Universidad de Lima

Facultad de Ingeniería y Arquitectura

Carrera de Ingeniería de Sistemas

PREDICCIÓN DE POSTULANTES QUE COMETERAN FRAUDE INTERNO CON ALGORITMO DE APRENDIZAJE SUPERVISADO

Tesis para optar el Título Profesional de Ingeniero de Sistemas

Sergio Ernesto Espinoza Montalvo

Código 20142424

Asesor

José Antonio Taquía Gutiérrez

Lima – Perú

Abril de 2020

UNIVERSIDAD DE LIMA
MCMLXV
SCIENTIA ET PRAXIS

PREDICTION OF APPLICANTS THAT WILL COMMIT INTERNAL FRAUD WITH SUPERVISED LEARNING ALGORITHMS

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

El fraude interno es un gran problema para las empresas, ocasionando pérdidas monetarias importantes. Diversas investigaciones han propuesto mejoras al proceso de selección de personal utilizando minería de datos. El presente trabajo propone utilizar la información histórica de postulantes a una empresa para predecir si cometerán fraude durante su estadía. Se encuentran modelos con un nivel de precisión alto, pero que tienen un error de clasificación mayor para encontrar los casos de fraude. Se realizaron modelos con los algoritmos de C-45, Random Forest y redes neuronales y se evaluó el aporte de las características al resultado. Utilizando un algoritmo genético se determinó que 13 variables eran las más relevantes para el problema. Algunas de estas variables coinciden con variables mencionadas en la literatura encontrada sobre trastornos antisociales. Se concluye que hay valor en información de postulantes para determinar si cometerán fraude interno durante su estadía en la empresa.

Palabras Clave: Aprendizaje supervisado, Predicción de Fraude, Minería de datos, Trastorno antisocial de la personalidad, Fraude Interno

ABSTRACT

Internal fraud is a big issue for companies, resulting in relevant monetary losses. Several investigations have proposed improvements to the personnel selection process making using of Data Mining. The present work proposes to use past information of applicants to a company to predict if they will commit fraud during their stay. We find models with high precision, but that have a bigger classification error to find the fraud cases. After several experiments, we find around 13 features of this universe that are most relevant to the model. Some of these features match with features mentioned in literature about antisocial disorders. We conclude that there is value in applicant information to predict if they will commit internal fraud during their stay in the company.

Keywords: Supervised Learning, Fraud predictions, Data Mining, Antisocial personality disorder, internal fraud

INTRODUCCIÓN

En la presente investigación se analizan casos de fraude interno en una empresa de suministros de cómputo en Perú. La empresa en estudio sufrió un robo en el año 2015 y múltiples daños a causa del fraude en los últimos años. Además de este caso se buscó en la literatura el impacto económico del fraude interno en las empresas y estos se evalúan en los miles de dólares (EY, 2017). Se describió por la empresa en estudio que muchas de las personas que cometían estos actos habían sido problemáticas en la oficina y podían haber tenido un historial o ser reincidentes en el futuro. En base a este se propone solucionar el problema identificando a este tipo de personas durante el proceso de selección, de modo que se evite su ingreso a la compañía y los daños económicos asociados. Se usa la información histórica de los postulantes para predecir cuales cometerán fraude en base a una clasificación con algoritmos de aprendizaje supervisado. Esto es con el objetivo de que con el modelo la empresa pueda prevenir que postulantes riesgosos ingresen y ocasionen nuevos casos de fraude. En el capítulo I se muestran las consecuencias que el fraude interno ha ocasionado en las empresas, se aprecia que este mismo tiene un costo económico bastante alto. En el capítulo II se resume la literatura estudiada, se encuentran diversos artículos que analizan problemas relacionados a fraude y a predicción de comportamiento de postulantes utilizando minería de datos. En el capítulo III se resume el marco teórico del tema en cuestión, donde se incluyen investigaciones sobre el trastorno de personalidad antisocial y la base teórica de los modelos de análisis predictivo a utilizar. En el capítulo IV se describe la implementación, que incluirá el procesamiento de la información de los postulantes. Además, se presentan los resultados obtenidos tras diversas experimentaciones. En el capítulo V se presentan validaciones de los resultados obtenidos en el capítulo IV. Finalmente, en las conclusiones y recomendaciones se da la discusión de los resultados obtenidos, comparando estos con la teoría y la literatura estudiada. Se concluye que hay valor en la información de postulantes para predecir el fraude interno, y por ende para la empresa para implementar un modelo de este tipo.

CAPITULO I: PLANTEAMIENTO DEL PROBLEMA

1.1 Formulación del problema

Uno de los procesos más importantes del área de recursos humanos es el proceso de selección de personal, el cual consiste en seleccionar al personal idóneo para incorporar a la organización. El resultado de este proceso serán las personas que construyen la organización. Los profesionales de recursos humanos presentan varios desafíos durante el proceso de selección ya que requieren tomar muchas decisiones de tipo gerencial (Jantan, 2009).

Uno de los indicadores que se desea sobre postulantes a una empresa es cómo éstos se integrarán a la organización. Por esta razón se evalúa el perfil psicológico de los mismos (Scroggins, 2008).

La mayor parte de fraudes que ocurren en las organizaciones son cometidos por sus propios empleados o en colusión con éstos. Sin embargo, las empresas enfocan la mayor parte de sus recursos en amenazas externas. (Rich, 2009). Los empleados generalmente tienen acceso y conocimiento sobre las aplicaciones y bases de datos de la empresa. Asimismo, los empleados tienen la capacidad de poder evitar medidas física y electrónicas de la organización (Randazzo, 2004).

El fraude interno desde una perspectiva de Ingeniería de Sistemas es tan peligroso como cualquier daño que pueda venir del exterior. (Mills 2017). Esto se debe a que alguien del interior puede infringir en el sistema sin tener que vulnerar este primero, ya que puede tener acceso libre (Bensing, 2009). La diferencia del fraude interno con algún daño que puedan originarse de errores humanos, es que estos segundos pueden mitigarse con la implementación de controles y de capacitación a empleados. Sin embargo, el fraude solo depende de la intención maliciosa del empleado, lo cual no se puede controlar (Silowash, 2012)

El fraude interno es uno de los problemas que está presente en todas las organizaciones y ocasiona pérdidas económicas importantes a las mismas (Kroll, 2012); así como pérdida de reputación, pérdida de información, robos de equipos, entre otros. (Smith, 2005).

El fraude interno afecta a las empresas financieramente, (Junqué, 2014) además la organización debe asumir el costo de controles y auditorías para monitorear o prevenir el fraude interno (Willson, 2009). En Perú, en el año de 2013, el 22% de los casos de fraude ocasionaron más de 100 mil dólares en pérdidas para las grandes empresas. El fraude interno compone el 80% de estos casos. (EY, 2017)

La organización se beneficia al saber cuáles son los empleados que serán positivos para el ambiente laboral. Si una organización no puede conservar a sus mejores empleados, probablemente pierda su ventaja competitiva, y, en consecuencia, pérdida de calidad y resultados. (Chang, 2009).

Para el personal de recursos humanos es difícil detectar qué postulantes tienen la intención de cometer fraude. El postulante puede ser consciente que durante el proceso será evaluado en este aspecto e intentar manipular al evaluador a pensar que es un candidato seguro (Kroll, 2012).

Los artificios tradicionales utilizados para el proceso de selección han demostrado no ser del todo efectivos (Varshney, 2014), debido a que el fraude interno sigue prevaleciendo en las organizaciones del país. Cabe mencionar el reciente caso de la cajera del BCP que robo 5 millones de soles (El Comercio, 2017), o el caso del robo a la empresa Compudiskett en el año 2015, del cual fue responsable una empleada de la compañía que brindó información a los delincuentes (PERU, 2017).

Reconocer el patrón de comportamiento de los empleados es uno de los activos más importantes para las empresas actualmente en el mercado. (Rashid, 2016). Se está usando la minería de datos para apoyar al área de recursos humanos en las empresas. (Horesh, 2016) El vicepresidente de eQuest (empresa que ayuda a organizaciones a encontrar empleados) en minería de datos para recursos humanos, David Bernstein, dijo en 2013 que se necesitan modelos predictivos que se integren con el planeamiento estratégico de la organización en lo que concierne a la gestión del talento humano (Bersin, 2013). Lamentablemente, un estudio del 2014 indica que solo cerca de 4% de las organizaciones utiliza este tipo de herramientas (Varshney, 2014). Prevenir la intención de fraude mediante patrones de comportamiento puede ser extremadamente útil. (Bersin, 2013)

1.2 Hipótesis de la Investigación

En base al problema presentado, se propone:

H_o: Existen variables en información de postulantes a una empresa que permiten determinar con mayor efectividad si estos cometerán fraude interno o no.

Al ser verdadera esta hipótesis, se demostraría que se puede usar minería de datos para predecir qué posibles ingresantes a una empresa serían perjudiciales para esta, y, en consecuencia, evitar su ingreso.

1.3 Objetivo de la investigación

1.3.1 Objetivo General.

Predecir qué postulantes cometerán fraude interno utilizando algoritmos de aprendizaje supervisado para el proceso de selección.

1.3.2 Objetivos Específicos

- Identificar cuáles son los factores que influyen para que una persona cometa fraude interno.
- Elegir las características a usar en la información de los postulantes basado en conocimiento sobre factores de influencia en fraude interno
- Construir el modelo de predicción basado en diferentes algoritmos de aprendizaje supervisado.
- Comparar los modelos y determinar cuál es el más efectivo para el problema
- Validar los resultados obtenidos mediante técnicas encontradas en la literatura
- Optimizar los modelos con selección de características e hiperparámetros

1.4 Justificación

Una de las tareas más relevantes de la ciencia es hacer y comparar modelos con la información que se recolecta (MacKay, 1992). Al incrementar efectividad en los métodos para detectar el fraude interno se reduce la intención futura de otras personas de cometer el mismo, ya que hay una percepción de mayor riesgo en intentar realizar tales acciones criminales. (Junqué, 2014) Como se comentó previamente, hay cifras alarmantes respecto al fraude interno en las compañías. Si consideramos que 81% de los casos de fraude vienen del interior y que 22% de estos tuvieron un costo de más de 100mil dólares,

entonces nos veríamos frente a grandes pérdidas anuales para el PBI del país a causa de estas conductas. (EY, 2017)

La literatura existente nos muestra que hay esfuerzos y tendencias por utilizar técnicas de Machine Learning para mejorar los procesos de selección de personal, ya sea por el lado del talento o del comportamiento. David Bernstein, vicepresidente de eQuest afirma que el uso de estas técnicas será necesario para que las empresas se mantengan competitivas; sin embargo, el uso actualmente es muy bajo (Varshney, 2014).

Dentro de este contexto se propone predecir el comportamiento de los empleados de una manera más efectiva a través del uso de técnicas de minería de datos. El personal de recursos humanos podría utilizar esta información al seleccionar un candidato, de modo que se puedan reducir las incidencias de fraude interno en la empresa, y, por ende, los costos y los daños asociados con estas mismas.

CAPÍTULO II: ESTADO DEL ARTE

Diversas investigaciones han tratado el problema de selección de personal en las empresas. Varshney (2014) utilizó minería de datos para predecir la aptitud de los postulantes para su puesto. La información de recursos humanos fue la más relevante para este modelo, donde se utilizaron los algoritmos de SVM (Support Vector Machine) y regresión lineal. (Varshney, 2014) Varshney nos indica también que se espera tener un valor de significación menor al 5% en este tipo de problemas. (Varshney, 2014)

Una investigación de 2016 utilizó múltiples fuentes de datos sobre los empleados de una organización para poder predecir expertise de estos mismos en tiempo real. La técnica utilizada fue Ordinal Regression Clustering, la cual consiste en utilizar regresión para agrupar datos. (Horesh, 2016) Este estudio utilizó la técnica de compleción de matriz para obtener los datos faltantes (alrededor de 80%) con buenos resultados (Horesh, 2016). Esta técnica utiliza el método de minimización de matriz para asumir valores faltantes. (Candès, 2009)

En el estudio de Rashid (2016) se predijo comportamiento de empleados utilizando métodos con minería de datos tales como Naive Bayes, FRNN, Árboles de decisión y redes neuronales. El mejor resultado fue de 98.12% de precisión usando redes neuronales (Rashid, 2016). Se pueden apreciar las características usadas en el modelo en la figura 2.1 (f*igura 2.1*).

En el estudio de Jantan se compara la efectividad de los Algoritmos de C4.5, Random Forest, MLP y redes neuronales para determinar el impacto positivo que tendrá el empleado en la empresa. Se analizaron 53 características de las Bases de Datos, dividas en diversas categorías (*tabla 2.1*). El algoritmo C4.5/J48 demostró ser el más efectivo, con 95.14% de precisión. (Jantan, 2011).

En un estudio previo, Jantan (2009) comentó sobre las características que se pueden analizar de una persona durante su selección: divididas entre variables de personalidad y de aptitud académica *(figura 2.2*).

Figura 2.1

Variables usadas en experimento de Rashid

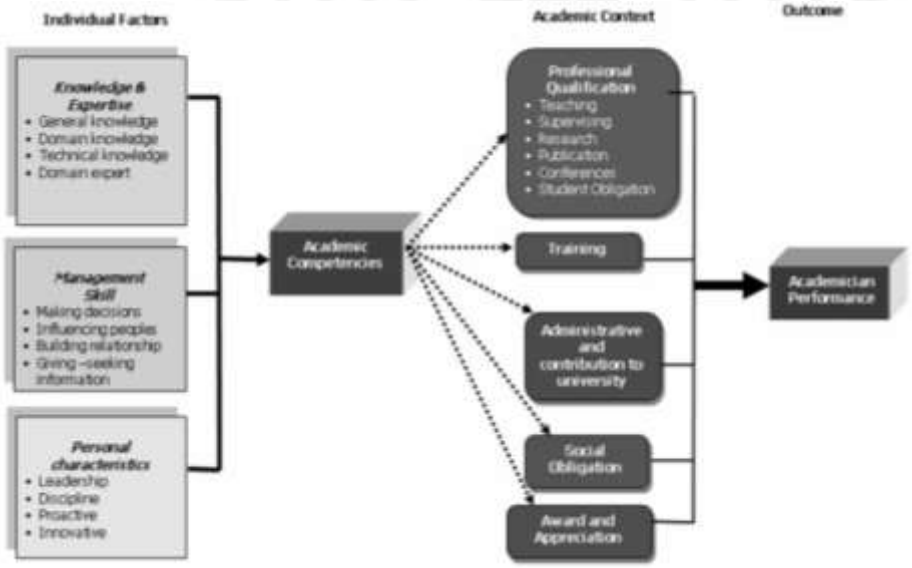
This section is filled by staffs			
Number	Features	Number	Features
1	gender	17	position
2	age	18	department
3	qualification	19	computer skills
4	language	20	job security
5	marriage	21	smoking
6	partner working	22	transportation
7	number of children	23	vacation days
8	average age of their children	24	nationality
11	resident	25	employment type
12	job time		
13	hours of work		
This section is filled by superintends			
14	salary	26	number of activities
15	years of service	27	number of penalties
16	social assurance	28	absence days

Nota: Variables usadas en el experimento de Rashid para predecir comportamiento de empleados, se aprecia que están relacionadas con información de recursos humanos.

Fuente: Rashid, 2016

Figura 2.2

Aptitud académica



Nota: Variables que definen aptitud académica de acuerdo a Jantan.

Fuente: Jantan, 2016

Se ha observado el uso de análisis de textos para determinar las intenciones de una persona, para esto se utilizan algoritmos tales como el de "Porter Stemming" u otras técnicas para normalizar los bloques de texto, de modo que se puedan analizar para detectar patrones. (Edwards, 2017)

Esta técnica se utilizó por Edwards (2017) para detectar estafadores en conversaciones de correo electrónico. Para esto Edwards utilizó el conjunto de datos de Enron. El algoritmo de Naive Bayes con la técnica de BOW (Bag of Words) presentó los mejores resultados (96.3%). El estudio nos demostró que se pueden encontrar distintas fases en una conversación transcrita con alguien con intenciones fraudulentas. (*figura 2.3*)

Tabla 2.1

Predecir talento laboral

Factor	**Atributos**
Información Personal	7
Previo rendimiento	15
Conocimientos	17
Habilidades Blandas	8
Características de personalidad	6

Nota: Variables del análisis de Jantan para predecir talento en la fuerza laboral.
Fuente: Jantan, 2011

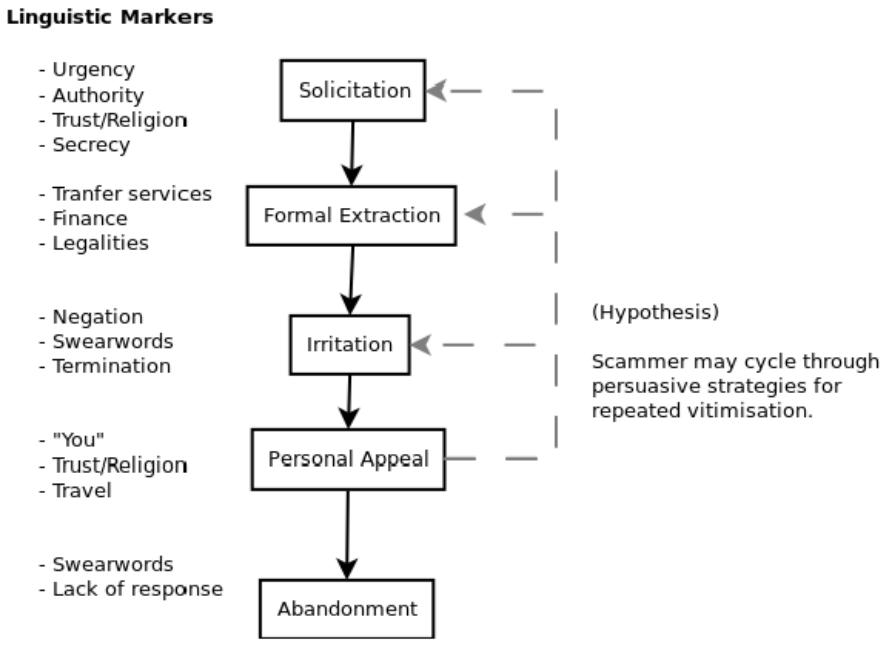
Previamente ya se ha utilizado minería de datos para detectar amenazas internas (Young, 2014). Algunos estudios utilizan algoritmos tales como redes Bayesianas y regresión logística para estos modelos (Machado, 2012) Mills comprobó que hay diferencias entre el comportamiento de usuarios maliciosos y usuarios comunes en un sistema. Algunos de estos pueden ser duraciones de sesiones sospechosas, total de

entradas al sistema y comportamientos sospechosos durante la sesión (Mills, 2017) Se ha encontrado una relación entre trastorno antisociales e individuos que cometan este tipo de actos (Aquino, 2003).

Según Bhattacharyya, una manera de diferenciar usuarios normales de fraudulentos es considerar todas las transacciones que se alejen de la norma como sospechosas. (Bhattacharyya, 2011). En su estudio se obtuvo 86% de precisión para detectar transacciones fraudulentas utilizando Random Forest.

Figura 2.3

Fases de interacción con un estafador



Nota: Fases de interacción con un estafador de acuerdo a Edwards.

Fuente: Edwards, 2017

Sobre la detección de fraude podemos decir lo siguiente (*tabla 2.2*).

Tabla 2.2

Falsos positivos

	Amenaza Real	Amenaza Falsa
Alarma Positiva	Verdadero Positivo	Falso Positivo
Alarma Negativa	Verdadero Negativo	Falso Negativo

Nota: Efectos en la empresa dependiendo del resultado de la predicción y la precisión de esta. El color rojo marca el resultado a evitar, donde la compañía es dañada, mientras que el verde los resultados ideales

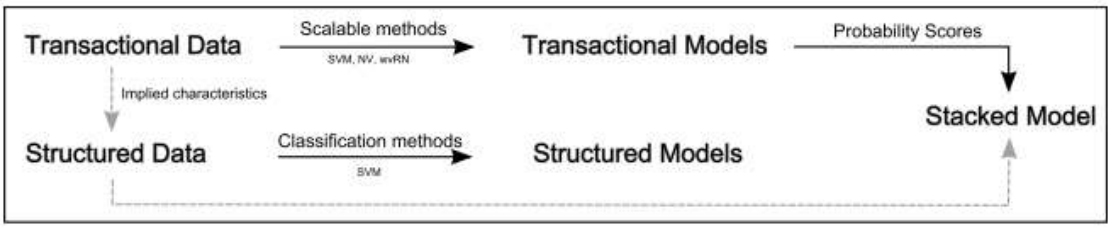
Fuente: Aquino, 2003

Los verdaderos positivos o falsos negativos son las situaciones ideales, donde el sistema predice correctamente lo que ocurrirá y se puede actuar acorde. El falso positivo no es ideal, pero es preferible a un verdadero negativo. En un verdadero negativo la compañía sufre daños por fraude por no poder detectarlo a tiempo, que es la que se desea evitar. (Aquino, 2003). Por esto mismo, los falsos positivos son preferibles a los verdaderos negativos ya que estos segundos si van a terminar en daños para la compañía. (Bolin, 2001).

Para este tipo de problemas se debe realizar un modelo combinado, que utilice varios tipos de información (Junqué, 2014). La figura 2.4 describe cómo combinar diferentes modelos de información (*figura 2.4*).

Figura 2.4

Modelos combinados



Nota: Guía para crear un modelo con información combinada de acuerdo a Junqué.

Fuente: Junqué, 2014

Asimismo, Zhang considera que es importante siempre incluir información sobre el campo en el modelo, ya que incrementa el éxito del sistema. (Zhang, 2011) Algunos datos resaltantes son que el género y la clase social eran buenos predictores de fraude, con una mayor prevalencia mayor en hombres y clases más bajas (Mills, 2017).

CAPÍTULO III: MARCO TEÓRICO

3.1 Selección de personal en compañías

La selección de personal es el proceso metódico para elegir a las personas que trabajaran para una compañía. El resultado de un arduo proceso de selección son varios de años de una relación exitosa entre el empleado y el empleador. (Hr-guide, 2017)

Este proceso tiene miles de años de antigüedad, siendo los casos más antiguos documentados en China para selección de trabajadores para el gobierno. En el mundo moderno, desde hace 100 años aproximadamente, los psicólogos se han enfocado en encontrar métodos científicos más efectivos para seleccionar personal. (Wang, 2011) Esto se debe a que los recursos humanos son el recurso más importante de la empresa (Chang, 2009) y poder reconocer su comportamiento es una necesidad para que una empresa sea competitiva (Rashid, 2016).

