Laboratorio de Aprendizaje Estadístico

Profesor Gabriel Alejandro Morales Ruiz

ITESO

Universidad Jesuita
de Guadalajara

Proyecto 1

Análisis estadístico del agotamiento en trabajo remoto mediante modelos de regresión

Fecha: 23 Febrero, 2026

Integrantes:

Julieta Madrigal Flores
Gibrán Leonardo Chávez González
Diana Fernanda Barbosa Dueñas

Índice

Objetivos **3**
Marco Teórico **4**
Regresión lineal **4**
Regresión polinomial **4**
Interacción de factores **5**
Significancia de factores **5**
Regularización (Ridge, Lasso y Elastic Net) **6**
Análisis de Dataset **6**
Pipeline **9**
Conclusiones **10**
Referencias **11**

Objetivos

Generales

Analizar los factores que se asocian al nivel de agotamiento en trabajadores que se desempeñan de manera remota a través de la construcción y comparación de modelos de regresión con y sin regulación, con la finalidad de identificar variables significativas y evaluar la capacidad predictiva del modelo sobre el *burnout score*.

Específicos

1. Analizar la estructura y las características del dataset Work From Home Burnout Dataset para identificar las variables disponibles y las transformaciones necesarias para poder utilizarlas en modelos de regresión.

2. Construir tres modelos de regresión distintos que incluyen combinaciones e interacciones entre variables relevantes para el burnout score.

3. En cada modelo, implementar cuatro versiones : sin penalización, con regularización de Ridge, con regularización de Lasso y con regularización de Elastic Net; además de un proceso apropiado de limpieza, transformación, escalamiento y partición de train-test para el dataset.

4. Comparar el desempeño predictivo de cada versión de los modelos a través de R^2 y explorar la selección de coeficientes bajo cada técnica de regularización.

5. Evaluar la significancia estadística de los predictores de cada modelo sin penalización para determinar qué variables e interacciones muestran evidencia relevante sobre el burnout score.

Marco Teórico

Regresión lineal

La regresión lineal es un procedimiento estadístico que modela la relación entre la variable dependiente Y y las o la variable independiente X, asumiendo que dicha relación es lineal en los factores. El modelo más común es $Y = \beta 0 + \beta 1 X + \varepsilon$, donde $\beta 0$ es el intercepto, $\beta 1$ es la pendiente y ε es el error aleatorio. Está técnica permite describir, explicar y estimar fenómenos reales desde un enfoque más sencillo y formal.

Para definir los parámetros del modelo, se estiman mediante el método de Mínimos Cuadrado Ordinarios (OLS), el cual minimiza la suma de los errores al cuadrado entre los valores observados y los valores predichos del modelo. Los estimadores OLS son sin sesgo y eficientes, ya que funcionan bajo supuestos como linealidad, independencia, homocedasticidad y normalidad de los errores; esto permite que las inferencias estadísticas resultantes sean válidas.

Finalmente, la ventaja principal de este procedimiento es la interpretación directa de los coeficientes, donde β1 hace referencia al cambio esperando en la variable dependiente por cada incremento unitario en la variable independiente y a su vez, mantiene constantes a las demás variables. Sin embargo, una de las limitaciones se presenta cuando las relaciones no son lineales, existen valores atípicos o colinealidad entre los predictores; en este caso se debe de hacer un análisis a profundidad de los residuos y diagnósticos del modelo para poder validar los resultados del modelo.

Regresión polinomial

La regresión polinomial es un derivado de la regresión lineal que permite modelar relaciones no lineales entre la variable dependiente y la independiente a través del uso de términos polinomiales como x^2, x^3, etc. pero se mantiene lineal en los parámetros, lo cual permite el uso de OLS para hacer estimaciones. Por ejemplo, un modelo de segundo grado se puede expresar como $Y = \beta 0 + \beta 1 X + \beta 2 X2 + \varepsilon$, el cual permite capturar datos que una regresión lineal simple no podría capturar por la curvatura. En ingeniería, economía y análisis de procesos sirve para aproximar relaciones no lineales dentro del intervalo de datos analizado.

Sin embargo, una de las desventajas es el sobreajuste cuando el modelo utiliza grados muy altos, ya que puede afectar la capacidad predictiva del modelo. A partir de esto, se recomienda seleccionar el grado del polinomio con base en criterios estadísticos y validación cruzada; es decir, elegir el modelo que sea simple, entendible y con un buen ajuste.

