Serie Documentos de Trabajo

Superintendencia de Seguridad Social
Santiago - Chile

DOCUMENTO DE TRABAJO N°20

Modelos predictivos por aprendizaje automatizado de accidentes laborales y factores psicosociales del trabajo

Unidad de Estudios y Estadísticas
Evelyn Benvin
Carlos Soto

Unidad de Prevención y Vigilancia
Macarena Candia
Juan Manuel Pérez

2020

SUPERINTENDENCIA DE SEGURIDAD SOCIAL

SUPERINTENDENCE OF SOCIAL SECURITY

La Serie Documentos de Trabajo corresponde a una línea de publicaciones de la Superintendencia de Seguridad Social, que tiene por objetivo divulgar trabajos de investigación y estudios realizados por profesionales de esta institución, encargados o contribuidos por terceros. El objetivo de estas publicaciones es relevar temas de interés para las políticas de seguridad social, difundir el conocimiento adquirido e incentivar el intercambio de ideas.

Los trabajos aquí publicados tienen carácter preliminar y están disponibles para su discusión y comentarios. Los contenidos, análisis y conclusiones expresados son de exclusiva responsabilidad de su(s) autor(es), y no reflejan necesariamente la opinión de la Superintendencia de Seguridad Social.

Si requiere de mayor información, o desea contactarse con el equipo editorial, escriba a: publicaciones@suseso.cl.

Si desea conocer otras publicaciones, artículos de investigación y proyectos de la Superintendencia de Seguridad Social, visite nuestro sitio web: www.suseso.cl.

The Working Papers Series of the Superintendence of Social Security disseminates research and policy analysis conducted by its staff, outsourced or contributed by third parties. The purpose of the series is to discuss issues of interest for the social security policies, expose new knowledge and encourage the exchange of ideas.

These papers are preliminary research reports intended for discussion and comments. The contents, analysis and conclusions presented are solely the responsibility of the author(s), and do not necessarily reflect the position of the Superintendence of Social Security.

For further information, or to contact the editors, please write to: publicaciones@suseso.cl.

For other publications, research papers and projects of the Superintendence of Social Security, please visit our website: www.suseso.cl.

Superintendencia de Seguridad Social
Huérfanos 1376
Santiago, Chile

Modelos predictivos por aprendizaje automatizado de accidentes laborales y factores psicosociales del trabajo

Superintendencia de Seguridad Social

Resumen

En este documento se han aplicado técnicas de regresión logística y *Random Forest* con el objeto de evaluar los principales determinantes de los accidentes laborales sufrido por trabajadores cubiertos por la ley de accidentes N°16.744 y explorar métodos que mejoren la capacidad predictiva de un modelo clásico, como la regresión logística, gracias al uso de técnicas más propias del aprendizaje automatizado (*machine learning*). Para realizar dicho análisis se utilizó la base de datos del cuestionario SUSESO/ISTAS 21 en su versión completa para los años 2017 y 2018.

Muchas veces los modelos con variables de estudio categóricas tienden a clasificar correctamente la clase mayoritaria, pero clasifican mal la clase minoritaria cuando los datos se encuentran desbalanceados. Una muestra de datos desbalanceada ocurre cuando una de las clases de respuesta está en un número muy inferior a otra. Con el objetivo de abordar este problema se utilizó el método de *Synthetic Minority Oversampling* (SMOTE) para balancear la muestra y obtener resultados con un mejor rendimiento.

Las variables más relevantes que explican los accidentes laborales son aquellas relacionadas con la auto percepción de la salud y bienestar personal como son la salud general, salud mental, vitalidad y estrés. Asimismo, las variables relacionadas con el número de días de ausencia laboral debido a una licencia médica de origen común muestran una fuerte correlación entre el grupo de trabajadores con accidentes laborales. Sin embargo, dicha relación no es significativa.

Finalmente, las variables de riesgo psicosocial medidos en el cuestionario SUSESO/ISTAS 21, no mostraron una asociación consistente con la probabilidad de tener un accidente laboral. Es probable que los factores de riesgo psicosocial no ejerzan su influencia directamente sobre la probabilidad de tener un accidente, sino que a través del estrés como mediador, incluso a través de la autopercepción de salud general y mental.

Introducción

Los accidentes del trabajo (AT) han sido objeto de preocupación mundial. La OIT [1] estimó para el año 2014 una tasa mundial de 11.096 accidentes del trabajo por 100.000 personas en la fuerza de trabajo y 380.500 muertes anuales vinculadas a los AT. El costo económico de los AT y enfermedades profesionales varía en cada país, pero va desde el 1,8 al 6% del PIB, con una media de 4% [2].

En Chile, la tasa de AT ha estado en una declinación permanente y en 2019 se encontraba en 3,0 por cada 100.000 trabajadores protegidos (217.800 accidentes), en tanto los accidentes fatales en 2019 llegaron a 197 trabajadores fallecidos [3].

Variados estudios confirman que existe una relación entre los factores psicosociales del trabajo y los AT [4-11]. Johannessen et al. [12] siguió una cohorte de 6.745 trabajadores noruegos por 3 años, detectando un exceso de riesgo atribuible de AT en los casos de alta presión laboral (*high job strain*, *Odds Ratio* = 2,31), alto conflicto de rol (*Odds Ratio* = 3,01) y altas exigencias emocionales (*Odds Ratio* = 1,96). Julià et al. [13], por su parte, en una cohorte de 16.693 trabajadores españoles, a un año de seguimiento medida con la versión breve del COPSOQ español (ISTAS21), determinaron que las altas exigencias psicológicas (*Relative Ratio* = 2,22), la calidad del liderazgo (*Relative Ratio* = 1,87), el trabajo activo y las posibilidades de desarrollo (*Relative Ratio* = 1,83), y la baja estima (*Relative Ratio* = 1,28) eran variables que se asociaban a la posibilidad de tener un accidente en el trabajo. Se ha sugerido, adicionalmente, que las condiciones de agotamiento [14] y el estrés psicosocial [10], al disminuir la capacidad de los trabajadores para evitar el peligro y reducir la percepción de riesgo, se vinculan a los AT.

De los trabajos que en Chile han examinado la asociación entre factores de riesgo psicosocial laboral y los AT o las enfermedades profesionales [15-20] es difícil extraer conclusiones sólidas debido a las diferentes metodologías utilizadas y en algunos casos el tamaño de efecto es mínimo. Solo el trabajo de Ibáñez [18] muestra alguna correlación clara entre riesgo psicosocial laboral (RPSL) (medido con SUSESO/ISTAS21, versión corta) y accidentabilidad, con una rho de Spearman igual a -0,29 (exigencias psicológicas) y rho igual a -0,34 (trabajo activo), ambas significativas ($p < 0,05$).

Desde 2013 el cuestionario SUSESO/ISTAS21 [21,22] en su versión breve (VB), es un cuestionario de auto-reporte, aplicado a cada trabajador de manera confidencial y anónima. Su aplicación es obligatoria en todas las empresas e instituciones chilenas que tengan trabajadores contratados (se excluyen de la obligatoriedad las empresas con menos de 10 trabajadores) [23]. En su versión completa (VC) [22] debe ser aplicado en algunas situaciones especiales, como cuando el cuestionario VB, utilizado como tamizaje, ha arrojado un nivel elevado de riesgo, o cuando se ha detectado un trabajador con una enfermedad mental de origen laboral. No obstante, también puede aplicarse de manera voluntaria.

Aprovechando el gran volumen de los datos recogidos y conservados en la Superintendencia de Seguridad Social (SUSESO) del cuestionario en su versión completa, el presente trabajo tiene como objetivo estimar la relación que existe en Chile

entre variables sociodemográficas, de salud, de empleo y de riesgo psicosocial en el trabajo y los AT y construir un modelo predictivo de los accidentes a través de diversas técnicas estadísticas y de aprendizaje automatizado (*machine learning*), lo que debiera permitir focalizar los esfuerzos en la disminución de las tasas de accidentabilidad. La capacidad predictiva de algunos modelos, como la regresión logística, es baja, aún cuando su capacidad explicativa puede ser elevada. Con un modelo explicativo, lo que se desea es conocer y fijar los factores que tienden a favorecer un evento, o a estar asociados a este aumentando la probabilidad de que suceda, y conocer principalmente la fuerza de esa asociación [24]. Un modelo predictivo busca, por otra parte, estimar la aparición de un evento, aún sin que podamos darnos una explicación clara de cómo fue que el evento se produjo. Esta contradicción entre capacidad explicativa y capacidad predictiva es fuente de frecuente polémica, dado que muchos algoritmos funcionan de manera más o menos opaca, generando el fenómeno conocido como "caja negra" (*black box*), y a pesar de eso pueden encontrarse en la base de decisiones de todo tipo sin que exista una explicación comprensible para todos los interesados [25].

Nuestro interés es doble. Por una parte, generar un análisis de los factores que se asocian y explican la probabilidad de un accidente laboral, pero por otra parte explorar métodos que mejoren la capacidad predictiva de un modelo clásico, como la regresión logística, gracias al uso de técnicas más propias del aprendizaje automatizado (*machine learning*). Para quienes deben tomar decisiones, tanto la explicación como la predicción son elementos críticos, y deben mostrar una suficiente transparencia en sus procedimientos y resultados.

Las variables incorporadas en nuestros modelos más importantes fueron las relacionadas a la auto percepción de salud y bienestar personal, es decir, salud general, salud mental, vitalidad y la escala de estrés de Setterlind, las cuales fueron estadísticamente significativas. Esto nos indica que el sentido de auto percepción de los trabajadores sobre su salud y bienestar es un aspecto relevante a la hora de intentar explicar un accidente laboral. De igual manera, otras variables importantes fueron la Clasificación Internacional Uniforme de las Ocupaciones (CIUO) de los trabajadores y la Clasificación Internacional Industrial Uniforme (CIIU) del centro de trabajo, que también resultaron estadísticamente significativas. Otras variables adicionales que resultaron ser importantes estadísticamente fueron la edad, el salario líquido, días de licencia médica reportadas en el cuestionario, tiempo trabajando en la empresa y su relación contractual, y también, la tasa de ausencia laboral de la empresa asociada a diagnósticos de enfermedades mentales y otros diagnósticos.

Con respecto a las variables de riesgo psicosocial, todas las subdimensiones se concentraron en la parte baja del ranking de importancia y no mostraron una asociación consistente con la probabilidad de tener un accidente laboral, lo que también ha sido observado en algunos estudios previos. Es probable que los factores de riesgo psicosocial no ejerzan su influencia directamente sobre la probabilidad de tener un accidente, sino que a través del estrés como mediador, incluso a través de la autopercepción de salud general y mental.

Otro aspecto relevante que consideramos fue comparar los distintos modelos de predicción y la modelización en aprendizaje automatizado por Random Forest, donde este último resultó tener una mejor capacidad predictiva que la regresión logística.

El presente informe se divide en cinco capítulos, el primero es la introducción, luego se detalla la metodología utilizada (estructura del cuestionario, descripción de variables, muestra utilizada, modelos, balanceo de datos, técnicas estadísticas y de aprendizaje automatizado, evaluación de la predicción, software), en el tercer capítulo se presentan los resultados (algunas estadísticas descriptivas, resultados de regresión logística y de otras técnicas). Por último, se valoran las técnicas y se discuten los resultados.

Método

Este trabajo es un estudio de corte transversal que utilizó la base de datos innominada conservada por la SUSESO de las respuestas al cuestionario SUSESO/ISTAS 21, versión completa (VC), recogidas en 2017 y 2018.

El cuestionario SUSESO/ISTAS 21 es un instrumento diseñado para medir, examinar y analizar los riesgos psicosociales en el trabajo y es implementado en dos versiones (VC y VB). Nosotros utilizamos la VC ya que tiene la ventaja de considerar un mayor número de dimensiones [22], por lo tanto, nos proporcionó una herramienta más útil para una investigación en salud ocupacional. En cualquiera de sus dos versiones, el cuestionario se debe aplicar a todos los trabajadores que prestan servicios en un mismo lugar o centro de trabajo, sea este una empresa, faena, sucursal o agencia (incluye a los trabajadores temporales) [22,23].

Además de las variables contenidas en el cuestionario SUSESO/ISTAS 21, que están asociadas a cada trabajador, se consideró también algunas variables asociadas a la empresa en la que fue aplicado el cuestionario, que fueron utilizadas como variables *proxy* de los centros de trabajo[1], y que nos permitieron calcular las tasas de ausentismo laboral anual por licencias médicas de las empresas, separándolas en diagnósticos de enfermedades mentales, osteomusculares y otros.

El cuestionario SUSESO/ISTAS21

El cuestionario es de carácter confidencial y anónimo, y se responde luego de un período de información y sensibilización de los trabajadores. Según norma [22], solo se admiten tasas de respuesta de un 60% o más. Los procesos de medición por lo general deben repetirse cuando no se alcanza este límite.

[1] Estas variables fueron extraídas de registros administrativos de la Superintendencia de Seguridad Social. En particular utilizamos el Sistema de Gestión de Reportes e Información para la Supervisión de las Mutuales (GRIS Mutuales) y el Sistema de Información de Licencias Médicas y Subsidios por Incapacidad Laboral (SILMSIL).

El cuestionario VC (ver Tabla Nº1) posee 142 preguntas distribuidas en 19 subdimensiones de riesgo psicosocial y variables demográficas (sexo -binaria- y edad -menos de 26 años; entre 26 y 35 años; entre 36 y 45 años; entre 46 y 55; y más de 55 años), salud y bienestar personal (escalas SF-36 de salud general, salud mental y vitalidad; una escala de estrés de Setterlind), variables de trabajo actual, salario y elementos vinculados (contrato y horario de trabajo, condiciones generales del trabajo, endeudamiento, días de licencia médica, entre otras más). Además, se pregunta si el trabajador tuvo un accidente de trabajo en los últimos 12 meses, excluyendo los accidentes de trayecto. La respuesta por accidente de trabajo es nuestra variable de salida o dependiente (binaria).

Tabla Nº1. Resumen de la estructura del Cuestionario SUSESO/ISTAS21, versión completa.