De la misma manera una empresa debe ser capaz de mantener a su personal talentoso para mantener una fuerza de trabajo de élite. La complejidad de esta tarea crea la necesidad de poseer un departamento de recursos humanos que utilicé todos sus esfuerzos en crear una fuerza de trabajo impecable. (Rashid, 2016). Escoger personal requiere bastantes decisiones de tipo gerencial y es un reto para cualquier profesional de Recursos Humanos (Jantan, 2010). El 40% de las empresas considera que la selección de personal es un hecho de suma importancia (Perrin, 2005).

3.2 Fraude en compañías

El delito más temido en países desarrollados es el fraude (Cano C., 2017). Este termina siendo muy costoso para las empresas, acumulando costos no solo de las pérdidas a causa de las incidencias, sino también para su detección y prevención (Ai, 2013).

El acto de fraude requiere de 3 elementos: una necesidad en la persona para realizar el acto, la oportunidad de cometer el acto y la habilidad del criminal de racionalizar el acto cometido (Kroll, 2012)

3.2.1 Fraude interno

Se estima que un 80% de los casos de fraude en Perú vienen del interior de la compañía, además 50% de las compañías en Latinoamérica consideran que el fraude interno es un gran problema (EY, 2017).

Algunas teorías que explican las motivaciones empleado para cometer fraude son la Anomie y la teoría de expectativas. La primera indica que empleados de muy bajo nivel jerárquico y que tienen aspiraciones muy altas intentarán alcanzarlas como sea, y, por ende, cometerán fraude. La segunda indica que los empleados realizarán acciones en base a cómo perciben su ambiente, y como interactúa esta percepción con sus deseos de alcanzar el éxito y su base moral (Smith, 2005).

3.3 Minería de datos

La minería de datos consiste en el análisis de grandes cantidades de información para encontrar patrones. Los patrones se usarán para obtener conocimiento que se utilizará para las necesidades del ser humano (Rashid, 2016). La popularidad de la minería de datos se debe a la explosión de la información, donde varios de estos métodos se necesitan para encontrar conocimiento o información relevante (Chang, 2009). La minería de datos se ha aplicado para detección de fraude en varios dominios, tales como telecomunicaciones, banca, seguros, tráfico web, el gobierno, etc. (Junqué, 2014)

Hay gran interacción con la probabilidad en este tipo de estudios. Cualquier intento de predecir comportamiento futuro es inherentemente probabilístico ya que se analizan patrones (Jensen, 2005). La minería de datos utiliza información histórica para poder predecir mediante probabilidades un resultado futuro.

Al procesar información para minería de datos siempre podrán existir datos faltantes, tienen que reemplazar los valores vacíos para poder analizar la información. Se pueden usar promedios o técnicas de análisis de datos para reemplazar estos valores. (Horesh, 2016)

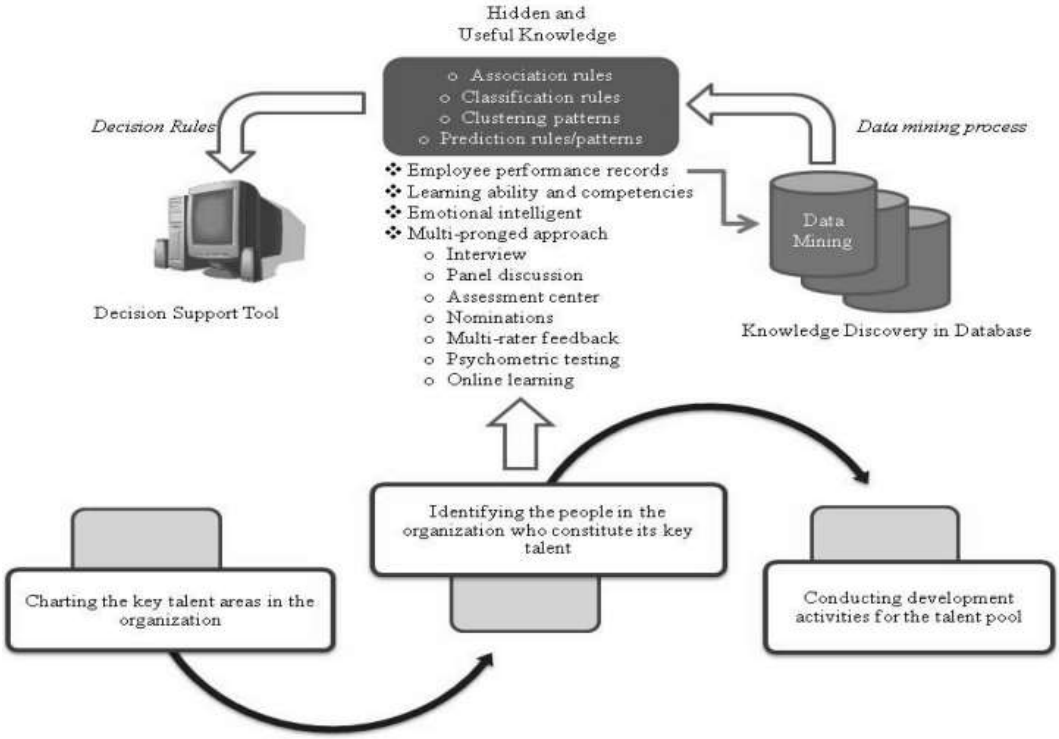
3.3.1 Minería de datos en recursos humanos

En vista de las exigentes demandas crecientes de calidad, costo y tiempos en las compañías, la minería de datos para Recursos Humanos se ha vuelto una necesidad. (Horesh, 2016) Información de los empleados que se puede explotar incluyen productos del trabajo, actividad en las redes sociales, currículos vitae, información de recursos humanos, evaluaciones, entre otros (Horesh, 2016).

La información del área humana es una gran fuente para aplicar minería de datos. (Jantan, 2010). Jantan describió las siguientes áreas a obtener información para obtener información de recursos humanos con minería de datos (*figura 3.1*).

Figura 3.1

Predecir talento humano



Nota: Predecir talento humano utilizando un árbol de decisión C4.

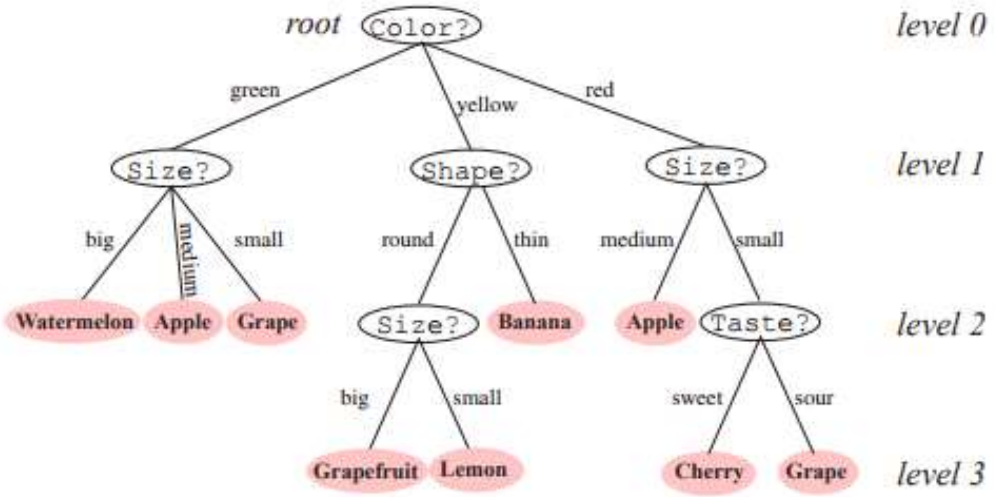
Fuente: Jantan, 2010

3.4 Árboles de decisión

Los árboles de decisión son un algoritmo de minería de datos que muestran un resultado en base a preguntas (*figura 3.2*). (Duda, 2000) Suelen ser construidos de arriba hacia abajo. Para obtener cada nodo hacen una evaluación de cada sub-set resultante. El árbol obtiene la partición más eficiente y procede a repetir el mismo proceso con el siguiente nodo hasta que ya no se puede obtener una mejor división. (Blockeel, 2002)

Figura 3.2

Árbol de decisión



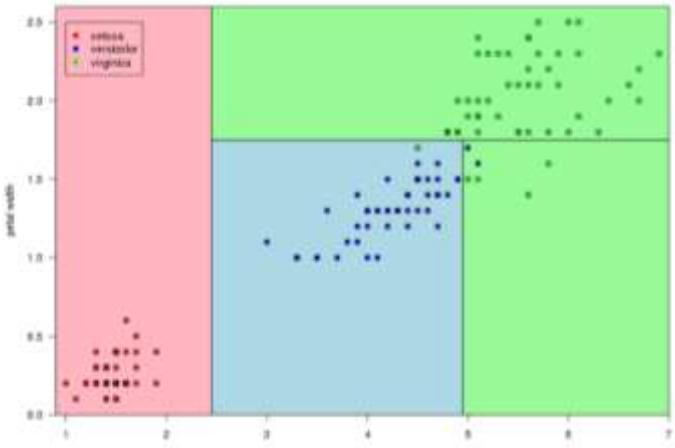
Nota: Ejemplo de árbol de decisión CART.

Fuente: Duda, 2000

El algoritmo coloca los datos en un plano cartesiano y segrega los registros en base a su etiqueta (o el atributo por el que se va a realizar la clasificación). Las variables que se utilizarán para realizar la clasificación son las coordenadas del plano. Después se obtienen las divisiones óptimas para clasificar la data (Breiman, 1984). La figura 3.3 (*figura 3.3*) muestra un ejemplo de cómo se realizan estas divisiones, y la figura 3.4 (*figura 3.4*) grafica el árbol de decisión resultante. El set de datos ejemplificado en las figuras es “Iris” del repositorio UCI de minería de datos. (Merz, 1996)

Figura 3.3

Particiones de información en árbol de decisión

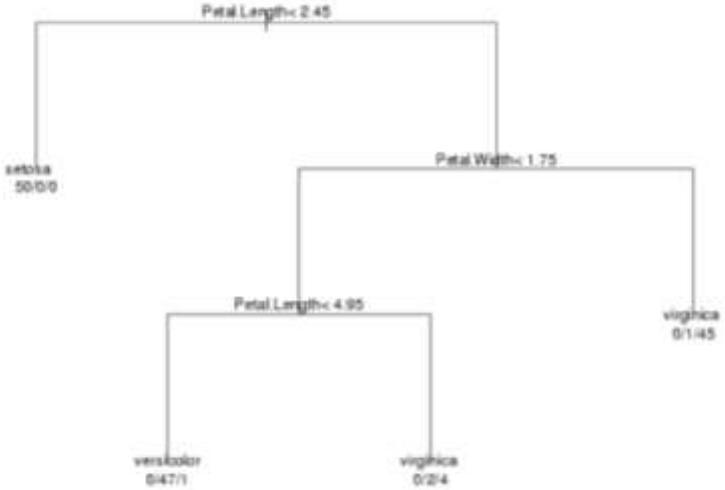


Nota: Divisiones que realiza el algoritmo de árbol de decisión CART en un plano cartesiano de dos variables. Los colores corresponden a la etiqueta que es el tipo de flor.

Fuente: Breiman, 1984

Figura 3.4

Árbol de decisión “Iris”



Nota: Árbol de decisión resultante de la división de información en “Iris”. Podemos ver que las decisiones para elegir el tipo de flor se hacen en los puntos de corte del plano cartesiano..

Fuente: Breiman, 1984

3.4.1 Random Forest

La combinación de modelos de autoaprendizaje incrementa el éxito de la predicción.

El algoritmo de Random Forest consiste en varios árboles de decisión de baja complejidad que se agruparán en una especie de "bosque". Estos después predecirán el resultado y se elegirá por votación el resultado final. Los árboles de decisión se crean en base a una partición de data y de variables. Esto último es lo que se conoce como "bagging". (Breiman, 2001).

Un árbol de decisión determina mediante ciertas variables si un registro se comportará de cierta manera u otra. Random Forest parte de 2 premisas, que los árboles de decisión generados aleatoriamente serán en su mayoría correctos con sus predicciones y que si se usan los votos de estos se tendrá una mejor predicción. Los árboles de decisión se forman a partir de variables y conjuntos de datos de la población de data aleatorias. Finalmente se realiza promedio de los resultados de los árboles de decisión para obtener el resultado final.

Los resultados mejoran con el uso de variables aleatorias y entrenando árboles poco complejos, ya que por la ley de números grandes los árboles de decisión promediados tendrán poca varianza. (Breiman, 2001) Asimismo, la selección aleatoria de la información mejora mucho los tiempos de procesamiento. (Breiman, 2001)

El Random Forest tiene buenos resultados con información no tan consistente, lo que le permite adaptarse a varios sets de datos, y además se beneficia de la inserción de ruido y de variables donde la dependencia es incierta o desconocida (Breiman, 2001).

3.4.1.1 Bagging

Bagging consiste en seleccionar conjuntos de datos de data para alimentar un modelo (Breiman, 2001). Es una técnica que se utiliza dentro cross-validation, para poder validar los resultados con información que no se usó para entrenar. Wolpert demostró en modelos de regresión que el bagging permite obtener resultados estimados más acertados que con información de test. (Wolpert, 1997)

3.4.1.2 Variables latentes

Las variables latentes son aquellas que no pueden ser observadas o medidas directamente. Algunos ejemplos de variables latentes pueden ser resultados de exámenes,

CI y sexo. El concepto de variables latentes parte del hecho que el investigador no puede medirlas con confiabilidad y se basa en un juicio por parte de un tercero. Este es el caso de exámenes donde el evaluado puede haber tenido suerte al responder o que estén siendo sesgados por la persona que lo corrigió. En el caso del sexo se basa en un juicio de la persona que lo observa. Estas terminan siendo conexiones hacia verdades que se pueden asumir sobre los objetos en estudio. (Borsboom, 2008)

3.4.1.3 "Registros fuera de caja" (Out of Bag Estimate)

Los "registros fuera de caja" se utiliza para calcular el error estimado de un modelo. Un estimado utilizando "registros fuera de caja" es el resultado de la técnica de bagging. Breiman calculó que en un caso estándar de bagging aproximadamente 37% de los registros no son usados en el algoritmo de entrenamiento. (Breiman, 1997) Estos registros se pueden utilizar para obtener estimados acertados del desempeño del algoritmo en un escenario real. De acuerdo a múltiples experimentos realizados por Breiman se determina que el nivel de precisión de un predictor con registros "Out of Bag" es bastante alto (Breiman, 2001).

Esto sería en contraste con utilizar la data con la que se entrenó el algoritmo para obtener un porcentaje de precisión. Esto puede estar bastante sujeto a "overfitting", al adaptarse mucho a la información de entrenamiento. (Breiman, 1997). La principal aplicación de Out of Bag estimates sería en el algoritmo de Random Forest, ya que este utiliza bagging para construir los múltiples árboles entrenados. Finalmente, se utilizan los registros fuera de caja en los algoritmos entrenados para obtener el porcentaje de precisión estimado del modelo (Breiman, 1997).

3.5 Redes Neuronales Bayesianas (BRNN)

Las redes neuronales se basan en el funcionamiento del cerebro humano, tienen el propósito de modelar la manera en la que la plasticidad de las neuronas genera nuevos aprendizajes. (Haykin, 1999) Estas funcionan como una "Caja Negra", aprenden del modelo desde cero, lo que significan que no necesitan información previa del problema (Ahmad, 2017).

Hay un principio conocido como la navaja de Ockham, que indica que un método más simple es preferido a uno complejo que no puede generalizar bien (McKay, 1992). Aplicado a redes neuronales, esto implicaría aplicar una función que mejore la

generalización de estas mismas. La generalización hace referencia a que una red neuronal no solo tendrá buenos resultados con datos de entrenamiento, sino también con datos reales (Foresee, 1997). Las redes neuronales con regularización bayesiana agregan un factor a la fórmula utilizada para calcular el error medio cuadrático, de modo que al reducir el error la red neuronal aun tendrá buena generalización (Foresee, 1997). En otras palabras, consiste en reducir el "overfitting".

Los métodos bayesianos fueron introducidos por primera vez en 1939 (McKay, 1992). Estos fueron desarrollados a lo largo de los años y son un fundamento en la teoría de probabilidad. (Jaynes, 1986).

Para realizar esto, el algoritmo utiliza una función objetivo donde se introducen un valor alfa y beta (Foresee, 1997). El valor beta determinará cuanto énfasis se dará en reducir el error de la red neuronal, y el valor alfa cuando énfasis se otorgará en mejorar la generalización la red (Foresee, 1997). Para optimizar la función objetivo se utiliza una función de Bayes, donde se maximiza la probabilidad que los pesos de la red neuronal sean los más óptimos. Para maximizar la probabilidad se usa el método de Gauss-Newton para solucionar una matriz hessiana (Foresee, 1997).

3.6 Redes Neuronales con Backpropagation

Las redes neuronales con Backpropagation permiten encontrar relaciones entre diferentes factores. (Goh, 1995). Este es uno de los métodos más investigados para entrenar redes neuronales (Heerman, 1992). La optimización con Backpropagation se da gracias a la técnica de optimización no-lineal de la gradiente descendiente (Heerman, 1992). Esto permite que las redes neuronales sean funcionales no-lineales (Pineda, 1987). En este sistema primero se generan pesos aleatorios, se optimizan estos en base a los valores de entrada y valores de salida deseados; el error o diferencia detectada entre estos dos valores se propaga hacia atrás en la red neuronal para ajustar los pesos (Vogl, 1988).

3.7 Selección de Características

Para tener éxito utilizando minería de datos se requiere procesar previamente la información. Se tiene que aplicar la minería de datos de forma cuidadosa, eligiendo las variables correctas. Esto es para obtener mejores resultados de entrenamiento en los algoritmos. (Rashid, 2016) La selección de características es el método más común (Chang, 2009).

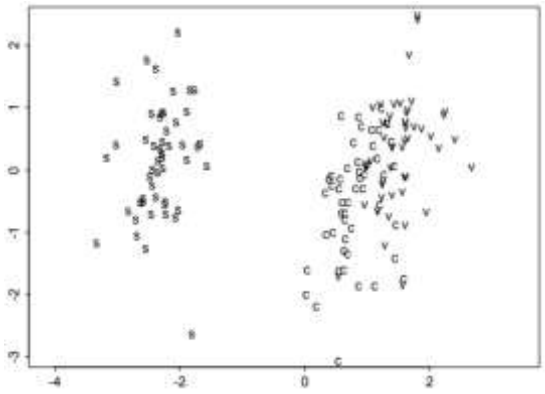
Los sets de datos pueden tener características que no mejoran el éxito del sistema y por ende son irrelevantes. La selección de características consiste en escoger el conjunto de variables más óptimo dentro de un set de datos. Esto se hace con el objetivo de tener la mejor predicción posible (Rashid, 2016). Se realiza reemplazando de forma iterativa los candidatos a variables por otros candidatos analizados en el grupo que generan modelos más exitosos. (Rashid, 2016) El experimento de Rashid mostró una mejora en promedio de 1.75% en precisión utilizando selección de características con las técnicas de Information Gain, Gain Ratio, OneR y Fuzzy RST.

3.7.1 Análisis de componentes (PCA)

PCA o Análisis de componentes principales es una técnica estadísticas multivariante que consiste en hallar la variabilidad máxima entre las variables de un modelo. (Venables, 2002). Sus objetivos son reducción de información e interpretación (Johnson, 2007). Esto se logra buscando la máxima variación en combinaciones lineales de las variables (Venables, 2002). Estas combinaciones lineales se multiplican por un vector para obtener la matriz de covarianza (Johnson, 2007). El primer componente siempre es el que muestra la mayor variación en los datos (Venables, 2002). Algunos de los usos comunes de PCA son: encontrar que variables están relacionadas en set de datos, encontrar que variables son relevantes para el modelo que se está creando y determinar si se puede reducir el set de datos para el modelo sin perder información relevante (SPSS, 2020).

Figura 3.5

Ejemplo de Análisis de componentes principales



Nota: Dos principales componentes para el set de datos de "iris". Podemos ver como se agrupan los diferentes tipos de flores en la matriz de covarianza.

Fuente: Venables, 2002

3.8 Cross Validation

Cross Validation o Validación cruzada consiste realizar "n" particiones un set de datos dado. Después se entrena el algoritmo elegido en toda la información de la base de datos con la excepción de la partición elegida, y se utiliza la partición para evaluar el algoritmo. Es una de las técnicas más usadas en aprendizaje de máquina, ayudando en tareas tales como selección de características, calibración de parámetros y estimar precisión de un modelo dado. (Blockeel, 2002)

La desventaja del este método es que es un cálculo pesado para una computadora, ya que ejecuta un algoritmo "n" veces en un grupo de datos. (Blockeel, 2002)

3.9 Trastorno Antisocial de la personalidad

El trastorno de personalidad antisocial (TPA) es un trastorno de la personalidad perteneciente al Cluster B. (APA, 2013) El grupo B constituye trastornos de dramatización, imprevisibilidad y variabilidad emocional. (PSISE, 2017). Un trastorno antisocial se caracteriza por un patrón de desconsiderar los sentimientos y bienestar de los demás. Estas personas también se pueden conocer como sociópatas o psicópatas (Căndel, 2017).

De acuerdo al DSM-5, manual sobre trastornos psiquiátricos, una persona con TPA se caracteriza por incumplir las reglas sociales, ser una persona que miente seguido, no tener consideración por la seguridad ajena, tener una falta de remordimiento ante

acciones poco éticas, tener una conducta irresponsable y comportamiento impulsivo (APA, 2013). Algunas prognosis para personas con TPA son suicidio, abuso de sustancias, violencia, crimen, desempleo y estar sin hogar (Le Corff, 2014). Por ello, una persona con un trastorno antisocial estará mucho más inclinada a cometer actos como fraude y actividades criminales en general (Căndel, 2017).