Interacción de factores

La interacción de factores ocurre cuando el efecto que tiene una variable independiente sobre la variable dependiente se ve influenciada por el nivel de otra variable; es decir, la variable independiente no actúa sola, su efecto depende de otra. Ahora bien, en modelos de regresión está interacción se puede representar con términos multiplicativos como X_1 y X_2, y el modelo podría representarse de esta manera : $\mathbf{Y = \beta 0 + \beta 1 (X_1) + \beta 2 (X_2) + \beta 3 (X_1 \times X_2)}$**.** Esta interacción es muy utilizada en casos reales, ya que muchas variables no funcionan de manera aislada, se combinan, se potencian y se afectan entre sí. Es por eso que incluirlas permite capturar relaciones más realistas, modelar efectos en conjunto y tener un mejor entendimiento de los datos.

Sin embargo, la interacción aumenta la complejidad del modelo y podría dificultar la interpretación de los coeficientes. Al hacer una inclusión excesiva, se puede generar sobreajuste; por ello, es fundamental recordar que solo se debe incorporar términos de interacción cuando exista respaldo teórico o evidencia clara y siempre respetando el principio de jerarquía del modelo, es decir, que primero se tome en cuenta el efecto individual de cada variable y después se ve el efecto en conjunto.

Significancia de factores

La significancia de factores es la evaluación estadística de sí un coeficiente de regresión es significativamente diferente de cero, es decir, verificar si la variable tiene o no efecto sobre la variable dependiente. Este análisis se hace a través de pruebas de hipótesis, donde la hipótesis nula establece que la variable NO tiene efecto sobre la variable dependiente y la hipótesis alterna menciona que la variable SI tiene efecto sobre la variable dependiente. Ahora bien, en esta evaluación el estadístico t y el p-value, donde el estadístico t mide qué tan grande es el coeficiente comparado con su error , si t es grande en valor absoluto significa que hay evidencia fuerte de que el coeficiente NO es cero pero si t es pequeño, significa que el coeficiente podría ser básicamente cero.

Por otro lado, el p-value hace el estadístico un poco más entendible, de tal forma que si el p-value es < 0.05 indica que hay muy poca probabilidad de que el coeficiente sea 0 o que el valor asignado haya sido al azar y por ende, esta variable ES significativa. Esto implica que existe suficiente información para afirmar que la variable es relevante para el modelo y en modelos de múltiples variables, esta significancia se interpreta considerando las variables incluidas. Aunque, es importante mencionar que a pesar de que sea un coeficiente significativo puede tener un impacto pequeño en términos reales.

Regularización (Ridge, Lasso y Elastic Net)

Regularización es una técnica que pretende mejorar la capacidad de generalizar los modelos de regresión, principalmente cuando se tiene muchos predictores o existe una alta correlación entre ellos. Esta técnica añade un término de penalización a la función de error del modelo para reducir el peso asignado a los coeficientes y controlar la complejidad del modelo.

La regresión Ridge utiliza la penalización de tipo L2, la cual reduce el peso asignado a los coeficientes pero sin reducirlos completamente a cero. Mientras que Lasso emplea una penalización L1, permitiendo que algunos coeficientes se vuelvan exactamente cero y seleccionado de manera automática las variables significativas. Ahora bien, Elastic Net combina las penalizaciones de L1 y L2 para lograr un balance entre Ridge y Lasso. Este método selecciona grupos de variables en conjunto mediante la validación cruzada y priorizando el desempeño predictivo del modelo.

Análisis de Dataset

¿De dónde viene?

La página Kaggle es una plataforma de competencias de ciencia de datos que funciona como una comunidad en línea donde los usuarios pueden utilizar distintos datasets creados por los propios integrantes, generalmente sobre diversos temas, promoviendo que los creadores compartan sus proyectos en formatos accesibles.

Además, permite publicar conjuntos de datos, desarrollar modelos dentro de un entorno web de ciencia de datos, colaborar con otros científicos de datos e ingenieros de aprendizaje automático, y participar en concursos diseñados para resolver distintos desafíos dentro de la ciencia de datos.

¿Qué contiene?

Nuestro dataset está basado en una recopilación de datos sobre el desempeño laboral de un conjunto de trabajadores, busca reflejar cómo es realmente el día a día de las personas que trabajan desde casa, mostrando cómo sus hábitos laborales se relacionan con su nivel de cansancio y su productividad.