Sección general			
Unidades	**Conceptos**	**Carácter**	**Preguntas**
Datos demográficos	Sexo y edad	Obligatorias	2
Salud y bienestar personal	Salud General	Obligatorias	5
	Salud Mental	Obligatorias	5
	Vitalidad	Obligatorias	4
	Escala de estrés de Setterlind	Obligatorias	12
Trabajo actual, salario y elementos vinculados	Segmentación: unidad geográfica, ocupación (CIUO-08) y departamentos	Editables y Optativas	3 5
	Condición general de trabajo	Obligatorias	6
	Jornada de trabajo	Obligatorias	3
	Contrato de trabajo y salario	Obligatorias	2
	Endeudamiento	Obligatorias	2
	Licencias médicas	Obligatorias	1
	Accidentes laborales	Obligatorias	1
	Enfermedades profesionales	Obligatorias	2
	Carga de trabajo doméstico	Obligatorias	
Total			**53**
Sección específica de riesgo psicosocial			
Dimensiones	**Subdimensiones**	**Carácter**	**Preguntas**
Exigencias psicológicas	Exigencias cuantitativas	Obligatorias	7
	Exigencias cognitivas	Obligatorias	8
	Exigencias sensoriales	Obligatorias	4
	Exigencias emocionales	Obligatorias	2
	Exigencias de esconder emociones	Obligatorias	2
Trabajo activo y posibilidades de desarrollo	Influencia	Obligatorias	7
	Control sobre el tiempo de trabajo	Obligatorias	4
	Posibilidades de desarrollo en el trabajo	Obligatorias	7
	Sentido del trabajo	Obligatorias	3
	Integración en la empresa	Obligatorias	4
Apoyo social en la empresa y calidad de liderazgo	Claridad de rol	Obligatorias	4
	Conflicto de rol	Obligatorias	5
	Calidad de liderazgo	Obligatorias	6
	Calidad de la relación con superiores	Obligatorias	5
	Calidad de la relación con compañeros de trabajo	Obligatorias	6

Compensaciones	Estima Inseguridad respecto del contrato de trabajo Inseguridad respecto de las características del trabajo	Obligatorias Obligatorias Obligatorias	5 5 3
Doble presencia	Preocupación por tareas domésticas	Obligatorias	2
Total			**89**
Total del cuestionario			**142**

Fuente: Manual del Método del Cuestionario SUSESO/ISTAS21, noviembre 2020 (22)

Variables de ocupación, empleo y otros elementos vinculados

- **Ocupación del trabajador**

El cuestionario posee una variable de ocupación que, aunque no es obligatoria, es editable por cada lugar de trabajo y permite extraer y codificar las ocupaciones (con la Clasificación Internacional Única de Ocupaciones, CIUO). Dado que la ocupación puede ser una variable importante para pronosticar los AT, se seleccionaron los cuestionarios con ocupación CIUO claramente definida. Se consideraron nueve grupos principales: directores y gerentes; profesionales, científicos e intelectuales; técnicos y profesionales de nivel medio; personal de apoyo administrativo; trabajadores de los servicios y vendedores de comercios y mercados; agricultores y trabajadores calificados agropecuarios, forestales y pesqueros; oficiales, operarios y artesanos de artes mecánicas y de otros oficios; operadores de instalaciones y máquinas y ensambladores; y ocupaciones elementales. No se considera el grupo principal ocupaciones militares ya que no existen estas ocupaciones en la muestra.

- **Sueldo Líquido del trabajador**

El sueldo líquido mensual considera las siguientes categorías: $200.000 o menos; entre $200.001 y $500.000; entre $500.001 y $800.000; entre $800.001 y $1.000.000; entre $1.000.001 y $2.000.000; y más de $2.000.000.

- **Antigüedad en la empresa**

Tiempo en la empresa como variable categórica, que mide el tiempo en que cada trabajador ha prestado servicios en la empresa o institución. Las categorías son las siguientes: de 0 hasta 6 meses; más de 6 meses y hasta 2 años; más de 2 años y hasta 5 años; más de 5 años y hasta 10 años; y más de 10 años.

- **Horario de trabajo**

Horario de trabajo, que considera las siguientes categorías: horario diurno, turnos exclusivamente nocturnos y turnos rotatorios.

- **Relación laboral**

Relación laboral, variable que refleja la relación laboral con la empresa o institución. Las categorías son: contrato indefinido, contrato temporal y otro o sin contrato.

- **Dos o más secciones en la empresa**

Variable binaria que refleja si los trabajadores han trabajado en más de una sección o departamento al mismo tiempo.

- **Dificultad para pagar deudas**

Una variable vinculada al empleo es la dificultad para pagar deudas con las siguientes categorías ordinales: no tengo deudas; tengo deudas, pero no tengo dificultades para pagarlas; tengo deudas, y tengo ocasionales dificultades para pagarlas; tengo deudas, y tengo siempre dificultades para pagarlas; y tengo deudas, y tengo permanentes y graves dificultades para pagarlas.

- **Número de días de licencias médica**

Consideramos los días de licencia médica reportados por los trabajadores en los últimos doce meses como variable categórica. Se excluyeron licencias médicas por prenatal y postnatal, y por permiso por enfermedad grave de niño menor de 1 año. Los tramos en días de licencia médica fueron los siguientes: 0 días; entre 1 y 7 días; entre 8 y 14 días; entre 15 y 30 días; entre 31 y 60 días; entre 61 y 90 días; y más de 90 días.

Variables de salud y estrés

- **Autopercepción de salud general, salud mental, vitalidad y estrés**

Las variables de salud y bienestar personal (salud general, salud mental y vitalidad) del SF-36 [26], que están en el cuestionario, se transformaron en variables binarias, "caso" de mala salud y "no caso", donde el "caso" de mala salud se definió como aquella observación con un puntaje igual o menor a dos desviaciones estándar bajo la media. Un proceso similar se usó en la variable de estrés, considerando "alto estrés" aquellos puntajes iguales o superiores a dos desviaciones estándar sobre la media. El objetivo de esta transformación es facilitar la lectura e interpretación de los resultados.

Puntaje de riesgo psicosocial

En la Sección Específica de Riesgo Psicosocial del cuestionario se busca medir el riesgo en diecinueve subdimensiones. Todas las preguntas tienen respuestas en una escala Likert con puntuaciones que van desde 0 a 4 puntos, donde el mayor puntaje siempre indica mayor riesgo.

Cada subdimensión o escala tiene diferente número de preguntas y de puntos, y por lo mismo, y de acuerdo con el manual del instrumento [22], los puntajes brutos fueron transformados a porcentajes para realizar comparaciones entre escalas en base a una misma puntuación estandarizada. Para la transformación, se sumaron los puntajes obtenidos de cada subdimensión, se dividió por el máximo puntaje posible que se puede obtener en dicha subdimensión, y luego se multiplicó por 100 conservando 2 decimales. Una vez que se obtuvo el puntaje de cada trabajador para cada subdimensión, se comparó con los valores estandarizados para Chile y de esa forma, se categorizó en alto, medio y bajo el riesgo de cada individuo en cada subdimensión, de acuerdo con lo señalado en el manual del instrumento.

Por último, se dicotomizaron las respuestas en "alto riesgo" y el resto. El objetivo fue similar al de las variables de salud, esto es, permitir una interpretación más sencilla de los resultados.

Variables asociadas a la empresa del trabajador y centro de trabajo

- **Tasa de ausencia laboral**

Desde los registros administrativos de la Superintendencia de Seguridad Social se utilizó una variable que caracteriza a la empresa en cuanto a la tasa de ausencia laboral debido a las licencias de origen común de sus trabajadores.

La tasa anual de ausencia laboral por diagnósticos mentales, osteomusculares y otros diagnósticos se obtuvo a partir del registro administrativo del Sistema de Información de Licencias Médicas y Subsidios por Incapacidad Laboral (SILMSIL). Se utilizó el número de días de licencias médicas de origen común por empresa junto con el código de diagnóstico CIE-10 (Clasificación Internacional de Enfermedades) para clasificarlas en tres grupos: Trastornos mentales y del comportamiento (Código CIE-10 F), Enfermedades osteomusculares (Código CIE-10 M) y Otros diagnósticos (resto de códigos CIE-10).

Al mismo tiempo en el registro del Sistema de Gestión de Reportes e información para la supervisión (GRIS Mutuales), se obtuvo el número promedio de protegidos del régimen Ley 16.744, por empresa. Con ambos insumos se construyó la tasa de ausencia laboral anual considerando estos tres grupos de diagnósticos por separado. Es decir, el número de días que se ausentaron los trabajadores debido a una licencia médica de origen común sobre el número de días laborales de este grupo de trabajadores en un año. Para estimar la tasa anual de ausencia laboral por diagnóstico se utilizó la siguiente fórmula:

$$Tasa\_AL_{e,t,d} = \left(\frac{\sum_{i=1}^{N} Días\_LM_{e,t,d}}{Protegidos_{e,t} \cdot 20 \cdot 12} \right) \cdot 100$$

Donde $Tasa\_AL_{e,t,d}$ es la Tasa de ausencia laboral de la empresa e, en el año t ($t = \{2017, 2018\}$), en la familia de diagnósticos CIE-10 agrupada en d ($d = \{Mentales, Osteomusculares, Otros\}$). Por su parte, $\sum_{i=1}^{N} Días\_LM_{e,t,d}$ es la sumatoria de los días de ausencia laboral de origen común del trabajador i ($i = \{1, \dots, N\}$) en la empresa e, en el año t, de la familia de diagnósticos d.

En el denominador se encuentra $Protegidos_{e,t}$ que es el promedio mensual de trabajadores protegidos por la Ley 16.744 en la empresa e en el año t, que se multiplica por 20 que son los días hábiles aproximado por mes (se excluyen fin de semanas y feriados legales) y por 12 que es el número de meses. Por lo tanto, el resultado obtenido refleja el porcentaje de días de ausencia laboral de origen común en un año, considerando el universo de días laborales de los trabajadores de la empresa.

- **Clasificación Internacional Industrial Uniforme (CIIU) del centro de trabajo**

La actividad económica se obtuvo del trabajo que realiza la Superintendencia con el cuestionario SUSESO/ISTAS21 y corresponde a las actividades económicas del centro de trabajo, las cuales se agruparon en ocho sectores económicos: Agricultura y pesca; Minería; Industria manufacturera; Electricidad, Gas y Agua (EGA); Construcción; Comercio; Transporte y comunicaciones; y Servicios.

La muestra utilizada

Para este estudio se trabajó con los resultados obtenidos del cuestionario completo para los años 2017 y 2018 a nivel nacional. Los datos disponibles del cuestionario no son obtenidos de una muestra aleatoria y tampoco tienen el carácter de un censo, por lo tanto, no es factible realizar inferencias estadísticas sobre el riesgo psicosocial en ningún nivel de agrupación.

También debe considerarse que, de acuerdo con el Protocolo del Ministerio de Salud [23], a partir del 2013, ciertos centros de trabajo deben aplicar de manera obligatoria la VC. Estos casos son:

a. Cuando se determina la existencia de un trabajador con una enfermedad profesional de salud mental.
b. Cuando la medición con el cuestionario breve arroja riesgo alto
c. Por iniciativa propia de una organización, que prefiere realizar la versión completa del cuestionario, en vez de la breve.

Por lo tanto, se puede observar un sesgo de selección a la hora de analizar los datos. Por lo mismo, no es de extrañar que los resultados tiendan a estar sobre representados en la categoría de riesgo alto en alguna de las dimensiones o subdimensiones.

En consecuencia, es una muestra no probabilística y por conveniencia, y no puede considerarse representativa de la población trabajadora chilena. Sin embargo, tiene

varias ventajas. La primera es su tamaño. Durante 2017 y 2018 se recogieron 224.800 cuestionarios en VC, que representan alrededor del 3,7% de la fuerza laboral protegida por la Ley de Accidentes del Trabajo y Enfermedades Profesionales [3]. Esto la convierte en una fuente importante de datos sobre riesgo psicosocial.

Por otra parte, aun cuando es una muestra sesgada, dado que la mayoría de los centros de trabajo que aplican la VC están obligados a hacerlo porque se encuentran en riesgo y/o tienen trabajadores con enfermedades mentales de origen laboral, lo que no permite estimar tasas, sí es posible establecer correlaciones entre sus parámetros.

Para el estudio, se excluyeron de la muestra todos los cuestionarios con respuestas vacías en las variables seleccionadas en nuestro modelo, por lo que se consideró finalmente un total de 113.556 cuestionarios.

Los modelos utilizados

Dado que el objetivo del estudio es explorar una relación funcional mediante la predicción o pronóstico de una variable binaria, utilizamos dos modelos predictivos para el análisis de los datos. El primero fue una regresión logística (RL) [27], que permite una observación clara de la fuerza del efecto de cada variable independiente o de entrada sobre la probabilidad de que la variable de salida sea de la clase de interés o criterio (tener un accidente). El segundo modelo fue *Random Forest* (RF) [28], un algoritmo de clasificación que toma en cuenta las variables independientes o de entrada para pronosticar la clase que adoptará la variable de salida, pero no permite conocer con el mismo grado de exactitud la fuerza del efecto de cada variable de entrada.

Ambos modelos se utilizan en los análisis de aprendizaje automatizado supervisado. El aprendizaje automatizado supervisado es un algoritmo que toma un vector de entrada compuesto por *n*-variables y lo proyecta o mapea sobre un valor o etiqueta de clase que es el objetivo, salida o *target*. El término "supervisado" significa que el entrenamiento del algoritmo se realiza con conjuntos de datos en los que se conoce previamente el contenido de la variable de salida. Esto es, el algoritmo procesa el vector de entrada e intenta modelar la distribución de probabilidades para predecir la salida "*y*" en función del vector de entrada "*x*" [29]. El rendimiento relativo del aprendizaje automatizado depende en gran medida del algoritmo empleado, y no es sencillo saber a priori cuál tipo de algoritmo arrojará la mejor clasificación en un conjunto de datos en particular [29]. Por lo general, cuando se trata de datos de baja dimensionalidad (con pocas variables independientes), y cuando el objetivo es observar el tamaño del efecto de las variables sobre el criterio, la regresión logística sigue siendo el modelo más adecuado [30]. Pero cuando la intención es principalmente predictiva (se quiere predecir un resultado sin conocer en detalle la participación de cada variable) y, sobre todo, cuando se trata de datos de alta dimensionalidad, tiene un mejor rendimiento *Random Forest* [30]. Por lo mismo, la regresión logística permite una clara interpretación de los resultados, lo que es menos claro con *Random Forest* y mucho menos con otros modelos de aprendizaje automatizado tal como las redes neurales, situación que ha generado no poca controversia [25]. Una interpretación clara es deseable para quien necesite diseñar alguna estrategia de intervención a partir de sus resultados.

