Basado en Algoritmos Gen´eticos,» Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificia, pp. 65-75, 2008.
[21] J. L. Alba Castro, «Maquinas de vercotes de soporte,» 1 08 2014. [En línea]. Available: https://web.archive.org/web/20140801145654/http://www.gts.tsc.uvigo.es/~jalba/doctorado/SVM.pdf.
[22] J. K. Y. Yoon, «A practical approach to bankruptcy prediction For small businesses: substituting the unavailable financial data for credit card sales information,» Expert systems with Applications, pp. 3624-3629, 2010.
[23] E. W. Steyerberga, «Medidas del rendimiento de modelos de predicción y marcadores pronósticos: evaluación de las predicciones y clasificaciones,» Revista española de cardiologia, p. 788–794, 2011.
[24] S. Schroedl, «AUROC - Area under Receiver Operating Characteristic,» 07 04 2008. [En línea]. Available:https://www.mathworks.com/matlabcentral/fileexchange/19468-auroc-area-under-receiveroperating-characteristic.
[25] López de Ullibarri Galparsoro I, «Curvas ROC,» Cad Aten Primaria, pp. 229-235, 2001.
[26] H. y. McNeil, «Curvas ROC,» 1 08 2016. [En línea]. Available: http://www.hrc.es/bioest/roc_1.html#Mcneil.
[27] W. Dwinnell, «Data Mining in Matlab» 20 06 2007. [En línea]. Available: http://matlabdatamining.blogspot.com/2007/06/roc-curves-and-auc.html.

.

Publicación Facultad de Ingeniería y Red de Investigaciones de Tecnología Avanzada – RITA

REVISTA

TIA