Básicamente, ayuda a entender cómo influyen cosas como las horas trabajadas, el tiempo frente a la pantalla, la cantidad de reuniones, los descansos, las horas de sueño y el trabajo fuera del horario normal en qué tan bien una persona logra terminar sus tareas y qué tan cerca puede estar del agotamiento.

El dataset tiene alrededor de 1800 registros diarios de distintos usuarios, donde cada fila representa un día específico de trabajo, incluyendo tanto días entre semana como fines de semana para reflejar los modelos de trabajo flexibles actuales.

Al juntar datos de carga laboral con indicadores de bienestar, nos permite analizar el riesgo de burnout, comparar productividad con bienestar, encontrar patrones de comportamiento e incluso desarrollar modelos que detecten el agotamiento de forma temprana. Por eso, es útil para temas como aprendizaje supervisado, análisis de tendencias en el tiempo y estudios de recursos humanos o salud laboral.

¿Qué información dan las muestras?

Estamos trabajando con un conjunto de aproximadamente 1800 muestras distintas, lo que nos permite observar una gran variedad de rutinas laborales reales. Gracias a esta cantidad de datos, el modelo puede aprender patrones mucho más cercanos a la vida diaria de los trabajadores y no solo basarse en unos pocos casos aislados.

El archivo de Excel funciona como una especie de fotografía general del comportamiento laboral, donde podemos ver diferentes formas en las que una persona organiza su día: cuánto trabaja, cuánto tiempo pasa frente a la pantalla, cómo distribuye sus descansos y cómo todo eso impacta en su rendimiento. Con toda esta información, podemos calcular promedios que nos ayudan a entender cómo sería, en términos generales, la rutina típica de un trabajador.

Al enfocarnos en los distintos datos recopilados, hemos podido estimar cómo se desempeña en promedio un trabajador dentro de este contexto. Por ejemplo, el análisis sugiere que un trabajador promedio suele trabajar al menos alrededor de 7 horas al día, lo cual nos da una base para comparar otros factores como productividad, nivel de cansancio o eficiencia en completar tareas.

Más allá del número como tal, lo importante es que estos promedios nos permiten identificar tendencias generales, entender qué hábitos podrían estar ayudando o perjudicando el desempeño, y usar esa información para hacer análisis más profundos o incluso predicciones sobre el comportamiento laboral y el riesgo de agotamiento en el futuro.

¿Qué se quiere analizar?

Lo que se quiere analizar con este Excel es entender cómo es realmente la rutina diaria de las personas que trabajan desde casa y cómo esa rutina influye en su rendimiento y en su nivel de agotamiento. No solo se busca ver números sueltos, sino encontrar relaciones entre distintos hábitos, como cuántas horas trabaja una persona, cuánto tiempo pasa frente a la computadora, cuántas reuniones tiene al día, si toma descansos suficientes y cuántas horas

duerme. La idea es observar cómo todos estos factores juntos afectan qué tan bien logra completar sus tareas y qué tan propensa puede ser a sufrir burnout.

También se quiere identificar patrones generales que ayuden a describir cómo es el comportamiento promedio de un trabajador remoto. Por ejemplo, ver si trabajar más horas siempre significa ser más productivo, o si pasar demasiado tiempo frente a la pantalla aumenta el cansancio. Con este análisis, se pueden generar conclusiones útiles para entender mejor el equilibrio entre productividad y bienestar.