Balanceo de datos

En cualquiera de los dos métodos clasificatorios de aprendizaje automatizado, uno de los problemas habituales es el desbalance del número de datos de las categorías dicotómicas de respuesta, en el que estos muestran una gran asimetría (*skewness*) de distribución [29]. Una muestra de datos está desbalanceada si una de las clases de respuesta (que habitualmente es la que nos interesa) está en un número muy inferior a otra. Cuando se trabaja con datos desbalanceados, los modelos de clasificación adquieren un sesgo y tienden a clasificar correctamente la clase mayoritaria, pero clasifican mal la clase minoritaria [30]. Incluso, un modelo podría clasificar todos los casos como si simplemente pertenecieran a la clase mayoritaria ignorando por completo la clase minoritaria y aun así obtener una gran exactitud, equivalente a la proporción de casos mayoritarios en la muestra, dejando la falsa impresión de que el modelo funciona razonablemente bien. En nuestro caso, debido al desbalance entre la clase de interés minoritaria (si declara haber tenido un accidente del trabajo) (n = 15.051; 13,25%) y la clase complementaria mayoritaria (si declara no haber tenido accidente del trabajo) (n = 98.505; 86,75%), ambos algoritmos de aprendizaje (*Random Forest* y regresión logística) tienden a focalizar su clasificación en la clase mayoritaria, en tanto ignoran o mal clasifican la clase minoritaria. De hecho, si consideramos todo el conjunto de datos, el algoritmo podría clasificar todo como "no AT" alcanzando una exactitud de un 86,9%, lo que parece falsamente elevado (solo clasificó bien la clase mayoritaria). Nosotros, además, lo probamos con regresión logística y nos arrojó una exactitud de un 86,9%, en tanto la sensibilidad (capacidad de detectar los casos positivos, "AT") fue apenas de un 0,63%, y la especificidad (capacidad de detectar los casos negativos, "no AT") fue de un 99,92%.

Existen diversos procedimientos para afrontar este problema [29, 30-34]. Por ejemplo, se puede asignar un costo diferencial a los errores de clasificación, mayor para la clasificación incorrecta de la clase de interés. Un segundo procedimiento consiste en agregar un sesgo interno en el proceso de discriminación para compensar el desbalance. Otro método consiste en extraer nuevas muestras de los datos, ya sea sobre muestreando la clase minoritaria y/o sub muestreando la clase mayoritaria hasta que ambas clases sean más o menos equivalentes. No obstante, este procedimiento tiene una serie de problemas, por ejemplo, deja afuera del análisis a un grupo importante de datos que podrían ser críticos, o agregar datos inadecuados en el sobre muestreo provocando un sobreajuste (*overfitting*) del modelo. Nosotros utilizamos este último procedimiento y mostró el problema mencionado y se obtuvo un rendimiento mediocre en la clasificación final, de manera que fue descartado.

Un procedimiento algo diferente para abordar este problema fue creado por Chawla et al. [35] al que llamaron *Synthetic Minority Oversampling Technique* (*SMOTE*), en el que se sobre muestrea generando observaciones "sintéticas" de la clase minoritaria en vez de sobre muestrear con reemplazo (*bootstrapping*). La técnica *SMOTE* en esencia consiste en generar una serie de registros sintéticos (esto es, artificiales) a partir de los registros minoritarios reales y sus K-vecinos más cercanos (*K-Nearest Neighbors, KNN*). En el método *KNN* [36,37] las observaciones reales (vectores) generan un espacio multidimensional (*feature space*) que puede dividirse o particionarse en "regiones de

cercanía". En estas regiones se pueden introducir puntos (registros) sintéticos (artificiales) de manera aleatoria a lo largo de los segmentos de línea que unen puntos cercanos (observados) en el espacio definido así. Los nuevos puntos (artificiales) tendrán características altamente probables y se les puede considerar como nuevos "k vecinos más cercanos" (*KNN*), tal como si fueran registros observacionales reales. El algoritmo SMOTE que utilizamos [38] calcula hasta 5 de estos posibles vecinos artificiales por cada punto observacional real de la clase minoritaria, pero en la práctica utiliza menos dependiendo del tamaño requerido de la nueva muestra. Por ejemplo, si se requiere que la nueva muestra sea 200% mayor, genera 2 k vecinos por cada observación real y construye una nueva muestra minoritaria que contendrá 2 veces más observaciones sintéticas. Simultáneamente, el algoritmo SMOTE permite sub muestrear la clase mayoritaria adaptándola a la nueva clase minoritaria, en un porcentaje que se puede establecer como híper parámetro (esto es, qué nueva relación tendrá la clase mayoritaria con respecto a la minoritaria). En este trabajo, se eligió un híper parámetro con sobre muestra de 200% para la clase minoritaria (es decir, habrá 3 veces el número original de observaciones minoritarias) y 150% para la clase mayoritaria (es decir, habrá 1 vez y media observaciones mayoritarias sobre la cantidad de *nuevas* observaciones minoritarias). Con esto se logra que el modelo ahora trabaje, bien sobre una base sesgada a la inversa (cuando es la clase minoritaria la que se convierte en mayoritaria), o bien sobre dos clases que están balanceadas (si se elige un híper parámetro que haga equivalentes ambas clases). Con el parámetro submuestra ajustado a 150%, SMOTE muestrea una cantidad igual de casos mayoritarios que minoritarios, de manera que trabajamos sobre una base balanceada. La base así preparada evita los problemas de sobreajuste (*overfitting*) en la clase mayoritaria que tienen las técnicas tradicionales de sobre muestreo. Además, es posible considerar muchas características de los casos mayoritarios que se perderían si solo se tomara una muestra con un número de casos mayoritarios similar a los minoritarios originales.

Regresión logística

La regresión logística (27) es una técnica mucho más conocida que *Random Forest* o SMOTE, de modo que no creemos necesario describirla con detalle. La regresión logística entrega el logaritmo de la razón entre la probabilidad de que un evento ocurra sobre la probabilidad de que el mismo evento no ocurra (*odds*), en presencia de una o más variables independientes:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$$

donde π es la probabilidad de que ocurra un evento (por ejemplo, que exista un accidente de trabajo), 1-π es la probabilidad de que no ocurra el evento y los β_i son los coeficientes de la regresión asociados a las variables x_i que suelen obtenerse por el método de máxima verosimilitud [27]. En suma, la regresión logística entrega el logaritmo de la probabilidad (entendida como razón entre dos probabilidades) de que

un evento ocurra cuando cambian las variables independientes. La clasificación misma implica elegir un umbral de probabilidad arbitrario (normalmente es 0,5) por sobre el cual el evento será clasificado como perteneciente a una clase (por ejemplo, accidente del trabajo).

Conociendo el comportamiento de la regresión logística con la base no balanceada, probamos a comparar los resultados sobre una base balanceada. Con ambos cálculos (sobre base no balanceada y sobre base balanceada), probamos su capacidad predictiva con una validación cruzada y el área bajo la curva ROC.

Random Forest

Luego se utilizó *Random Forest* como modelo de predicción. Esta metodología está basada en un conjunto de árboles de decisión sobre muestras aleatorias del conjunto de datos y del conjunto de variables [28,38]. La salida del "bosque aleatorio" es por la moda, es decir, por la respuesta más frecuente entre todos los árboles, lo que suele compararse con un sistema de "votación" (cada árbol "vota" por su respuesta). Como en todos los modelos de estimación en datos de alta dimensionalidad, existen dos problemas relevantes (e inseparables) que son el sesgo y la varianza [29,30]. El sesgo se observa comparando los valores predichos del modelo con los valores efectivos de la variable de estudio. Los árboles de clasificación individuales se adaptan o ajustan con gran precisión a los datos con los que se entrenan y presentan por lo tanto un alto sesgo (por eso predicen bien los datos de entrenamiento), pero este gran ajuste a los datos actuales tiene al mismo tiempo por contrapartida que en la medida que el modelo se complejiza se vuelve muy sensible a las pequeñas variaciones de los datos nuevos, es decir, posee una escasa varianza, y por este motivo tienen un mal rendimiento al clasificar datos nuevos [30]. Este problema se le conoce como dilema (*trade-off*) sesgo/varianza [29,30].

Random Forest procura resolver el dilema sesgo/varianza utilizando no uno, sino que un conjunto (*ensemble*) de árboles (un "bosque"), en el que cada árbol se entrena para la clasificación con solo una muestra del total de datos disponibles y solo una parte aleatoria de todas las variables. Este proceso busca de-correlacionar los árboles y de esa manera disminuir para el "bosque" la tasa de error de clasificación. La forma en que se muestrean los datos es por *bagging* (*boostrap aggregating*), que toma una muestra con reemplazo de los datos diferente para cada uno de los árboles, pero además toma solo una cantidad aleatoria de todas las variables (siempre menor que el total de variables), que se aplica a cada nodo de decisión de cada árbol. La cantidad de datos de la muestra y la cantidad de variables seleccionada se mantienen constantes durante el proceso y es uno de los híper parámetros del modelo. Cada árbol, en consecuencia, se entrena con solo una parte de los datos (alrededor de un 67,5% de ellos), lo que deja aproximadamente un tercio de los datos sin utilizar en cada árbol. Estos datos no utilizados constituyen el *Out-Of-Bag* (*OOB*) y con ellos el modelo puede evaluar su precisión de clasificación, lo que se muestra a través de la tasa de error del OOB (*OOBE*), pero también sirven para estimar la importancia de cada variable. Nosotros efectuamos una segunda validación cruzada, además del cálculo del *OOB*.

Una de las ventajas adicionales de *Random Forest* es que a partir del cálculo del error *OOB* puede construir medidas que reflejan la importancia de las variables en el modelo. Existen dos indicadores de la importancia de las variables que vienen con el algoritmo "randomForest" en R [38], la reducción media de exactitud (*MDA*, por sus siglas en inglés "*Mean Decrease Accuracy*") y la reducción media del índice Gini, o reducción de impureza (MDG, por sus siglas en inglés "*Mean Decrease Gini*"). El *MDA* refleja el impacto o pérdida que tiene sobre el rendimiento de la predicción la modificación del estado de una variable en particular en la base de datos de entrenamiento, esto es, cuando una variable cambia su estado (por ejemplo, "mujer" a "hombre" en el caso de una variable binaria). Esta medida nos permitiría encontrar las variables que tienen mayor poder predictivo en nuestro modelo, aun cuando no fueran dicotómicas. Por su parte, el *MDG* es una medida que entrega información sobre la impureza de un nodo. Un alto Gini (cercano a 1) quiere decir que una observación en particular, con un estado de las variables en particular, puede quedar clasificada de modo incorrecto con alta probabilidad. La disminución del Gini indica que la probabilidad de error en la clasificación disminuye. Por lo tanto, mide la capacidad de las variables en los nodos de dividir las observaciones en una categoría homogénea. Una alta reducción del Gini provocada por una variable implica que, al no considerarla, la calificación sufre en precisión.

Validación cruzada

La validación cruzada consiste simplemente en una comparación de las salidas de clasificación proyectadas por el modelo (la predicción) y los resultados observados (esto es, los resultados reales o etiquetas que contiene la base de prueba). La predicción se efectúa con un valor de probabilidad arbitrario, que normalmente se fija en 0,5 (por ejemplo: $\tilde{y} = 1$ si $\pi > 0{,}5$; $\tilde{y} = 0$ si $\pi \leq 0{,}5$). Esto se lleva a una matriz de confusión con los cuatro resultados posibles, lo que permite obtener los estadísticos de rendimiento habitual.

Para ejecutar la validación cruzada, el modelo (sea regresión logística, balanceada y no balanceada, o sea *Random Forest*) se entrena solo sobre una parte de los datos (utilizamos el 70%, lo que llamamos base *train*) y se prueba sobre el 30% restante (que llamamos base *test*).

Área bajo la curva (ROC)

Por último, los resultados fueron procesados a través del procedimiento *Receiver Operating Characteristics* (*ROC*) [42] que calcula el grado de rendimiento de cada modelo (la tasa de verdaderos positivos sobre la tasa de falsos positivos, o 1-especificidad) que se expresa a través del área bajo la curva (*AUCROC*) y la estimación de su punto de corte óptimo, y además permite observar de manera gráfica el rendimiento de cada modelo. El área bajo la curva es un buen estimador de la capacidad predictiva de un modelo, mejor aún que la validación cruzada, dado que calcula la fuerza predictiva de todos los posibles puntos de corte (no solo 0,5).

El software utilizado

Todo el proceso se llevó a cabo con el software R y RStudio, diferentes paquetes estadísticos (glm, DMwR, ROCR, Caret, randomForest) y gráficos asociados [38–42].

Descripción estadística

La muestra de datos se constituyó de 113.556 trabajadores que completaron el cuestionario y tenían una ocupación bien definida. Dentro de este grupo hubo 98.505 (86,75%) que declararon no haber tenido un accidente laboral y 15.051 (13,25%) que declararon haber tenido un accidente laboral en los últimos 12 meses[2]. En el Anexo Nº1 se presenta un resumen de las distribuciones de todos los campos que fueron utilizados para caracterizar los trabajadores con y sin accidentes. Adicionalmente, se incluye la diferencia de medias de las variables utilizadas con sus respectivos test de significancia con el objetivo de visualizar las diferencias estadísticas en ambos grupos de acuerdo con la característica observada.

Al realizar una comparación entre la población que declaró haber tenido accidente laboral con aquella que declaró no haber tenido uno, se observa una distribución muy similar entre hombres y mujeres, siendo la participación de los hombres levemente superior. Los tramos de edad presentan una distribución parecida entre ambos grupos, salvo el tramo menor a 26 años que presentó una mayor frecuencia entre quienes declararon un accidente.

En relación con la ocupación, los "Operadores de instalaciones y máquinas y ensambladores" y "Trabajadores de los servicios y vendedores de comercios y mercados" presentan una diferencia de 5 puntos porcentuales mayor entre quienes declararon un accidente en relación con aquellos que no lo hicieron. De esto modo, ambas ocupaciones representan el 38% de los trabajadores que declararon un accidente.