De acuerdo a un experimento realizado por Yavuz, algunos factores que serían más comunes en pacientes con TPA serían menor probabilidad de estar casado, mayor probabilidad de estar desempleados, mayor uso de alcohol y de sustancias psicotrópicas, uso del cigarrillo, historia familiar con trastornos antisociales, violencia familiar, intentos de suicidio y migración. (Yavuz, 2016) En aspectos como interacción social y funcionamiento personal estos individuos no parecían tener diferencia con el grupo de control; sin embargo, tenían peores puntajes en ira (aunque igual capacidad de controlarla) y de flexibilidad psicológica. (Yavuz, 2016)

Algunos aspectos como signos neurológicos suaves (NSS) han demostrado tener comorbilidad con trastornos psiquiátricos, incluyendo trastornos antisociales. Demirel midió integración sensorial, coordinación motora, ejecución de actividades motoras complejas y otros medidores NSS; donde todos los pacientes con TPA tuvieron peores resultados. De igual manera existe una mayor prevalencia del abuso de sustancias en la familia, desempleo, rotación frecuente de trabajos y divorcio o soltería (Demirel, 2016). Los signos y medidores neurológicos suaves indican diversos tipos de disfunción cerebral para realizar actividades motoras. Se dice que estos son causados por deficiencia de conexiones en algunas estructuras cerebrales o ausencia de estas. (Dazzan, 2002).

Varios trastornos psiquiátricos tienen comorbilidad entre ellos, lo que significa que la presencia de uno generalmente se da con la presencia del otro igualmente (CAT, 2017). El trastorno antisocial también tiene comorbilidad con el trastorno de conducta desafiante en la adolescencia y con el trastorno AHDH (Déficit de atención e hiperactividad) (Le Corff, 2014). Algunas características de personalidad que se pueden encontrar en individuos con TPA son alta hostilidad, alta impulsividad, búsqueda de emoción, baja responsabilidad, baja contemplación, bajo altruismo, baja genuinidad y valores poco formados. (De Clercq, 2003). Una matriz que relaciona diferentes factores de personalidad con el trastorno antisocial basada en el experimento de De Clercq se

puede ver en la tabla 3.1. De verde están los factores/dominios con una alta correlación positiva y de rojo los factores/dominios con una alta correlación negativa (*tabla 3.1*).

Figura 3.1

Factores de personalidad

Dominio	Faceta	Correlación
Neuroticismo	Ansiedad	-0.09
Neuroticismo	Hostilidad	0.33
Neuroticismo	Depresión	0.1
Neuroticismo	timidez	0.01
Neuroticismo	Impulsividad	0.25
Neuroticismo	Vulnerabilidad	0.11
Extraversión	Afecto	-0.02
Extraversión	Sociabilidad	-0.2
Extraversión	Asertividad	-0.06
Extraversión	Actividad	0.09
Extraversión	Búsqueda de emoción	0.08
Extraversión	Emociones positivas	0.2
Apertura a la experiencia	Fantasía	-0.16
Apertura a la experiencia	Estética	-0.17
Apertura a la experiencia	Sentimientos	-0.2
Apertura a la experiencia	Acciones	0.04
Apertura a la experiencia	Ideas	-0.15
Apertura a la experiencia	Valores	-0.26
Amabilidad	Confianza	-0.3
Amabilidad	Franqueza	-0.47
Amabilidad	Altruismo	-0.42
Amabilidad	Conformidad	-0.33
Amabilidad	Modestia	-0.17
Amabilidad	Mentalidad tierna	-0.31
Responsabilidad	Competencia	-0.28
Responsabilidad	Orden	-0.25
Responsabilidad	Obediencia	-0.41
Responsabilidad	Búsqueda de logros	-0.25
Responsabilidad	Auto-disciplina	-0.33
Responsabilidad	Deliberación	-0.4

Nota: Factores de personalidad altamente correlacionados al trastorno de personalidad antisocial, de acuerdo al estudio de De Clercq. Representando el verde correlación positiva y el rojo correlación negativa.

Fuente: De Clercq, 2003

CAPÍTULO IV: IMPLEMENTACIÓN

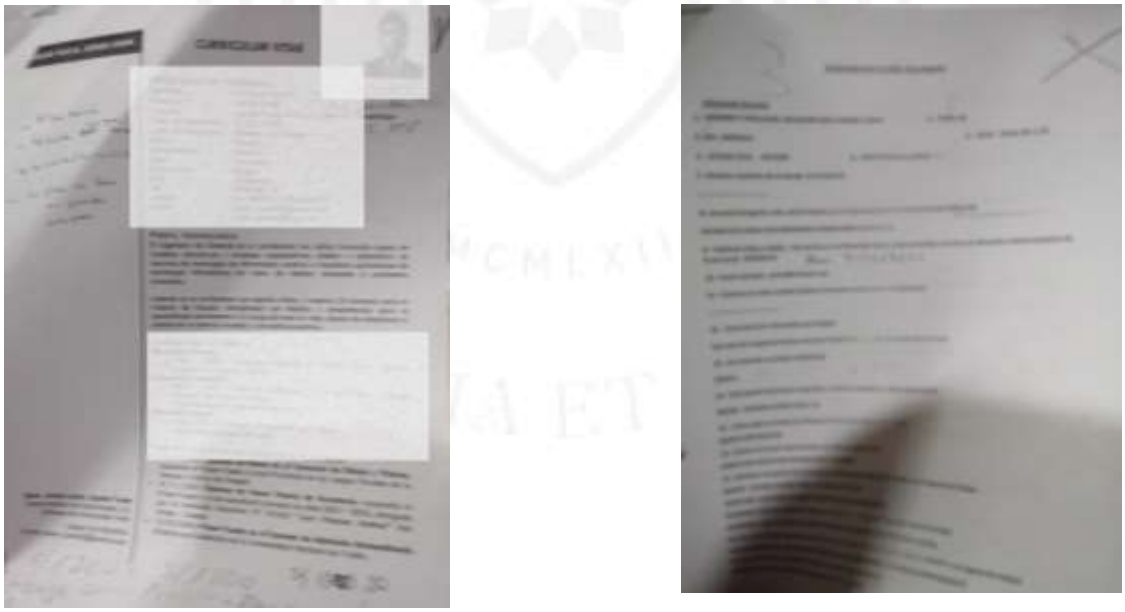
4.1 Datos

Para la muestra de la experimentación se utilizarán 150 registros de empleados de una empresa mayorista de suministros de cómputo, los cuales trabajaron entre 2013 y 2016. Desde el año 2011 aproximadamente se comenzaron a documentar y a realizar más pruebas a los postulantes a la organización, por ello mismo se eligió esta muestra para hacer la experimentación. Es necesario elegir empleados que ya hayan dejado la organización para que se pueda determinar si cometieron o no el ilícito durante su estadía. Dichos registros incluyen la hoja de vida del postulante, la hoja de información y pruebas psicológicas realizadas a los mismos.

La información fue proporcionada por la empresa de manera física. Adicionalmente se obtuvo la información que nos indica que empleados cometieron fraude en la empresa. Ejemplos de los documentos recibidos se pueden apreciar en la figura 4.1, datos sensibles censurados por motivos de protección de datos (*figura 4.1*).

Figura 4.1

Documentos postulantes



Nota: Ejemplo de documentos entregados por la empresa.

4.2 Preparación de datos

Los documentos fueron trasladados a una base de datos estructurada, que contendrá información de las 3 secciones: hoja de vida, hoja de información y pruebas psicológicas. *(figura 4.2)*.

Figura 4.2

Base de datos

	Estado civil	n hijos	N trabajos p	pretensione	Dispuesto a	horas	Sexo(M=0,F	Waartegg1	Waartegg2	Waartegg3	Waartegg4	Fr
8	0	0	2	0		1	0	2	3	8	4	NF
9	0	0	3	850		1	0	7	6	1	5	F
10	3	99	3	1800		5	0	1	7	5	2	F
11	2	1	7	1		5	0	3	7	6	7	NF
12	1	1	6	1000		1	0	8	5	5	2	NF
13	2	2	5	900		1	0	8	7	1	5	NF
14	0	0	6	875		1	0	7	2	3	4	NF
15	0	0	7	800		1	0	TODOS		3	NINGUNO	no lleno	F
16	0	NA	5	NA	NA		0	7	7	4	2	NF
17	0	0	5	1500	NA		0	TODOS	NINGUNO	1	NINGUNO	NF
18	0	0	4	1300	NA		0	1	4	3	5	NF
19	0	0	12	1500	NA		1	1	5	8	6	NF
20	0	0	4	1000	NA		1	5	7	2	NA	NF
21	0	NA	4	NA	NA		0	3	7	6	8	NF
22	0	0	3	1200	NA		0	6	2	8	7	NF
23	0	0	2	1200	NA		0	TODOS	NINGUNO	TODOS	NINGUNO	NF
24	2	0	6	2000	NA		0	8	3	8	4	NF

Nota: Base de datos con la información de los registros.

Algunas de las variables identificadas incluyen edad (en años), estado civil (soltero, conviviente o casado), vivienda alquilada o propia, número de convivientes en hogar, número de trabajos pasados, meses promedio en cada trabajo pasado, tiempo desempleado entre trabajos pasados, número de hijos, respuestas en la prueba Wartegg, resumen en CV, ubigeo geográfico de vivienda, pretensiones laborales, entre otras.

Para el caso de variables cualitativas se utilizarán números para describir las diferentes posibilidades, donde los equivalentes se indicarán en el encabezado de la columna. El identificador para cada uno de los registros en la base de datos será el nombre del empleado. Todos los valores de tipo texto almacenarán la transcripción exacta (incluyendo errores ortográficos) de los textos escritos por el postulante que se encuentren en los registros.

Para variables con datos vacíos o que no se pueden determinar se colocará la denominación “NA”, posteriormente se utilizará algún método de tratamiento de datos vacíos para estos mismos.

Para procesar los resultados de las pruebas psicológicas se cuenta con el apoyo de una psicóloga experta en la prueba de Wartegg. Se convertirán las respuestas en números y fracciones mediante fórmulas, de modo que se puedan utilizar en el modelo.

Como una depuración inicial, en la base de datos consolidada se descartarán todas las columnas tengan valores poco discriminantes o que tengan muchos valores vacíos. Esta versión modificada de la base de datos se utilizará para los algoritmos de predicción.

4.3 Instrumentos

La información se dividirá en 2 grupos.

- El primero contiene variables cuantitativas y cualitativas que puedan ser procesados directamente por los algoritmos predictivos.
- El segundo contiene las respuestas de los postulantes de las pruebas psicológicas, para analizar dicha información se cuenta con la ayuda de una psicóloga experta en el examen de Wartegg. Se utilizará el conocimiento ofrecido por la psicóloga para generar variables latentes para el modelo con las respuestas de los exámenes.

4.4 Procedimiento y diseño

Se utilizarán los algoritmos de Random Forest, Redes Neuronales con regularización Bayesiana (BRNN) y árboles de decisión para predecir qué postulantes a una empresa cometerán fraude interno.

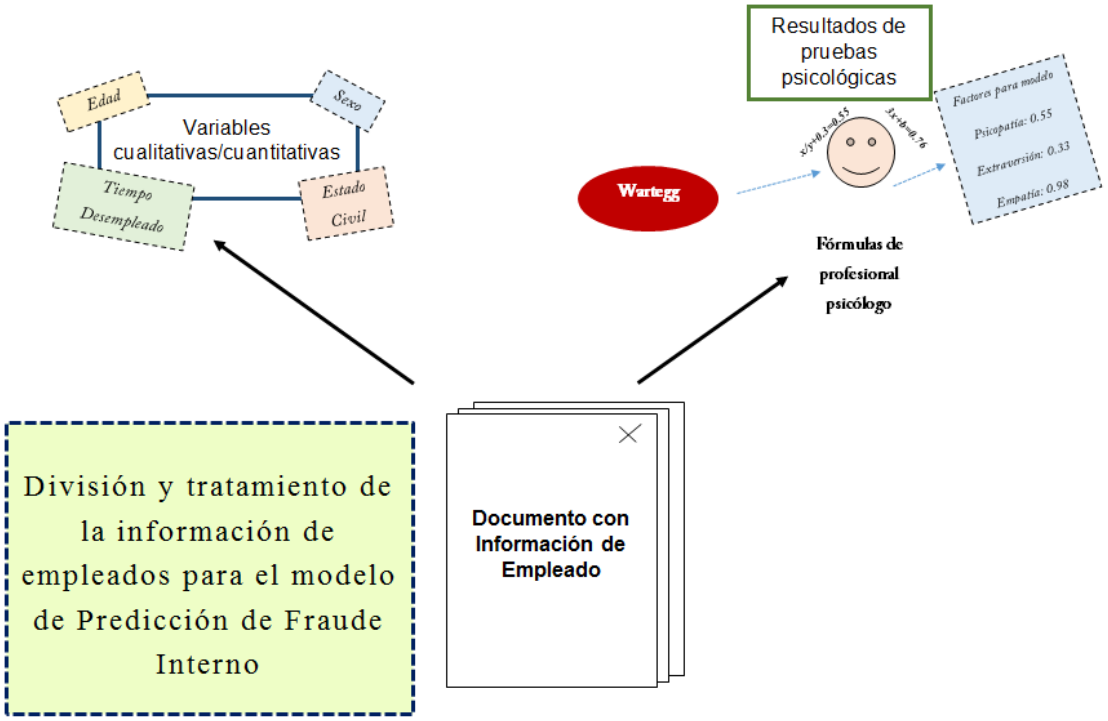
La información obtenida en la primera sección será utilizada para realizar las primeras pruebas de viabilidad en R y en Python, para lo cual se seleccionarán características aptas para el análisis basado en conocimiento teórico. Esta información tendrá que pasar un procesamiento final para que pueda ser ingresada, donde se inferirán los valores vacíos mediante técnicas imputación (minería de datos para predecir valores faltantes), promedios o la técnica de compleción de matriz encontrada en la literatura (Horesh, 2016).

Para la segunda sección se agregarán los valores calculados con las fórmulas del profesional en psicología al modelo. Los resultados describirán factores de personalidad en el postulante, los cuales se vieron que estaban correlacionados para determinar personas con perfiles antisociales de acuerdo a la literatura. (De Clercq, 2003).

Un resumen del procedimiento y diseño del experimento apreciar en la figura 4.3 *(figura 4.3)*.

Figura 4.3

Propuesta de trabajo



Nota: División y tratamiento de la información de empleados para el modelo de Random Forest de predicción de fraude.

4.5 Análisis de datos

Teniendo el modelo con el aporte de las 2 secciones y con las variables para análisis seleccionadas, se procederá a realizar el análisis de la información. Esto consistirá en la búsqueda de patrones mediante gráficos estadísticos en el software R (librería ggplot2). Además, se utilizará la librería de Random Forest de R para hallar algunas de las variables más relevantes estadísticamente (información ofrecida por el software). Posteriormente se creará el modelo final con las características validadas en R.

Para la selección de características se utilizarán métodos encontrados en la literatura (algoritmos evolutivos) y técnicas estadísticas (análisis de componentes, correlación). Se realizarán predicciones en información de entrenamiento, precisión con

información de test, porcentaje de precisión OOB (Out of Bag Estimate) y Cross-Validation como una guía para determinar que variables crean un modelo más apto.

Se comparará el OOB (Out of Bag Estimate) del Random Forest con el porcentaje de precisión obtenida de la data de prueba. Se espera que estos sean muy parecidos, de modo que concuerde con la teoría del algoritmo estudiada Breiman (Breiman, 2001).

4.6 Prueba de viabilidad

Como parte inicial de la implementación se utilizó una partición de la data y de las características para probar la viabilidad del modelo y los siguientes enunciados relacionados con la literatura estudiada.

- El OOB es similar a un porcentaje de precisión obtenido con data de test (Breiman, 2001)
- La cantidad de árboles en Random Forest favorece el resultado del modelo (Breiman, 2001)
- Hay variables relevantes en información de empleados para predecir fraude (tesis en cuestión)

4.7 Ejecución (prueba de viabilidad)

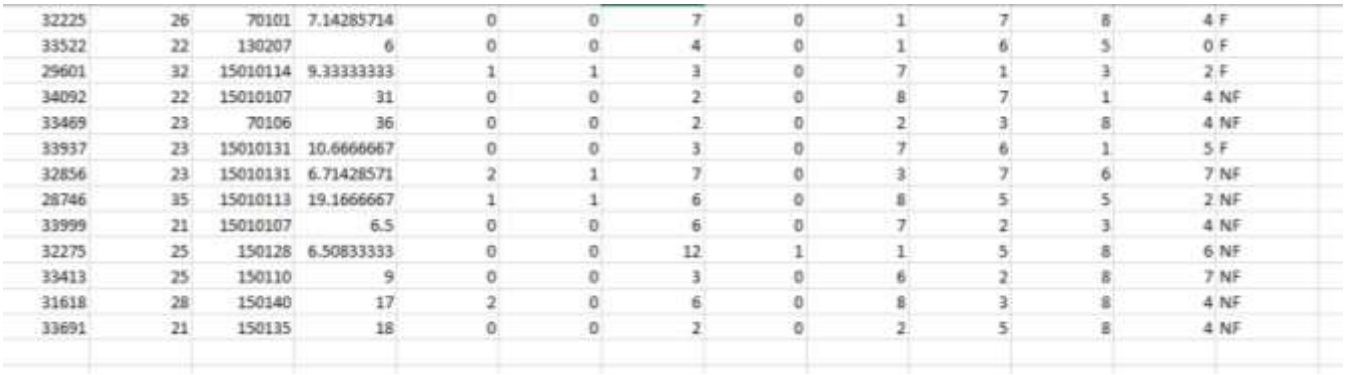
Se utilizó un código de Python que pueda ejecutar un Random Forest en un archivo csv. Para el conjunto de datos no se incluyeron datos vacíos y todas las variables fueron numéricas. Para validar el código, este previamente se probó en "Iris", un conjunto de datos bastante popular de minería de datos (UCI, 2017). Se obtuvieron resultados favorables (94.7% de precisión, ver anexo 1).

El prototipo imprimirá como resultado el número de árboles del bosque entrenado (parámetro) y los resultados de precisión en base al OOB del algoritmo. Este procesamiento se realizará 5 veces y el resultado final será el promedio de estos 5 resultados.

Se crearán 4 archivos de prueba en formato .csv que utilizarán diversos factores numéricos de la base de datos excluyendo los registros que contengan valores vacíos. Uno de los archivos de prueba se puede ver en la figura 4.5 *(figura 4.5)*. Algunas variables que incluyen son estado civil, meses promedio en trabajos pasados, ubigeo, n° de convivientes, n° de hijos y si cometieron fraude o no.

Figura 4.5

Archivo de prueba



32225	26	70101	7.14285714	0	0	7	0	1	7	8	4 F
33522	22	130207	6	0	0	4	0	1	6	5	0 F
29601	32	15010114	9.33333333	1	1	3	0	7	1	3	2 F
34092	22	15010107	31	0	0	2	0	8	7	1	4 NF
33469	23	70106	36	0	0	2	0	2	3	8	4 NF
33937	23	15010131	10.6666667	0	0	3	0	7	6	1	5 F
32856	23	15010131	6.71428571	2	1	7	0	3	7	6	7 NF
28746	35	15010113	19.1666667	1	1	6	0	8	5	5	2 NF
33999	21	15010107	6.5	0	0	6	0	7	2	3	4 NF
32275	25	150128	6.50833333	0	0	12	1	1	5	8	6 NF
33413	25	150110	9	0	0	3	0	6	2	8	7 NF
31618	28	150140	17	2	0	6	0	8	3	8	4 NF
33691	21	150135	18	0	0	2	0	2	5	8	4 NF

Nota: Archivo test2.csv para la prueba Python.

Finalmente, estos se procesarán con diferentes valores de número de árboles en el Random Forest: 1, 5, 50 y además con diferentes números de semilla. Se espera que los resultados mejoren con mayor cantidad de árboles (Breiman, 2001)

4.8 Resultados (prueba de viabilidad)

Los resultados de la prueba de viabilidad se detallan en la tabla 4.1 *(tabla 4.1)*.

La primera columna muestra el número de semilla (seed) usado y el número de árboles(n_trees) que son los parámetros del algoritmo. Las siguientes columnas muestran porcentaje de precisión promedio obtenido (utilizando el OOB) para cada una de las pruebas con los parámetros indicados, y además se incluye la desviación estándar de los promedios a la derecha. Estos resultados se muestran para cada una de las combinaciones de seed/n_trees y para uno de los archivos test creados.

Se puede apreciar que el archivo que tuvo mejores resultados fue test2.csv ya que tuvo los mejores promedios para los 50 árboles y los resultados máximos de todas las pruebas.

Esto se puede atribuir a que el archivo test2.csv era el que tenía más variables, lo que mostraría que más variables benefician un modelo de Random Forest (Breiman, 2001).

Asimismo, para todas las pruebas los resultados son mejores mientras más árboles se introducen para entrenar al modelo, lo que comprueba el segundo teorema descrito.