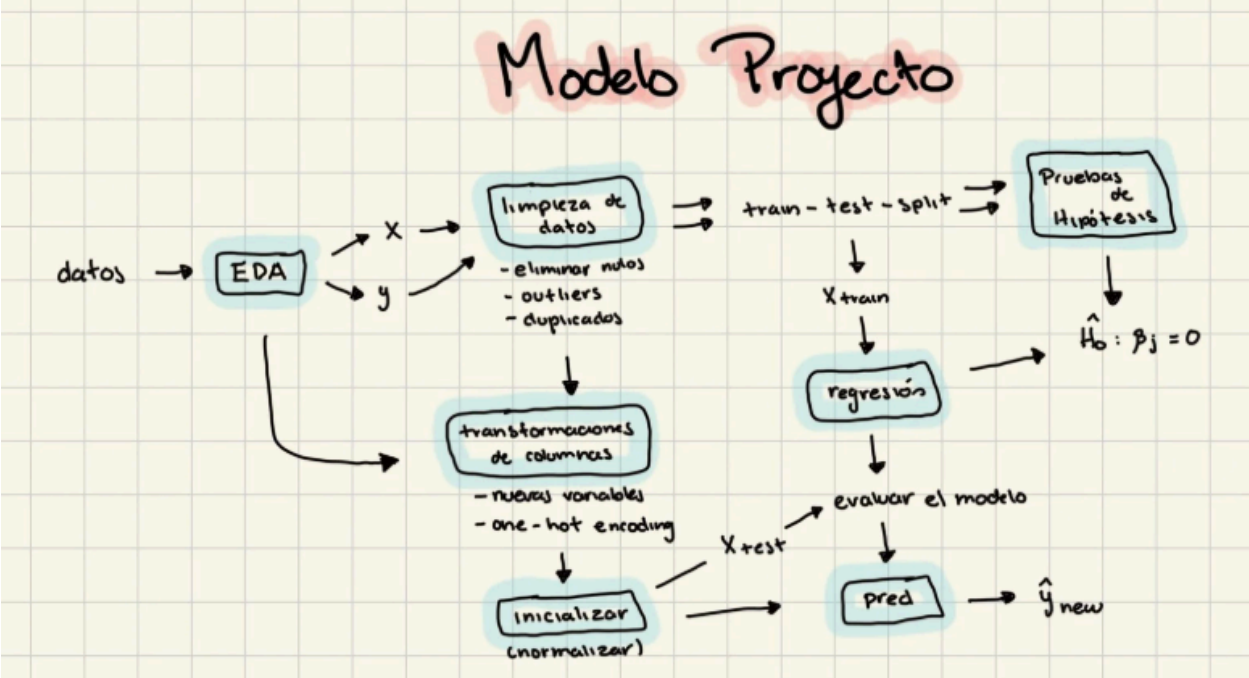
Al final, el objetivo es usar estos datos para detectar señales tempranas de agotamiento, mejorar la organización del trabajo y ayudar a que las personas tengan rutinas más saludables sin afectar su desempeño laboral.

¿Qué variables se tienen que transformar para poder usarse en un modelo de regresión?

Para poder usar este conjunto de datos en un modelo de regresión, primero es necesario transformar las variables que son categóricas o de texto, ya que los modelos matemáticos normalmente solo trabajan con números. En este caso, una de las principales variables que necesitaría transformarse es **day_type**, porque contiene valores como "*Weekday"* y "*Weekend"*. Esta variable se podría convertir, por ejemplo, en valores numéricos (0 y 1) o usando técnicas como *one-hot encoding*. Lo mismo pasaría con **burnout_risk**, ya que contiene categorías como *Low*, *Medium* o *High*, y el modelo necesita que estas categorías se representen numéricamente.

Por otro lado, las variables numéricas como **work_hours**, **screen_time_hours**, **meetings_count**, **breaks_taken**, **after_hours_work**, **sleep_hours**, **task_completion_rate** y **burnout_score** ya están listas para usarse en regresión, aunque en muchos casos sería recomendable escalarlas o estandarizarlas para que todas estén en rangos similares y el modelo funcione mejor. En resumen, lo más importante es convertir las variables categóricas a números y asegurarse de que las variables numéricas estén en una escala adecuada para que el modelo pueda analizarlas correctamente.

Pipeline



El diagrama describe los pasos para el análisis de regresión lineal:

1. Comienza con la Exploración de Datos (EDA) para entender que datos tenemos, qué tipo son y sus dominios.
2. Hacemos limpieza de datos, siguiendo por transformaciones de columnas y escalamiento.
3. Se realiza un train-test-split para dividir los datos en entrenamientos y prueba.
4. Se entrena el modelo para hacer predicciones y se realizan pruebas de hipótesis para verificar si los coeficientes β_j son significativamente de 0.

Conclusiones

En este proyecto se logró analizar de manera integral cómo distintos hábitos laborales dentro del trabajo remoto pueden influir directamente en el nivel de agotamiento de los trabajadores. A partir del uso de un dataset obtenido de Kaggle, fue posible estudiar aproximadamente 1800 registros que reflejan rutinas reales de trabajo, permitiendo observar patrones entre variables como horas laborales, tiempo frente a pantalla, descansos, sueño y carga fuera de horario de trabajo.

