Revista EIA

UNIVERSIDAD EIA

Predicción del Ausentismo Laboral por Incapacidad: Una Aproximación desde los Modelos Lineales Generalizados

Laura Margarita Guerrero Guerra[1]
Nicolás Moreno[1]
✉ Juan Daniel Molina[2]

1. Universidad EAFIT, Colombia
2. Institución Universitaria ITM, Colombia

Revista EIA
ISSN 1794-1237
e-ISSN 2463-0950
Año XIX/ Volumen 22/ Edición N.44
Julio - diciembre 2025
Reia4431 pp. 1-27

Publicación científica semestral
Universidad EIA, Envigado, Colombia

Para citar este artículo / To reference this article /
Guerrero Guerra, L. M.; Moreno, N. y Molina, J. D.
Predicción del Ausentismo Laboral por Incapacidad: Una Aproximación desde los Modelos Lineales Generalizados

Revista EIA, 22(44), Reia4431 pp. 1-27
https://doi.org/10.24050/reia.v22i43.1890

✉ *Autor de correspondencia:*
Molina, J. D.
Ingeniería Industrial, Doctorado en Ciencias - Estadística
Correo electrónico:
juanmolina1192@correo.itm.edu.co

Recibido: 30-04-2025
Aceptado: 20-06-2025
Disponible online: 01-07-2025

Resumen

El ausentismo laboral representa un desafío crítico para la gestión del talento humano, con implicaciones económicas relevantes, efectos negativos en la reputación organizacional y riesgos para la competitividad empresarial. Este artículo examina los factores que inciden en el ausentismo laboral, con un enfoque en la predicción de dos aspectos fundamentales: el número esperado de días de ausencia de un empleado en los próximos tres meses y la probabilidad de que registre al menos un día de ausencia en ese mismo período. Para ello, se emplean modelos de conteo y de respuesta binaria, enmarcados en los modelos lineales generalizados, logrando predicciones de alta precisión de la ocurrencia de incapacidades laborales, a nivel individual, en una institución financiera colombiana. Además, desde las dos perspectivas de modelación se encontró que los factores que mayor influencia tienen sobre la ocurrencia de incapacidades laborales son el mes del año, el sexo, el tipo de contrato, el área de trabajo y el histórico de la cantidad de meses en los que el empleado ya había presentado incapacidad.

Palabras clave: Incapacidad laboral, regresión, clasificación, predicción, modelos lineales generalizados, regresión logística, regresión de Poisson, métricas de desempeño, gestión del talento humano, gestión del riesgo.

Prediction of Labor Absenteeism due to Incapacity: A Generalized Linear Model Approach

Abstract

Workplace absenteeism poses a critical challenge for human talent management, with significant economic implications, negative impacts on organizational reputation, and risks to business competitiveness. This article examines the factors influencing absenteeism, focusing on the prediction of two key aspects: the expected number of absence days for an employee over the next three months and the probability of registering at least one absence day during the same period. To achieve this, count and binary response models framed within generalized linear models are employed, yielding highly accurate predictions of the occurrence of work incapacities, at an individual level, in a Colombian financial institution. In addition, from both modeling perspectives, it was found that the factors that have the greatest influence on the occurrence of work incapacities are the month of the year, sex, type of contract, work area, and the history of the number of months in which the employee had already presented incapacity.

Keywords: Work incapacity, regression, classification, prediction, generalized linear models, logistic regression, Poisson regression, performance metrics, human resource management, risk management.

1. Introducción

Las organizaciones se enfrentan a una batalla continua para reducir los costos y al mismo tiempo aumentar la productividad, situación en la cual un aspecto que a menudo se pasa por alto es la gestión de ausencias (Navarro y Bass, 2006; Sanchez, 2015). Sin embargo, el impacto del ausentismo puede conducir, no solo a pérdidas financieras, sino también a la pérdida del buen nombre de la organización (Raja y Gupta, 2019). Varios estudios han demostrado que los costos directos e indirectos del ausentismo pueden ascender a un porcentaje importante de la nómina (Navarro y Bass, 2006; Kocakulah et al., 2016), y en el caso específico del ausentismo por enfermedad, el costo puede llegar a ser mayor que el salario pagado al empleado enfermo que no se presenta al lugar de trabajo donde,

sin sustitutos disponibles, las implicaciones para la productividad son mucho mayores (Martiniano et al., 2012; Borda et al., 2017; Berg et al., 2008). Debido a lo anteriormente expuesto, las organizaciones necesitan entender el fenómeno del ausentismo e identificar los factores que inciden en él, con el fin de establecer acciones que contribuyan a disminuirlo (Araujo et al., 2019; Daza y Perez, 1997).

La Organización Internacional del Trabajo (OIT) define al ausentismo como la inasistencia al trabajo por parte de un empleado que se pensaba iba a asistir, excluyendo periodos vacacionales, huelgas o permisos sindicales; y refiere al ausentismo laboral de causa médica a todo aquel periodo de inasistencia atribuible a una incapacidad del individuo, exceptuando las licencias de maternidad o las atribuibles a prisión (ILO, 2020; ICONTEC, 1996). Son diversos los trabajos que se han llevado a cabo con relación a la predicción del ausentismo laboral. Gomes y Lopes (2022) presentan una metodología para clasificar el ausentismo a través de redes neuronales y agregación de propagación de relevancia por capas (LRP), con el fin de identificar las características más relevantes y asignar puntuaciones de relevancia por clase y entre todas las clases. El enfoque propuesto presentó las tasas de asertividad más altas entre los métodos comparados (Decision Tree, Gradient Boosted Tree, Random Forest y Tree Ensemble), con una precisión promedio de 0.83, identificando las características más relevantes para la clasificación del ausentismo, dentro de las cuales se destacaron la ocurrencia de consultas médicas, el mes de la ausencia, si el individuo es fumador social, la ocurrencia de consultas dentales, el cumplimiento de la meta por alcanzar a nivel laboral, entre otras. Trabajos similares a este se presentan en Montano et al. (2020), Qaisar (2020), Ali-Shah et al. (2020), Coussement y Van den Poel (2008) y Vafeiadis et al. (2015).

En otro estudio realizado para un proveedor belga de servicios de recursos humanos (RRHH) y bienestar (Lawrance et al., 2021), se describe un sistema de apoyo para la toma de decisiones, cuyo objetivo es identificar grupos de empleados con riesgo de presentar ausencia por enfermedad, a quienes luego se les pueden dirigir intervenciones destinadas a reducir o prevenir las ausencias. Modelan el ausentismo como un problema de clasificación binaria

con asimetría de pérdidas y conceptualizan una matriz de costos de clasificación errónea de las ausencias por enfermedad de los empleados. El algoritmo base utilizado es un clasificador de árbol de decisión, a partir del cual adoptan una amplia gama de ensambles de árboles de decisión, haciendo uso de datos reales de recursos humanos y nómina que contienen predictores no relacionados con la salud (características demográficas, características del ambiente de trabajo, patrones de ausencia histórica).