Entre quienes declararon un accidente, se observa una mayor frecuencia de trabajadores con una antigüedad en la empresa de 2 a 5 años, de trabajadores con contrato indefinido y con jornadas rotativas. Asimismo, entre quienes declararon un accidente existe una mayor participación de los trabajadores con remuneraciones más bajas. En particular, el 42% de los trabajadores que declaró no haber tenido un accidente tiene una remuneración menor o igual a $500.000, dicho porcentaje aumenta a 47% para el grupo que declaró un accidente.

La percepción de salud y bienestar del trabajador tiene una incidencia importante en el grupo que ha sufrido un accidente. Así, aquellos que declaran haber tenido un muy mal estado de salud ("caso") sufren más accidentes que el grupo con mejor percepción de su salud. En general, se observa que el grupo que declaró un accidente tiene una

[2] Corresponde a la pregunta: "En los últimos 12 meses, ¿ha tenido usted algún accidente de trabajo como golpe, caída, herida, corte, fractura, quemadura o envenenamiento? (excluya accidentes de trayecto)"

frecuencia de 4 puntos porcentuales promedios adicionales que el grupo que declaró no tener un accidente laboral.

Aunque es mayor la frecuencia de trabajadores que declaran un accidente en las empresas con una tasa de ausencia laboral más elevada, esta diferencia no fue significativa.

Finalmente, entre los trabajadores que declaran un accidente existe una mayor proporción de trabajadores en riesgo alto para todas las subdimensiones de riesgo psicosocial. De este modo, para subdimensiones catalogadas con riesgo alto, el grupo que declaró tener accidente tiene una frecuencia de 5,5 puntos porcentuales promedio mayor que el grupo sin accidentes.

Resultados

La base de datos contenía un total de 224.800 cuestionarios. Se excluyeron todos los que tenían un vacío en las variables seleccionadas en el modelo, lo que dejó un total de 113.556 cuestionarios disponibles para el estudio. De ellos, 39.712 cuestionarios se respondieron en 2017, provenientes de trabajadores de 290 empresas u organizaciones, y otros 73.844 cuestionarios provenían de 505 empresas u organizaciones y se respondieron en 2018. Utilizando el Clasificador Chileno de Actividades Económicas 2007 (CIIU.CL 2007), el sector económico con mayor presencia en la muestra fue el de servicios (que incluye a los organismos estatales, salud y educación), y tuvo alrededor de un 60% de presencia en la muestra en ambos años (Tabla Nº2).

Tabla Nº2. Número total de empresas y trabajadores según actividad económica.

Actividad Económica	2017				2018			
	Empresas		Trabajadores		Empresas		Trabajadores	
	N	%	N	%	N	%	N	%
Agricultura y Pesca	8	2,8	341	0,9	6	1,2	784	1,1
Comercio	28	9,7	8.824	22,2	82	16,2	14.849	20,1
Construcción	9	3,1	209	0,5	11	2,2	364	0,5
Electricidad, Gas y Agua	5	1,7	480	1,2	6	1,2	195	0,3
Industria Manufacturera	22	7,6	1.443	3,6	32	6,3	1.937	2,6
Minería	1	0,3	64	0,2	3	0,6	181	0,3
Servicios	196	67,6	26.708	67,3	326	64,6	43.012	58,3
Transporte y Comunicaciones	21	7,2	1.643	4,1	39	7,7	12.522	17,0
Total	**290**	**100,0**	**39.712**	**100,0**	**505**	**100,0**	**73,844**	**100,0**

En la Tabla Nº3 se muestra la proporción de hombres y mujeres en el total de cuestionarios. No existe una diferencia significativa en las proporciones por sexo por tramo etario, aunque puede verse que las mujeres tienden a ser mayoritarias en los

tramos intermedios (26 a 45 años) y los hombres en los tramos extremos (<26 y >45 años).

Tabla Nº3. Número de trabajadores según tramo de edad y sexo.

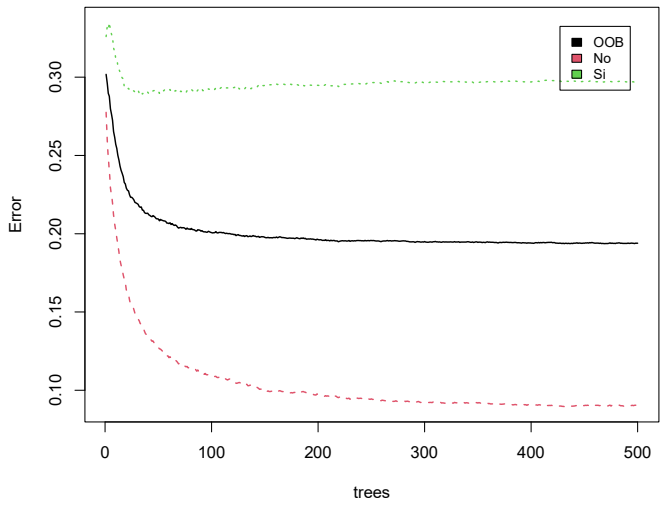
Edad	Hombres		Mujeres		Total	
	N	%	N	%	N	%
< 26	6.379	11,4	6.083	10,6	12.462	11,0
26 - 35	17.256	30,8	19.889	34,6	37.145	32,7
36 - 45	14.423	25,7	15.871	27,6	30.294	26,7
46 - 55	10.575	18,9	10.474	18,2	21.049	18,5
> 55	7.411	13,2	5.195	9,0	12.606	11,1
Total	**56.044**	**100,0**	**57.512**	**100,0**	**113.556**	**100,0**

Como ya fue mencionado, la proporción de observaciones con accidente laboral fue de un 13,25% del total y las observaciones sin accidente laboral fueron el 86,75% restante.

Al balancear la muestra con el procedimiento SMOTE se obtuvo una nueva muestra balanceada con 90.306 observaciones (50% en cada clase de la variable de salida) y que fue el insumo para el procedimiento de clasificación en *Random Forest* y también para las predicciones con la regresión logística.

Por *default*, el algoritmo randomForest de R utiliza 500 árboles (38). Contrastamos la tasa de error OOB contra el número de árboles, buscando si el modelo pudiera mejorar con un número mayor de árboles, pero puede verse que la tasa se vuelve casi asintótica a partir aproximadamente de los 200 árboles, de modo que mantuvimos el default (Figura 1):

Figura 1: Tasa de error estimada OOB contra el número de árboles



A partir del *OOBE Random Forest* entrega una estimación de la importancia de las variables sobre el modelo, *MDA* y *MDG*.

Como se puede observar en la Figura 2, las cuatro dimensiones más importantes según la reducción media de exactitud (MDA) son el estrés, la salud mental, la salud general y vitalidad. Las tres variables de riesgo psicosocial más importantes fueron claridad de rol, influencia y esconder emociones.

Por otro lado, las variables que aportaron menos a la predicción del modelo fueron la subdimensiones exigencias cognitivas, inseguridad respecto del contrato de trabajo e integración en la empresa.

En la Figura 2 también aparece la reducción media del índice Gini (*MDG*). Dentro de las variables más importantes bajo este criterio siguen siendo estrés, salud mental y vitalidad, sin embargo, las tres tasas de ausencia laboral de las empresas por diagnósticos mentales, osteomusculares y otros diagnósticos entran en la parte alta del ranking de importancia. Con respecto a las subdimensiones relacionadas con el riesgo psicosocial, las tres más importantes fueron exigencias sensoriales, sentido del trabajo y claridad de rol. Por último, las variables con menos importancia fueron la Subdimensión integración en la empresa, si el trabajador ha estado en dos o más secciones y la Subdimensión estima.

El último análisis realizado con la herramienta de *Random Forest* fue validar el modelo utilizando la base de datos de testeo, y para ello, construimos la matriz de confusión que se muestra en la Tabla Nº4. El 4,3% de las observaciones testeadas fueron mal clasificadas como no tiene accidente laboral, cuando en realidad, si habían declarado tener uno; y un 14,9% fueron mal clasificados como tiene accidente laboral, cuando se observó que no habían tenido uno. Además, el modelo predijo de manera correcta un 34,9% para aquellos que declararon haber tenido un accidente laboral, mientras que un 46,0% para aquellos que declararon no haber tenido accidente laboral. Por último,

nuestro modelo tiene un error promedio de clasificación igual a 17,7%. En la Figura 3 se presenta la curva ROC que tiene un área bajo la curva igual a 0,89.

Tabla Nº4. Matriz de confusión de la base de datos de testeo.

Accidente de trabajo	Observados			
		Si	No	Error de clasificación
Proyectados	Si	9.442	1.167	11,0
	No	4.023	12.460	24,4

Figura 2: Importancia de las variables

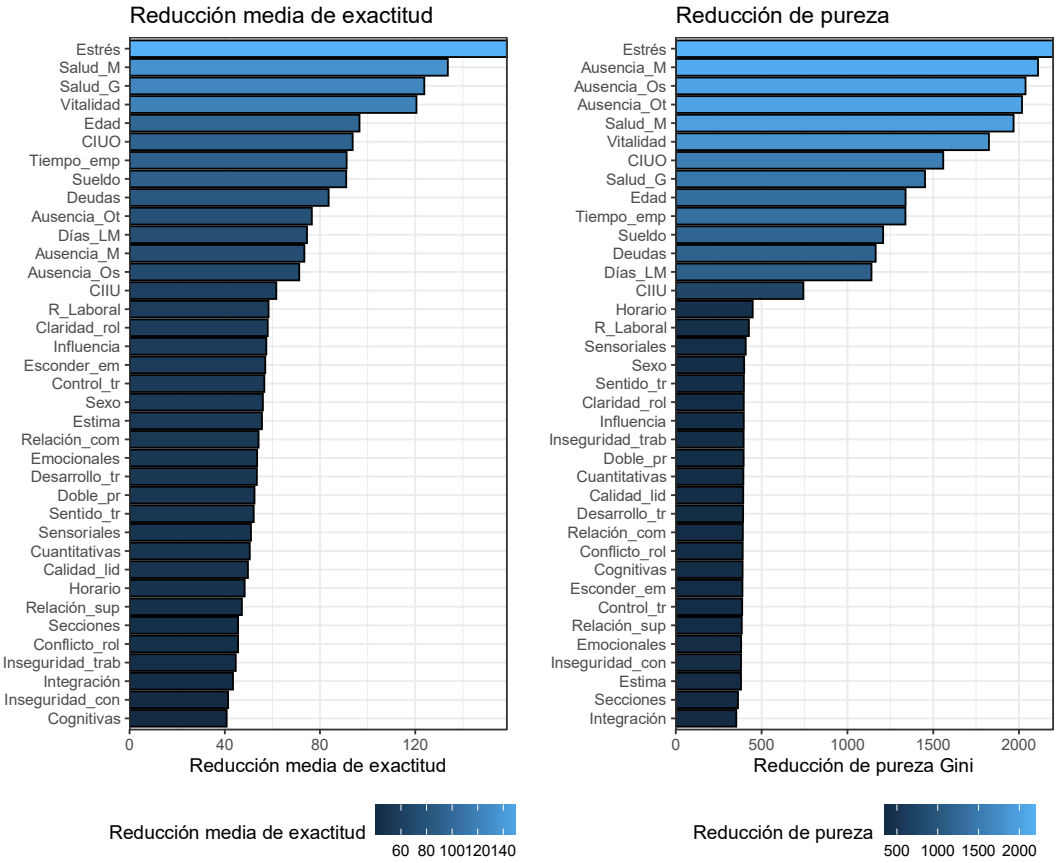
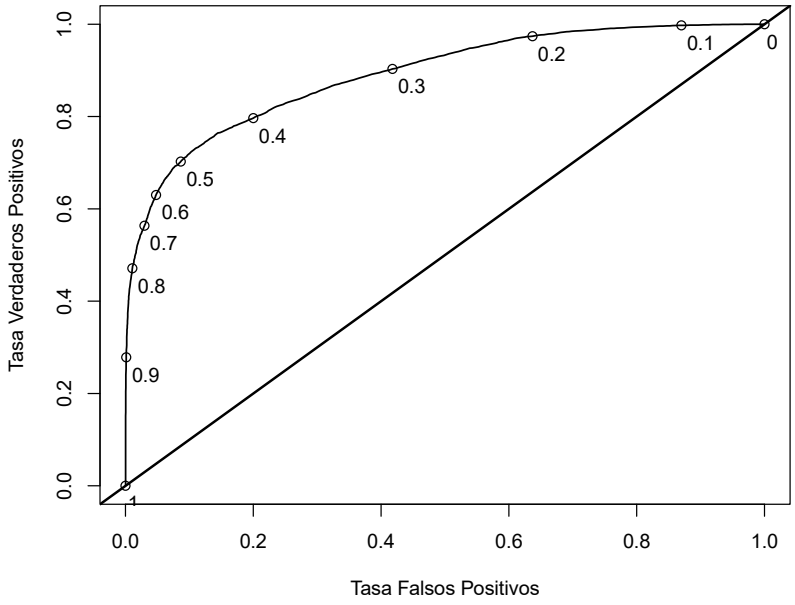


Figura 3: Curva ROC



Resultados utilizando regresión logística

Adicionalmente, utilizando la base de datos balanceada completa y procurando una explicación (más que una predicción), estimamos modelos de regresión logísticos, tomando como variable dependiente si los trabajadores declaran haber tenido un accidente laboral o no, y como variables independientes las 5, 10 y 15 primeras variables más importantes, que resultaron del criterio de importancia Reducción Media de Exactitud (*MDA*). De igual manera en el Anexo Nº2 de este documento podrá encontrar la estimación del modelo considerando todas las variables incorporadas en el *Random Forest*.

En la estimación (modelo 1) de la Tabla Nº5, consideramos como variables independientes la escala de estrés de Setterlind, salud mental, salud general, vitalidad y edad. Todos los coeficientes estimados resultaron significativos estadísticamente. En la Tabla Nº5 se muestran los *Odds Ratio* y en paréntesis los respectivos intervalos de confianza al 95%.