El modelo 1 funcionó como base al incluir las 8 variables disponibles, obtuvo un R^2 de 0.92 pero a su vez, evidenció multicolinealidad ya que solo task_completion_rate fue estadísticamente significativa. En el modelo 2 se exploró si trabajar fuera del horario laboral modificaba el efecto de completar tareas sobre el burnout, siendo after_hours_work una variable no significativa pero con el t-statistic más alto del modelo 1; sin embargo ni la variable ni la interacción resultaron significativas y el R^2 se mantuvo igual que el modelo 1. Finalmente en el modelo 3, se incorporó el término cuadrático de task_completion_rate y resultó ser más informativo, es decir, la variable de task_completion_rate en orden 1 y en orden 2 lograron mejorar el R^2 a 0.93, confirmando que la relación entre completar tareas y el burnout no es lineal sino que varía según el nivel de la variable.

Es importante mencionar que la regularización con Ridge, Lasso y Elastic Net no lograron mejorar significativamente el desempeño de ningún modelo, lo que sugiere que task_completion_rate domina casi toda la capacidad predictiva del modelo y las demás variables predictivas aportan muy poca información sobre el burnout, de tal manera que penalizar los coeficientes no resuelve el problema. En conclusión, la tasa de completación de tareas es el factor más determinante del burnout en el trabajo remoto y su efecto no es lineal, lo que significa que aumentar la productividad no reduce el agotamiento de forma constante.

Referencias

ApX Machine Learning. (2025). *Common Split Ratios* en ApX Machine Learning. Recuperado de: Common Train-Test Split Ratios

FasterCapital . (2025). *Efectos de interacción e interpretación de los efectos de interacción en escenarios de regresión múltiple* en FasterCapital. Recuperado de: Efectos de interaccion interpretacion de los efectos de interaccion en escenarios de regresion multiple - FasterCapital

FasterCapital.(2025). *Variables predictivas: identificación de factores influyentes clave: el papel de las variables predictoras en la regresión lineal múltiple* en FasterCapital. Recuperado de: Variables predictivas identificacion de factores influyentes clave el papel de las variables predictoras en la regresion lineal multiple - FasterCapital

GeekorGeeks. (2026). What is Lasso Regression en Geeks for Geeks. Recuperado de: What is Lasso Regression - GeeksforGeeks

GeeksforGeeks. (2025). *Interpreting the results of Linear Regression using OLS Summary* en Geeks for Geeks. Recuperado de: Interpreting the results of Linear Regression using OLS Summary - GeeksforGeeks

GeeksforGeeks. (2025). *Lasso vs Ridge vs Elastic Net - ML* en Geeks for Geeks. Recuperado de: Lasso vs Ridge vs Elastic Net - ML - GeeksforGeeks

GeeksforGeeks. (2025). *Ridge Regression* en Geeks for Geeks. Recuperado de: Ridge Regression - GeeksforGeeks

Gisbert Juárez, M. (2026). *Regresión polinomial en Probabilidad y Estadística.net*. Recuperado de: ▹ Regresión polinomial

Great Learning Editorial Team. (2024). *A Complete understanding of LASSO Regression* en Great Learning. Recuperado de: What is LASSO Regression Definition, Examples and Techniques

JavierCara. (2025). *Regresión lineal con términos de interacción* en JavierCara.html. Recuperado de: Regresión lineal con términos de interacción

Matematrix. (2025). *Qué es la regresión polinómica y cómo se aplica y deriva* en Matematrix. Recuerdo de: Qué es la REGRESIÓN POLINÓMICA y su APLICACIÓN VALIOSA

MD TOUHIDUL ISLAM.(2025). *Elastic Net Regression Explained with Example and Application* en Statical Aid. Recuperado de: Elastic Net Regression Explained with Example and Application

Minitab Blog Editor. (2019). *Cómo Interpretar los Resultados del Análisis de Regresión: Valores P y Coeficientes* en Minitab. Recuperado de: Cómo Interpretar los Resultados del Análisis de Regresión: Valores P y Coeficientes

Montgomery, D.C., Peck, E. A. & Vining, G. G. (2012). *Introduction to linear regression analysis* (5th edition). Wiley. Introduction to Linear Regression Analysis

Murel, J. & Kavlakoglu. (2025). *What is ridge regression?* en IBM. Recuperado de: What Is Ridge Regression? | IBM

Nau, R. (2020). *Statistical forecasting: notes on regression and time series analysis* en Fuqua School of Business Duke University. Recuperado de: Introduction to linear regression analysis

Shinde, S. (2026). *Work From Home Employee Burnout Dataset* en Kaggle. Recuperado de: https://www.kaggle.com/datasets/sonalshinde123/work-from-home-employee-burnout-dataset