Con la participación de los empleados de una compañía fabricante de automóviles alemana se realizó el estudio que presentan Fischer et al. (2020), cuyo objetivo es estimar la fracción excesiva de ausencia potencialmente evitable atribuible a las características psicosociales del grupo de trabajo. Se predicen las tasas de ausencia por enfermedad de grupos de trabajo durante 12 meses, utilizando datos provenientes de un examen de salud integral de referencia (evalúa las características laborales, el comportamiento de la salud y los factores de riesgo biomédicos), y aplicando modelos lineales generalizados con una función de enlace logit binomial. Un modelo de predicción de 7 características psicosociales a nivel de grupo de trabajo explicó el 70% de la variación de las tasas de ausencia futuras, a partir de lo cual se concluye que las características psicosociales a nivel del grupo de trabajo representan una proporción importante de todas las ausencias por enfermedad, de manera que las intervenciones de promoción de la salud deberían abordar dichas características.

El estudio desarrollado por Svärd et al. (2024) buscaba establecer la posible influencia del sobrepeso/obesidad y las condiciones físicas y mentales del trabajo sobre las ausencias por enfermedad, tanto de corta duración (1 a 7 días) como de larga duración (mayores a 8 días), entre los empleados finlandeses de mediana edad. A partir de datos relacionados con el índice de masa corporal y las condiciones laborales, utilizan modelos de regresión binomial negativa para calcular la razón de tasas (RRs por sus siglas en inglés) y el intervalo de confianza del 95% para los periodos de ausencia por enfermedad, concluyendo que la combinación de sobrepeso/obesidad y las condiciones de trabajo extenuantes contribuyen a las ausencias por enfermedad y deben

 https://doi.org/10.24050/reia.v22i44.1890

ser consideradas a la hora de buscar estrategias para reducirlas. De manera similar, Salonsalmi et al. (2023) exploran la posible relación entre la educación parental, las adversidades vividas durante la niñez y las ausencias por enfermedad entre empleados de mediana edad, a través de la implementación de un modelo de regresión de Poisson. Con base en el cálculo de la razón de tasas y los intervalos de confianza del 95%, se concluye que un nivel bajo de educación parental y las adversidades vividas durante la infancia (dificultades económicas, divorcio de los padres, alcoholismo en los padres, etc.) sí contribuyen a la generación de ausencias por enfermedad en la adultez. De manera que promover el bienestar de las familias con niños podría ayudar en la mitigación de esta situación.

La situación analizada en el marco de este artículo corresponde al de una institución financiera colombiana, en la cual se busca predecir las incapacidades laborales de sus empleados a nivel individual, con el objetivo de brindar a la organización herramientas para gestionar y controlar este tipo de ausentismos, a partir de la identificación de variables que impacten su ocurrencia, lo que permitiría orientar los planes de intervención hacia aquellos aspectos que son relevantes. Se examinaron los factores que inciden en el ausentismo laboral, con un enfoque en la predicción de dos aspectos fundamentales: el número esperado de días de ausencia de un empleado en los próximos tres meses y la probabilidad de que registre al menos un día de ausencia en ese mismo período. Para ello, se emplearon modelos de conteo y de respuesta binaria, enmarcados en los modelos lineales generalizados, logrando predicciones de alta precisión y un análisis detallado de las variables asociadas al fenómeno.

El presente artículo está estructurado de la siguiente forma: en la sección 2 se presentan los materiales y métodos utilizados en el desarrollo del artículo, en la sección 3 se presentan los resultados y se interpreta el alcance e implicaciones de los mismos, y finalmente, en la sección 4 se presenta una discusión a partir de la metodología y resultados del artículo, además de las principales conclusiones que de éstas se desprenden.

2. Materiales y Métodos

Se recopilaron registros mensuales de 29 variables correspondientes a 3,000 empleados de una institución financiera en Colombia, abarcando el periodo de enero de 2022 a diciembre de 2023. Los datos incluyen el número de días no laborados, diferenciando entre ausencias por vacaciones y por incapacidad médica. A partir de esta información, se calculó el número total de días que un empleado se ausentó por incapacidad médica en un periodo de tres meses y se definió una variable binaria que indica si el empleado estuvo ausente al menos un día por incapacidad médica durante ese intervalo.

La metodología empleada en este estudio utiliza modelos de regresión aplicados a las variables mencionadas, con el objetivo de identificar los factores que inciden en las incapacidades laborales y predecir tanto el número de días no laborados como la probabilidad de que un empleado tome una incapacidad médica en los próximos tres meses. Específicamente, para modelar el número de días no laborados por incapacidad se ajustaron modelos de conteo, mientras que para la variable binaria se empleó un modelo de regresión logística.

Se evaluaron modelos de conteo, incluyendo la regresión de Poisson y la regresión binomial negativa. Debido a la alta concentración de ceros en los datos, ya que la mayoría de los empleados no solicitaban licencias médicas en este intervalo de tiempo, también se probaron las variantes infladas en ceros de ambos modelos (Gil-Bellosta, 2018). Por otro lado, se evaluó modelos de regresión binaria con penalización Lasso para disminuir el efecto de multicolinealidad entre las variables.

2.1. Modelos considerados

2.1.1 Regresión Poisson

La regresión Poisson está enmarcada en los modelos lineales generalizados y puede describirse de la siguiente manera, considere $X = (x'_1,x'_2,...,x'_n) \in R^{nxp}$ una matriz de diseño, $Y = (y_1,y_2,...,y_n) \in R^n$ el vector de respuestas de conteo y $\beta = (\beta_1,\beta_2,...,\beta_p) \in R^{1xp}$ un vector de

Revista EIA https://doi.org/10.24050/reia.v22i44.1890

parámetros. Así, en el modelo de regresión Poisson se considera es un vector de variables independientes con Y∼Poisson(λ), la media dada por (McCullagh, 2019; Isaac y Rivera, 2021)

$$\log(\lambda) = \log(E(Y|X)) = \beta'X.$$

En otras palabras, la media de la distribución Poisson está dada por

$$\lambda = E(Y|X) = e^{\beta'X}.$$

Dada la alta cantidad de ceros en la variable de respuesta también fue considerada la variante inflada en ceros de la regresión Poisson, la cual considera que el resultado de una realización y_i es cero con probabilidad p_i o es una variable aleatoria con distribución Poisson con parámetro λ_i=β'Xi. El parámetro p_i es, a su vez, obtenido por medio de una regresión probit, es decir,

$$p_i = \phi(\beta' x_i),$$

donde ϕ es la función de distribución de la distribución normal estándar.

2.1.2 Regresión logística

El modelo de regresión binaria utilizado en este trabajo es el modelo de regresión logística, en este modelo de regresión lineal se establece que Y∼Bernoulli(p), donde (McCullagh, 2019; James et al., 2013)

$$p = \frac{1}{1 + e^{\beta X}}$$

El rendimiento de los modelos de regresión para variables de conteo se evalúa con base en el error cuadrático medio, absoluto y mediano; para cuantificar la variabilidad explicada por los modelos propuestos se consideró el pseudo-R^2 y por último, las predicciones

fueron evaluadas utilizando el método de validación cruzada con un 70% de ajuste y un 30% para el testeo. En el caso del modelo de respuesta binaria se utilizaron los estadísticos de exactitud, precisión, recuperación y puntuación F1, también considerando un 70% de observaciones para el ajuste y un 30% para el testeo.