Tabla 4.1

Resultados Random Forest en Python

Seed,n_trees	**test1**		**test2**		**test3**		**test4**	
	OOB	***s***	**OOB**	***s***	**OOB**	***s***	**OOB**	***s***
2, 1	0.6500	0.1369	0.3000	0.2739	0.4667	0.1826	0.5000	0.3536
2, 5	0.7000	0.2092	0.2000	0.2739	0.4000	0.2789	0.6000	0.2236
2, 50	**0.6500**	0.1369	**0.8000**	0.2739	**0.6667**	0.2357	**0.6500**	0.2236
10, 1	0.4500	0.4108	0.5000	0.3536	0.7333	0.2789	0.5000	0.2500
10, 5	0.4000	0.2850	0.7000	0.2739	0.3333	0.2357	0.4500	0.2092
10, 50	**0.6500**	0.2236	**0.4000**	0.4183	**0.6667**	0.3333	**0.5500**	0.2092
933, 1	0.4000	0.1369	0.5000	0.3536	0.8000	0.1826	0.5500	0.1118
933, 5	0.4000	0.3354	0.4000	0.2236	0.5333	0.1826	0.6000	0.1369
933, 50	**0.5500**	0.3260	**0.8000**	0.4472	**0.5333**	0.1826	**0.6000**	0.2236
744, 1	0.6500	0.2850	0.1000	0.2236	0.6000	0.2789	0.6000	0.2236
744, 5	0.6000	0.1369	0.5000	0.5000	0.7333	0.1491	0.5500	0.2092
744, 50	**0.5000**	0.1768	**0.4000**	0.2236	**0.7333**	0.1491	**0.5000**	0.3062
21, 1	0.3000	0.1118	0.5000	0.3536	0.5333	0.3801	0.3500	0.1369
21, 5	0.5500	0.2092	0.7000	0.2739	0.7333	0.2789	0.6000	0.2850
21, 50	**0.6000**	0.1369	**0.9000**	0.2236	**0.6000**	0.4346	**0.6000**	0.2236
1234, 1	0.4000	0.2850	0.5000	0.0000	0.4667	0.3801	0.6000	0.1369
1234, 5	0.5500	0.2092	0.6000	0.2236	0.6000	0.2789	0.5500	0.2092
1234, 50	**0.6500**	0.1369	**0.7000**	0.2739	**0.5333**	0.3801	**0.5000**	0.2500
total	**0.5361**	**0.2160**	**0.5278**	**0.2882**	**0.5926**	**0.2668**	**0.5472**	**0.2179**
1	**0.4750**	**0.2278**	**0.4000**	**0.2597**	**0.6000**	**0.2805**	**0.5167**	**0.2021**
5	**0.5333**	**0.2308**	**0.5167**	**0.2948**	**0.5556**	**0.2340**	**0.5583**	**0.2122**
50	**0.6000**	**0.1895**	**0.6667**	**0.3101**	**0.6222**	**0.2859**	**0.5667**	**0.2394**

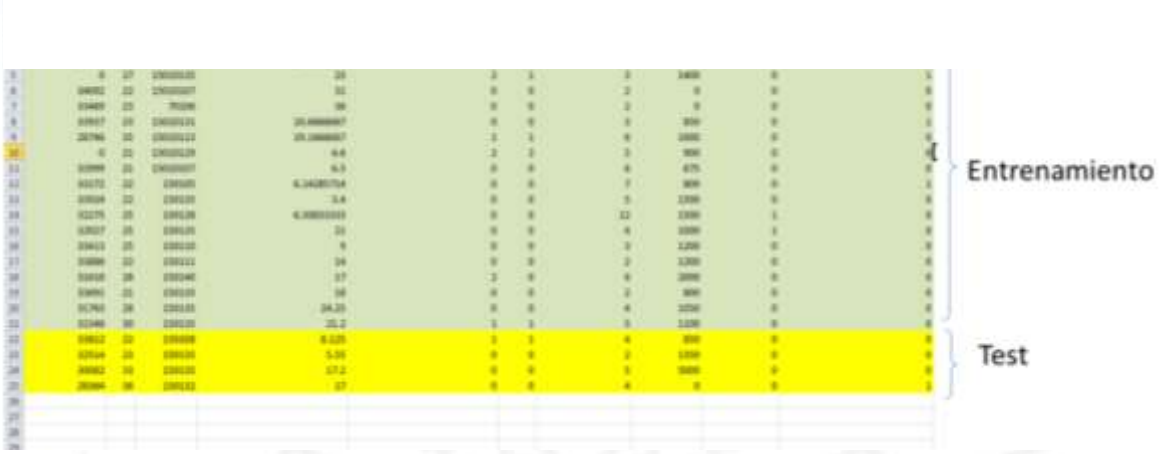
Nota: Resultados de Random Forest en código Python en los archivos de prueba test1, test2, test3 y test4. La columna "OOB" muestra el promedio obtenido del OOB para cada prueba y la columna "s" la desviación estándar para estos resultados

Se hizo una segunda prueba en R, para probar el primer y tercer teorema. Para esto se utilizaron más datos, eligiendo las variables asumidas como más importantes según la literatura en psiquiatría estudiada (*ver Capítulo 3.9*). Una sección de estos se utilizará a manera de prueba; de modo que podamos comparar la precisión del OOB con la de la información de prueba y con una prueba de Cross-Validation.

Con esta segunda prueba se espera comprobar que existen variables relevantes en la información de los postulantes para predecir cuales de estos cometerá fraude interno.

Tabla 4.6

Información de entrenamiento y prueba



Nota: Archivo a ser usado en prueba de R, la información en amarillo denota la información a ser usada para probar el modelo, mientras que el resto para entrenarlo

De acuerdo a la teoría de conductas antisociales (*ver Capítulo 3.9*), el estado marital y los meses de promedio entre trabajos deberían ser buenos predictores de un trastorno de personalidad antisocial, y por ende de fraude. Otras variables escogidas incluyen ubigeo, pretensiones laborales, sexo, nº de hijos y edad. 20 de los registros serán de entrenamiento y los otros 4 de prueba (figura 4.6).

La última columna indica si el empleado cometió fraude, 1 indicando que esto fue positivo.

Los resultados en R se pueden observar en la tabla 4.2 *(Tabla 4.2).*

Tabla 4.2

Random Forest en R

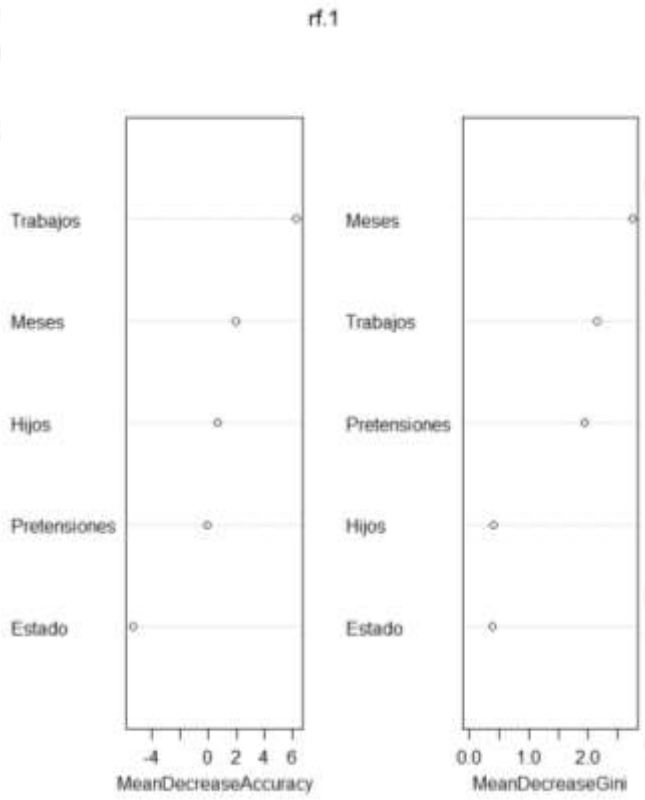
Matriz de Confusión			
OOB		60%	
	0	1	*Error de clase*
0	11	3	**21.43%**
1	5	1	**83.33%**

Nota: Resultados de ejecutar el Random Forest en la información de entrenamiento en la consola de R

Asimismo, se puede ver la valoración de variables del algoritmo en la figura 4.7 (*figura 4.7)*

Figura 4.7

Valoración Variables RF



Nota: Valoración de variables del algoritmo Random Forest tras ejecución del modelo

Como se puede denotar, el Random Forest entrenado obtuvo un error de 40% para el estimado OOB, lo que nos da una precisión de 60% para nuestro modelo. Sin embargo, en la matriz de confusión podemos ver que el modelo es efectivo prediciendo empleados que no cometerán fraude (78.57% de precisión), pero muy deficiente para predecir empleados que cometerán fraude (16.67% de precisión). Al ser nuestro tema de interés detectar los empleados que cometen fraude, nuestra prueba nos muestra que debemos enfocarnos en reducir el error de clase para los empleados que cometieron fraude.

En los gráficos de importancia obtenidos, podemos apreciar que las variables más importantes para nuestro modelo fueron “Trabajos” y “Meses”. La primera describía el número de trabajos pasados y la segunda la cantidad de meses promedio en estos trabajos. Esto se relaciona con uno de los atributos encontrados en la investigación sobre perfiles antisociales. “Job-Hopping” o cambiar frecuentemente de trabajos era una característica común de personas con este trastorno (Yavuz, 2016). Por otro lado, el estado civil, el cual se relacionaba con este trastorno, demostró tener muy poca relevancia en nuestro modelo; esto es contrario al resultado esperado (Demirel, 2016).

Para validar los resultados de precisión obtenidos con OOB se realizó una prueba de Cross-Validation con el mismo modelo de Random Forest. Se utilizaron los parámetros estándar para esta prueba (10 folds, repetido 10 veces).

Tabla 4.3

Cross-Validation en R

mtry	Precisión	Kappa
2	0.58%	-0.09459
3	0.60%	-0.01486
5	**0.61%**	**0.03816**

Nota: Prueba de Cross-Validation en el modelo de Random Forest en R. Los mejores resultados fueron utilizando todas las variables (mtry=5)

En la tabla 4,3 se puede apreciar que el mejor resultado (con mtry=5) obtuvo una precisión de 61% (t*abla 4.3*). Esto es de acuerdo a lo esperado, ya que es muy similar al porcentaje obtenido del OOB (60%).

El uso de todas las variables favoreció a la obtención de esta predicción, y la similitud con el OOB nos muestra que nuestro modelo aún no está muy sujeto a “overfitting”.

Los resultados para las predicciones con la información de prueba se pueden ver en la tabla 4.4 *(tabla 4.4).*

Tabla 4.4

Resultados en archivo de prueba en R

Registro	Resultado	Equivalencia	Real	¿Acertó?
1	1	fraude	0	no
2	0	no-fraude	0	si
3	0	no-fraude	0	si
4	0	no-fraude	1	no

Correctos	2
Total	4
Precisión	50%

Matriz de Confusión			
	0	1	precisión
0	2	1	66.60%
1	1	0	0%

Nota: Resultados de predicciones de modelo de Random Forest entrenado sobre información de prueba.

Al acertar en la mitad de los registros, tendríamos una precisión de 50%. El modelo no pudo predecir correctamente el caso de fraude que se encontraba en nuestra información de prueba, lo que sigue mostrando que el modelo entrenado no es muy eficiente para detectar las personas que cometieron fraude en la empresa.

Los resultados finales de cada prueba serían los siguientes:

- OOB: 60%
- Cross-Validation: 61%
- Test: 50%

Hay similitud y entre los porcentajes y el error de clase en todas las pruebas, lo que probaría el primer teorema definido para la prueba de viabilidad: El OOB tiene un nivel de precisión acertado. (Breiman, 2001).

4.9 Construcción de los modelos

Una vez realizada la prueba de viabilidad se procedió a entrenar los algoritmos con las diferentes secciones de información para aprender más sobre el modelo y encontrar algunas de las variables más importantes.

Se utilizaron 25 registros no relacionados a los registros de la prueba de viabilidad para entrenar un nuevo Random Forest y comparar los resultados. El archivo se puede apreciar en la figura 4.8 *(figura 4.8)*.

Figura 4.8

Datos para entrenar segundo RF

	Edad	Meses	Estado	Trabajos	Sexo	Desempleo	Fraude
1	24	13.800000	0	5	0	0.8000000	0
2	22	3.400000	0	5	0	2.0000000	0
3	25	6.500333	0	12	1	1.0944444	0
4	25	21.000000	0	4	1	0.2500000	0
5	35	40.129000	0	4	0	0.4033333	0
6	25	9.000000	0	3	0	1.6666667	0
7	22	14.000000	0	2	0	0.0000000	0
8	28	17.000000	2	6	0	4.8333333	0
9	21	18.000000	0	2	0	0.5000000	0
10	28	24.250000	0	4	0	0.0000000	0
11	27	13.371429	1	7	0	0.1428571	0
12	31	17.000000	1	4	0	6.7500000	1
13	30	21.200000	1	5	0	4.6000000	0
14	22	8.125000	1	4	0	0.0750000	0
15	23	5.550000	0	2	0	0.0000000	0

Nota: Registros utilizados para entrenar un segundo Random Forest y validar los resultados de la prueba de concepto.

El modelo más exitoso entrenado fue solo utilizado 2 variables, las cuales coinciden con las 2 variables más relevantes encontradas en el experimento previo. *(Tabla 4.5)*

Tabla 4.5

Resultados segundo RF

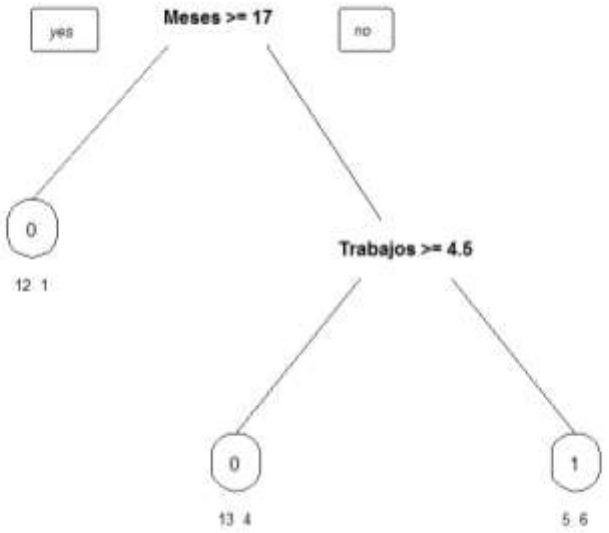
Matriz de Confusión			
OOB		**80%**	
	0	1	*Error de clase*
0	18	3	**14.26%**
1	2	2	**50.00%**

Nota: Resultados del segundo Random Forest entrenado, podemos apreciar mejor precisión global, llegando ésta a ser de 80% y con un error de clase de 50% para predecir los casos de fraude.

Los resultados mejoraron considerablemente, ya que se obtuvo solo un error global de 20% (precisión de 80%) y 50% de error para la clase de predecir fraude. Esto validaría el hecho que las variables más relevantes para predecir fraude en la sección de la información relacionada a información de los postulantes estarían relacionadas con el atributo de "Job-Hopping". (*Ver Capítulo 3.9*)

Figura 4.9

Árbol de decisión 2 variables



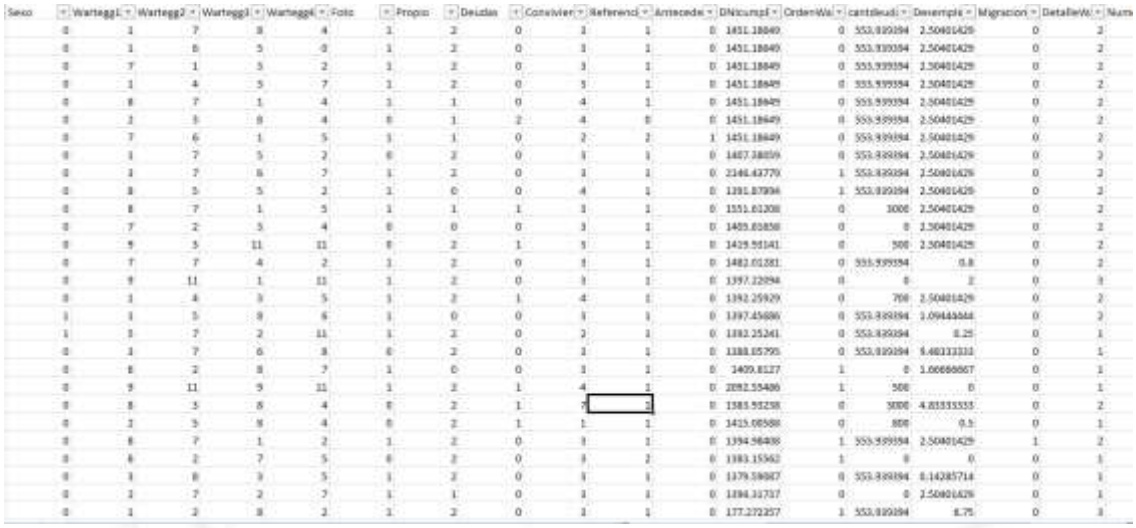
Nota: Árbol de decisión entrenado en un conjunto de datos de 41 registros con 2 variables, se puede ver que la mayor concentración de casos de fraude se da en postulantes que han rotado seguido entre trabajos, lo que se relaciona con perfiles antisociales en la literatura estudiada.

Para poder observar de manera gráfica cómo aportan estas dos variables para clasificar a los empleados entre casos de fraude y casos de no-fraude, entrenamos un árbol de decisión CART con un set de 41 datos de registros de empleados. El árbol de decisión entrenado se puede apreciar en la figura 4.9 *(figura 4.9)*.

Se puede ver que la variable de meses promedio en trabajos determina con más de 80% de precisión en el primer nodo varios casos de no-fraude, ya que si el postulante paso en promedio igual o más a 17 meses en sus trabajos pasados entonces es muy probable que no vaya a cometer fraude interno. En el siguiente nodo el algoritmo determinar que la siguiente división se da cuando la cantidad de trabajos pasados es mayor que 4. Vemos que el árbol puede hallar en esta división los casos de fraude con 53% de precisión. Cabe resaltar que el hecho que el árbol de decisión muestre mejores resultados que el Random Forest puede deberse a "overfitting". Sin embargo, nos valida el hecho que hay bastante capacidad de clasificación en el hecho que un postulante esté cambiando de trabajos muy rápidos, una cualidad asociada a perfiles antisociales, que tanto en la teoría como en los registros de postulantes de la empresa está asociada a cometer fraude interno.

Figura 4.10

Base de datos sin valores vacíos



Sexo	Wartegg1	Wartegg2	Wartegg3	Wartegg4	Foto	Propio	Deudas	Conviven	Referenci	Antecede	DNIcumpl	OrdenWa	cantdeud	Desempleo	Migracion	DetalleWa	Num
0	1	7	8	4	1	2	0	1	1	0	1451.18849	0	553.939394	2.50401429	0	2	
0	1	6	5	0	1	2	0	1	1	0	1451.18849	0	553.939394	2.50401429	0	2	
0	7	1	5	2	1	2	0	1	1	0	1451.18849	0	553.939394	2.50401429	0	2	
0	1	4	5	7	1	2	0	1	1	0	1451.18849	0	553.939394	2.50401429	0	2	
0	8	7	1	4	1	1	0	4	1	0	1451.18849	0	553.939394	2.50401429	0	2	
0	2	5	8	4	0	1	2	4	0	0	1451.18849	0	553.939394	2.50401429	0	2	
0	7	6	1	5	1	1	0	2	2	1	1451.18849	0	553.939394	2.50401429	0	2	
0	1	7	5	2	0	2	0	1	1	0	1407.38859	0	553.939394	2.50401429	0	2	
0	1	7	6	7	1	2	0	1	1	0	2146.43779	1	553.939394	2.50401429	0	2	
0	8	5	5	2	1	0	0	4	1	0	1291.87994	1	553.939394	2.50401429	0	2	
0	8	7	1	5	1	1	1	1	1	0	1551.61208	0	3000	2.50401429	0	2	
0	7	2	5	4	0	0	0	1	1	0	1405.61858	0	0	2.50401429	0	2	
0	9	5	11	11	0	2	1	1	1	0	1429.55141	0	500	2.50401429	0	2	
0	7	7	4	2	1	2	0	1	1	0	1482.01281	0	553.939394	0.8	0	2	
0	9	11	1	11	1	2	0	1	1	0	1397.22094	0	0	2	0	3	
0	1	4	3	5	1	2	1	4	1	0	1392.25929	0	700	2.50401429	0	2	
1	1	5	8	6	1	0	0	1	1	0	1397.45686	0	553.939394	1.09444444	0	2	
1	1	7	2	11	1	2	0	1	1	0	1382.25241	0	553.939394	0.25	0	1	
0	1	7	6	8	0	2	0	1	1	0	1388.65795	0	553.939394	5.48333333	0	1	
0	8	2	8	7	1	0	0	1	1	0	1409.8127	1	0	1.66666667	0	1	
0	9	11	9	11	1	2	1	4	1	0	2092.55486	1	500	0	0	1	
0	8	5	8	4	0	2	1	1	1	0	1383.55258	0	3000	4.83333333	0	2	
0	2	5	9	4	0	2	1	1	1	0	1415.08588	0	800	0.5	0	1	
0	8	7	1	2	1	2	0	1	1	0	1394.98408	1	553.939394	2.50401429	1	2	
0	6	2	7	5	0	2	0	1	2	0	1383.15562	1	0	0	0	1	
0	1	8	3	5	1	2	0	1	1	0	1379.59847	0	553.939394	0.14285714	0	1	
0	1	7	2	7	1	1	0	1	1	0	1394.31737	0	0	2.50401429	0	1	
0	1	2	8	2	1	2	0	1	1	0	177.272257	1	553.939394	8.75	0	3	

Nota: Base de datos generada con 63 registros de postulantes, se convirtieron todas las variables a numéricas, de modo que se puedan hallar promedios para reemplazar los valores vacíos.