Como se puede observar, las variables de salud y bienestar personal resultan ser muy relevantes a la hora de explicar un accidente laboral. Aquellos trabajadores que tienen un nivel alto en la escala de estrés de Setterlind, manteniendo todo lo demás constante, tienen 5,92 veces más probabilidades de tener un accidente laboral que aquellas personas que no tienen un nivel alto en la escala de estrés. Por su parte para aquellos trabajadores que tienen un nivel alto (de manera independiente) en salud mental, salud general y vitalidad tienen 4,26; 5,08; y 4,78 veces más probabilidades de tener un accidente, respectivamente. Por último, al analizar por edad de los trabajadores, consideramos como categoría base de comparación aquellos trabajadores que tienen

menos de 26 años. Los *Odds Ratio* de todos los tramos tienen menos probabilidades de tener un accidente laboral que aquellos trabajadores que tienen menos de 26 años, destacando el tramo de edad entre 46 años o más que tienen 1,43 veces menos probabilidades de tener un accidente laboral.

En la estimación (modelo 2) de la Tabla Nº5, agregamos las cinco variables adicionales en el orden de importancia según el criterio *MDA* que son: la clasificación de ocupaciones (CIUO), tiempo de antigüedad en la empresa, sueldo líquido del trabajador, dificultad de pagar las deudas y tasa de ausencia laboral anual asociadas a otros diagnósticos. Al estimar este modelo no cambian la significancia estadística de los coeficientes del modelo 1. Por su parte, las nuevas variables incorporadas presentan significancia estadística, con excepción de la variable tiempo de antigüedad en la empresa (la categoría entre 5 y 10 años en la empresa tuvo una significancia estadística marginal) y dos categorías de la variable que mide la dificultad de pagar las deudas ("Tengo deudas, pero no tengo dificultades para pagarlas" y "Tengo deudas, y tengo ocasionales dificultades para pagarlas"). En el caso de la variable ocupación de los trabajadores, consideramos como categoría base de comparación las ocupaciones asociadas a Directores y gerentes ya que asumimos que estos trabajadores cumplen funciones, en su mayoría, poco riesgosas en cuanto a accidentabilidad. Las ocupaciones Agricultores y trabajadores calificados agropecuarios, forestales y pesqueros; Operadores de instalaciones y máquinas y ensambladores; y Oficiales, operarios y artesanos de artes mecánicas y de otros oficios tuvieron los mayores *Odds Ratios* estimados, lo que significa que tienen 2,82; 2,50; y 2,24 veces, respectivamente, más probabilidades de tener accidentes laborales que aquellos trabajadores que tienen ocupaciones de Directores y Gerentes. En el caso de la variable de sueldo líquido del trabajador, todos los tramos tienen menos probabilidades de tener accidentes laborales que aquellos trabajadores que tienen un sueldo líquido de $200.000 o menos. Los trabajadores que están en el tramo de sueldos líquidos adyacente que va entre $200.001 y $500.000, tienen 1,24 veces menos probabilidades de tener un accidente laboral, manteniendo todo lo demás constante. Con respecto a la variable de deuda, las dos categorías que resultaron significativas "Tengo deudas, y tengo permanentes y graves dificultades para pagarlas" y "Tengo deudas, y tengo siempre dificultades para pagarlas" tienen 1,39 y 1,18 veces, respectivamente, más probabilidades de tener un accidente laboral que aquellos trabajadores que declararon no tener deudas. Por último, la variable asociada a las empresas donde se realizó el cuestionario, la tasa de ausencia laboral considerando otros diagnósticos, tuvo un *Odds Ratio* estimado de 1,09.

El último modelo que estimamos, (modelo 3) de la Tabla Nº5, considera todas las variables de los modelos 1 y 2 y agregamos las cinco variables que continúan en importancia según el criterio MDA. Estas variables son: días de licencias médicas reportadas por el trabajador en el cuestionario SUSESO/ISTAS 21, las tasas de ausencia laboral anual asociadas a diagnósticos mentales y osteomusculares, la Clasificación Internacional Industrial Uniforme (CIIU) del centro de trabajo y la relación laboral que tienen los trabajadores con la empresa o institución.

En el modelo 3 la variable tiempo de antigüedad en la empresa, el coeficiente estimado para aquellos trabajadores que llevan entre 5 y 10 años deja de ser significativo, no

obstante, el coeficiente estimado para aquellos trabajadores que llevan entre 2 y 5 años pasa a tener una significancia marginal. Por su parte, la variable tasa de ausencia laboral asociada a otros diagnósticos deja de ser significativa. El resto mantienen su significancia estadística. De las nuevas variables incorporadas, todas tienen significancia estadística, con la excepción de la tasa de ausencia laboral asociada a diagnósticos osteomusculares, y los sectores económicos EGA y minería. Con respecto a la variable número de días de licencias médicas reportadas por los trabajadores, todos los tramos tienen una mayor probabilidad de tener un accidente laboral comparado con aquellos trabajadores que declararon no tener días de licencias médicas en los últimos doce meses. Los tramos que presentaron un mayor *Odds Ratio* fueron "Más de 90 días" y "Entre 61 y 90 días" con 1,88 y 1,81 veces, respectivamente, mayor probabilidad de tener un accidente laboral comparado con aquellas personas con cero días de licencias médica.

Por su parte, las variables asociadas a las empresas, los *Odds Ratio* estimados de las tasas de ausencia laboral de diagnósticos mentales y osteomusculares fueron cercanos a 1, por lo cual nos refleja que existe poca relación estadística con la ocurrencia de los accidentes laborales. En la siguiente variable, CIIU, consideramos como categoría base de comparación el sector de Agricultura y Pesca. Los sectores económicos de Comercio; Industria Manufacturera; y Transporte y Comunicaciones tuvieron los mayores *Odds Ratios* estimados, lo que significa que tienen 1,79; 1,66; y 1,64 veces, respectivamente, más probabilidades de tener accidentes laborales que aquellos trabajadores que pertenecen al sector económico de Agricultura y Pesca. Por último, la variable relación laboral, que tiene como base de comparación aquellos trabajadores que tienen contrato indefinido tuvieron un *Odds Ratio* de 1,26 para aquellos trabajadores con contrato temporal y 1,49 para los que no tienen contrato o tienen otro tipo de contrato.

Tabla Nº5. Resultados estimación Regresión Logística con base de datos balanceada

	Variable dependiente: Accidente laboral		
	(Modelo 1)	(Modelo 2)	(Modelo 3)
Escala de estrés de Setterlind - Riesgo alto	5,92*** (5,56; 6,30)	5,65*** (5,31; 6,02)	5,49*** (5,15; 5,85)
Salud mental – "Caso"	4,26*** (4,01; 4,52)	4,21*** (3,96; 4,48)	4,11*** (3,86; 4,37)
Salud general - "Caso"	5,08*** (4,78; 5,40)	4,75*** (4,47; 5,06)	4,63*** (4,35; 4,93)
Vitalidad - "Caso"	4,78*** (4,49; 5,09)	4,66*** (4,37; 4,97)	4,55*** (4,26; 4,85)
Menor a 26 años - Categoría base			
Entre 26 y 35 años	0,69*** (0,66; 0,73)	0,74*** (0,71; 0,78)	0,76*** (0,72; 0,80)
Entre 36 y 45 años	0,64*** (0,61; 0,67)	0,69*** (0,65; 0,73)	0,70*** (0,66; 0,74)
Entre 46 y 55 años	0,70*** (0,66; 0,74)	0,73*** (0,69; 0,78)	0,75*** (0,71; 0,79)
Más de 55 años	0,70*** (0,66; 0,74)	0,73*** (0,68; 0,78)	0,75*** (0,70; 0,80)
Directores y gerentes - Categoría base			
Agricultores y trabajadores calificados agropecuarios, forestales y pesqueros		2,82*** (1,56; 5,06)	3,45*** (1,84; 6,41)
Ocupaciones elementales		2,17*** (1,94; 2,44)	2,08*** (1,85; 2,34)
Oficiales, operarios y artesanos de artes mecánicas y de otros oficios		2,24*** (1,99; 2,53)	2,19*** (1,93; 2,48)
Operadores de instalaciones y máquinas y ensambladores		2,50*** (2,25; 2,80)	2,16*** (1,93; 2,43)
Personal de apoyo administrativo		1,35*** (1,22; 1,50)	1,31*** (1,18; 1,46)
Profesionales, científicos e intelectuales		1,41*** (1,27; 1,56)	1,39*** (1,25; 1,54)
Técnicos y profesionales de nivel medio		1,85*** (1,66; 2,06)	1,81*** (1,63; 2,03)
Trabajadores de los servicios y vendedores de comercios y mercados		1,95*** (1,76; 2,16)	1,85*** (1,66; 2,06)

Menos de 6 meses - Categoría base			
Entre 6 meses y 2 años en la empresa		0,97 (0,92; 1,03)	0,99 (0,93; 1,06)
Entre 2 años y 5 años en la empresa		1,05 (0,99; 1,11)	1,05* (0,99; 1,12)
Entre 5 años y 10 años en la empresa		0,94* (0,89; 1,00)	0,96 (0,90; 1,02)
Más de 10 años en la empresa		0,96 (0,91; 1,02)	1,01 (0,95; 1,08)
Menos de $200.000 de sueldo líquido - Categoría base			
Entre $200.001 y $500.000 de sueldo líquido		0,81*** (0,75; 0,87)	0,84*** (0,78; 0,90)
Entre $500.001 y $800.000 de sueldo líquido		0,75*** (0,70; 0,81)	0,81*** (0,75; 0,87)
Entre $800.000 y $1.000.000 de sueldo líquido		0,75*** (0,69; 0,81)	0,80*** (0,74; 0,87)
Entre $1.000.001 y $2.000.000 de sueldo líquido		0,66*** (0,61; 0,71)	0,71*** (0,65; 0,77)
Más de $2.000.000 de sueldo líquido		0,65*** (0,59; 0,73)	0,74*** (0,66; 0,83)
No tiene deuda - Categoría base			
Tengo deudas, pero no tengo dificultades para pagarlas		0,97 (0,92; 1,03)	0,96 (0,91; 1,02)
Tengo deudas, y tengo ocasionales dificultades para pagarlas		1,01 (0,95; 1,07)	0,97 (0,92; 1,03)
Tengo deudas, y tengo permanentes y graves dificultades para pagarlas		1,39*** (1,28; 1,51)	1,31*** (1,20; 1,42)
Tengo deudas, y tengo siempre dificultades para pagarlas		1,18*** (1,10; 1,26)	1,12*** (1,05; 1,20)
Tasa de ausencia - Otros diagnósticos		1,09*** (1,08; 1,11)	1,02 (1,00; 1,05)
Sin licencia médica - Categoría base			
Entre 1 y 7 días de licencia médica			1,48*** (1,41; 1,54)
Entre 8 y 14 días de licencia médica			1,62*** (1,53; 1,73)
Entre 15 y 30 días de licencia médica			1,58*** (1,50; 1,66)
Entre 31 y 60 días de licencia médica			1,65*** (1,52; 1,80)
Entre 61 y 90 días de licencia médica			1,81*** (1,58; 2,07)
Más de 90 días de licencia médica			1,88*** (1,64; 2,15)
Tasa de ausencia - Diagnósticos mentales			1,05*** (1,03; 1,07)
Tasa de ausencia - Diagnósticos osteomusculares			1,03 (0,99; 1,07)
Agricultura - Categoría base			
Comercio			1,79*** (1,47; 2,18)
Construcción			1,58*** (1,18; 2,12)
EGA			0,83 (0,60; 1,15)
Industria Manufacturera			1,66*** (1,35; 2,05)
Minería			0,89 (0,57; 1,38)
Servicios			1,30*** (1,07; 1,58)
Transporte y Comunicaciones			1,64*** (1,34; 2,00)
Relación laboral - Indefinido - Categoría base			
Relación laboral - Temporal			1,26*** (1,22; 1,31)
Relación laboral - Otro o sin contrato			1,49*** (1,29; 1,72)
Constante	0,71*** (0,68; 0,74)	0,39*** (0,3; 0,45)	0,21*** (0,17; 0,27)
Observaciones	90.306	90.306	90.306

Nota: *p<0,1; **p<0,05; ***p<0,01

Área bajo la curva (Curvas ROC)

Por último, comparamos el rendimiento de las estimaciones de los modelos utilizados; el *Random Forest* y la regresión logística utilizando la base de datos original (no balanceada) y la base de datos balanceada. Con esta comparación podremos conocer la capacidad de predicción de estos modelos. A continuación, se presenta la Tabla Nº5 con las principales estadísticas:

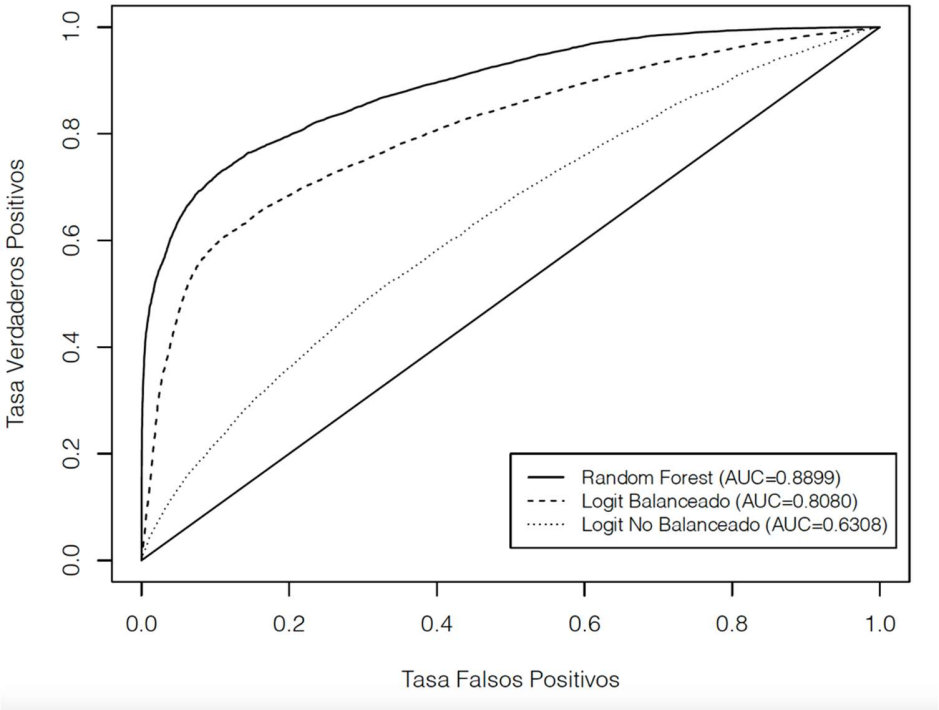
Tabla Nº5. Estadísticas predictivas modelos utilizados.