3. Resultados

En esta sección se presentan los principales resultados obtenidos tras aplicar la metodología propuesta en los materiales y métodos de este artículo. La tabla 1 presenta los factores y/o características que se evaluó si tiene algún tipo de impacto sobre el ausentismo laboral por incapacidad médica. Y la tabla 2 presenta la variable objetivo en dos escenarios, uno de clasificación, pare el modelo logístico, en el que se evalúo la probabilidad de que registre al menos un día de ausencia en los siguientes tres meses y otro de regresión, para el modelo Poisson, en el que se analizó el número esperado de días de ausencia de un empleado para el mismo período de tiempo.

Revista EIA https://doi.org/10.24050/reia.v22i44.1890

Tabla 1. *Descripción de características o factores de incidencia*

Características			
#	**Nombre**	**Tipo**	**Descripción**
1	Mes	Categórica	Mes al que corresponde el registro
2	Año	Categórica	Año al que corresponde el registro
3	Sexo	Categórica	Sexo del empleado
4	AñosEdad	Numérica	Edad del empleado
5	EstadoCivil	Categórica	Estado Civil del empleado (Soltero/Casado)
6	TipoContrato	Categórica	Tipo de contrato del empleado (Indefinido/Definido)
7	AñosAntiguedad	Numérica	Años de antigüedad en la organización
8	Cargo	Categórica	Cargo que ocupa en la organización (Auxiliar, Analista, Diseñador, etc.)
9	TipoCargo	Categórica	Tipología del cargo (Operativo/Profesional)
10	Area	Categórica	Área a la que pertenece el empleado (Productos, Canales, etc.)
11	Ciudad	Categórica	Ciudad de residencia (Medellín/ Bogotá/ Cali/ Barranquilla/ Otras)
12	Region	Categórica	Región en la que se encuentra (Caribe, Pacífico, etc.)
13	Ocupacion	Numérica	Porcentaje de ocupación del empleado durante el mes
14	Ausencia	Numérica	Porcentaje de ausencia del empleado durante el mes
15	Adherencia	Numérica	Porcentaje de adherencia durante el mes, donde la adherencia indica el esfuerzo (en términos de tiempo) realizado por el empleado para cumplir con sus responsabilidades.
16	FTE MenorAdherencia	Numérica	Menor esfuerzo realizado por el empleado, respecto a su jornada laboral. Unidad de medida: FTE (Full Time Equivalent).
17	FTE MayorAdherencia	Numérica	Mayor esfuerzo realizado por el empleado, respecto a su jornada laboral. Unidad de medida: FTE (Full Time Equivalent).
18	RiesgoPsicosocial	Categórica	Nivel de riesgo psicosocial del empleado según las mayores/menores adherencias vividas durante el mes (0: Óptimo, 1: Satisfactorio, 2: Tolerable, 3: Crítico, 4: Severo)
19	MesesPermanencia	Numérica	Número de meses consecutivos con riesgo psicosocial deteriorado (en nivel Tolerable, Crítico o Severo).
20	CondicionEspecial	Numérica	Porcentaje de condición especial del empleado, donde la condición especial se refiere a una condición de salud que impide que el colaborador trabaje con su capacidad al 100%.
21	DiasIncapacidad	Numérica	Número de días de incapacidad laboral durante el mes
22	DiasVacaciones	Numérica	Número de días en los que el empleado estuvo en vacaciones durante el mes
23	DiasOtrasAusencias	Numérica	Número de días en los que el empleado estuvo ausente por otros motivos.
24	VariacionIPC	Numérica	Variación del IPC correspondiente al mes (indicador macroeconómico, fuente: DANE)
25	TasaDesempleo	Numérica	Tasa de desempleo correspondiente al mes (indicador macroeconómico, fuente: DANE)
26	MesesCE	Numérica	Número de meses (durante el último cuatrimestre) en los que empleado tuvo condición especial
27	MesesIncapacidad	Numérica	Número de meses (durante el último cuatrimestre) en los que empleado presentó incapacidad laboral.
28	MesesAusencia	Numérica	Número de meses (durante el último cuatrimestre) en los que empleado tuvo ausencia, independiente del motivo.

Tabla 2. *Descripción de la variable objetivo*

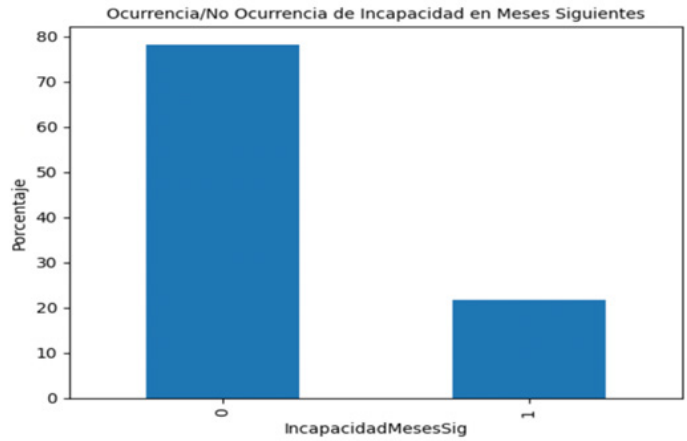
Variable Objetivo			
Enfoque	**Nombre**	**Tipo**	**Descripción**
Clasificación	IncapacidadMesesSig	Categórica binaria	Ocurrencia o no ocurrencia de incapacidad laboral en los 3 meses siguientes
Regresión	DiasIncapacidadMesesSig	Numérica discreta	Número de días de incapacidad laboral durante los 3 meses siguientes

3.1 Resultados asociados con el modelo logístico

Se realizó un análisis exploratorio de datos, basado en la comprensión de la variable objetivo y en la visualización de posibles relaciones/dependencias con las características, además de la identificación de altas correlaciones. Respecto al análisis asociado con el modelo de clasificación o logístico, se observó, como se evidencia en la Figura 1 que el comportamiento de la variable objetivo presenta un desbalance importante en las clases, dado que el 78% de los registros corresponden a la clase 0, es decir, a la No Ocurrencia de Incapacidad, mientras que solo el 22% de los registros corresponden a la clase 1, Ocurrencia de Incapacidad. Esta situación hace necesaria la aplicación de métodos de balanceo con el fin de gestionar el desequilibrio de las clases.