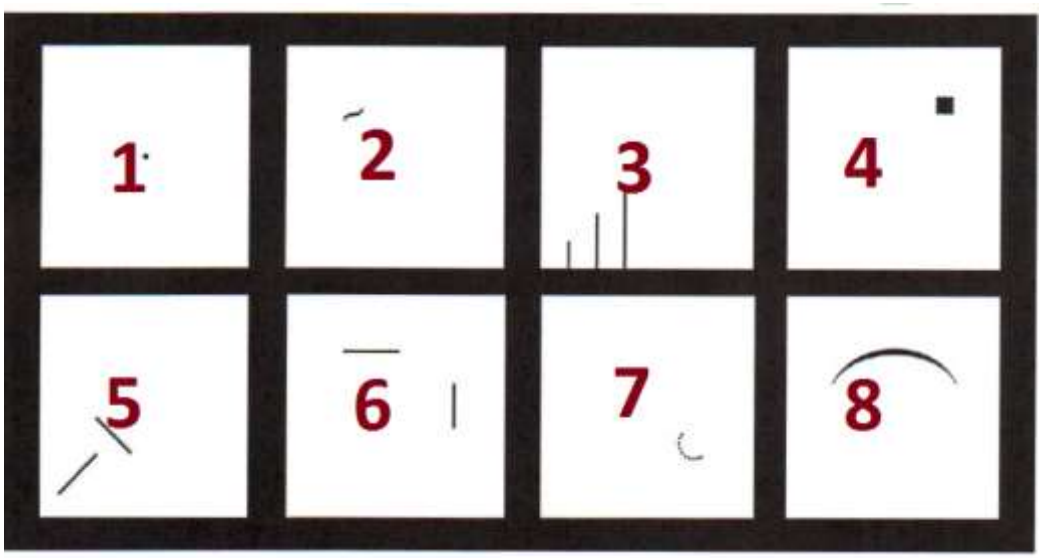
Se continuó la exploración con otras variables del modelo. Para esto se utilizaron 63 registros de los empleados. Se convirtieron todos los valores a numéricos en una hoja de cálculo y se reemplazaron los valores vacíos como promedios de la misma columna. Esta nueva base de datos se detalla en la figura 4.10. *(figura 4.10)*

Se generó nuevamente un Random Forest y un árbol de decisión con esta información. La variable “Trabajos” siguió demostrando ser la más relevante, sin embargo, como variables secundarias se obtuvieron otras dos que no se tenían anteriormente. Una de estas fue “Wartegg4”. Esta característica era la respuesta del postulante a una pregunta en el examen psicológico de Wartegg, donde se le preguntaba que dibujo le pareció más difícil. Se ingresó a la base de datos para esta columna el número que corresponde al dibujo que el postulante eligió. Los cuadros del examen y sus números correspondientes se detallan a continuación. (*Figura 4.11*)

Si el postulante colocaba algo como “TODOS”, “NINGUNO” o dejaba el renglón vacío se asignaba el valor de 9 y 11 respectivamente. El árbol de decisión muestra la relación entre estas dos variables. Como se puede ver una mayor cantidad de trabajos está relacionado con la probabilidad de cometer fraude (57%). Si el postulante eligió a su dibujo más difícil como el 6 o el 2, entonces hay un 56% de posibilidades

Figura 4.11

Cuadros de la prueba de Wartegg



Nota: Numeración de los cuadros con los estímulos de la prueba de Wartegg, en ella los postulantes tienen que dibujar en los cuadros de forma libre y responder preguntas.

que vaya a cometer fraude. Los postulantes que eligieron otros números son casi en su totalidad empleados que no cometieron este acto (92%). *(Figura 4.12)*

Se entrenó un Random Forest con estas dos variables y se obtuvo precisión global de 76.7% y un error de clase para predecir fraude de 47% (53% de precisión para detectar casos de fraude), lo que presenta una mejora frente al modelo obtenido previamente con 25 registros. *(Tabla 4.6)*

Tabla 4.6

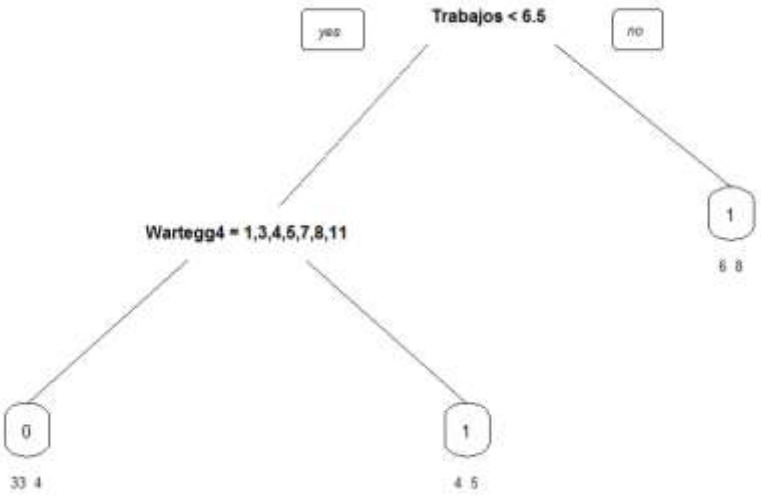
Número de trabajos y Wartegg4

Matriz de Confusión			
OOB		**77%**	
	0	1	*Error de clase*
0	37	6	**13.95%**
1	8	9	**47.06%**

Nota: Árbol de decisión generado con 63 registros de empleados, que muestra la relación entre el número de trabajos del empleado, su respuesta en la última pregunta del examen de Wartegg y el cometido de fraude interno.

Figura 4.12

Random Forest N°trabajos y Wartegg4

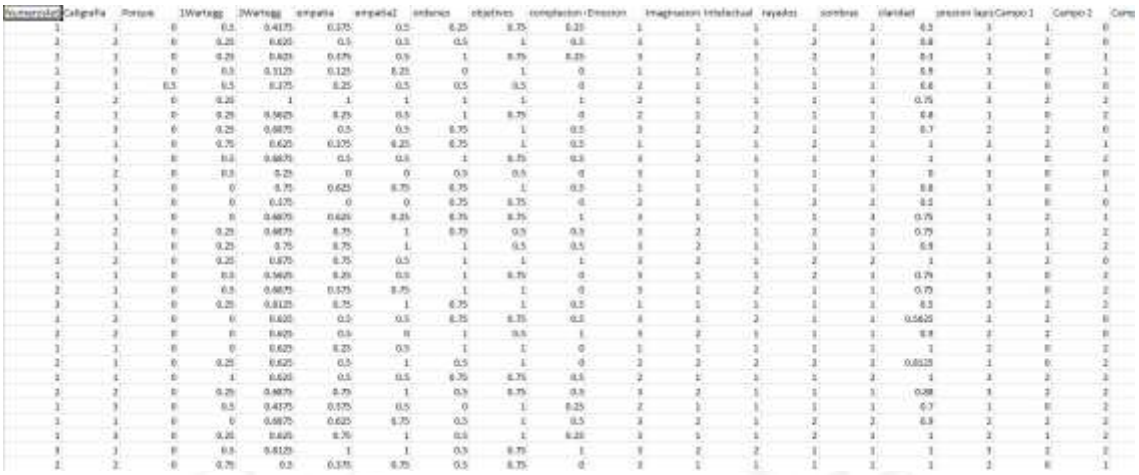


Nota: Random Forest generado con las variables de “Trabajos” y “Wartegg4”, entrenado con 63 registros. Se obtuvo un OOB estimate de 76.7% y un error de clase de 47% para predecir fraude.

Después de esto se realizaron pruebas con la sección 2 de la información tras procesar las pruebas de Wartegg con las fórmulas entregadas por la psicóloga. Se obtenían los resultados analizando los objetos dibujados, las respuestas a las preguntas y el tipo de trazo en los dibujos. Cada campo tiene una evaluación categórica del 0 al 2, siendo esta última la mayor puntuación. Cada campo de acuerdo por la teoría facilitada por la especialista psicóloga está relacionado con alguna faceta de la personalidad. Cada una de estas facetas fue evaluada entre 0 y 1 en base a los resultados de los campos. Los más relevantes a tomar atención serían el campo 2 y 7, relacionados con la empatía. La base de datos resultante (38 registros) se puede ver en la figura 4.13 (*figura 4.13*). La sección 2 de la información contendría lo que se conoce como variables latentes (*Ver Capítulo 3.4.1.2*)

Figura 4.13

Sección 2 de información

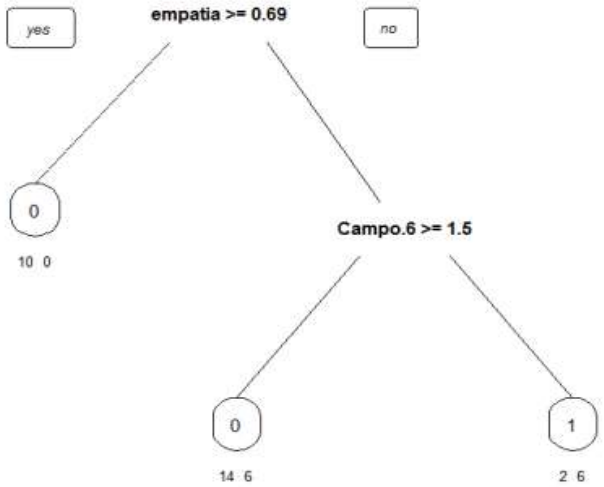


Nota: 38 registros de la sección de la información. Algunas variables incluyen empatía, %Wartegg, claridad, objetivos, Campo 1-8, etc.

Se procesaron estos registros con el algoritmo de C.45 (*figura 4.14*) y un random forest (*tabla 4.7*).

Figura 4.14

C.45 sección 2 de información



Nota: Árbol de decisión CART obtenido con el procesamiento de registros únicamente que incluyen variables de la sección 2 de la información. La variable empatía es la más relevante.

Tabla 4.7

Random Forest (solo sección 2 inf.)

Matriz de Confusión			
OOB		**74%**	
	0	1	*Error de clase*
0	24	2	**7.69%**
1	8	4	**66.67%**

Nota: Random Forest procesado con solo la sección 2 de la información. Se obtuvo un porcentaje de precisión de 73. 68% aproximadamente con 1000 árboles.

Los resultados muestran que las variables obtenidas de la prueba psicológica pueden aportar al modelo. El resultado del árbol de decisión muestra que la variable más relevante es la empatia, donde incluso se puede segmentar con un grupo de 100% de personas que no cometieron fraude cuando la calificación fue mayor a 0.69.

4.10 Tratamiento de la información

Para depurar las variables se utilizaron diversos métodos. Algunos de estos incluyen Information Gain, Análisis de componentes, valoración de variables del random forest, análisis visual con árboles de decisión, análisis visual con gráficos estadísticos , entre otros. Algunas de estas se detallaran a continuación.

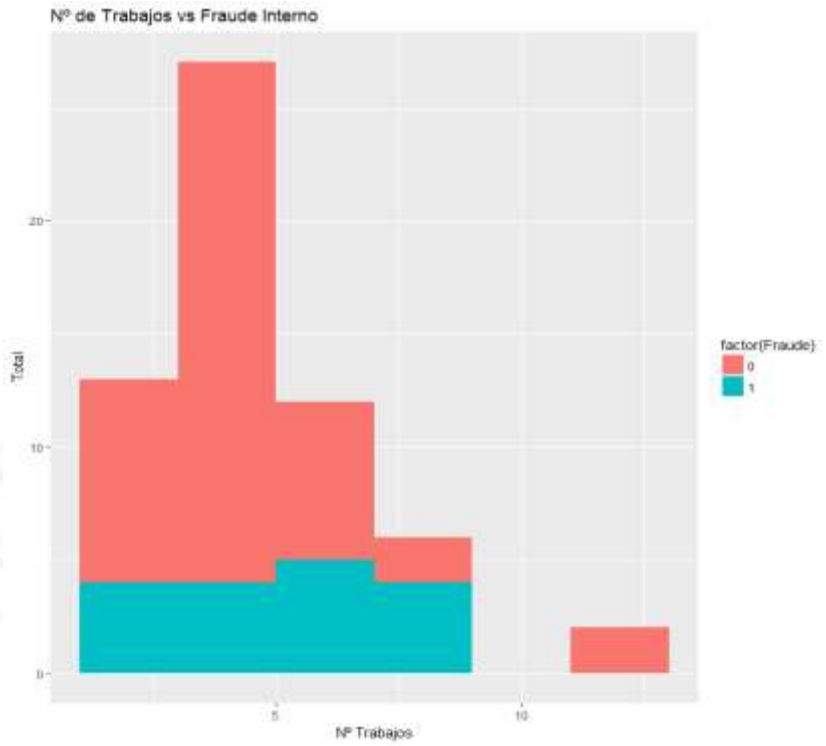
La literatura de psiquiatratía estudiada en TPA, nos indicaba que "Job-Hopping" era una de las características de las personas con este trastorno. Esto hace referencia a un individuo que cambia de trabajos frecuentemente y sin mucho tiempo de estadía en cada uno. Por ende, se esperaría que la cantidad de trabajos pasadas esté de alguna manera relacionado con la probabilidad que una persona tenga un trastorno antisocial (*ver capítulo 3.9*), o en otras palabras, con la probabilidad que una persona vaya a cometer fraude interno si ingresa a una compañía.

Se obtuvo un gráfico (*figura 4.15*) con los 63 registros, donde se colocaba en el eje Y el numero de postulantes, y en el eje X intervalos de 5 en 5 de el Nº de trabajos pasados. El gráfico con columnas apiladas muestra que cantidad de postulantes hay en cada intervalo. La columna roja represanta postulantes que no cometieron fraude, mientras que la azul postulantes que cometieron fraude.

El gráfico obtenido muestra una clara correlación entre un nº alto de trabajos pasados y mayor probabilidad que el postulante haya cometido fraude, ya que la columna azul tiene mayor peso comparado con la columna roja en el intervalo 5-10. Esto también se puede observar en varias de las pruebas realizadas, donde el Nº de trabajos tiene un peso importante en la valoración del random forest, o varios de los árboles de decisión generados aplican los primeros split con esta variable.

Figura 4.15

Nº de Trabajos y proporción Fraude Interno

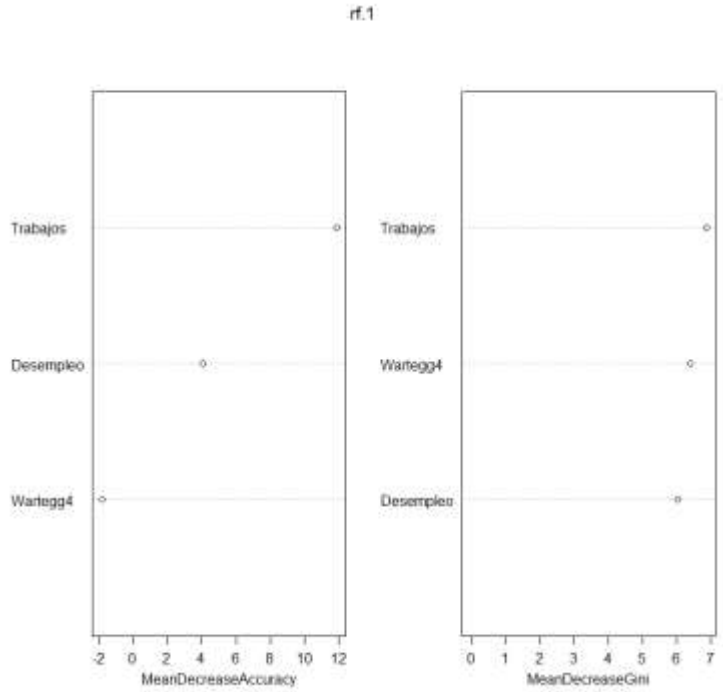


Nota: Nª de postulantes por comparado con una variable que describe la cantidad de trabajos pasados. Las columnas azules representan los postulantes que cometieron fraude. Se puede apreciar que la proporción de fraude va a aumentando conforme aumenta en el Nº de trabajos.

Otra herramienta útil para el análisis de variables fue el uso de la valoración del random forest. Esta crea un ranking de las variables en base a dos indicadores. El primero calcula cuanta precisión se pierde en el modelo si se retira alguna variable. El segundo es el coeficiente de Gini propuesto por Breiman, el cual calcula cuanta variación aportan las variables al modelo y que tan probable es que una en específico sea el primer split de alguno de los árboles aleatoriamente generados.

Figura 4.16

Valoración de variables más importantes primera sección



Nota: Valoración de Random Forest generado con algunas de las variables más importantes de la primera sección de la información. Se puede ver que las variables que aportan más a la precisión obtenida son las relacionadas a "Job-Hopping", que confirma lo encontrado en la literatura sobre perfiles antisociales y fraude interno.

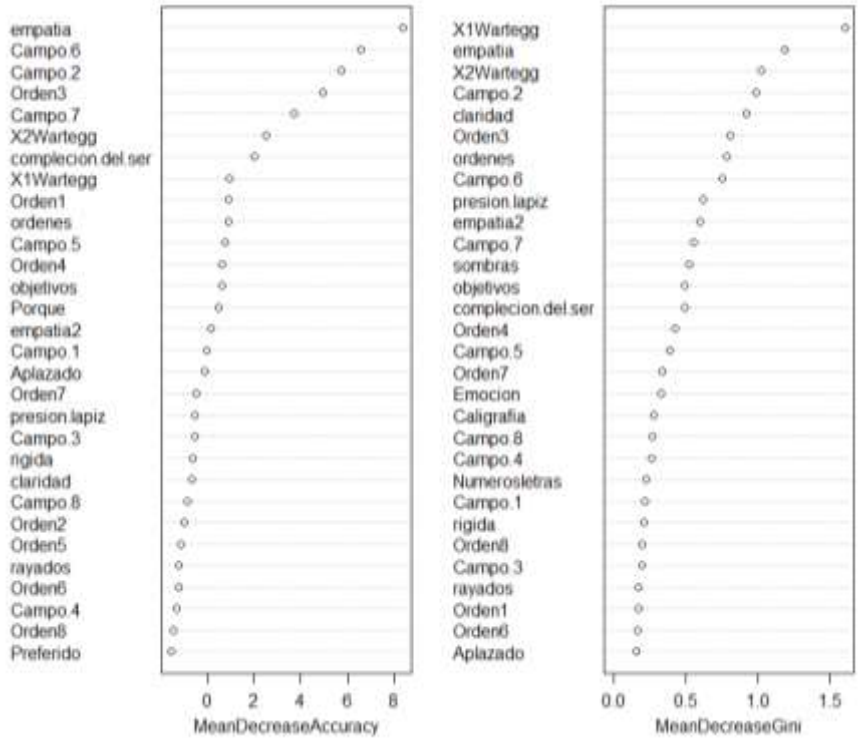
En la figura 4.16 (*figura 4.16*) podemos ver la valoración obtenida solo con las variables más importantes de la primera sección. Las mejor valoradas son "Trabajos" y "Desempleo", las cuales están relacionadas con atributos relacionados a la empleabilidad y "Job-Hopping". Como se vió en el capítulo 3.9 (*Capítulo 3.9*), una persona antisocial tiene más tendencia a rotar rapidamente de trabajo o estar más tiempo desempleado. El algoritmo nos muestra también que estas carácterísticas también son buenos predictores para cometer fraude, lo que comprobaría lo encontrado en la literatura estudiada.

En la figura 4.17 (*figura 4.17*) se procesó todas las variables de la segunda sección de la información con Random Forest. Se puede ver que varias de las características del test psicológico están aportando al algoritmo, lo que muestra que se pueden obtener variables relevantes con este recurso. Además, comparando los rankings en los dos

indicadores, podríamos concluir que las variables más relevantes son X1Wartegg y X2Wartegg (evaluaciones globales el examen), empatia y Campo.2 y Campo.6 (Evaluaciones específicas del campo 2 y 6). Siendo las dos primeras las más importantes evaluaciones del examen, y las tres siguientes atributos directamente relacionados con la empatía del evaluado. Como se vió previamente (*Capítulo 3.9*) la empatía estaba indirectamente correlacionada a las posibilidades que una persona tenga un perfil antisocial, o en consecuencia que cometa fraude interno en una empresa. De esta manera se puede ver concordancia entre la literatura encontrada, la información de los postulantes y los resultados obtenidos.

Figura 4.17

Valoración de variables segunda sección



Nota: Valoración de Random Forest que incluye todas las variables de la segunda sección. Se puede ver que algunas de las variables más relevantes son X1Wartegg y X2Wartegg (evaluaciones globales el examen), empatia y Campo.2 y Campo.6 (Evaluaciones específicas del campo 2 y 6. Esto muestra que los resultados del examen aportan al modelo, y que además los indicadores más importantes son los relacionados de la empatía (que concuerda con la teoría estudiada sobre TPA).

Tabla 4.6

Information Gain entre Fraude y variables del modelo

Nº	Var	IG Con Fraude	Nº	Var	IG Con Fraude	Nº	Var	IG Con Fraude	Nº	Var	IG Con Fraude
1	Campo.7	*0.091425421*	17	ordenes	*0.016680845*	33	Deudas	*0.006205621*	49	OrdenWart	*0.001475227*
2	complecion.del.ser	*0.053865307*	18	DNIcumple	*0.015032676*	34	Migrante	*0.005976484*	50	Wartegg4b	*0.001430037*
3	Trabajos	*0.051876842*	19	Edad	*0.014987125*	35	Orden3	*0.005731791*	51	Campo.8	*0.001293128*
4	Campo.2	*0.049653928*	20	Orden5	*0.014522957*	36	Wartegg2b	*0.005622095*	52	presionLap	*0.000231368*
5	empatia	*0.047660552*	21	Wartegg1	*0.013283822*	37	Pretensiones	*0.00515447*	53	claridad	*0.000222708*
6	sombras	*0.038449536*	22	Imaginacion	*0.012547367*	38	Wartegg2	*0.005077455*	54	Propio	*0.000222708*
7	Convivientes	*0.033551561*	23	Wartegg3b	*0.012297392*	39	Cantdeuda	*0.004464514*	55	Orden4	*8.81E-05*
8	EvaluacWar	*0.031876476*	24	Intelectual	*0.012163681*	40	Campo.4	*0.004143434*	56	Orden2	*8.81E-05*
9	Estado	*0.031501976*	25	NumerosLetrasWar	*0.012092493*	41	rigida	*0.004077245*	57	Orden8	*0*
10	Hijos	*0.029648918*	26	PerfectoWar	*0.010888492*	42	ConcIntWar	*0.004077245*	58	Aplazado	*0*
11	Desempleo	*0.02853149*	27	rayados	*0.010092147*	43	Referencias	*0.002599598*	59	Campo.3	*0*
12	Orden6	*0.028050147*	28	Campo.6	*0.008791356*	44	Orden1	*0.001985954*	60	Emocion	*0*
13	Meses	*0.022947546*	29	objetivos	*0.008576413*	45	Preferido	*0.001985954*	61	Letras	*0*
14	Campo.5	*0.02249697*	30	Wartegg3	*0.007992301*	46	Sexo	*0.001717197*	62	Foto	*0*
15	empatia2	*0.020799299*	31	Wartegg4	*0.007530264*	47	Campo.1	*0.001691977*			
16	Wartegg1b	*0.020206387*	32	DetalleWar	*0.007384802*	48	Orden7	*0.001656491*			

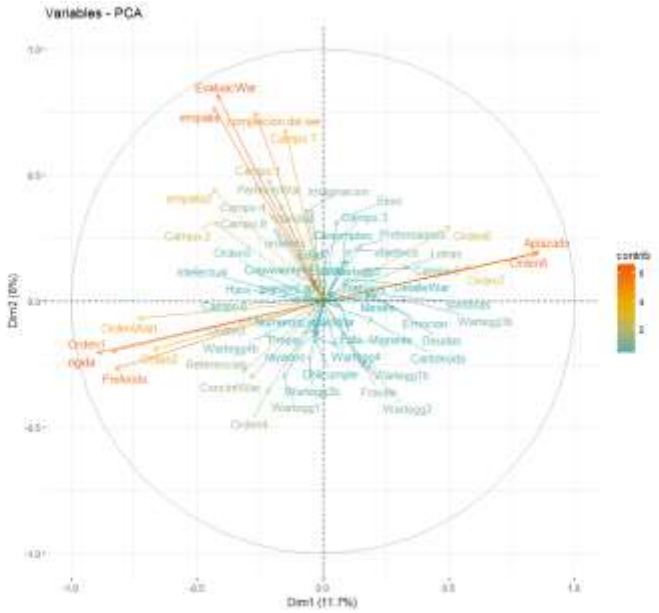
Nota: Coeficientes de entropía obtenidos con la librería infotheo en R, comparando las características del modelo con la variable a predecir de fraude.