Estadísticas	Logit No Balanceado	Logit Balanceado	Random Forest
Exactitud	0,8688	0,7470	0,8084
Sensibilidad	0,0063	0,6362	0,7012
Especificidad	0,9992	0,8565	0,9144
Área bajo la curva ROC	0,6308	0,8080	0,8899

Como se puede observar la sensibilidad de los modelos, que representa los casos correctamente predichos con accidente laboral (verdaderos positivos) y la especificidad, que representa los casos correctamente predichos sin accidente laboral (verdaderos negativos), mejora sustancialmente al utilizar técnicas de balanceo de las bases de datos, incluso se puede observar que en el modelo *Logit No Balanceado* la sensibilidad es cercano a cero y la especificidad cercano a uno, lo que refleja que las predicciones se concentran fuertemente en la categoría mayoritaria que en este caso es no haber sufrido de un accidente laboral durante los últimos doce meses.

Con respecto a la exactitud, que mide el valor predicho correctamente con respecto al total de casos, el *Logit No Balanceado* está sesgado por lo comentado anteriormente, y los otros dos métodos tienen valores similares. No obstante lo anterior, el método de *Random Forest* tiene mejor capacidad predictiva que la regresión logística. Lo mismo ocurre con las estadísticas de exactitud y sensibilidad. A continuación, se presenta la Figura 4 con las Curvas ROC y sus áreas bajo la curva de las tres estimaciones:

Figura 4: Curvas ROC



Como se puede observar en la Figura 4, el área bajo la curva de la estimación por *Random Forest* fue de 0,8899 que fue el modelo que tuvo mejor rendimiento de los tres utilizados, lo cual nos muestra que en conjunto con una herramienta de balanceo de base datos tiene una calidad superior para predecir los accidentes laborales.

Discusión

La posibilidad de buscar un modelo capaz de predecir los factores de riesgo para un accidente laboral puede indicar vías de control de esos factores, si están bien identificados.

La manera tradicional de establecer los riesgos ocupacionales, cuando lo que se modela es una variable dicotómica, como tener o no un accidente del trabajo, suele ser con modelos del tipo regresión logística, que son comprensibles en sus resultados, pero tienen baja capacidad predictiva cuando existe una gran dimensionalidad (existen muchas variables explicativas).

En este documento se explora un modelo de aprendizaje automatizado como una forma para mejorar la capacidad predictiva. El problema de estos modelos es que, aunque tienen mayor capacidad predictiva, no son fácilmente comprensibles. No obstante, procuramos buscar una solución más bien práctica, que consiste en establecer un buen modelo predictivo con un algoritmo de *Random Forest*, y utilizar sus resultados para alimentar un segundo algoritmo de regresión logística.

La variable que se busca predecir es tener un accidente del trabajo y buscar la asociación con diversas características personales del trabajador (como salud mental, estrés, edad, haber tenido reposo médico por diversas enfermedades, entre otras) y variables vinculadas al trabajo mismo (como ocupación, antigüedad en el empleo, sueldo, entre otras). Adicionalmente, buscamos establecer una asociación con variables de riesgo psicosocial en el trabajo.

Con este procedimiento, identificamos que las variables que más fuertemente se asocian a una alta probabilidad de tener un accidente laboral son de tipo personal, y entre ellas la variable que se asoció con mayor fuerza (*Odds Ratio*=5,49, Intervalo de Confianza: 5,15 – 5,85) [3] fue el alto estrés, manteniendo su lugar predominante en todas las variantes que utilizamos (*MDA* y *MDG* de *Random Forest*). No cabe duda de que tener un alto estrés, tiene una fuerte asociación con tener un accidente laboral, de manera que habría que considerar qué factores son los que determinan con mayor fuerza esta variable, que posiblemente se comporte como un mediador entre otros factores.

Esto nos indica algo relevante, en el sentido de que la auto percepción que tienen los trabajadores sobre su salud y bienestar (salud general y mental, vitalidad y estrés) es un aspecto importante a la hora de predecir accidentes laborales, por lo que entrega un insumo relevante para las empresas e instituciones cuando crean mecanismos de prevención de accidentes en sus organizaciones.

Por otro lado, al analizar la importancia de las variables utilizando *Random Forest*, todas las subdimensiones de riesgo psicosocial se concentraron en la parte baja del ranking de importancia y no mostraron una asociación consistente con la probabilidad de tener un accidente laboral, lo que también ha sido observado en algunos estudios [43]. Esto se puede observar en el Anexo Nº2, utilizando la base de datos balanceada, los *Odds*

[3] Resultado obtenido en el modelo 3 de la Tabla Nº5.

Ratios de las subdimensiones son todos cercanos a uno, por lo que refleja que no hay mayores diferencias en probabilidades de tener accidentes laborales, si los trabajadores tienen o no riesgo alto en alguna de las subdimensiones.

Es probable que los factores de riesgo psicosocial no ejerzan su influencia directamente sobre la probabilidad de tener un accidente, sino que a través del estrés como mediador, incluso a través de la autopercepción de salud general y mental. Esta es una hipótesis que habría que explorar.

Por último, al analizar el rendimiento predictivo de los modelos utilizados, proponemos utilizar *Random Forest* como herramienta ya que presenta mejores resultados de predicción que la regresión logística. De igual manera es relevante evaluar balancear la base de datos ya que si existe el problema de que la variable categórica de estudio está fuertemente concentrada en una categoría, *Random Forest* presentará el mismo problema que la regresión logística.

Conclusión

El presente estudio tiene por objetivo encontrar los principales determinantes de los accidentes laborales sufridos por trabajadores cubiertos por la ley de accidentes N°16.744. Se utilizó la base de datos del cuestionario SUSESO/ISTAS 21 en su versión completa para los años 2017 y 2018.

Se aplicaron distintas herramientas predictivas utilizando los modelos de aprendizaje automatizado para la regresión logística y *Random Forest* usando una muestra balanceada y no balanceada. El fin de balancear la base de datos es solucionar un problema frecuente que ocurre cuando la variable de estudio o interés es categórica y una proporción importante de alguna de sus categorías se encuentra fuertemente concentrada. En este caso la variable de interés es la ocurrencia de un accidente laboral (excluyendo accidentes de trayecto) durante los últimos doce meses declarado por los trabajadores que respondieron el cuestionario, donde un 86,75% de ellos declararon no haber tenido accidentes. Al intentar predecir la accidentabilidad a través de una regresión logística o *Random Forest* sin balancear la base de datos, los resultados fueron pobres ya que las predicciones se concentraron en la clase mayoritaria.

Para solucionar este problema, se utilizó el método para balancear la base de datos denominado *Synthetic Minority Oversampling* (SMOTE), donde se sobre muestrea la categoría minoritaria generando observaciones "sintéticas". Se definieron los parámetros del balanceo de manera tal que el número de casos con accidentes laborales sea idéntico a los que no tuvieron accidentes.

Al utilizar una base de datos balanceada la capacidad predictiva tanto de la regresión logística como *Random Forest* mejoran notoriamente, donde este último presenta un mejor rendimiento tanto en exactitud, sensibilidad y especificidad.

Random Forest, además de tener mejor capacidad predictiva, puede construir medidas que reflejan la importancia de las variables, a través de la reducción media de exactitud (MDA) y la reducción media del índice Gini (MDG). El MDA permite saber que variables tienen mayor poder predictivo y el MDG mide la probabilidad de clasificar mal una observación. Se utilizó el MDA para poder cuantificar la relación existente entre la ocurrencia de un accidente laboral y las variables más importantes entregadas por este criterio.

Se encuentra que las variables más importantes para explicar los accidentes laborales están relacionadas con la auto percepción en la salud de los trabajadores como son la salud general, salud mental, vitalidad y, sobre todo, el estrés.

Adicionalmente existen variables asociadas al trabajo que también juegan un rol importante como es el salario líquido, la clasificación en ocupaciones y actividad económica, la relación laboral, tiempo o antigüedad en la empresa, entre otras adicionales.

Otro hallazgo importante fue que las variables de riesgo psicosocial son menos relevantes a la hora de medir la importancia en los modelos utilizados, por lo que podría significar que estos no ejercen su influencia directamente en la probabilidad de tener

un accidente laboral, sino más bien sean un posible antecedente del estrés y se refleja en la percepción de salud de los trabajadores.

Anexo Nº1: "Distribuciones de las variables explicativas por ocurrencia de accidentes laborales"

	Sin Accidentes	Accidente	Test de diferencia de medias	
			Mean Diff	Test t
N	**98.505**	**15.051**		
Sexo				
Mujeres	0,51	0,49	0,0213***	[4,86]
Hombres	0,49	0,51	-0,0213***	[-4,86]
Edad				
Menor a 26 años	0,11	0,14	-0,0386***	[-14,13]
Entre 26-35 años	0,33	0,33	-0,0000542	[-0,01]
Entre 36-45 años	0,27	0,24	0,0326***	[8,42]
Entre 46-55 años	0,19	0,19	0,000221	[0,07]
Más de 55 años	0,11	0,11	0,00588*	[2,14]
Ocupación				
Agricultores y trabajadores calificados agropecuarios, forestales y pesqueros	0,00	0,00	-0,0000101	[-0,04]
Directores y gerentes	0,04	0,02	0,0184***	[11,18]
Ocupaciones elementales	0,05	0,06	-0,0127***	[-6,58]
Oficiales, operarios y artesanos de artes mecánicas y de otros oficios	0,04	0,05	-0,0117***	[-6,76]
Operadores de instalaciones y máquinas y ensambladores	0,08	0,13	-0,0486***	[-19,43]
Personal de apoyo administrativo	0,20	0,17	0,0384***	[11,03]
Profesionales, científicos e intelectuales	0,30	0,23	0,0730***	[18,41]
Trabajadores de los servicios y vendedores de comercios y mercados	0,19	0,24	-0,0500***	[-14,31]
Técnicos y profesionales de nivel medio	0,09	0,10	-0,00684**	[-2,70]
Antigüedad laboral				
Menos de 6 meses	0,09	0,08	0,0134***	[5,31]
Entre 6 meses y 2 años en la empresa	0,18	0,18	-0,00299	[-0,90]
Entre 2 años y 5 años en la empresa	0,24	0,27	-0,0237***	[-6,28]
Entre 5 años y 10 años en la empresa	0,22	0,22	0,000674	[0,19]
Más de 10 años en la empresa	0,27	0,26	0,0126**	[3,24]
Relación Laboral				
Contrato indefinido	0,70	0,74	-0,0333***	[-8,38]
Contrato temporal	0,29	0,25	0,0342***	[8,69]
Otro tipo de contrato	0,01	0,01	-0,000845	[-0,97]
Jornada				
Diurno	0,78	0,74	0,0386***	[10,50]
Nocturno	0,02	0,02	-0,00108	[-0,96]
Rotatorio	0,21	0,25	-0,0375***	[-10,48]
Trabaja en más de una sección en la empresa				
Sí	0,26	0,29	-0,0327***	[-8,48]
No	0,74	0,71	0,0327***	[8,48]
Remuneración				
Menos de $200.000 de sueldo líquido	0,04	0,05	-0,0101***	[-5,59]
Entre $200.001 y $500.000 de sueldo líquido	0,38	0,42	-0,0410***	[-9,63]
Entre $500.001 y $800.000 de sueldo líquido	0,26	0,29	-0,0233***	[-6,02]
Entre $800.000 y $1.000.000 de sueldo líquido	0,12	0,11	0,0105***	[3,67]
Entre $1.000.001 y $2.000.000 de sueldo líquido	0,15	0,11	0,0461***	[14,93]

Más de $2.000.000 de sueldo líquido	0,04	0,02	0,0178***	[10,63]
Situación de deuda				
No tiene deuda	0,08	0,07	0,00800***	[3,39]
Tiene deudas, pero no tiene dificultades para pagarlas	0,41	0,37	0,0329***	[7,66]
Tiene deudas y tiene ocasionales dificultades para pagarlas	0,36	0,37	-0,0123**	[-2,94]
Tiene deudas y tiene permanentes y graves dificultades para pagarlas	0,05	0,06	-0,0102***	[-5,47]
Tiene deudas y tiene siempre dificultades para pagarlas	0,11	0,13	-0,0183***	[-6,65]
Número de días licencias médicas en los últimos 12 meses (ISTAS)				
Sin licencia médica	0,05	0,06	0,0852***	[20,34]
Entre 0 y 7 días de licencia médica	0,11	0,13	-0,0153***	[-4,82]
Entre 8 y 14 días de licencia médica	0,65	0,57	-0,0163***	[-7,92]
Entre 15 y 30 días de licencia médica	0,15	0,17	-0,0291***	[-11,43]
Entre 31 y 60 días de licencia médica	0,06	0,07	-0,0135***	[-8,99]
Entre 61 y 90 días de licencia médica	0,09	0,12	-0,00474***	[-5,38]
Más de 90 días de licencia médica	0,03	0,04	-0,00639***	[-7,09]
Salud y bienestar personal				
Percepción general de salud				
En riesgo alto	49%	55%	-0,0439***	[-26,43]
Sin riesgo alto	51%	45%	0,0439***	[26,43]
Percepción de salud mental				
En riesgo alto	3%	8%	-0,0351***	[-21,13]
Sin riesgo alto	97%	92%	0,0351***	[21,13]
Percepción de la vitalidad				
En riesgo alto	3%	7%	-0,0316***	[-19,98]
Sin riesgo alto	97%	93%	0,0316***	[19,98]
Percepción de salud asociado al stress				
En riesgo alto	3%	6%	-0,0497***	[-30,16]
Sin riesgo alto	97%	94%	0,0497***	[30,16]
Actividad Económica del centro de trabajo				
Agricultura	0,01	0,01	0,00384***	[4,43]
Comercio	0,21	0,23	-0,0196***	[-5,52]
Construcción	0,01	0,00	0,000838	[1,35]
EGA	0,01	0,00	0,00371***	[5,52]
Industria	0,03	0,03	-0,00092	[-0,62]
Minería	0,00	0,00	0,00141***	[3,48]
Servicios	0,62	0,57	0,0564***	[13,26]
Transporte	0,12	0,16	-0,0457***	[-15,82]
Indicador de ausencia laboral por licencia médica de origen común (714 empleadores - 90 sin accidentes - 624 con accidentes)				
Tasa de ausencia - Diagnósticos mentales	1,64	1,85	-0,216	[-1,23]
Tasa de ausencia - Diagnósticos osteomusculares	0,97	0,97	-0,00827	[-0,09]
Tasa de ausencia - Otros diagnósticos	2,09	2,36	-0,269	[-1,72]