Revista EIA https://doi.org/10.24050/reia.v22i44.1890

Figura 1. *Participación de la Ocurrencia / No Ocurrencia de Incapacidad (variable objetivo)*

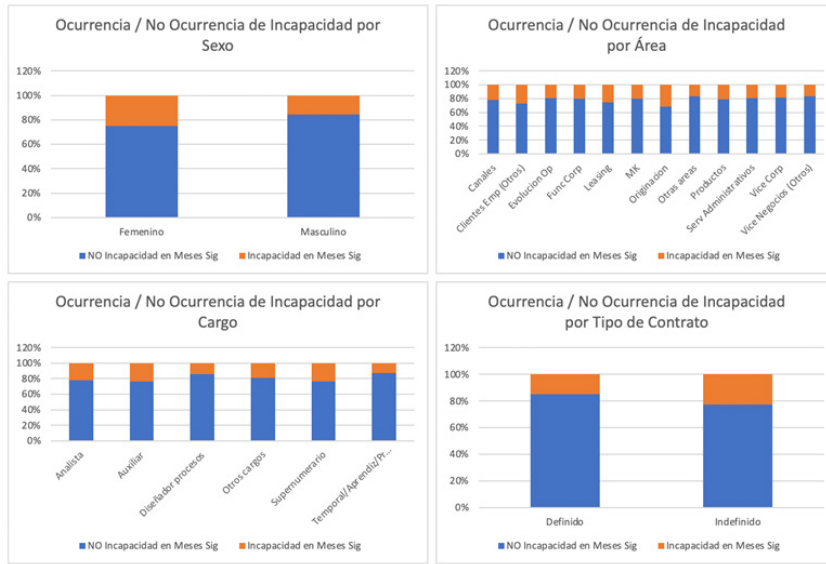


Adicionalmente, se exploraron posibles relaciones significativas entre las variables predictoras categóricas y la variable objetivo, a través de la prueba chi-cuadrado. Esta prueba permite analizar la relación entre variables categóricas. Se obtuvo un valor p menor a 0.05 para todas las características, a excepción de las variables Año y Tipo de Cargo, permitiendo rechazar la hipótesis de que las variables comparadas son independientes. En la Figura 2 se presenta el comportamiento de las variables: Sexo, Área, Cargo y Tipo de Contrato respecto a la ocurrencia o no de incapacidad laboral, evidenciando variaciones en las proporciones para cada categoría. Curiosamente, se observa que, en el caso del Cargo, los empleados con cargo de temporal/aprendiz/préstamo son los que menos se incapacitan, y respecto al Tipo de Contrato, los empleados con contrato a término definido se incapacitan menos que los de contrato a término indefinido.

En cuanto a la relación entre las variables predictoras numéricas y la variable objetivo, se aplicaron pruebas como la Mann-Whitney U. Esta es una prueba no paramétrica que puede detectar diferencias entre las medianas de dos grupos, para indagar sobre la igualdad o desigualdad de sus distribuciones. Se obtuvo valores p menores a 0.05 en todos los casos, excepto en el de la variable Meses de Permanencia. Los diagramas de caja de la Figura 3 evidencian la diferencia en las distribuciones de los grupos de las variables

Años Antigüedad, Ausencia, Meses Incapacidad y Meses Ausencia, definidos con base en la ocurrencia o no de incapacidad.

***Figura 2**. Características categóricas versus variable objetivo binaria*



En la Figura 4 se analiza la matriz de correlación de las variables numéricas, con el fin de verificar las correlaciones existentes entre ellas. Teniendo en cuenta que correlaciones muy altas pueden resultar en problemas de multicolinealidad. Entre la mayoría de las parejas de variables que se presentan una alta correlación se tiene también una relación o causalidad trivial, como es el caso de las variables Años de Antigüedad y Edad, Días de vacaciones y Ausencias, y FTE de Mayor Adherencia y Adherencia. El único par de variables que presentan una alta correlación, de 0.75, y no existe una relación trivial entre ellas son Riesgo Psicosocial y FTE Mayor Adherencia.

Para el entrenamiento o ajuste de los modelos, se exploraron diversos escenarios, en función de las variables predictoras, incluyendo el 100% de las variables versus excluyendo las variables que presentaron alta correlación con otras, el método de normalización, MinMaxScaler versus StandardScaler, el método de balanceo en el enfoque de clasificación, submuestreo, sobremuestreo, generación de muestras sintéticas con SMOTE-NC, la sintonización

Revista EIA https://doi.org/10.24050/reia.v22i44.1890

de hiperparámetros correspondientes a cada modelo y el umbral de decisión en el caso de la Regresión Logística. Además, el conjunto de datos se divide aleatoriamente en conjunto de entrenamiento y conjunto de prueba, donde para el entrenamiento se toma el 70% de los datos, y para la prueba, el 30% restante.

***Figura 3**. Características numéricas versus variable objetivo binaria*

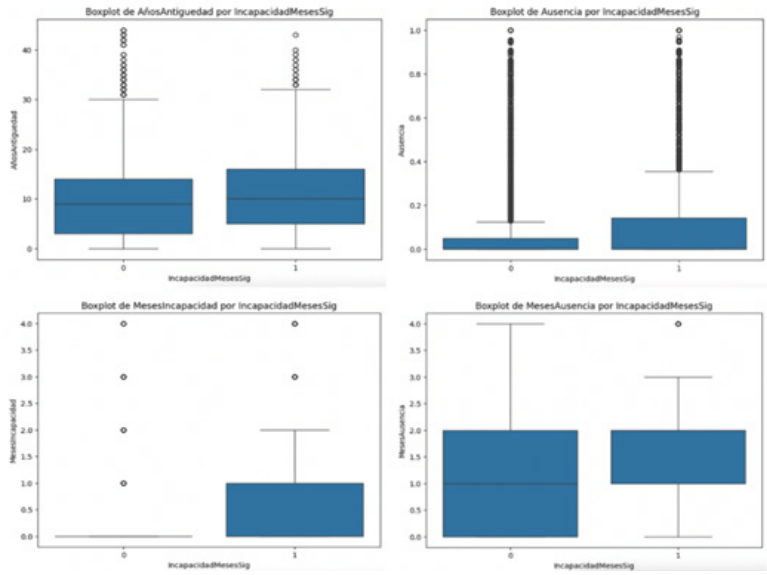
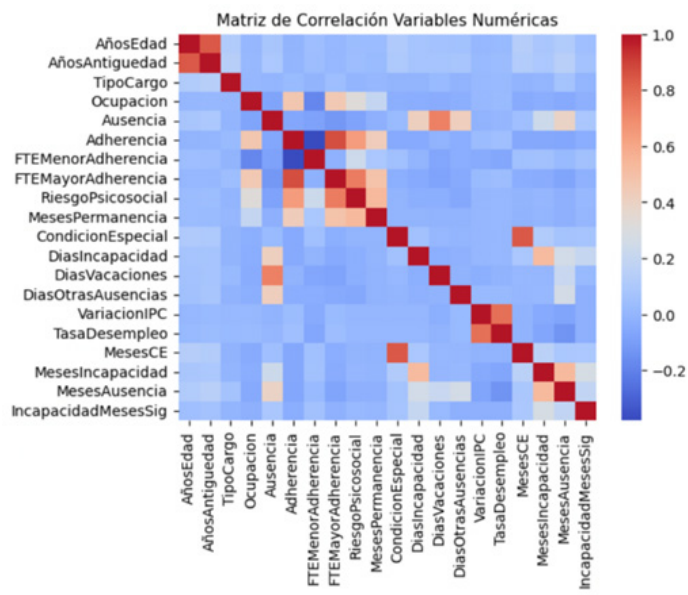