De la misma manera también se hicieron uso de otras técnicas encontradas en la literatura como Information Gain (Rashid, 2016). Para esto se utilizó la librería infotheo de R; donde se halló la información común entre todas las variables del modelo y la variable de fraude. Mientras mayor sea el nivel de entropía o información común entre ambas variables, estas estarán más correlacionadas. Se puede apreciar que algunos de los valores más altos se obtuvieron de las características del examen de Wartegg o variables relacionadas con un TPA (*Tabla 4.8*); vamos también características con valores muy bajos, las cuales probablemente no pueden aportar mucho al modelo.

Finalmente, otro de los métodos a utilizar fue el de Análisis de Componentes (PCA) para analizar los atributos del modelo. Se realizó un análisis de componentes de las 62 características del modelo (sección 1 y sección 2). Los resultados se pueden apreciar en la figura 4.18 (*figura 4.18*). Se incluyeron en el formato del PCA colores y flechas para poder visualizar más facilmente los resultados. Como se puede ver, hay carácterísticas que están aportando más al análisis (las que tienen más variación, o están más alejadas del centro), mientras que las variables más cercanas al centro son las que aportan poco al modelo.

Figura 4.18

PCA (63 características)

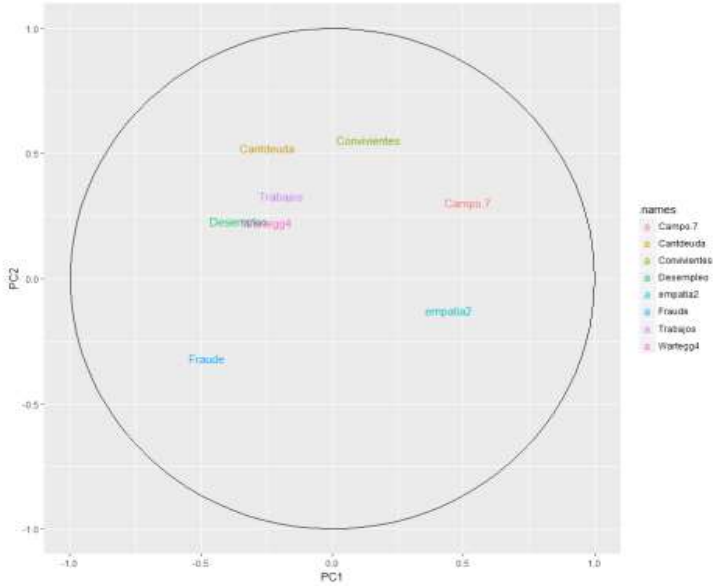


Nota: Análisis de componentes realizado en R con las 63 características del modelo, que incluye información de la sección 1 y sección 2 (de acuerdo a las definiciones del capítulo 4.4).

Retirando iterativamente las características que aportaban menos al modelo (según el PCA), llegamos a un modelo que contenía solo los siete atributos más relevantes. (*figura 4.19 y figura 4.20*). Como se puede observar varias de estas están relacionadas con el examen de Wartegg (empatia2, Campo.7,Wartegg4); y las otras están relacionadas con atributos que se vieron en el estudio de perfiles antisociales (*ver Capítulo 3.9*). El aspecto más llamativo es que la variable "Campo.7" está en un plano opuesto a la variable "Fraude" (variable a predecir), esto nos dice que son inversamente proporcionales. Esto significaría que mientras mayor es la evaluación del postulante en el campo 7 del examen de Wartegg (relacionado con la empatia de acuerdo a la información proporcionada por la experta psicóloga), menor es la probabilidad que el postulante vaya a cometer Fraude interno, o viceversa. Lo anterior mencionado muestra una relación relevante entre la información ofrecida de la empresa, lo encontrado en la literatura y la información proporcionada por la psicóloga.

Figura 4.19

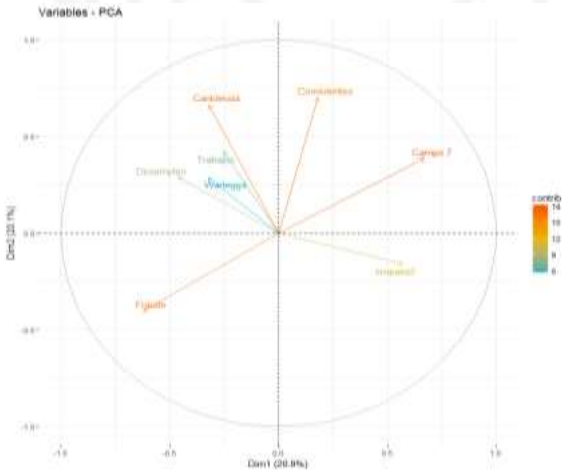
PCA (más relevantes, formato 1)



Nota: Análisis de componentes realizado en R con las características más relevantes del modelo, formato 1 (sin flechas ni colores de variación). Se puede ver qué Campo.7 y Fraude son inversamente proporcionales

Figura 4.20

PCA (más relevantes, formato 2)



Nota: Análisis de componentes realizado en R con las características más relevantes del modelo, formato 2 (con flechas y colores de variación). Se puede ver qué Campo.7 y Fraude son inversamente proporcionales

El detalle de las variables más relevantes obtenidas por el PCA sería el siguiente:

Wartegg4 (*Cuadro elegido en respuesta a la última pregunta del examen de Wartegg, "¿Cuál dibujo te pareció más difícil?")*

Desempleo (*Promedio de tiempo desempleado entre trabajos del postulante, obtenido con los meses de estadía expuestos en los trabajos pasados en el CV. De acuerdo a la teoría del TPA [ver Capítulo 3.9], las personas con TPA tienden a estar más tiempo desempleados*)

Trabajos (*Número de trabajos pasados del postulante, obtenido contando número de trabajos pasados expuestos en el CV. De acuerdo a la teoría del TPA [ver Capítulo 3.9], las personas con TPA tienden a rotar más entre trabajos.)*

Campo.7 (*Evaluación del Campo 7 del examen de Wartegg utilizando las fórmulas facilitadas por la experta psicóloga, penúltimo campo del examen que está cercanamente relacionado con la empatía*)

Convivientes (*Número de personas con la que el postulante convive, obtenido de hoja de información donde se le hacen diversas preguntas al postulante. De acuerdo a la teoría del TPA [ver Capítulo 3.9], las personas con TPA tienden más a vivir solos*)

Cantdeuda (*Cantidad de deuda que tiene el postulante, obtenido de hoja de información donde se la hacen diversas preguntas al postulante y también de evaluaciones de riesgo financieras. De acuerdo a la teoría del TPA [ver Capítulo 3.9], las personas con TPA a ser más impulsivas y endeudarse más*)

empatia2 (*Ponderación de evaluación de los campos de empatía del examen [Campos 2 y 7], ambos cercanamente relacionados con la empatía de acuerdo a la información proporcionada por la experta psicóloga*)

4.11 Resultados de implementación

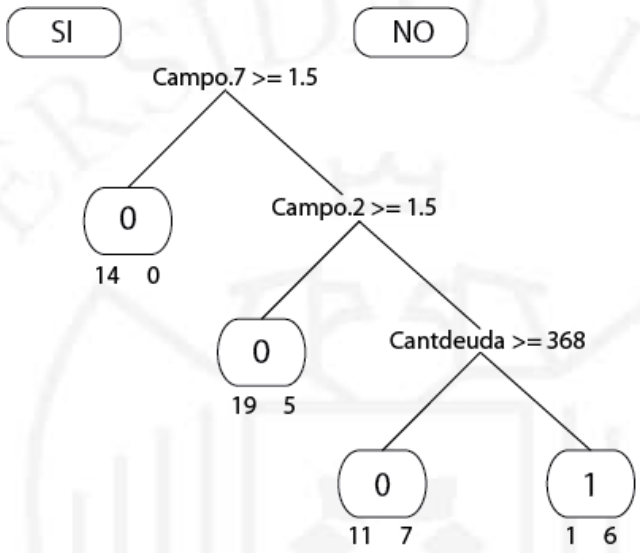
Una vez hechas pruebas con la dos primeras secciones de la información se consolidaron 63 registros con ambas secciones para realizar una prueba y comparación final entre los algoritmos de precisión.

Se compararon resultados con C.45 (*figura 4.21*), random forest (t*abla 4.9*) y con una red neuronal bayesiana (t*abla 4.10*) en R. Para este dataset todas las características se convirtieron a una forma numérica y todos los valores vacíos se reemplazaron con promedios. Los resultados finales se pueden apreciar la tabla 4.11. (t*abla 4.11*)

Se puede observar que el arbol de decisión hizo las principales divisiones con las variables “Campo.7” y “Campo.2”. Estos campos son las evaluaciones de los campos 7 y 2 respectivamente del examen de Wartegg. De acuerdo a lo indicado por la psicóloga, ambos de estos campos están directamente relacionados con la empatía.

Figura 4.21

C4.5 (63 registros)



Nota: Árbol de decisión final generado con los 63 registros y todas las variables. Se puede ver que las variables que aportaban más eran Campo.7, Campo.2 y Cantdeuda.

En el caso de BRNN, obtenemos el mejor RMSE de 0.4561 utilizando 6 neuronas. Los valores R cuadrado bajos indican que la data está muy dispersa, lo que es algo negativo para predicción. Al haber utilizado todas las características del modelo, se presume que realizando una selección de características se obtendrían mejores resultados.

Tabla 4.9

Resultados finales RF

Matriz de Confusión			
OOB		**76%**	
	0	1	*Error de clase*
0	39	6	**13.33%**
1	9	9	**50.00%**

Nota: Resultados obtenidos utilizando el algoritmo de Random Forest, se tiene un ratio de 50% de precisión solo para los postulantes que cometieron fraude.

Como se puede ver los mejores resultados se obtuvieron del random forest. El C.45 tuvo una precisión global ligeralmente más alta que el random forest, sin embargo este último fue mejor para poder predecir los postulantes que cometieron fraude. Se debe considerar también que el C.45 (modelo que tiene tendencia a hacer "over-fitting") obtuvo la precisión sobre los mismos datos que entrenó; mientras que el random forest utilizó el OOB. En el caso del modelo de red neuronal con regularización bayesiana, el porcentaje de precisión estimado de 60% es menor al de los otros métodos; además el valor de R cuadrado demuestra que los datos estan muy dispersos. Sin embargo, no se descarta que el algoritmo pueda tener resultados más favorables realizando selección de características y optimización de parámetros. Tomando en consideración todo esto el ranking final de los métodos para este problema sería el siguiente:

1. Random Forest
2. C.45
3. BRNN

Tabla 4.10

Resultados BRNN (63 registros)

neuronas	RMSE	R cuadrado	MAE
1	0.46675	0.1656	0.40528
2	0.45795	0.1822	0.39931
3	0.46050	0.1912	0.40481
4	0.45398	0.1853	0.39732
5	0.45359	0.1860	0.39640
6	**0.45061**	**0.1835**	**0.39555**
7	0.45246	0.1846	0.39647
8	0.45163	0.1856	0.39748
9	0.45213	0.1846	0.39800
10	0.45135	0.1853	0.39702

Nota: Resultados de red neuronal con regularización bayesiana con Cross-Validation aplicada en los 63 registros. Se puede ver qué se obtiene un error cuadrático promedio de 0.456 aproximadamente, con la mejor iteración del modelo que fue utilizando 6 neuronas. El valor de R cuadrado bajo representa alta dispersión en la data.

Tabla 4.11

Resultados finales

N° registros	sección información	Método	Variables más aportantes	Variables usadas	Precisión	Precisión (solo para personas que cometieron fraude)
60	1	RF	Trabajos, Desempleo	Trabajos, Desempleo	76.67%	52.95%
60	1	C.45	Trabajos, Wartegg4	Todas	76.66%	76.47%
38	2	RF	empatia, Campo 6	empatia, Campo 6	73.68%	33.34%
38	2	C.45	empatia, Campo 6	Todas	68.42%	50.00%
63	**1 y 2**	**RF**	**Trabajos, Wartegg4**	**Trabajos, Wartegg4**	**76.19%**	**50.00%**
63	**1 y 2**	**C.45**	**Campo.7, Campo.2, Deudas**	**Todas**	**79.36%**	**33.33%**
63	**1 y 2**	**BRNN**	**por determinar**	**Todas**	**RMSE (0.4647)**	**por determinar**

Nota: Resultados finales incluyendo diversas secciones de la información y 3 algoritmos de análisis predictivo.

Globalmente se puede apreciar que las variables más relevantes fueron relacionadas a Job-Hopping y a indicadores de empatía del test de Wartegg. Esto se correlaciona directamente con la literatura estudiada en psiquiatratía (Căndel, 2017) (Le Corff, 2014) (Yavuz, 2016) (De Clercq, 2003), ya que estas variables están relacionadas a una persona con un trastorno antisocial, cosa que está directamente relacionada a detectar una persona que potencialmente cometerá fraude intenro si ingresa a una empresa (o cometer actos criminales en general).

Finalmente, en vista de las necesidades del modelo de elegir las características de manera más óptima y de obtener una mejor precisión para los registros con instancias de

fraude, se utilizó un algoritmo genético para optimizar la selección de características en los algoritmos usados. Para esto se definió como cromosoma, un vector de 0 y 1s que definían si una característica del set de datos iba a usarse para entrenar el modelo o no. Un pseudocódigo que describe el algoritmo genético utilizado se puede visualizar en la figura 4.12 (*figura 4.12).*

Figura 4.12

Pseudocódigo Algoritmo Genético

pseudocódigo algoritmo genético:

setvariables = universo de posibles variables de entrenamiento en el modelo

cromosoma=subset de variables del universo de setvariables

Población inicial (definir 200 cromosomas aleatorios)

Población = Población inicial

Repetir por 100 generaciones

    Para cada cromosoma en Población

        Entrenar Modelo de Aprendizaje Supervisado

        Predecir con modelo entrenado en data set de entrenamiento

        Contar predicciones positivas en los registros de empleados que cometieron fraude

        Fitness=número de predicciones positivas en registros de empleados que cometieron fraude / total de registros de empleados que cometieron fraude

    Fin Para

    Mutar cromosomas de la generación con mayor fitness

    Población = 200 nuevos cromosomas mutados

Fin Repetir

Mejor cromosoma = mejor cromosoma de la generación 100

Retornar mejor cromosoma

Nota: Pseudocódigo del algoritmo genético utilizado para optimizar la selección de características.

CAPÍTULO V: VALIDACIÓN Y DISCUSIÓN DE RESULTADOS

Se hizo un análisis de los diferentes métodos de aprendizaje supervisado con diferentes números de características. La selección de características se realizó utilizando el método de análisis de componentes. En la tabla 5.1 se pueden apreciar los resultados (*tabla 5.1*). La columna Nº representa el número de características en el modelo, para esto se escogieron solo las mejores variables según el algoritmo de PCA. Los resultados de precisión fueron obtenidos con cross-validation (10 divisiones, 10 folds). La división entre Test y entrenamiento fue de 52 registros de aprendizaje (82% aprox.) y el resto para testeo. La letra F significa que el resultado de Precisión o de Test es solo para los registros donde los postulantes cometieron fraude. Con esta métrica analizamos que tan bien performa nuestra modelo para evitar los falsos negativos.

Tabla 5.1

Pruebas con selección de características usando PCA

	C.45				Random Forest				BRNN		
Nº	Precisión	Precisión F	Test	Test F	Precisión	Precisión F	Test	Test F	RMSE	Test	Test F
50	67.50%	0.00%	81.82%	0.00%	59.62%	0.00%	63.64%	0.00%	0.5032	72.73%	50.00%
40	66.91%	0.00%	81.82%	0.00%	59.62%	0.00%	63.64%	0.00%	0.4882	72.73%	50.00%
30	60.72%	43.75%	63.64%	0.00%	61.54%	0.00%	63.64%	0.00%	0.4921	90.91%	50.00%
20	**65.03%**	**56.25%**	**54.55%**	**0.00%**	63.46%	18.75%	54.55%	0.00%	0.4887	90.91%	50.00%
15	62.02%	50.00%	45.45%	0.00%	69.23%	18.75%	54.55%	0.00%	0.4919	90.91%	50.00%
10	64.05%	43.75%	45.45%	0.00%	**69.23%**	**25.00%**	**45.45%**	**0.00%**	0.4758	81.82%	0.00%
8	68.25%	0.00%	81.82%	0.00%	65.38%	0.00%	63.64%	0.00%	0.4771	81.82%	0.00%
5	66.92%	0.00%	81.82%	0.00%	51.92%	0.00%	63.64%	0.00%	**0.4689**	**90.91%**	**50.00%**
3	67.90%	0.00%	81.82%	0.00%	65.38%	0.00%	81.82%	0.00%	0.4671	81.82%	0.00%
2	69.63%	0.00%	81.82%	0.00%	65.38%	0.00%	81.82%	0.00%	0.4716	81.82%	0.00%
2	64.60%	0.00%	81.82%	0.00%	59.62%	0.00%	63.64%	0.00%	0.4636	81.82%	0.00%
2	67.90%	0.00%	81.82%	0.00%	59.62%	0.00%	81.82%	0.00%	0.4654	81.82%	0.00%

Nota: Resultados de prueba hecha con tres algoritmos de aprendizaje supervisado y diferente número de características. La selección de características se utilizó únicamente basándose en el método de análisis de componentes.

Como se puede ver, la selección de características influye en los resultados que se obtienen. Los modelos muestran mejoras mientras más variables se van reduciendo, sin embargo esto solo ocurre hasta un punto de quiebre, donde los resultados empeoran si se siguen retirando características. Se puede ver que este punto de quiebre es diferente para cada uno de los algoritmos. Los mejores resultados para cada método fueron

resaltados en la tabla. La red neuronal con regularización bayesiana (BRNN) se benefició más de la reducción de; y además obtuvo los mejores resultados, donde solo clasificó incorrectamente un registro en el set de datos de prueba.

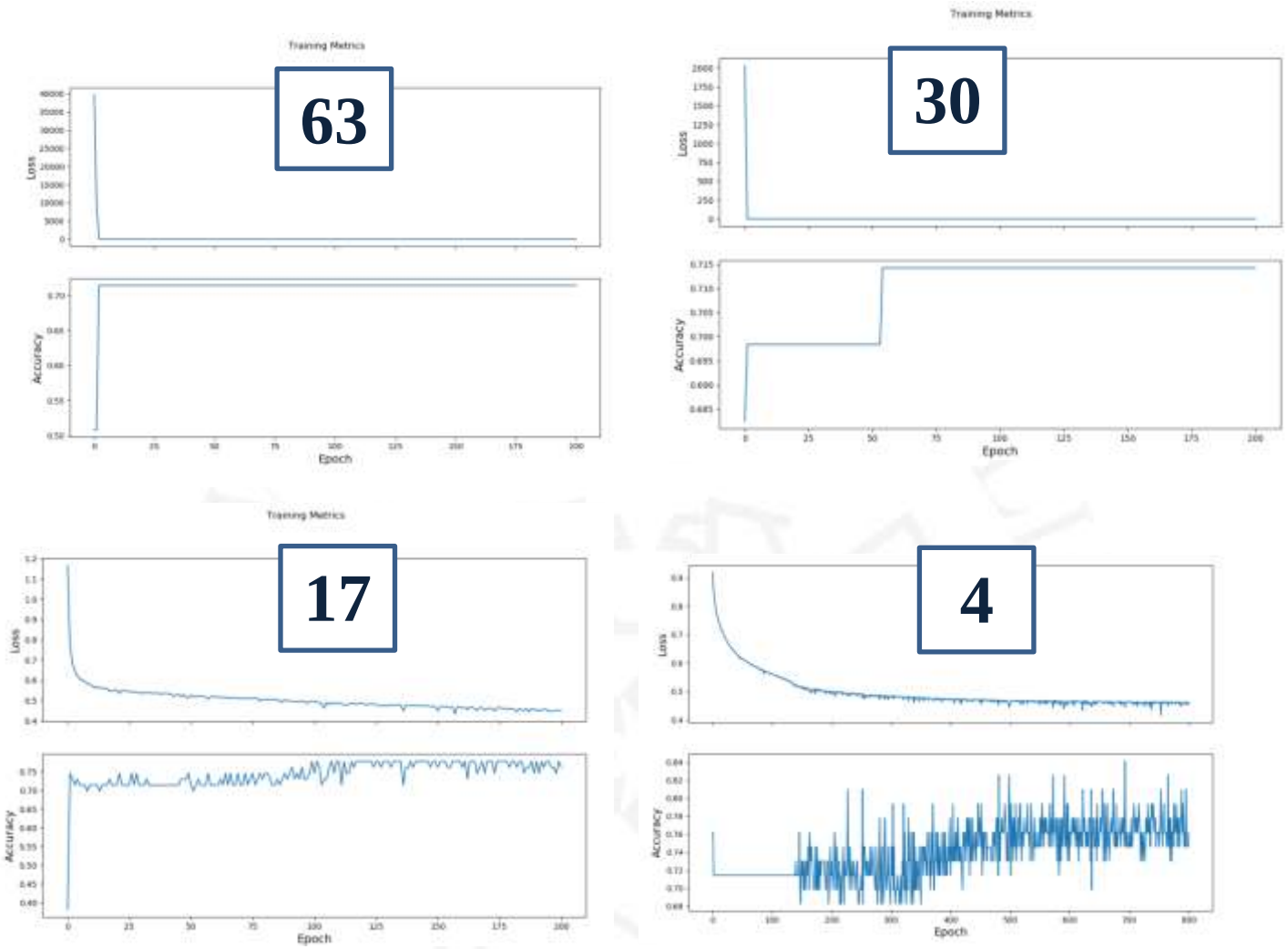
De este experimento podríamos obtener dos conclusiones. La primera siendo que la selección de características beneficia el modelo pero solo hasta cierto punto. La segunda siendo que las redes neuronales parecen ser el algoritmo más apto para este problema.

Para probar estas dos hipótesis, se simularon nuevamente los resultados del experimento pero esta vez solo usando redes neuronales con Backpropagation. Este algoritmo obtuvo los mejores resultados en el experimento de Rashid (2016), donde se predecía comportamiento de empleados. Se generaron los modelos en Tensorflow, el cual nos mostraba el porcentaje de precisión mientras el modelo se entrenaba en cada iteración. De la misma manera, se replicó el modelo con diferente número de características. Los resultados se pueden apreciar en la figura 5.1 (*Figura 5.1*).