	Sin accidente	Con accidente	Test de diferencia de medias	
			Mean Diff	Test t
****** Dimensión Exigencias psicológicas**				
Subdimensión Exigencias psicológicas cuantitativas (CU)				
Riesgo Alto	0,38	0,44	-0,0518***	[-12,15]
Riesgo Bajo/Medio	0,62	0,57	0,0518***	[12,15]
Subdimensión Exigencias psicológicas cognitivas (CO)				
Riesgo Alto	0,31	0,34	-0,0321***	[-7,93]
Riesgo Bajo/Medio	0,70	0,66	0,0321***	[7,93]
Subdimensión Exigencias psicológicas emocionales (EM)				
Riesgo Alto	0,57	0,65	-0,0737***	[-17,09]
Riesgo Bajo/Medio	0,43	0,35	0,0737***	[17,09]
Subdimensión Exigencias psicológicas de esconder emociones (EE)				
Riesgo Alto	0,53	0,60	-0,0749***	[-17,17]
Riesgo Bajo/Medio	0,47	0,40	0,0749***	[17,17]
Subdimensión Exigencias psicológicas sensoriales (ES)				
Riesgo Alto	0,53	0,55	-0,0222***	[-5,09]
Riesgo Bajo/Medio	0,47	0,45	0,0222***	[5,09]
****** Dimensión Trabajo activo y desarrollo de habilidades (D2)**				
Subdimensión Influencia (IN)				
Riesgo Alto	0,58	0,63	-0,0505***	[-11,75]
Riesgo Bajo/Medio	0,42	0,37	0,0505***	[11,75]
Subdimensión Posibilidades de desarrollo en el trabajo (PD)				
Riesgo Alto	0,52	0,58	-0,0579***	[-13,26]
Riesgo Bajo/Medio	0,48	0,42	0,0579***	[13,26]
Subdimensión Control sobre los tiempos de trabajo (CT)				
Riesgo Alto	0,40	0,45	-0,0550***	[-12,81]
Riesgo Bajo/Medio	0,60	0,55	0,0550***	[12,81]
Subdimensión Sentido del trabajo (ST)				
Riesgo Alto	0,39	0,42	-0,0354***	[-8,29]
Riesgo Bajo/Medio	0,61	0,58	0,0354***	[8,29]
Subdimensión Integración en la empresa (IE)				
Riesgo Alto	0,24	0,28	-0,0417***	[-11,13]
Riesgo Bajo/Medio	0,76	0,72	0,0417***	[11,13]
****** Dimensión Apoyo social en la empresa y calidad de liderazgo (D3)**				
Subdimensión Claridad de rol (RL)				
Riesgo Alto	0,55	0,59	-0,0399***	[-9,18]
Riesgo Bajo/Medio	0,45	0,41	0,0399***	[9,18]
Subdimensión Conflicto de rol (CR)				
Riesgo Alto	0,33	0,40	-0,0723***	[-17,51]
Riesgo Bajo/Medio	0,67	0,60	0,0723***	[17,51]
Subdimensión Calidad de liderazgo (CL)				
Riesgo Alto	0,38	0,46	-0,0714***	[-16,72]
Riesgo Bajo/Medio	0,62	0,55	0,0714***	[16,72]

Subdimensión Calidad de la relación con superiores (RS)				
Riesgo Alto	0,44	0,52	-0,0769***	[-17,68]
Riesgo Bajo/Medio	0,56	0,48	0,0769***	[17,68]
Subdimensión Calidad de la relación con sus compañeros/as de trabajo (RC)				
Riesgo Alto	0,41	0,46	-0,0580***	[-13,47]
Riesgo Bajo/Medio	0,59	0,54	0,0580***	[13,47]
****** Dimensión Compensaciones (D4)**				
Subdimensión Estima (ET)				
Riesgo Alto	0,49	0,58	-0,0836***	[-19,13]
Riesgo Bajo/Medio	0,51	0,42	0,0836***	[19,13]
Subdimensión Inseguridad respecto a las condiciones generales del contrato (IC)				
Riesgo Alto	0,28	0,31	-0,0315***	[-7,99]
Riesgo Bajo/Medio	0,72	0,69	0,0315***	[7,99]
Subdimensión Inseguridad respecto a las características específicas del trabajo (IT)				
Riesgo Alto	0,38	0,45	-0,0695***	[-16,35]
Riesgo Bajo/Medio	0,62	0,56	0,0695***	[16,35]
****** Dimensión Doble presencia (D5)**				
Subdimensión Preocupación por tareas domésticas (DP)				
Riesgo Alto	0,49	0,55	-0,0557***	[-12,73]
Riesgo Bajo/Medio	0,51	0,45	0,0557***	[12,73]

Anexo Nº2: "Estimación Regresión Logística utilizando base de datos balanceada y no balanceada"

	Variable dependiente: Accidente laboral	
	(BBDD no balanceada)	(BBDD balanceada)
Escala de estrés de Setterlind - Riesgo alto	1,88*** (1,73, 2,04)	5,14*** (4,83; 5,48)
Salud mental - Riesgo alto	1,21*** (1,10; 1,32)	3,86*** (3,63; 4,10)
Salud general - Riesgo alto	1,70*** (1,58; 1,84)	4,46*** (4,19; 4,75)
Vitalidad - Riesgo alto	1,22*** (1,11; 1,34)	4,33*** (4,06; 4,62)
Menor a 26 años - Categoría base		
Entre 26 y 35 años	0,78*** (0,74; 0,83)	0,77*** (0,73; 0,81)
Entre 36 y 45 años	0,70*** (0,66; 0,75)	0,72*** (0,68; 0,76)
Entre 46 y 55 años	0,77*** (0,72; 0,83)	0,77*** (0,73; 0,82)
Más de 55 años	0,74*** (0,68; 0,79)	0,78*** (0,73; 0,84)
Directores y gerentes - Categoría base		
Agricultores y trabajadores calificados agropecuarios, forestales y pesqueros	2,84*** (1,36; 5,46)	3,17*** (1,69; 5,94)
Ocupaciones elementales	2,02*** (1,76; 2,32)	2,05*** (1,82; 2,31)
Oficiales, operarios y artesanos de artes mecánicas y de otros oficios	2,29*** (1,97; 2,66)	2,16*** (1,90; 2,45)
Operadores de instalaciones y maquinas y ensambladores	2,17*** (1,87; 2,53)	2,16*** (1,92; 2,43)
Personal de apoyo administrativo	1,31*** (1,16; 1,48)	1,32*** (1,18; 1,46)
Profesionales, científicos e intelectuales	1,31*** (1,16; 1,48)	1,39*** (1,26; 1,55)
Técnicos y profesionales de nivel medio	1,67*** (1,46; 1,90)	1,83*** (1,63; 2,04)
Trabajadores de los servicios y vendedores de comercios y mercados	1,78*** (1,57; 2,02)	1,79*** (1,61; 1,99)
Menos de 6 meses - Categoría base		
Entre 6 meses y 2 años en la empresa	1,10** (1,02; 1,19)	0,97 (0,91; 1,03)
Entre 2 años y 5 años en la empresa	1,14*** (1,06; 1,23)	1,01 (0,95; 1,08)
Entre 5 años y 10 años en la empresa	1,05 (0,97; 1,14)	0,93** (0,87; 0,99)
Más de 10 años en la empresa	1,10** (1,02; 1,20)	0,99 (0,93; 1,06)
Menos de $200.000 de sueldo líquido - Categoría base		
Entre $200.001 y $500.000 de sueldo líquido	0,93* (0,85; 1,01)	0,85*** (0,79; 0,91)
Entre $500.001 y $800.000 de sueldo líquido	0,86*** (0,78; 0,94)	0,81*** (0,75; 0,87)
Entre $800.000 y $1.000.000 de sueldo líquido	0,81*** (0,73; 0,89)	0,80*** (0,73; 0,87)
Entre $1.000.001 y $2.000.000 de sueldo líquido	0,72*** (0,64; 0,79)	0,71*** (0,65; 0,77)
Más de $2.000.000 de sueldo líquido	0,67*** (0,58; 0,78)	0,72*** (0,64; 0,81)
No tengo deudas - Categoría base		
Tengo deudas, pero no tengo dificultades para pagarlas	1,06 (0,98; 1,14)	0,97 (0,91; 1,03)
Tengo deudas, y tengo ocasionales dificultades para pagarlas	1,03 (0,96; 1,11)	0,94** (0,88; 1,00)
Tengo deudas, y tengo permanentes y graves dificultades para pagarlas	0,99 (0,90; 1,10)	1,14*** (1,04; 1,24)
Tengo deudas, y tengo siempre dificultades para pagarlas	1,04 (0,95; 1,13)	1,02 (0,95; 1,09)
Tasa de ausencia - Otros diagnósticos	1,03* (1,00; 1,06)	1,03** (1,00; 1,06)
Sin licencia médica - Categoría base		
Entre 1 y 7 días de licencia médica - ISTAS	1,29*** (1,23; 1,36)	1,45*** (1,39; 1,52)
Entre 8 y 14 días de licencia médica - ISTAS	1,40*** (1,30; 1,50)	1,58*** (1,48; 1,68)
Entre 15 y 30 días de licencia médica - ISTAS	1,36*** (1,28; 1,44)	1,52*** (1,45; 1,60)
Entre 31 y 60 días de licencia médica - ISTAS	1,46*** (1,33; 1,60)	1,60*** (1,47; 1,74)
Entre 61 y 90 días de licencia médica - ISTAS	1,44*** (1,23; 1,67)	1,73*** (1,51; 1,98)
Más de 90 días de licencia médica - ISTAS	1,53*** (1,32; 1,77)	1,86*** (1,62; 2,14)
Tasa de ausencia - Diagnósticos mentales	1,03*** (1,01; 1,06)	1,04*** (1,02; 1,06)
Tasa de ausencia - Diagnósticos osteomusculares	1,02 (0,97; 1,06)	1,00 (0,96; 1,04)
Agricultura - Categoría base		
Comercio	1,51*** (1,20; 1,91)	1,69*** (1,39; 2,07)
Construcción	1,31 (0,93; 1,85)	1,50*** (1,12; 2,02)
EGA	0,85 (0,57; 1,25)	0,79 (0,56; 1,09)

Industria Manufacturera	1,46*** (1,15; 1,87)	1,55*** (1,26; 1,92)
Minería	0,81 (0,43; 1,41)	0,82 (0,52; 1,28)
Servicios	1,34** (1,07; 1,70)	1,24** (1,02; 1,51)
Transporte y Comunicaciones	1,34** (1,06; 1,71)	1,55*** (1,27; 1,90)
Relación laboral - Indefinido - Categoría base		
Relación laboral - Temporal	1,01 (0,96; 1,06)	1,23*** (1,18; 1,28)
Relación laboral - Otro o sin contrato	1,11 (0,93; 1,31)	1,35*** (1,16; 1,56)
Subdimensión Claridad de rol - Riesgo alto	0,96* (0,92; 1,00)	0,90*** (0,87; 0,93)
Subdimensión Influencia - Riesgo alto	1,02 (0,98; 1,07)	0,90*** (0,87; 0,93)
Subdimensión Exigencias psicológicas de esconder emociones - Riesgo alto	1,06*** (1,02; 1,11)	0,97* (0,93; 1,00)
Subdimensión Control sobre los tiempos de trabajo - Riesgo alto	1,01 (0,97; 1,05)	1,10*** (1,07; 1,14)
Sexo - Hombre	0,98 (0,94; 1,02)	1,00 (0,96; 1,03)
Subdimensión Estima - Riesgo alto	1,04* (0,99; 1,09)	0,92*** (0,89; 0,96)
Subdimensión Calidad de la relación con sus compañeros/as de trabajo - Riesgo alto	1,04* (1,00; 1,08)	1,11*** (1,07; 1,15)
Subdimensión Exigencias psicológicas emocionales - Riesgo alto	1,07*** (1,02; 1,12)	0,90*** (0,86; 0,93)
Subdimensión Posibilidades de desarrollo en el trabajo - Riesgo alto	1,04* (1,00; 1,08)	0,97* (0,94; 1,00)
Subdimensión Preocupación por tareas domésticas - Riesgo alto	1,02 (0,99; 1,06)	1,01 (0,98; 1,04)
Subdimensión Sentido del trabajo - Riesgo alto	0,98 (0,94; 1,03)	1,07*** (1,04; 1,11)
Subdimensión Exigencias psicológicas sensoriales - Riesgo alto	0,97* (0,93; 1,01)	0,97 (0,94; 1,01)
Subdimensión Exigencias psicológicas cuantitativas - Riesgo alto	1,08*** (1,03; 1,12)	1,12*** (1,08; 1,16)
Subdimensión Calidad de liderazgo - Riesgo alto	1,02 (0,98; 1,08)	1,09*** (1,05; 1,13)
Horario de trabajo - Diurno - Categoría base		
Horario de trabajo - Nocturno	0,91 (0,79; 1,04)	1,12** (1,00; 1,26)
Horario de trabajo - Rotatorio	1,00 (0,96; 1,04)	1,25*** (1,20; 1,29)
Subdimensión Calidad de la relación con superiores - Riesgo alto	1,03 (0,98; 1,08)	1,00 (0,96; 1,04)
Dos o más secciones - Si	1,15*** (1,10; 1,19)	1,40*** (1,35; 1,45)
Subdimensión Conflicto de rol - Riesgo medio	1,04* (0,99; 1,08)	1,14*** (1,10; 1,18)
Subdimensión Inseguridad respecto a las características específicas del trabajo - Riesgo alto	1,04* (1,00; 1,09)	1,10*** (1,06; 1,14)
Subdimensión Integración en la empresa - Riesgo alto	0,99 (0,94; 1,03)	1,20*** (1,16; 1,25)
Subdimensión Inseguridad respecto a las condiciones generales del contrato - Riesgo alto	1,01 (0,97; 1,06)	1,25*** (1,21; 1,30)
Subdimensión Exigencias psicológicas cognitivas - Riesgo alto	1,11*** (1,06; 1,16)	1,33*** (1,28; 1,38)
Constante	0,05*** (0,04; 0,07)	0,17*** (0,13; 0,21)
Observaciones	113.556	90.306

Nota: *p<0,1; **p<0,05; ***p<0,01

Anexo Nº3: "Definiciones dimensiones y subdimensiones Riesgo Psicosocial"

D1) Las subdimensiones de la Dimensión de Exigencias Psicológicas: esta dimensión considera aspectos cualitativos y cuantitativos desde el punto de vista del trabajador, asociados principalmente al volumen y presión del trabajo, ritmo de la tarea y los tiempos para realizarla. Todas las subdimensiones son variables categóricas que miden el nivel de riesgo de los trabajadores y se clasifican en riesgo bajo, riesgo medio y riesgo alto.