***Figura 4.** Matriz de correlación, enfoque clasificación*

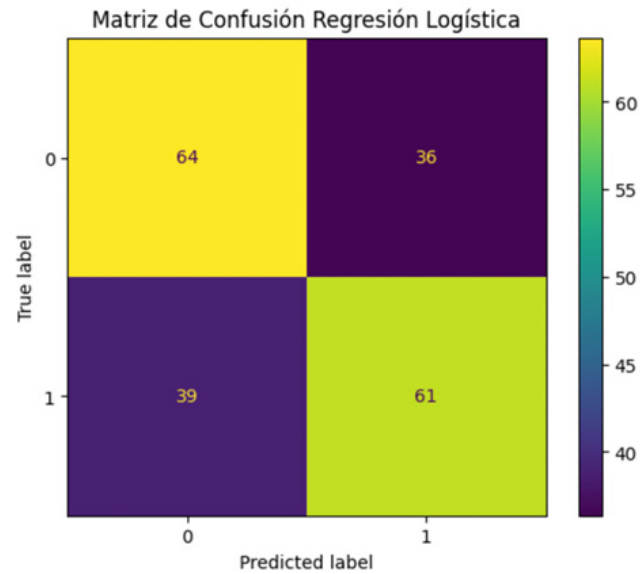


El mejor escenario obtenido para el modelo de clasificación o logístico se dio con la exclusión de las variables que presentaban alta correlación con otras variables predictoras, usando el método de normalización StandardScaler, el cual estandariza las características eliminando la media y escalando a la varianza unitaria, y con el método de balanceo Sobre-muestreo, igualando el número de muestras en las clases de la variable objetivo. Bajo este escenario las métricas de desempeño que el modelo obtuvo fueron: Accuracy del 63%, Precision del 32%, Recall del 61% y F1-Score del 42%. Es de recalcar que, dado que se gestionó el desbalanceo en los datos de manera previa, y que el interés principal es identificar en qué medida el modelo es capaz de clasificar como positivos los casos que realmente lo son, es decir, donde sí ocurre una incapacidad, se dio mayor importancia a los resultados de Accuracy y Recall. Respecto a estas métricas la Regresión Logística alcanza valores de 63% y 61%, respectivamente, que son valores adecuados. Además, la Regresión Logística tiene las ventajas que es un modelo simple, que suele entrenarse y entregar predicciones sin un gran costo computacional, y en especial, es un modelo naturalmente interpretable (MS, 2023), en el que los coeficientes para cada una de las variables pueden interpretarse directamente en términos de probabilidades, facilitando la comprensión de cómo cada característica afecta la probabilidad de un resultado.

En la Figura 5 se presenta la matriz de confusión, en la cual se evidencia que el 61% de los casos positivos reales, donde realmente hay incapacidad, fueron clasificados correctamente, y el 64% de los casos negativos reales, donde no hay incapacidad, fueron clasificados como tal, estos son los resultados correspondientes a la métrica Recall.

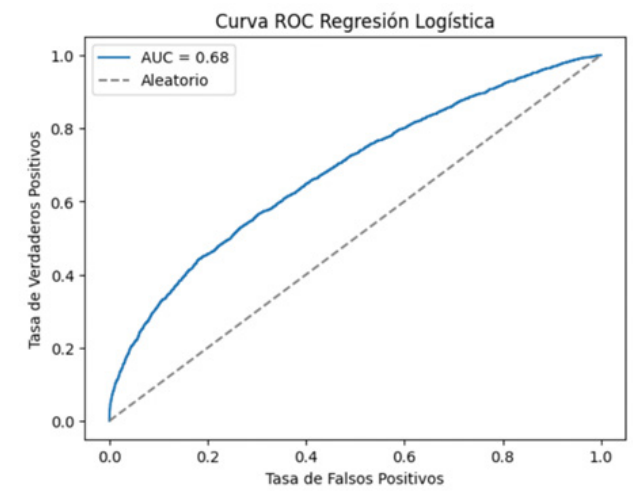
Revista EIA https://doi.org/10.24050/reia.v22i44.1890

Figura 5. *Matriz de confusión para Regresión Logística*



En la Figura 6 se encuentra la curva ROC, que es una herramienta gráfica utilizada para evaluar el rendimiento de un modelo de clasificación binaria, y el AUC (Área bajo la curva), a partir de los cuales se puede concluir que hay un 68% de probabilidad de que el modelo de Regresión Logística pueda distinguir entre la clase positiva y la clase negativa (Ocurrencia de Incapacidad versus No Ocurrencia de Incapacidad).

Figura 6*. Curva ROC y AUC para Regresión Logística*



Respecto al modelo de clasificación o logístico se encontró que, las variables con los coeficientes positivos de mayor magnitud, son los meses octubre y noviembre, resultantes de la codificación a variables binarias de la variable original Mes, y la variable numérica Meses de Incapacidad. Por otra parte, las variables con los coeficientes negativos de mayor magnitud son el mes de enero y el sexo Masculino, resultante de la codificación de la variable original Sexo. Teniendo en cuenta que se trata de variables significativas dentro del modelo y que tienen los coeficientes de mayor magnitud, se concluye que las variables mencionadas anteriormente son las que mayor impacto tienen sobre la ocurrencia o no de incapacidades laborales, donde el aumento en las variables con coeficiente positivo aumenta la probabilidad de ocurrencia de incapacidades laborales, mientras que el aumento en las variables con coeficiente negativo disminuye la probabilidad de ocurrencia de incapacidades.

Para determinar la magnitud en la que se aumenta o disminuye dicha probabilidad, dado que en la Regresión Logística los coeficientes representan el efecto de las variables sobre la probabilidad logarítmica de que ocurra el evento de interés, generalmente llamada “log-odds” o “logit”, es necesario convertir el coeficiente, tomando su exponente (McCullagh, 2019). En la Tabla 3 se realiza esta transformación y se presenta el efecto final que las variables más importantes tienen sobre la probabilidad de ocurrencia de las incapacidades laborales. De acuerdo con lo anteriormente expuesto, se puede interpretar que, durante el mes de enero, se reduce en un 61% las posibilidades que ocurran incapacidades, por el contrario, durante el mes de octubre las posibilidades de que ocurran incapacidades aumentan en un 125%, y en noviembre aumentan en un 98%. Además, los hombres presentan un 39% menos de posibilidad de incapacitarse que las mujeres. Y si se tiene que un aumento en un mes en el tiempo que una persona lleva incapacitada aumenta en un 54% la posibilidad que para el siguiente trimestre continúe incapacitada.

Revista EIA https://doi.org/10.24050/reia.v22i44.1890

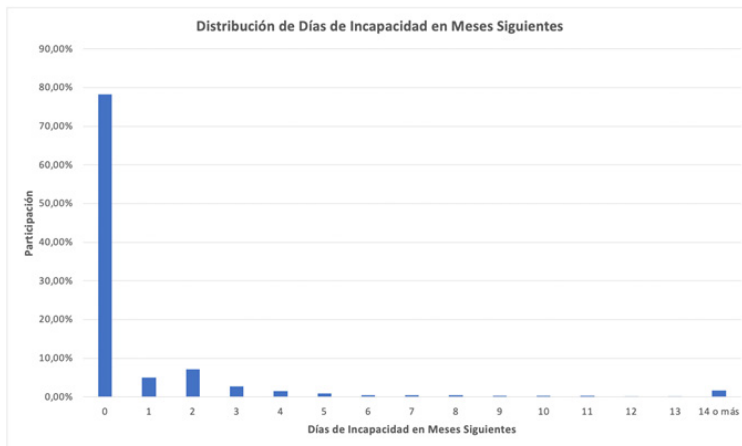
Tabla 3. *Efecto de características sobre probabilidad de ocurrencia de incapacidades*