De izquierda a derecha y de arriba abajo se va entrenando el algoritmo nuevamente con menor número de características. Representando el primero un modelo con todas las variables, y el último, un modelo con 4 variables. El número de variables usado para cada modelo se coloca en un recuadro de texto encima del gráfico. Se puede apreciar que la selección de características mejora los resultados considerablemente, sin embargo similarmente al experimento previó, cuando se reducen a cierto nivel los resultados comienzan a empeorar. En el último gráfico se puede observar que el aprendizaje de la red neuronal tiene alta varianza entre cada iteración y que no llega los porcentajes de precisión que llegó el modelo previo (con 7 características). Los resultados son favorables, obteniendo un promedio de 80% de precisión en la mejor iteración del experimento. Con esto podríamos probar las dos hipótesis previamente planteadas.

Figura 5.1

Efectos de la selección de características en el aprendizaje de los modelos



Nota: Gráficos que muestran cómo se va generando el aprendizaje de una red neuronal con cada iteración utilizando cierto número de características de nuestro modelo. De izquierda a derecha y de arriba abajo se van reduciendo el número de características utilizando PCA. Se puede apreciar que la selección mejora el aprendizaje y los resultados, sin embargo cuando se llega a un número muy bajo de atributos el modelo se vuelve muy inestable.

El algoritmo optimizó la solución para hallar el set de características donde se tenga el mejor porcentaje de precisión para los casos de fraude. Se realizó el método en cada uno de los algoritmos de aprendizaje supervisado utilizados en la investigación, los resultados se detallan en la tabla 5.2 (*Tabla 5.2*).

Tabla 5.2

Porcentajes de precisión tras selección de características con algoritmo genético

Algoritmo	Precisión	Precisión F	N° Caracteristicas usadas
Random Forest	79.4%	**66.7%**	2
Decision Tree	71.4%	**0.0%**	3
Red Neuronal con Regularización Bayesiana	96.8%	**100.0%**	13
Red Neuronal con Back-Propagation	98.4%	**100.0%**	13

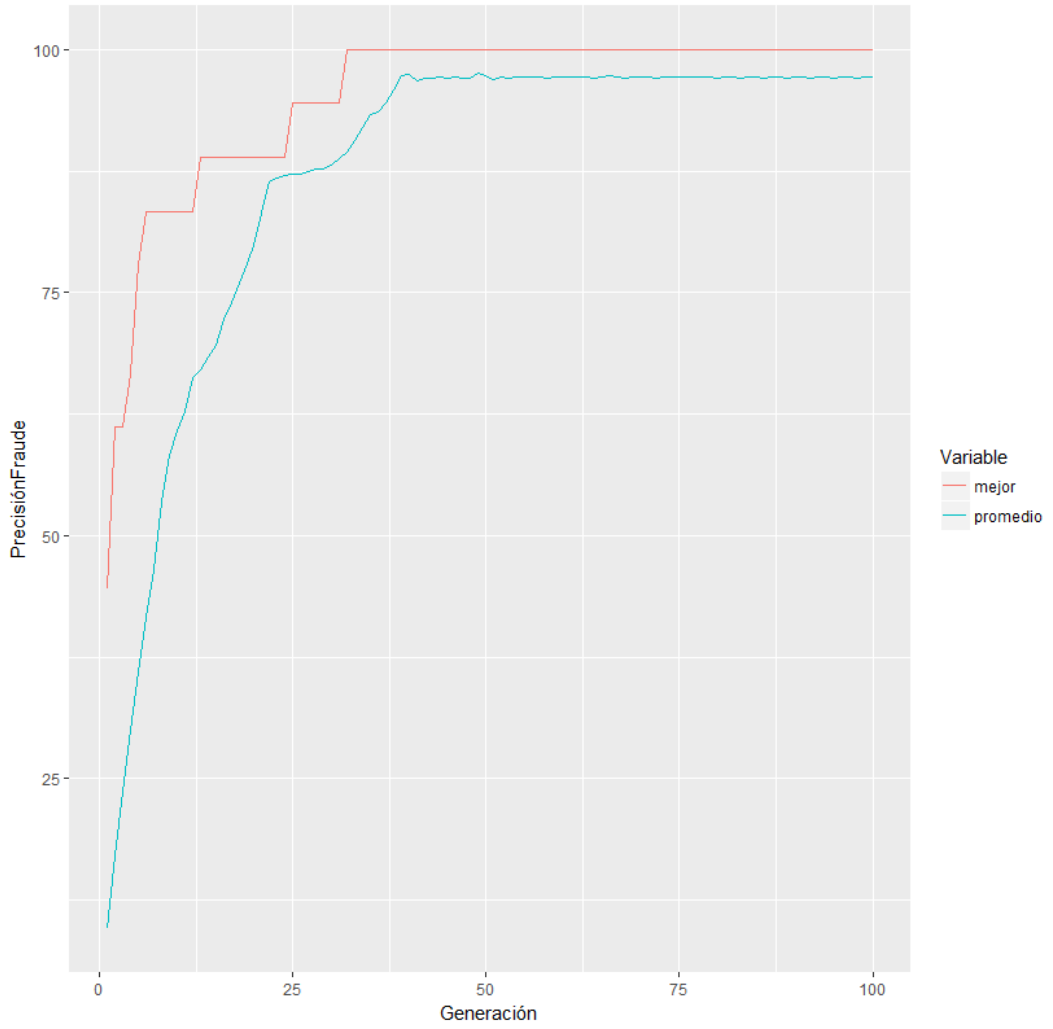
Nota: Resultados obtenidos tras entrenar los algoritmos de aprendizaje supervisado con el set de características optimizado con un algoritmo genético.

Se observa que los resultados mejoran considerablemente para todos los métodos a excepción del árbol de decisión. En este último caso vemos que el algoritmo no pudo optimizar efectivamente la solución deseada. En parte esto se puede deber a que el árbol de decisión ya realiza selección de características al entrenarse. En el caso de las redes neuronales vemos resultados muy superiores a los tenidos previamente, llegando incluso a tener 100% de precisión para los casos de Fraude. Esto terminaría de comprobar que las redes neuronales son el algoritmo que mejor se adapta a este problema.

Para inicializar el algoritmo genético se utilizó una población aleatoria de 200 cromosomas. Para cada cromosoma se calculaba el fitness prediciendo con el algoritmo de aprendizaje escogido la data de entrenamiento. Cabe mencionar que las características usadas para entrenar el algoritmo se determinaban en base a los genes del cromosoma. Después se comparaban las predicciones con los registros de empleados que cometieron fraude. El fitness se obtenía dividiendo el número de predicciones para estos registros que fueron positivas para los registros de empleados que habían cometido fraude entre el número total de estos registros. El algoritmo escogía a los mejores candidatos de cada generación y usaba un ratio de mutación de 0.1 para crear la siguiente generación. Se configuró el algoritmo evolutivo para que tuviera un total de 100 generaciones. La evolución del fitness por generación para el algoritmo genético que optimizó la red neuronal se puede apreciar en la figura 5.2 (*Figura 5.2*). Se puede observar que alrededor de la generación 30 se encontró la solución óptima.

Figura 5.2

Fitness de Cromosomas por generación en algoritmo genético



Nota: Fitness de cromosomas por generación del algoritmo genético. Se puede ver que alrededor de la generación 30 se alcanzó el resultado óptimo.

Respecto a las características elegidas, para el caso de Random Forest fueron “Trabajos” y “Orden5”. La primera siendo la que define el número de trabajos pasados, y la segunda siendo el orden en el que los postulantes realizaron el Campo 5 (Energía Vital) del examen de Wartegg. De acuerdo al material facilitado por la experta en este examen, este campo investiga la agresividad e impulsividad. Estas características estarían relacionadas con un trastorno antisocial (*Capítulo 3.9)*.

Para las redes neuronales, las siguientes fueron las variables en la solución óptima:

-Hijos *(Número de hijos de postulantes. De acuerdo a la teoría de trastornos antisociales [Capítulo 3.9], las personas con un trastorno antisocial tienden a tener menos hijos)*

-Trabajos *(Número de trabajos previos)*

-Wartegg1 *(Respuesta a primera pregunta del examen de Wartegg, ¿Qué dibujo le gustó más?)*

-Wartegg3 *(Respuesta a tercera pregunta del examen de Wartegg, ¿Qué dibujo le pareció más fácil?)*

-Intelectual *(Indica si los postulantes describieron los dibujos a detalle o de manera concisa)*

-sombras *(Cantidad de sombra en dibujos de examen de Wartegg, de acuerdo a lo indicado por la experta en el examen esto está relacionado con la manera en los que los postulantes se relaciona emotivamente)*

-claridad *(Número de dibujos que se distinguen a primera vista/Total de dibujos del examen)*

-Campo.2 *(Evaluación del campo 2 del examen)*

-Campo.5 *(Evaluación del campo 5 del examen)*

-Campo.7 *(Evaluación del campo 7 del examen)*

-Orden.1 *(Orden en que postulante dibujó el campo 1 del examen)*

-Orden.2 *(Orden en que postulante dibujó el campo 2 del examen)*

-Orden.7 *(Orden en que postulante dibujó el campo 7 del examen)*

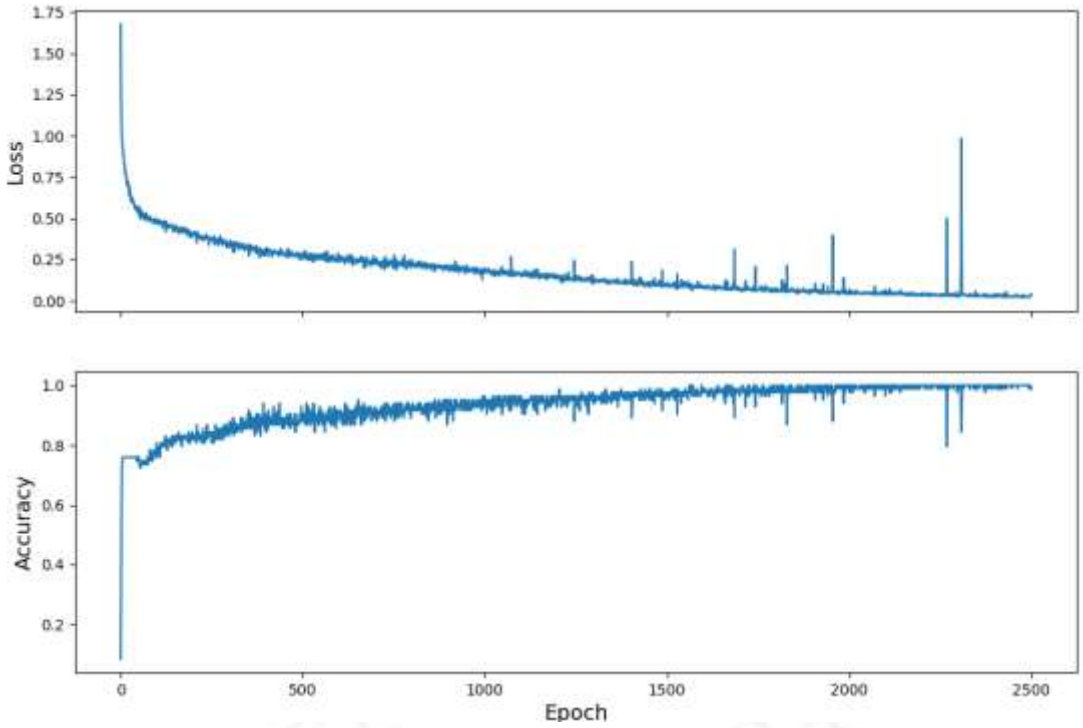
Se puede apreciar una gran presencia de características relacionadas con el examen de Wartegg. Asimismo, vemos algunas de las variables que se definieron previamente como las más importantes y la presencia de otras no vistas previamente. Varias de estas variables tenían puntajes altos en el análisis de entropía; sin embargo en el caso del PCA, muchas de estas no aportaban mucha variación.

Con estos resultados se habría cumplido el objetivo de tener un modelo fiable que tenga un sesgo a reportar falsos positivos. Teniendo 100% de precisión para los casos de fraude se podría hipotéticamente evitar todos los casos de fraude interno en la compañía

prediciendo con la información de los postulantes a ingresar. Asimismo, se puede ver la efectividad del algoritmo genético para optimizar la solución, ya que se pudieron obtener los mejores resultados en el menor tiempo posible. Para validar estos resultados se recolectaron algunos datos vacíos con ayuda de la empresa y se ingresaron más registros. Con este nuevo set de datos se entró nuevamente la red neuronal utilizando el set de características obtenido con el algoritmo genético. Los resultados se pueden observar en la figura 5.4 (*figura 5.4*). Se obtuvo un porcentaje precisión de 100%.

Figura 5.4

Red neuronal Back-Propagation con 78 registros



Nota: Red Convolucional entrenada utilizando 78 registros y nueva base de datos con menos valores vacíos. Se obtuvo 100% de precisión.

Utilizando el modelo optimizado se predijo si postulantes o trabajadores recientemente ingresados a la empresa en cuestión eran de riesgo para esta misma. Esta información fue facilitada a la empresa, la cual nos actualizó con el status de cada uno de los postulantes/ingresante. La empresa tomó decisiones sobre el personal en base a cada uno de los resultados obtenidos, se espera que con este feedback se pueda ver una

reducción de casos de fraude en la organización. Se pueden observar los resultados en la tabla 5.3 (*Tabla 5.3*). No se han incluído los nombres de las personas por motivos de protección de datos, la columna de status indica el status actual de la persona y la columna de "Predicción" indica si la predicción del algoritmo dio positiva o negativa.

Tabla 5.3

Predicción en postulantes/ingresantes de la empresa

Postulante	Tipo	Predicción	Status
1	Activo	1	Retirado, perdió billete de $100 y dio problemas con sus supervisores
2	Activo	0	Trabajando
3	Retirado	1	Abandono de trabajo
4	Postulante	0	Ingresó, trabajando
5	Activo	0	Trabajando
6	Activo	0	Trabajando
7	Activo	1	Renunció
8	Postulante	0	Trabajando
9	Postulante	1	No entró, antecedentes
10	Activo	0	Renunció
11	Activo	1	Retirado en periodo de prueba por predicción
12	Activo	1	Renunció
13	Activo	1	Abandono de trabajo
14	Activo	0	Abandono de trabajo
15	Postulante	1	No ingresó por predicción
16	Postulante	1	No ingresó por predicción
17	Postulante	1	No se ubicó, no entró
18	Postulante	1	Mintió sobre deudas, no entró
19	Postulante	0	No pasó entrevista, no entró
20	Postulante	1	Drogas, deudas,no entró
21	Postulante	1	No ingresó por predicción
22	Postulante	0	Deudas, no entró
23	Postulante	1	Se fue de viaje, no entró
24	Postulante	1	Experiencia limitada, no entró
25	Postulante	1	Poca experiencia, no entró
26	Postulante	1	Falta de experiencia, no entró
27	Postulante	1	No ingresó por predicción
28	Postulante	1	No ingresó por predicción
29	Postulante	1	No entró porque trabajaba en la noche en otro lado
30	Postulante	0	Trabajando
31	Postulante	1	No ingresó por predicción
32	Postulante	1	No entró, no se ubicó para examen médico a pesar de repetidas comunicaciones
33	Postulante	1	No ingresó por predicción
34	Postulante	1	No ingresó por predicción
35	Postulante	0	No entró, no fue a la prueba poligrafica
36	Postulante	1	No entró tenia antecedentes por robo agravado
37	Postulante	1	No entró, tenia antecedentes de robo y violación cuando era menor de edad
38	Postulante	1	No entró, no quizo ir a la Prueba poligráfica

Nota: Resultados tras predecir con la red convolucional entrenada en postulantes e ingresantes a la empresa. El status actual se ha actualizado en la columna "Status"

Finalmente, se hizo una evaluación económica del beneficio del uso del modelo predictivo en la empresa. Para esto se solicitó un histórico de los casos de fraude interno y su costo estimado total en dólares en tres periodos de tiempo. Los tres periodos de

tiempo fueron construidos en base a la fecha en la que se empezó a dar feedback sobre postulantes/ingresantes recientes a la organización. Con esto se definieron un periodo de post-feedback, un periodo de pre-feedback y un periodo de control previo al pre-feedback. Se puede ver que tanto en el periodo de control como en el de pre-feedback se tuvieron 2 casos de fraude respectivamente. Sin embargo, desde el punto en el que se empezó a dar información a la empresa aún no se registra ningún caso de fraude. Esto representaría en el universo de información actual una mejora económica para la misma utilizando el modelo predictivo. Se estimaría que el uso del algoritmo de predicción de fraude estaría ahorrando 1300 dólares por mes a la empresa. (*Tabla 5.4*)

Tabla 5.4

Evaluación del uso del modelo predictivo en la empresa

Periodo	Periodo 1 (control)	Periodo 2 (pre-feedback)	Periodo 3 (post-feedback)
Tiempo periodo	*19mar-27abr*	*28abr-4jun*	*05jun-12jul*
Casos Fraude Interno	**2**	**2**	**0**
Costo estimado ($)	**600**	**2000**	**0**

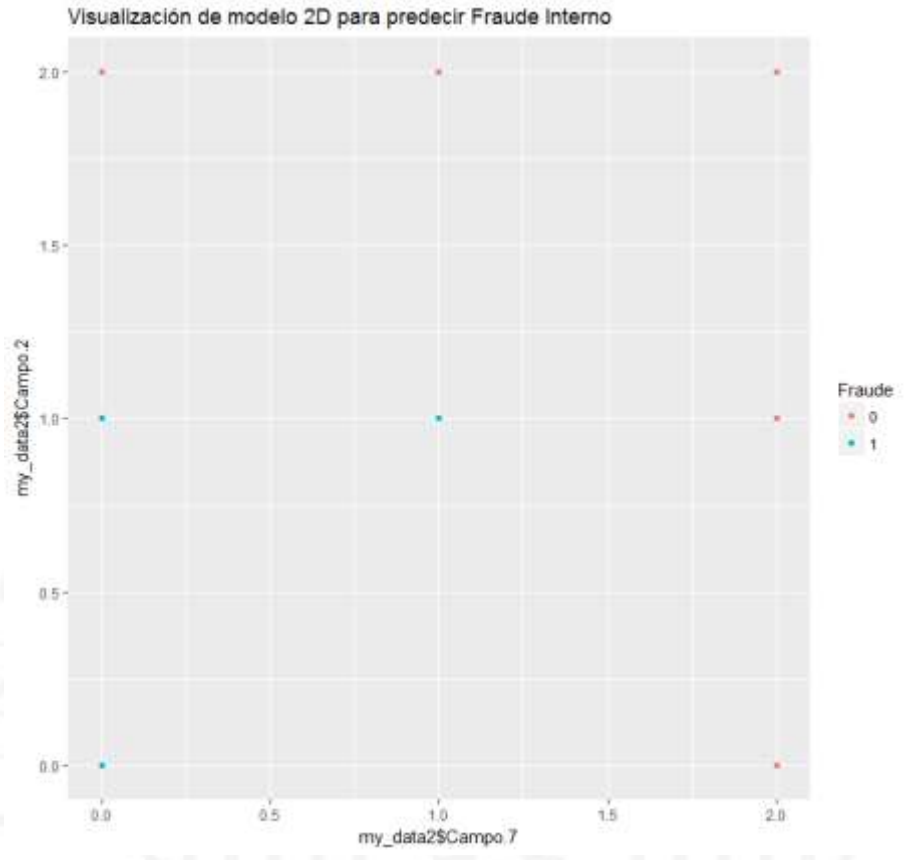
Nota: Análisis de casos históricos de fraude interno en la empresa tras implementación de modelo predictivo.

Para determinar la causa de la superioridad de las redes neuronales frente a los árboles de decisión/random forest para este modelo de predicción, graficamos nuestro modelo en un plano de dos dimensiones.

En la literatura estudiada (*Capítulo 3.4*) se vio que los árboles de decisión realizaban la segmentación de la información usando los valores de las variables y determinando la división óptima para poder clasificar los datos. En la figura 5.4 (*figura 5.4*) se pueden apreciar las dos primeras variable que uso el árbol de decisión para clasificar la información. Se puede observar que el modelo se adapta a un árbol de decisión, ya que fácilmente se pueden definir en bloques los casos de no fraude, cuando Campo.2 y Campo.7 son ambos menores o iguales 1 la mayor parte de los registros corresponden a personas que cometieron fraude.

Figura 5.4

Campo.7 vs Campo.2 Fraude interno

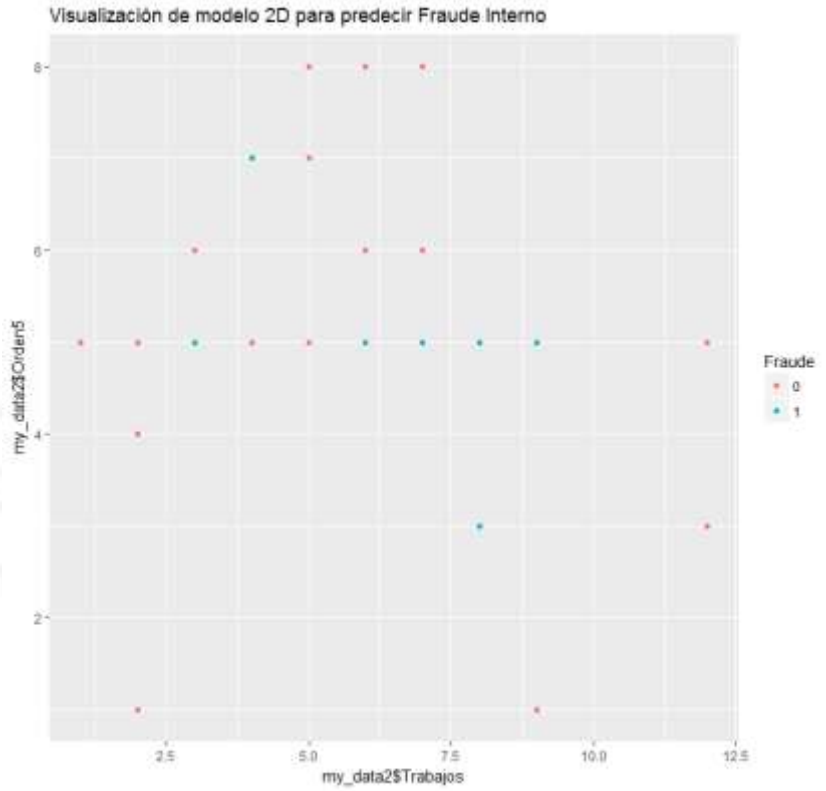


Nota: Variables Campo.2 y Campo.7 en un plano cartesiano, segmentadas por los casos de fraude.