1. *Exigencias psicológicas cuantitativas (CU):* mide la cantidad o volumen de trabajo exigido contrastado con el tiempo disponible para realizarlo.
2. *Exigencias psicológicas cognitivas (CO):* mide las exigencias sobre diferentes procesos mentales (atención, memoria, decisiones) y responsabilidad por las consecuencias de lo que se hace.
3. *Exigencias psicológicas emocionales (EM):* mide si el trabajador logra mantenerse emocionalmente distante de la tarea, sobre todo cuando hay que relacionarse a nivel personal con los usuarios que también se expresan emocionalmente, y ante lo que el trabajador puede reaccionar con agobio o desgaste emocional.
4. *Exigencias psicológicas de esconder emociones (EE):* mide la demanda de ocultamiento de las emociones que surgen en el transcurso del trabajo, normalmente ante la atención de personas.
5. *Exigencias psicológicas sensoriales (ES):* mide las exigencias laborales que significan utilizar los sentidos, en especial la visión, con una alta atención y alerta a los detalles.

D2) Las subdimensiones de la Dimensión Trabajo Activo y Desarrollo de Habilidades: esta dimensión considera la autonomía del trabajador en cuanto a la posibilidad de influir en sus horarios laborales, ritmos, métodos, variedad, iniciativa y calidad. Todas las subdimensiones son variables categóricas que miden el nivel de riesgo de los trabajadores y se clasifican en riesgo bajo, riesgo medio y riesgo alto.

1. *Influencia (IN):* mide el margen de decisión o autonomía respecto al contenido y las condiciones de trabajo (secuencia de la tarea, métodos a utilizar, tareas a realizar, cantidad de trabajo, horarios, elección de compañeros).
2. *Posibilidades de desarrollo en el trabajo (PD):* evalúa si el trabajo es fuente de oportunidades de desarrollo de las habilidades y conocimientos de cada persona.
3. *Control sobre los tiempos de trabajo (CT):* mide aspectos como la posibilidad de pausar o interrumpir momentáneamente la tarea, sea para

un descanso breve, por atender obligaciones personales o para tomar vacaciones.

4. *Sentido del trabajo (ST):* considera atributos más allá de los fines simplemente instrumentales que puede tener el trabajo (estar ocupado y obtener a cambio unos ingresos económicos), el sentido del trabajo consiste en relacionales con otros valores o fines trascendentes.

5. *Integración en la empresa (IE):* mide el grado de identificación de cada persona con la empresa o institución en general.

D3) Las subdimensiones de la Dimensión Apoyo Social en la Empresa y Calidad de Liderazgo: esta dimensión considera aspectos de apoyo social y de liderazgo. Todas las subdimensiones son variables categóricas que miden el nivel de riesgo de los trabajadores y se clasifican en riesgo bajo, riesgo medio y riesgo alto.

1. *Claridad de rol (RL):* mide el grado de definición de las acciones y responsabilidades del puesto de trabajo.

2. *Conflicto de rol (CR):* mide las exigencias contradictorias que se presentan en el trabajo y que pueden generar conflictos de carácter profesional o ético, cuando las exigencias de lo que hay que hacer son diferentes de las normas y valores personales.

3. *Calidad de liderazgo (CL):* evalúa conductas y atributos del jefe o supervisor directo que permiten juzgar su valor como líder.

4. *Calidad de la relación con superiores (RS):* mide los atributos tanto del jefe directo como de la organización en general que posibilita recibir el tipo de ayuda e información que se necesita y en el momento adecuado para realizar el trabajo.

5. *Calidad de la relación con sus compañeros/as de trabajo (RC):* mide las relaciones con los compañeros de trabajo que se expresan tanto en formas de comunicación como en la posibilidad de recibir el tipo de ayuda para realizar el trabajo en el momento adecuado, así como el sentido de pertenencia a un equipo.

D4) Las subdimensiones de la Dimensión Compensaciones: esta dimensión evalúa el desbalance que puede existir entre el esfuerzo y la recompensa asociada, así como también el control de estatus (estabilidad del empleo, cambios no deseados). Todas las subdimensiones son variables categóricas que miden el nivel de riesgo de los trabajadores y se clasifican en riesgo bajo, riesgo medio y riesgo alto.

1. *Estima (ET):* mide el reconocimiento y apoyo de los superiores y compañeros por el esfuerzo realizado para desempeñar el trabajo.

2. *Inseguridad respecto a las condiciones generales del contrato (IC):* mide la preocupación por las condiciones del contrato, estabilidad o renovación,

variaciones del sueldo, formas de pago del sueldo, posibilidades de despido y ascenso.

3. *Inseguridad respecto a las características específicas del trabajo (IT):* mide la inseguridad sobre las condiciones de trabajo tales como movilidad funcional (cambios de tareas) y geográfica, cambios de la jornada y horario de trabajo.

D5) Las subdimensiones de la Dimensión Doble Presencia: esta dimensión mide la preocupación por cumplir con las tareas domésticas, además de las tareas propias del trabajo. Todas las subdimensiones son variables categóricas que miden el nivel de riesgo de los trabajadores y se clasifican en riesgo bajo, riesgo medio y riesgo alto.

1. *Preocupación por tareas domésticas (DP):* mide la intranquilidad provocada por las exigencias domésticas que puedan afectar el desempeño laboral.

Referencias

[1] Hämäläinen P, Takala J, Kiat TB. Global Estimates of Occupational Accidents and Work-related Illnesses 2017. WSH Institute. Ministry of Health and Social Affairs, Finland. Singapore: Workplace Safety & Health Institute; 2017.

[2] Nenonen N, Saarela KL, Takala J, Kheng LG, Yong E, Ling LS, Manickam K, Hämäläinen P. Global estimates of occupational accidents and fatal work-related diseases in 2014. Tampere University. WSH Institute, VTT. Singapore: Workplace Safety & Health Institute; 2014.

[3] Superintendencia de Seguridad Social (SUSESO). (2019a). Estadísticas de la seguridad social 2019. https://www.suseso.cl/608/w3-article-592655.html; consultado el 12/oct/2020

[4] Lu ML, Nakata A, Park JB, Swanson NG. Workplace psychosocial factors associated with work-related injury absence: a study from a nationally representative sample of Korean workers, *Int J Behav Med*. 2014; 21: 42-52.

[5] Amick BC, McDonough P, Chang H, et al. (2002), Relationship between all-cause mortality and cumulative working life course psychosocial and physical exposures in the United States labor market from 1968 to 1992, *Psychosom Med*. 2002; 64(3): 370-81.

[6] Ramos AK, Carlo G, Grant K, Trinidad N, Correa A. Stress, Depression, and Occupational Injury among Migrant Farmworkers in Nebraska. Safety. 2016;2(4):23.

[7] Niedhammer I, Chastang JF, David S. Importance of psychosocial work factors on general health outcomes in the national French SUMER survey. Occup Med. 2008; 58: 15-24.

[8] Swaen GMH, van Amelsvoort LGPM, Bültmann U, Kant IJ. Fatigue as a risk factor for being injured in an occupational accident: results from the Maastricht Cohort Study. Occup Environ Med. 2003;60(Suppl 1): i88-i92.

[9] Lee SJ. Psychosocial work factors in new or recurrent injuries among hospital workers: a prospective study. Int Arch Occup Environm Health. 2015;88(8): 1141-1148.

[10] Salminen S, Kivimäki M, Elovainio M, Vahtera J. Stress factors predicting injuries of hospital personnel. Am J Ind Med. 2003;44: 32-36.

[11] Nakata A, Ikeda T, Takahashi M, Haratami T, Hojou M, Fujioka Y, Swanson NG, Araki S. Impact of Psychosocial Job Stress on Non-Fatal Occupational Injuries in Small and Medium-Sized Manufacturing Enterprises. Am J Ind Med. 2006; 49:658-669.

[12] Johannessen HA, Gravseth HM, Sterud T. Psychosocial factors at work and occupational injuries: A prospective study of the general working population in Norway. Am J Ind Med. 2015;58: 561-567.

[13] Julià M, Catalina-Romero C, Calvo-Bonacho E, Benavides FG. Exposure to psychosocial risk factors at work and the incidence of occupational injuries: A cohort study in Spain. JOEM. 2016;58: 282-286.

[14] Halbesleben JRB. The Role of Exhaustion and Workarounds in Predicting Occupational Injuries: A Cross-Lagged Panel Study of Health Care Professionals. J Occup Health Psychol. 2010;15(1): 1-16.

[15] Maturana R, Soto M, El-Far S, La Rivere C, Vega J, Cumsille MA. Factores psicosociales relacionados con accidentes del trabajo. Rev Chil Neuro-Psiquiat. 1994;32: 278-284.

[16] Rebolledo P. Accidentes Ocupacionales: Aspectos Psicosociales. Cienc trab. 2005;7(16): 61-66.

[17] Bravo C., Nazar G. Riesgo psicosocial en el trabajo y salud en conductores de locomoción colectiva urbana en Chile. Salud trab (Maracay). 2015;23(2): 105-114.

[18] Ibáñez JA. Influencia de fatiga laboral, riesgos psicosociales y conflicto trabajo-familia en la accidentabilidad de trabajadores forestales [tesis]. Concepción: Universidad de Concepción; 2016.

[19] Silva H, Lefio A, Marchetti N, Benoit P, 2014. Riesgos Psicosociales en Conductores de Transporte de Carga y Pasajeros Urbanos e Interurbanos, y su Asociación con la Autopercepción de Salud y Siniestralidad Laboral. Cienc Trab. May-Ago; 16 [50]: 67-74).

[20] Seguel K, Navarrete E, Bahamondes G, 2017. Explicación de la Accidentabilidad Laboral Basada en Factores de Riesgo Psicosocial y Rasgos de Personalidad en el Transporte Forestal. Cienc Trab. Sep-Dic; 19 [60]: 157-165).

[21] Alvarado R, Pérez-Franco J, Saavedra N, Fuentealba C, Alarcón A, Marchetti N, Aranda W. Validación de un cuestionario para evaluar riesgos psicosociales en el ambiente laboral en Chile. Rev Med Chile. 2012;140: 1154-63.

[22] Superintendencia de Seguridad Social (SUSESO). Redactado por M. Candia y J. Pérez-Franco. Manual del Método SUSESO/ISTAS21, 3a edición. Santiago: Superintendencia de Seguridad Social; noviembre, 2020. Disponible en: https://www.suseso.cl/606/w3-article-19640.html.

[23] Ministerio de Salud (MINSAL). Protocolo de Vigilancia de Riesgos Psicosociales en el Trabajo. Santiago: MINSAL; 2013; 2017.

[24] Ellis PD. The Essential Guide to Effect Sizes. Cambridge, UK: Cambridge University Press; 2010.

[25] Rudin, Cynthia. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* **1,** 206–215 (2019). https://doi.org/10.1038/s42256-019-0048-x

[26] Olivares P. Perfil del estado de salud de beneficiarios de ISAPREs: Informe preliminar. Documento de Trabajo. Marzo 2005. Superintendencia de ISAPREs. Santiago: Ministerio de Salud; 2005.

[27] Sperandei S. Understanding logistic regression analysis. Biochem Med (Zagreb). 2014;24(1):12–18. Published 2014 Feb 15. doi:10.11613/BM.2014.003.

[28] Breinman L. Random Forest. Machine Learning. 2001;45:5-32.

[29] Kirasich K, Smith T, Sadler B. Random Forest vs Logistic Regression: Binary Classification for Heterogeneus Datasets. SMUData Science Review 1(3), Articule 9. https://scholar.smu.edu/datasciencereview/vol1/iss3/9.

[30] Couronné, Raphael; Probst, Philipp; Boulesteix, Anne-Laure. Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics. 2018; 19:270; https://doi.org/10.1186/s12859-018-2264-5

[31] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. Springer, 2017.

[32] Weiss GM, Provost F (2001) The effect of class distribution on classifier learning: An empirical study. Department of Computer Science.

[33] Laurikkala J (2001) Improving identification of difficult small classes by balancing class distribution. 63–66.

[34] Estabrooks A, Jo T, Japkowicz N (2004) A multiple resampling method for learning from imbalanced data sets. Computational Intelligence 20: 18–36.

[35] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16:321-357.

[36] Silverman BW, Jones MC. E. Fix and Hodges JL. (1951): An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation. Int Stat Review, 1989, 57(3): 233 – 247.

[37] Altman NS. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. Am Stat. 1992; 46(3):175 – 185

[38] Liaw, Andy; Wiener, Matthew. Classification and Regression by randomForest. R 2002; News 2(3):18-22.

[39] Torgo, L. (2010). Data Mining with R, learning with case studies Chapman and Hall/CRC. URL: http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR

[40] R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

[41] RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

[42] Sing T, Sander O, Beerenwinkel N, Lengauer T (2005). ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**(20), 7881. http://rocr.bioinf.mpi-sb.mpg.de

[43] Osca, Amparo; López-Araujo, Blanca. Work stress, personality, and occupational accidents: Should we expect differences between men and women? *Safety Science*. 2020; 124. https://doi.org/10.1016/j.ssci.2019.104582.