Característica	Coeficiente	Transformación	Efecto sobre Probabilidad de Ocurrencia Incapacidad
(Mes) Enero	-0.93	$e^{-0,93}$	**0.39**
(Mes) Octubre	0.81	$e^{0,81}$	**2.25**
(Mes) Noviembre	0.68	$e^{0,68}$	**1.98**
(Sexo) Masculino	-0.50	$e^{-0,50}$	**0.61**
MesesIncapacidad	0.43	$e^{0,43}$	**1.54**

3.2 Resultados asociados con el modelo de regresión Poisson

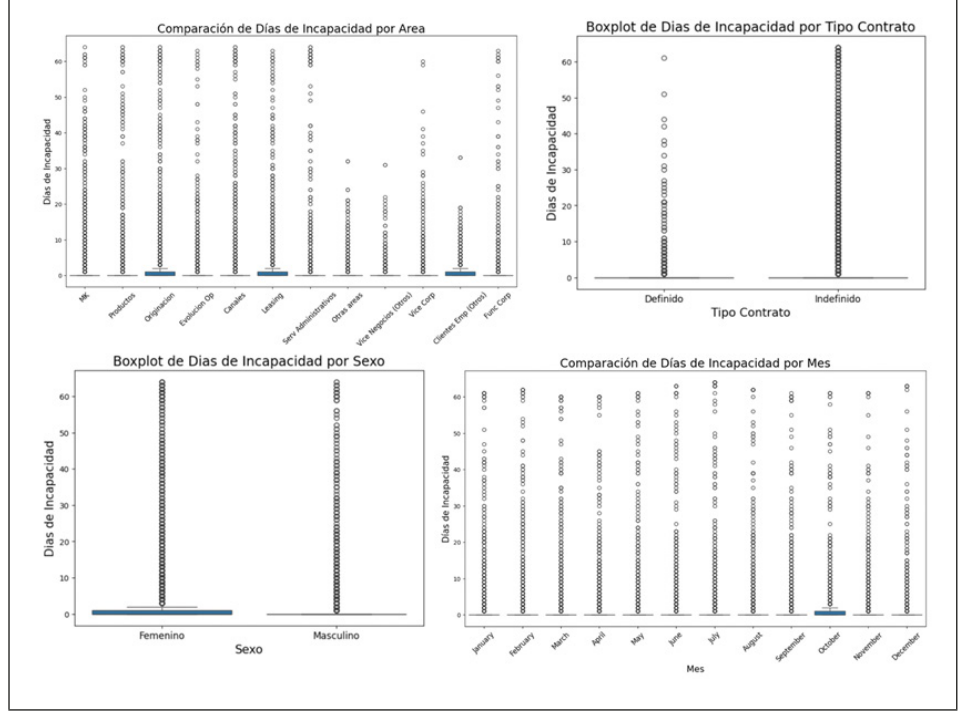
Otro análisis exploratorio de datos fue ejecutado, esta vez centrado en características de interés asociadas al enfoque de regresión, en este la variable objetivo es numérica y hace referencia al número de días de incapacidad que una persona tendrá durante los siguientes tres meses, la Figura 7 evidencia el predominio de la no ocurrencia de días de incapacidad en 49.660 registros (78%), seguida por la ocurrencia de 2 días de incapacidad en 4.540 registros (7%) y la disminución gradual de la cantidad de registros en valores cercanos, como 1, 3 y 4 días de incapacidad laboral. La alta frecuencia de cero días de incapacidad frente a los demás valores posibles, sugiere la evaluación de los modelos de regresión con inflación de ceros.

Figura 7. *Distribución de la cantidad de días de incapacidad (variable objetivo)*



Para la exploración de la relación entre las variables predictoras categóricas y la variable objetivo (ahora numérica), nuevamente se acude a la visualización a través de diagramas de caja, como los presentados en la Figura 8, en los cuales se evidencia la diferencia en las distribuciones de los grupos de la variable Días de Incapacidad (variable objetivo), definidos con base en las categorías de cada variable predictora, en este caso, de las variables Área, Sexo, TipoContrato y Mes.

Figura 8. *Características categóricas versus variable objetivo numérica*



En cuanto a la relación entre las variables predictoras numéricas y la variable objetivo, se analiza la matriz de correlación, que se presenta en la Figura 9, en la cual se evidencian mayores correlaciones con las variables DiasIncapacidad, MesesIncapacidad y Ausencia. La Tabla 4 resume el top 5 de los valores más altos de correlación con la variable objetivo.

Revista EIA https://doi.org/10.24050/reia.v22i44.1890

Figura 9. *Matriz de correlación (enfoque regresión)*

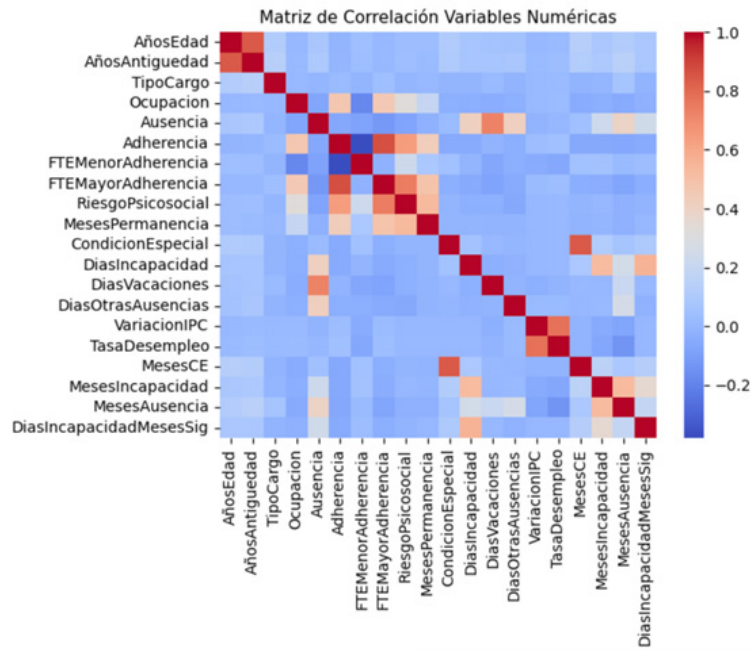


Tabla 4. *Top 5 correlaciones con la variable objetivo*

Top 5 Correlaciones con la Variable Objetivo	
Variable	**Valor Correlación**
DiasIncapacidad	0.56
MesesIncapacidad	0.36
Ausencia	0.24
MesesAusencia	0.19
MesesCE	0.12

Cabe recordar que, bajo este enfoque de regresión, se mantienen las mismas correlaciones altas entre algunas variables predictoras numéricas, que fueron detalladas anteriormente en la Figura 4. Esto es de gran importancia porque sugiere la exploración de escenarios en donde se eliminen ciertas características del modelo de predicción y/o se apliquen técnicas de regularización, con el fin de mejorar la capacidad de predicción y obtener mejores resultados.