De la misma manera, se graficó la variable “Orden5” y la variable “Trabajos” (*figura 5.5*). Las 2 variables, las cuales fueron elegidas como el set de características óptimas para el modelo de random forest, forman también un plano similar, donde se pueden distinguir divisiones claras entre los casos de fraude y no fraude. Cuando el valor de Orden5 es menor o igual a 5 y el valor de trabajos mayor a 5 se puede observar que hay gran predominancia de casos de fraude. Como se vio en la literatura (Capítulo 3.4.2) el algoritmo de Random Forest utiliza varios árboles de decisión para hacer lo predicción, lo que origina que un modelo donde se puedan apreciar divisiones claras en la información sea más óptimo.

Figura 5.5

Trabajos vs Orden5

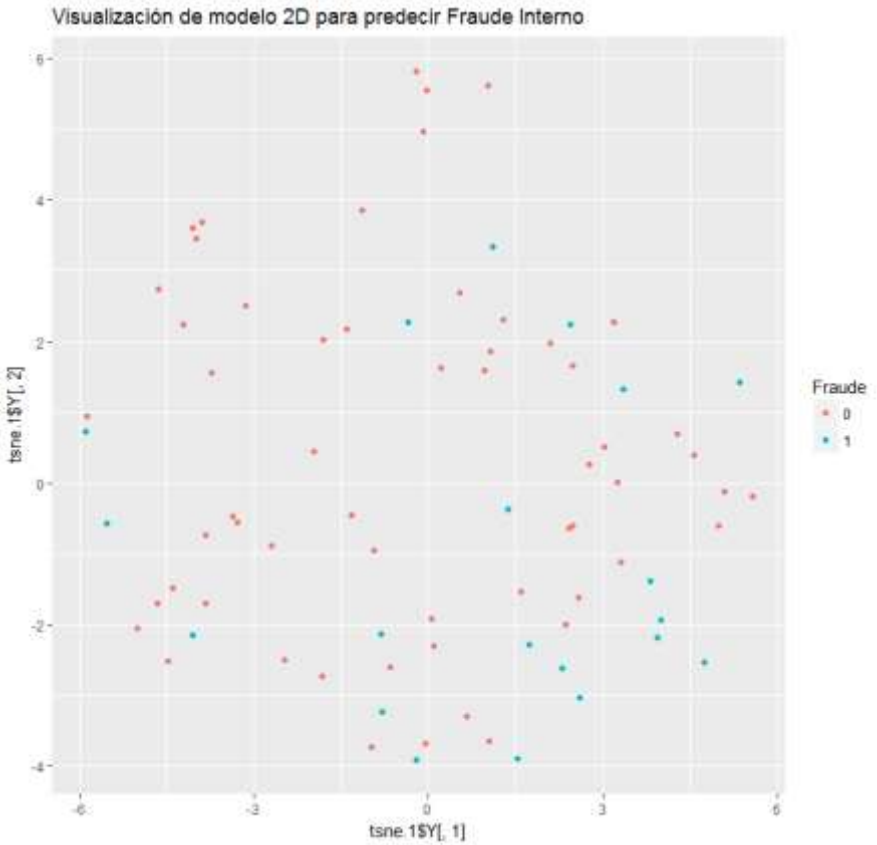


Nota: Variables Trabajos y Orden5 en un plano cartesiano, segmentadas por los casos de fraude.

Finalmente, se graficó el modelo óptimo encontrado por el algoritmo genético para la red neuronal. Para poder visualizar las 13 variables en 2 dimensiones se usó la técnica de t-SNE que permite graficar múltiples dimensiones de data en un mapa 2D. (van der Maaten, 2008). Los resultados se pueden apreciar en la figura 5.6 (*figura 5.6*). Se observa que ya no hay segmentaciones que favorezcan a los modelos de árbol de decisión y Random Forest.

Figura 5.6

Modelo de Fraude Interno con t-SNE



Nota: Visualización de modelo de predicción de Fraude Interno en un mapa 2D usando t-SNE.

Como validación final se realizó un cross-validation con el modelo de red neuronal. (10 divisiones, repetido 10 veces). Se obtuvo una precisión final de 71.25% utilizando las características óptimas. Este resultado mostraría que el modelo aun no estaría generalizando del todo bien. Quedaría pendiente para esto realizar más pruebas con una mayor cantidad de registros.

CONCLUSIONES

En el presente trabajo se investigaron métodos de clasificación para predecir fraude interno. En base a la teoría y literatura estudiada y los resultados obtenidos podemos llegar a las siguientes conclusiones.

- Se pudo comprobar que se pueden utilizar variables de registros de postulantes para predecir efectivamente cuales van a cometer fraude interno.
- Inicialmente, los modelos obtenidos tenían mayor porcentaje de error para predecir los empleados que cometieron fraude, sin embargo esto se pudo corregir utilizando selección de características
- De todos los métodos de selección de características utilizados, se pudo ver que el que obtuvo los mejores resultados y en menos tiempo fue el algoritmo genético.
- Varias de las variables relevantes para el modelo coincidieron con características relacionadas con personas con el trastorno antisocial de la personalidad. Esto muestra una correlación entre los resultados obtenidos con la información de los postulantes y la literatura estudiada.
- El modelo más efectivo para el modelo fueron las redes neuronales con Backpropagation, lo cual coincidió con el modelo que tuvo mejor resultado de precisión en uno de los artículos que investigaba un problema parecido al propuesto (predecir comportamiento de empleados con variables de recursos humanos).

RECOMENDACIONES

En base a los hallazgos, detallaremos las recomendaciones:

- Debido a que los resultados fueron obtenidos con una cantidad limitada de registros, se recomendaría experimentar si se pueden replicar los resultados obtenidos con más información y con información de otras empresas. Esto incluiría analizar si las variables más relevantes encontradas para estos modelos coinciden con las de este trabajo.
- Se recomendaría tomar en cuenta los factores que están correlacionados con el trastorno antisocial de la personalidad, ya que se observó que estaba muy relacionado con la probabilidad de cometer fraude interno. Los resultados muestras que las características descritas en la literatura sobre este trastorno también eran las más relevantes para el modelo predictivo.
- Se debería enfocar en obtener los mejores resultados de precisión para los casos de fraude. Esto se debe a que los falsos positivos son el resultado de error más favorable para la empresa. Asimismo, para poder optimizar esto resultado de precisión de la manera más eficiente se apreció que el algoritmo genético fue el método más apto.
- Cabría la posibilidad de realizar análisis de textos con el algoritmo de Porter-Stemming, BOW y Naive Bayes encontrado en la literatura. Esto se podría aplicar en los diversos textos escritos por los postulantes en las respuestas de la hoja de información y su CV, de modo que se pueda comprobar si hay una relación entre el output de este procesamiento y la probabilidad de que estos mismos hayan cometido fraude interno. De ser el caso se podría generar más variables relevantes para el modelo.
- La realidad en estudio puede variar frente a otras estudiadas en la literatura, en este caso estamos analizando una empresa mediana en la ciudad de Lima, Perú. Varios

aspectos de la cultura local pueden influir en cómo piensan o se comportan las personas que postulan a las empresas, y estos pueden diferir de otros contextos estudiados, donde algunas variables pueden interactuar de diferente manera o tener diferentes relevancias en los algoritmos para clasificar a los empleados.

- No se debería descartar el uso de pruebas psicológicas para obtener características para el modelo de predicción. En esto caso se procesó el examen de Wartegg con el apoyo de una psicóloga experta en este mismo. Se pudo ver en los resultados que utilizando las variables resultantes de este análisis el algoritmo obtuvo 100% de precisión para encontrar a los empleados que cometieron fraude. Se especula que se podrían procesar estos resultados con un algoritmo que pueda procesar los dibujos hechos por los postulantes. Esto se podría extender a otras pruebas psicológicas como el test de Markov, las cuales sin el uso de un algoritmo de este tipo tienen el riesgo de resultar en data subjetiva en el modelo.
- Determinar el costo computacional de nuestro modelo y posibles implementaciones en empresas reales. De modo que se puedan reducir a largo plazo las pérdidas económicas por este crimen al tener métodos efectivos en las empresas del país para evitar que postulantes que van a cometer fraude interno ingresen a trabajar.
- Se propone realizar investigaciones adicionales con mayor cantidad de registros, ya que se observa que el modelo a pesar de tener buena precisión aun no generaliza bien.

REFERENCIAS

Ahmad, M. W., Mourshed, M., y Rezgui, Y. (2017). Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. Energy and Buildings, 147, 77-89. doi: 10.1016/j.enbuild.2017.04.038

Ai, J., Patrick, L., Golden, Linda L., Guillén y Montserrat. (2013). A Robust Unsupervised Method for Fraud Estimation. The Journal of Risk and Insurance. 80 (1). 121-143: doi: 10.1111/j.1539-6975.2012.01467.x

Alraouji, Y., y Bramantoro, A. (2014). International Call Fraud Detection Systems and Techniques. Proceedings of the 6th International Conference on Management of Emergent Digital EcoSystems - MEDES 14. doi:10.1145/2668260.2668272

Aquino, K. y Douglas, S. (2003). Identity threat and antisocial behavior in organizations: The moderating effects of individual differences, aggressive modeling, and hierarchical status. Organizational Behavior and Human Decision Processes, 90(1), 195-208. doi: 10.1016/s0749-5978(02)00517-4

Bhattacharyya, S., Jha, S., Tharakunnel, K., y Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. Decision Support Systems, 50(3), 602-613. doi: 10.1016/j.dss.2010.08.008

Blockeel, H. y Struyf, Jan. (2002). Efficient Algorithms for Decision Tree Cross-validation. Journal of Machine Learning Research 3. 621-650

Bolin, A. y Heatherly, L. (2001). Predictors of Employee Deviance: The relationship between Bad Attitudes and Bad Behavior. Journal of Business and Psychology, 15(3), 405-418

Boorsboom, D. (2008). Latent Variable Theory. Measurement: Interdisciplinary Research and Perspectives. 6(1-2). 25-53. doi: 10.1080/15366360802035497

Breiman, L. (2001). Random Forests. Statistics Department University of California. Recuperado de https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf

Căndel, O. y Constantin, T. (2017). Antisocial and Schizoid Personality Disorder Scales: Conceptual bases and preliminary findings. Romanian Journal of Applied Psychology. 9(1). 10-16. doi: 10.24913/rjap.19.1.02

Candès, E. J. y Recht, B. (2009). Exact Matrix Completion via Convex Optimization. Foundations of Computational mathematics. 9(6). 717-772

Chang, H.Y. (2009). Employee turnover: a novel prediction solution with effective feature selection. WSEAS Trans. Inf. Sci. Appl. 3 (6), 417–426

Chen-Fu, C. y Li-Fei, C. (2008). Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. Expert Systems with Applications. 34(1). 280-290. doi: 10.1016/j.eswa.2006.09.003

Dazzan, P. y Murray, R. (2002). Neurological soft signs in first-episode psychosis: a systematic review. The British Journal of Psychiatry. 181 (43). 50-57. doi: 10.1192/bjp.181.43.s50

De Clercq, B. y De Fruyt, F. (2003). Personality disorder symptoms in adolescence: A five-factor model perspective. Journal of Personality Disorders. 17(4). 269-292. doi: https://doi.org/10.1521/pedi.17.4.269.23972

Demirel, O., Demirel, A., Kadak, M., Emül, M. y Duran, A. (2016). Neurological soft signs in antisocial men and relation with psychopathy. Psychiatry Research. 240. 248-252. doi: 10.1016/j.psychres.2016.04.094

Edwards, M., Peersman, C., y Rashid, A. (2017). Scamming the Scammers: Towards Automatic Detection of Persuasion in Advance Fee Frauds. Proceedings of the 26th International Conference on World Wide Web Companion, 1291-1299. doi:10.1145/3041021.3053889

Foresee, F. D., Hagan, M. T. (1997). Gauss-Newton approximation to Bayesian learning. Neural Networks,1997., International Conference on. doi: 10.1109/ICNN.1997.614194

Goh, A. T. (1995). Back-propagation neural networks for modeling complex systems. Artificial Intelligence in Engineering. 9(3). 143-151. doi: https://doi.org/10.1016/0954-1810(94)00011-S

Heerman, P. D. y Khazenie, N. (1992). Clasification of Multispectral Remote Sensing Data Using a Back-Propagation Neural Network. IEEE Transactions on GeoScience and Remote Sensing. 30 (1). doi: 10.1109/36.124218

Horesh, R., Varshney, K. R., y Yi, J. (2016). Information Retrieval, Fusion, Completion, and Clustering for Employee Expertise Estimation. 2016 IEEE International Conference on Big Data (Big Data)

Jantan, H., Hamdan, A.R. y Othman, A. (2009). Knowledge discovery techniques for talent forecasting in human resource application. Int. Sch. Sci. Res. Innov. Int. Sci. Index. 3(2), 178–186

Jantan, H. y Hamdan, A.R. (2010). Human Talent Prediction in HRM using C4.5 Classification Algorithm. International Journal on Computer Science and Engineering. 2(8). 2526-2534

Jantan, H., Hamdan, A.R. y Othman, A. (2011). Towards applying data mining techniques for talent management. Int. Conf. on Computer Engineering and Applications, IPCSIT, IACSIT Press, Singapore. 2

Jaynes, E. T. (1986). Bayesian Methods: General Background. In Maximum Entropy and Bayesian Methods in Applied Statistics. 1-25

Jensen, D., Palmer, K., Goldberg, H., Komoroske, J. y Neville. J. (2005). Using Relational Knowledge Discovery to Prevent Securities Fraud. Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining - KDD 05. doi: 10.1145/1081870.1081922

Junqué de Fortuny, J., Stankova, M., y Moeyersoms, J. (2014). Corporate Residence Fraud Detection. ACM SIGKDD international conference on Knowledge discovery and data mining, 1650-1659. doi:10.1145/2623330.2623333

Kroll, K. (2012). Keeping the Company Safe: Preventing and Detecting Fraud. Financial Executive Sep2012. 28(7). 20-23

Le Corff, Y. y Toupin, J. (2014). Overt versus covert conduct disorder symptoms and the prospective prediction of antisocial personality disorders. Journal of Personality Disorders. 28(6). 864-872. doi: 10.1521/pedi_2012_26_074

Machado, A. y Franca, R. (2012). Fraud detection in web transactions. Proceedings of the 18th Brazilian symposium on Multimedia and the web. 273-280. doi: 10.1145/2382636.2382695

MacKay, D. (1992). Bayesian Interpolation. Neural Computation. 4(3). 415-447. doi: https://doi.org/10.1162/neco.1992.4.3.415

Mills, J. U., Stuban, S. M., y Dever, J. (2017). Predict insider threats using human behaviors. IEEE Engineering Management Review. 45(1). 39-48. doi:10.1109/emr.2017.2667218

Pineda, J. F. (1987). Generalization of Back-Propagation to Recurrent Neural Networks. The American Physical Society. 59(19). 2229-2232

Rashid, Tarik, A. y Asia, L. J. (2016). Improvement on predicting employee behavior through intelligent techniques. 5(5). 136-142. doi: 10.1049/iet-net.2015.0106

Rich, E., Martinez-Moyano, I. J, Conrad, S., Andersen. D. F. y Stewart, T. R. (2008). A behavioral theory of insider-threat risks: A system dynamics approach. ACM Trans. Model. Comput. Simul. 18 (2)

Scroggins, W. A., Thomas, S. L. y Morris, J. A. (2008). Psychological Testing in Personnel Selection, Part II: The Refinement of Methods and Standards in Employee Selection. Public Personnel Management. 37(2). 185-198

Smith, A. D. (2005). Accountability in EDI Systems to Prevent Employee Fraud. Information Systems Management. Spring2005. 22(2). 30-38

van der Maaten, L. y Hinton, G. (2008). Visualizing Data using t-SNE. Journal of Machine Learning Research. 9. 2579-2605

Varshney, K. R., Chenthamarakshan, V., Fancher, S. W., Wang, J., Fang, D., y Mojsilović, A. (2014). Predicting employee expertise for talent management in the enterprise. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 14. doi:10.1145/2623330.2623337

Vogl, T. P., Mangis, J. K., Rigler, A. K., Zink, W. T. y Alkon, D. L. (1988). Accelerating the Convergence of the Back-Propagation Method. Biological Cybernetics. 59 (4). 257-263. doi: 10.1007/BF00332914

Wang, X. (2011). Characteristics of the human resource executives and companies that use more sophisticated employee selection methods. Emporia State University, ProQuest Dissertations Publishing, 2011. 1507482.

Willson, R., y Siponen, M. (2009). Overcoming the insider. Communications of the ACM, 52(9), 133-137. doi:10.1145/1562164.1562198

Wolpert, D. H. y MACREADY, W. G. (1997). An efficient method to estimate Bagging's Generalization Error. Machine Learning. 5. 1-16

Yavuz, K., ŞAHİN, O., Ulusoy, S., İpek, O., y Kurt, E. (2016). Experiential avoidance, empathy, and anger-related attitudes in antisocial personality disorder. Turkish Journal of Medical Sciences. 46(6). 1792-1800. doi: 10.3906/sag-1601-80

Young, W., Memory, A., Goldberg, H., y Senator T. (2014). Detecting unknown Insider Threats Scenarios. 2014 IEEE Security and Privacy Workshops. Recuperado de http://www.ieee-security.org/TC/SPW2014/papers/5103a277.PDF

Zhang, L., Yang, J., Chu, W., y Tseng, B. (2011). A machine-learned proactive moderation system for auction fraud detection. Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM 11. doi:10.1145/2063576.2064002

BIBLIOGRAFÍA

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). Arlington, VA: American Psychiatric Publishing

Bensing, R. G. (2009). An Assessment of Vulnerabilities for Ship-based Control Systems Naval Postgraduate School, Monterey, CA.

Bersin, J., Leonard, K. y Wang-Audia, W. (2013). High-Impact Talent Analytics: Building a World-Class HR Measurement and Analytics Function. Bersin by Deloitte. 131

Bmindful. (15 de 9 de 2017). Psychotherapy & group therapy. Obtenido de Acceptance & Commitment Therapy (ACT) Provided by Bmindful Psychotherapists in Ottawa: http://bmindful.ca/counselling-psychotherapy-coaching-ottawa/treatment-options/acceptance-and-commitment-therapy-act/

Breiman, L. (1997). OUT-OF-BAG ESTIMATION. Statistics Department: University of California

Breiman, L. (2003). Manual--Setting Up, Using, And Understanding Random Forests V4.0. https://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf

Breiman, L., Friedman, J. H., y Olshen, R. A. (1984). Classification and Regression Trees. Wadsworth, Belmont

Cano C., M. A. (8 de 9 de 2017). United States: InterAmerican Community Affairs. Obtenido de Fraude y Estafa en los Negocios: http://interamerican-usa.com/articulos/Auditoria/Fraud-Estaf-Neg.htm

CAT. (9 de 9 de 2017). Centre D' Assistència terapèutica Barcelona. Preguntas frecuentas. Obtenido de ¿Qué significa comorbilidad?: http://www.cat-barcelona.com/faqs/view/que-significa-comorbilidad

Duda, R. O., Hart, P. E., y Stork, D. G. (2000). Pattern Classification. Wiley-Interscience.

El Comercio. (12 de 10 de 2017). Cajera del BCP desvió 5 millones de soles usando un USB. Obtenido de sitio web de diario El comercio: http://elcomercio.pe/lima/policiales/cajera-desvio-5-millones-soles-banco-estilo-cromwell-galvez-437509

EY. (4 de 9 de 2017). Construyendo un ambiente ético: Estudio sobre el riesgo de fraude en el Perú. Obtenido de sitio web de EY: http://www.ey.com/pe/es/services/assurance/fraud-investigation---dispute-services/construyendo-un-ambiente-etico-estudio-sobre-el-riesgo-de-fraude-en-el-peru

EY. (11 de 9 de 2017). EY. Obtenido de EY 14th Global Fraud Survey 2016: http://www.ey.com/Publication/vwLUAssets/EY-14-global-fraud-survey/$FILE/EY-14-global-fraud-survey.pdf

Haykin, S. (1999). Neural Networks: A Comprehensive Foundation. Singapore: Pearson Education.

Hr-guide. (8 de 9 de 2017). Personnel Selection: Overview. HR Guide to the Internet: http://www.hr-guide.com/data/G300.htm

Johnson, R. y Wicherin, D. (2007). Applied Multivariate Statistical Analysis: Sixth Edition. Pearson Education

Körting, T. S. (4 de 4 de 2014). How Random Forest algorithm works. Obtenido de YouTube: https://www.youtube.com/watch?v=loNcrMjYh64

KROLL. (2009). Global Fraud Report. Estados Unidos: Kroll.

Merz, C. J. y Murphy, P. M. (1996). UCI repository of machine learning databases [http://www.ics.uci.edu/˜mlearn/mlrepository.html]. Department of Information and Computer Science, University of California, Irvine, California

PERU.com (24 de 9 de 2017). Jesús María: Trabajadora participó en gran robo de Compu Diskette. Obtenido de sitio web de peru.com: https://peru.com/actualidad/mi-ciudad/jesus-maria-trabajadora-participo-gran-robo-compu-diskette-noticia-337977

Psicólogos Infantiles Madrid (PSISE). (15 de 9 de 2017). Grupo B. Obtenido de Personalidad y trastornos de personalidad: https://psisemadrid.org/personalidad-y-trastornos-de-personalidad/

Randazzo, M. R., Keeney, M. M., Kowalski, E. F., Cappelli, D. M., y Moore, A. P. (2004). Insider Threat Study: Illicit Cyber Activity in the Banking and Finance Sector. U.S. Secret Service and CERT Coordination Center/Software Engineering Institute, Philadelphia, PA, 25.

SPSS. (5 de 4 de 2020). Principal Components Analysis (PCA) using SPSS Statistics: https://statistics.laerd.com/spss-tutorials/principal-components-analysis-pca-using-spss-statistics.php

Towers Perrin. (2005). Talent Management: The State of the Art. Connecticut: Towers Perrin HR Services.

UCI. (1 de 11 de 2017). Machine Learning Repository. Center for Machine Learning and Intelligent Systems: https://archive.ics.uci.edu/ml/index.php

Venables, W. N. y Ripley, B. D. (2002). Modern Applied Statistics with S. Springer-Verlag