Para el caso en que la variable objetivo es el número esperado de días de ausencia de un empleado en los siguientes tres meses, el mejor escenario obtenido para el modelo de regresión Poisson se dio con la exclusión de las variables que presentaban alta correlación con otras variables, o que resultaron no significativas en una primera implementación del modelo, y con el método de normalización StandardScaler. Bajo este escenario las métricas de desempeño que el modelo obtuvo fueron: Mean Squared Error (MSE) de 14.47, Maximum absolute error (MAE) de 1.42, R^2 del 84% y Median absolute error (MeAE) de 0.72. De estos resultados se recalca que un R^2 del 84% es un valor alto, lo cual representa un adecuado ajuste del modelo a los datos reales. Además, todas las medidas de error presentan valores relativamente bajos, lo cual confirma la pertinencia del modelo de regresión Poisson.

A partir del modelo de regresión Poisson se encontró que las covariables que tienen un mayor impacto sobre el número promedio de días que los empleados de la entidad financiera se incapacitan son los meses de octubre y noviembre, el sexo masculino, contrato a término indefinido, área diseño de procesos, contrato temporal / aprendiz, área de atención al cliente, área de vicepresidencia de negocios y meses de incapacidad.

Considerando que en el modelo de regresión Poisson al igual que en el modelo logístico la función liga es logarítmica, entonces para conocer el impacto de cada covariable sobre el promedio de días de incapacidad de los empleados es necesario calcular la función exponencial elevada a la estimación del coeficiente de la respectiva covariable. En la Tabla 5 se realiza esta transformación y se presenta el efecto que las variables más importantes tienen sobre el número promedio de días que los empleados se incapacitan.

De acuerdo con lo anteriormente expuesto, por ejemplo, se puede interpretar que, durante el mes de octubre el promedio global de días de incapacidad aumenta 1.24 veces respecto al promedio global, entre los empleados hombres el promedio de incapacidad es 0.82 veces el promedio global (lo que significa que es menor). Entre los empleados con contrato a término indefinido el promedio de días de incapacidad es 1.12 veces el global (es mayor), pero entre los empleados con contrato temporal / aprendiz el promedio

Revista EIA https://doi.org/10.24050/reia.v22i44.1890

es 0.54 veces el global (aproximadamente la mitad). En el Área de Diseño de Procesos el promedio de días de incapacidad es 0.67 veces el global (es menor) y dicho valor es similar al caso del Área de Vicepresidencia de Negocios. Y si se tiene que un aumento en un mes en el tiempo que una persona lleva incapacitada el promedio de días que la persona permanece incapacitada aumenta 1.31 veces.

Tabla 6. *Efecto de características sobre el promedio de días de incapacidades*

Característica	Coeficiente	Transformación	Efecto sobre el promedio de días de incapacidad
(Mes) Octubre	0.2184	$e^{0.2184}$	**1.2441**
(Mes) Noviembre	0.2133	$e^{0.2133}$	**1.2378**
Sexo Masculino	-0.1993	$e^{-0.1993}$	**0.8193**
Contrato Indefinido	0.1177	$e^{0.1177}$	**1.1249**
Área Diseño Procesos	-0.4012	$e^{-0.4012}$	**0.6695**
Contrato Temporal	-0.6224	$e^{-0.6224}$	**0.5367**
Área Atención Cliente	-0.266	$e^{-0.2660}$	**0.7664**
Área Vice Negocios	-0.4109	$e^{-0.4109}$	**0.6631**
Meses Incapacidad	0.2674	$e^{0.2674}$	**1.3066**

4. Discusión y Conclusiones

En el presente artículo se abordaron dos enfoques, de clasificación y de regresión para predecir la ocurrencia de incapacidades laborales, a nivel individual, en una institución financiera colombiana. En ambos enfoques los modelos seleccionados fueron regresión logística en el enfoque de clasificación, y regresión de Poisson en el enfoque de regresión, los cuales pertenecen a la familia de modelos lineales generalizados y tienen la ventaja de ser naturalmente interpretables y con baja complejidad computacional. Los dos modelos son coherentes y permiten abordar el problema de interés desde dos perspectivas distintas.

En el contexto de la Regresión Logística, los resultados obtenidos demuestran que el modelo es capaz de entregar aproximadamente un 63% de clasificaciones correctas frente al total de clasificaciones

hechas (accuracy), y en cuanto al total de ocurrencias reales de incapacidad laboral, o total de casos realmente positivos, es capaz de clasificar como tal al 61% (recall). Adicionalmente, según el AUC (Área bajo la curva) obtenido, existe un 68% de probabilidad de que el modelo pueda distinguir entre la clase positiva y la clase negativa (Ocurrencia de Incapacidad versus No Ocurrencia de Incapacidad), lo cual confirma que tiene una adecuada capacidad de discriminación, aunque esta es claramente susceptible de mejora.

El modelo de regresión Poisson demostró ser adecuado en el escenario en que la variable de interés es el número promedio de días que se incapacitan los empleados, esto en cuanto presentó un buen ajuste a los datos reales, con un R^2 del 84% y métricas de error bajas, MSE de 14.47, MAE de 1.42 y MeAE de 0.72. Además, con este modelo también es posible medir el impacto de cada una de las características o factores de interés sobre el fenómeno de ausentismo laboral por incapacidad.

De acuerdo con el análisis de características realizado tanto para la regresión logística como la regresión Poisson, el mes, el sexo, el tipo de contrato, el área de trabajo y el histórico de la cantidad de meses en los que el empleado ya había presentado incapacidad, son los factores que mayor influencia tienen sobre la ocurrencia o no ocurrencia de incapacidades laborales en los meses siguientes. En cuanto al mes, se identificó que, estando en los meses de octubre y noviembre, son mayores las probabilidades de que se presenten incapacidades en los tres meses siguientes, mientras que para el mes de enero, dichas probabilidades disminuyen en una medida importante. En términos del sexo, es más probable que se presenten incapacidades laborales en el sexo femenino que en el masculino. Para el tipo de contrato se encontró que las personas con contrato temporal/aprendiz se incapacitan mucho menos que aquellas que tienen un contrato a termino indefinido, además, existen áreas de gran impacto para la organización como Diseño de Procesos y Vicepresidencia de Negocias en las que se presentan mucho menos incapacidades que en el resto de la organización. Y frente al histórico de meses pasados en los cuales el empleado ya ha tenido incapacidad, en la medida en que esta cantidad de meses aumenta, la probabilidad de que el empleado continúe presentando incapacidades en los tres

Revista EIA https://doi.org/10.24050/reia.v22i44.1890

meses siguientes, también aumenta. Se tiene que los resultados de este análisis de características son similares y/o coherentes a los encontrados por Gomes y Lopes (2022), Lawrance et al. (2021), Montano et al. (2020) y Raja y Gupta (2019).

Los dos modelos considerados, regresión logística y regresión Poisson mostraron ser adecuados para analizar el fenómeno de ausentismo laboral por incapacidad. Ambos modelos pertenecen a la familia de modelos lineales generalizados, presentan una complejidad similar y la capacidad de interpretabilidad. Respecto al análisis de características las conclusiones obtenidas con cada modelo fueron similares, lo cual valida la pertinencia y coherencia de estas. Finalmente, se tiene que los dos modelos considerados se complementan mutuamente, pues permiten abordar una misma situación de interés desde dos perspectivas diferentes. Con la regresión logística se analiza la probabilidad de que los empleados se incapaciten o no en un determinado período y con la regresión Poisson se analiza el número promedio de días de incapacidad que tendrán los empleados para ese mismo período.

En términos generales, los resultados obtenidos para los modelos implementados en este artículo son susceptibles de mejora. Dada la naturaleza de las incapacidades laborales, la inclusión de variables biológicas podría mejorar significativamente las predicciones. Asimismo, variables relacionadas con el estilo de vida de la persona, como hábitos de alimentación, actividad física, etc. y otras dimensiones del riesgo psicosocial en el ambiente laboral (en este artículo solo se midió el riesgo psicosocial en función de una dimensión) podrían ser de gran valor para aumentar el rendimiento de los modelos. Adicionalmente, sería útil contar con mayor detalle de las incapacidades laborales vividas por los empleados, posiblemente una clasificación de estas, lo cual facilitaría la identificación de diferentes poblaciones que se pueden estar dando dentro de los datos.

5. Referencias

Ali Shah, S.A.; Uddin, I; Aziz, F; Ahmad, S; Al-Khasawneh, M.A.; Sharaf, M. (2020). An Enhanced Deep Neural Network for Predicting Workplace Absenteeism, Hindawi Complexity, 5843932. https://doi.org/10.1155/2020/5843932

Araujo, Vanessa S.; Rezende, Thiago S.; Guimaraes, Augusto J.; Silva Araujo, Vinicius J.; de Campos Souza, Paulo, V. (2019). A hybrid approach of intelligent systems to help predict absenteeism at work in companies. SN Applied Sciences, 1(6), 536. https://link.springer.com/article/10.1007/s42452-019-0536-y

Berg, T.; Elders, L.; Zwart, B.; Burdorf, A. (2008). The effects of work-related and individual factors on the work ability index: A systematic review. Occupational and Environmental Medicine, 66(4), 211–20. https://oem.bmj.com/content/66/4/211

Borda, M.C.; Rolón, E.; Díaz-Piraquive, F.N. y González, J. (2017). Ausentismo Laboral: Impacto en la Productividad y Estrategias de Control desde los Programas de Salud Empresarial. Universidad del Rosario. https://doi.org/10.48713/10336_13583

Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription

services: An application of support vector machines while comparing two parameter selection techniques. Expert systems with applications, 34(1), 313-327. https://doi.org/10.1016/j.eswa.2006.09.038

Daza, F.M.; Perez Bilbao, J. (1997). NTP 443: Factores psicosociales: metodología de evaluación. https://www.insst.es/documents/94886/326962/ntp_443.pdf/35f6978d-1338-43c3-ace4-e81dd39c11f0

Fischer, JE; Genser, B; Nauroth, P; Litaker, D; Mauss, D. (2020). Estimating the potential reduction in future sickness absence from optimizing group-level psychosocial work characteristics: a prospective, multicenter cohort study in German industrial settings. Journal of Occupational Medicine and Toxicology, 15(1), 33. https://doi.org/10.1186/s12995-020-00284-x

Gil Bellosta, Carlos J. (2018). Modelos con Inflación de Ceros y Separación Perfecta. https://www.datanalytics.com/2018/04/11/modelos-con-inflacion-de-ceros-y-separacion-perfecta/

Gomes, J.M.; Lopes, F.M. (2022). Interpretability with Relevance Aggregation in Neural Networks for Absenteeism Prediction. 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI). https://ieeexplore.ieee.org/document/9926870

International Labour Organization. (2020). Skilled Workers Matter. The business case for addressing absenteeism and turnover in Myanmar's garment sector. https://www.ilo.org/publications/skilled-workers-matter-business-case-addressing-absenteeism-and-turnover-0

Instituto Colombiano de Normas Técnicas y Certificación (ICONTEC). (1996). Norma Técnica Colombiana NTC 3793. Salud Ocupacional. Clasificación, Registro y Estadísticas de Ausetismo Laboral, 1-2. https://www.studocu.com/co/document/uninpahu-institucion-universitaria-sede-principal/administracion/pdfslide-dfs/28685112

Isaac P., Jaime; Rivera, Gerson. (2021). Modelos Lineales Generalizados con R. Chapter 8: Regresión de Poisson. https://bookdown.org/jaimeisaacp/bookglm/regresi%C3%B3n-de-poisson.html

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An introduction to statistical learning. New York: springer.

Kocakulah, M.C.; Kelley, A.; Mitchell, K.M.; Ruggieri, M.P. (2016). Absenteeism Problems And Costs: Causes, Effects And Cures. International Business & Economics Research Journal (IBER), 15(3). https://clutejournals.com/index.php/IBER/article/view/9673

Lawrance, N; Petrides, G; Guerry, M.A. (2021). Predicting employee absenteeism for cost effective interventions. Decision Support Systems, 147, 113539. https://doi.org/10.1016/j.dss.2021.113539

Martiniano, A.; Ferreira, R. P.; Sassi, R. J.; Affonso, C. (2012). Application of a neuro fuzzy network in prediction of absenteeism at work. 7th Iberian Conference on Information Systems and Technologies (CISTI). https://ieeexplore.ieee.org/document/6263151

McCullagh, P. (2019). Generalized linear models. Routledge.

Montano, IH; Marques, G; Alonso, SG; Lopez-Coronado, M; Diez, ID. (2020). Predicting Absenteeism and Temporary Disability Using Machine Learning: A Systematic Review and Analysis. Journal of Medical Systems, 44(9), 162. http://dx.doi.org/10.1007/s10916-020-01626-2

Navarro, C; Bass, C. (2006). The cost of employee absenteeism. Compensation & Benefits Review, 38(6), 26–30. https://doi.org/10.1177/0886368706295343

Qaisar, S. (2020). Predicting Absenteeism at Work Using Machine Learning Algorithms. MJPS, 7(1). https://www.researchgate.net/publication/350955612

Raja, H.; Gupta, R. (2019). The impact of employee absenteeism on organizational productivity with special reference to service sector. International Journal of Research in Humanities, Arts and Literature, 4(7), 581–594. https://oaji.net/articles/2019/488-1558087237.pdf

Salonsalmi, A; Rahkonen, O; Lahelma, E; Pietiläinen, O; Lallukka, T. (2023). Associations between low parental education, childhood adversities and sickness absence in midlife public sector employees. Scandinavian Journal of Public Health, 51(6), 953 – 962. https://journals.sagepub.com/doi/10.1177/14034948221087996

Sanchez, D.C. (2015). Ausentismo laboral: una visión desde la gestión de la seguridad y la salud en el trabajo. Revista Salud Bosque, 5(1), 182. http://revistas.unbosque.edu.co/index.php/RSB/article/view/182

Svärd, A; Lallukka, T; Oakman, J; Roos, E; Ervasti, J; Salmela, J. (2024). The Joint Contributions of Overweight/Obesity and Physical Mental Working Conditions to Short and Long Sickness Absence

among Young and Midlife Finnish Employees: A Register-Linked Follow-Up Study. Obesity Facts, 17(1), 37 – 46. https://doi.org/10.1159/000534525

Vafeiadis, T; Diamantaras K. I.; Sarigiannidis, G; Chatzisavvas, K. Ch. (2015). A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory, 55, 1-9. https://doi.org/10.1016/j.simpat.2015.03.003