Predicción de Burn Rate

COMISIÓN 29730 2022

EQUIPO

ANALIA COSTANZO - ALFREDO DELGADO - ARIEL FELDMAN

ÍNDICE

VERSIONADO

Versión	Fecha	Descripción
1.0	**01-08-22**	**EDA, Análisis Univariado, Bivariado y Multivariado**
1.1	**18-08-22**	**Comparación de algoritmos.**
1.2	**06-09-22**	**Evaluación de métricas de algoritmos específicos para nuestro modelo.**
1.3	**29-09-22**	**Presentación Ejecutiva del proyecto final**

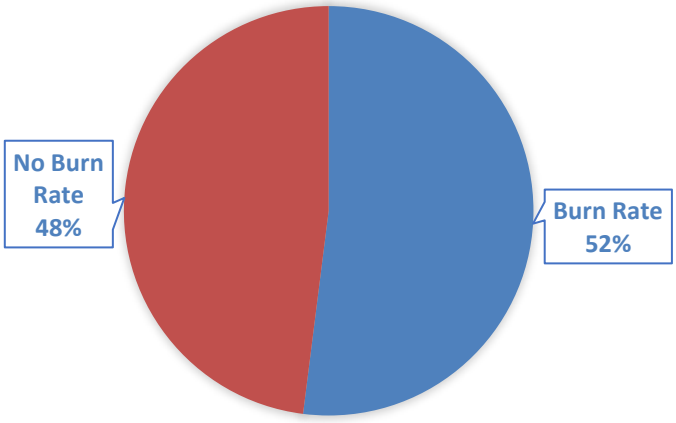
CONTEXTO

El burnout laboral, también denominado "síndrome del quemado" o "síndrome de estar quemado en el trabajo", es un estado de agotamiento físico, emocional y mental que está vinculado con el ámbito laboral, el estrés causado por el trabajo y el estilo de vida del empleado. Este síndrome puede tener impactos severos en el equilibrio emocional de una persona y, por consecuencia, una disminución en el desempeño laboral. Es un proceso en el que progresivamente el trabajador sufre una pérdida del interés por sus tareas y va desarrollando una reacción psicológica negativa hacia su ocupación laboral.

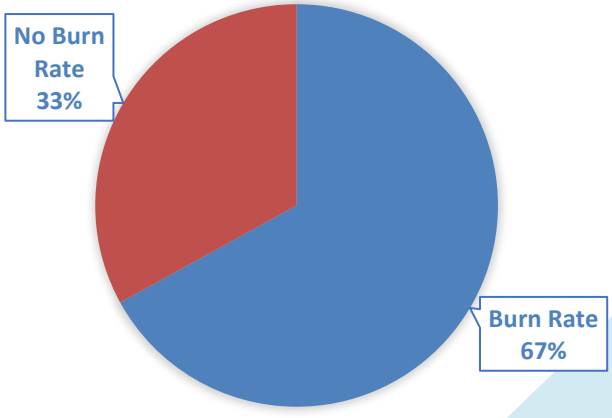
Los principales síntomas del burnout laboral son: Agotamiento físico y mental generalizado, despersonalización y cinismo (adopción de una actitud de indiferencia y desapego, reduciendo claramente su compromiso hacia el trabajo), descenso en la productividad laboral y desmotivación.

Según una encuesta internacional realizado por la compañía Indeed (1) (https://www.indeed.com/lead/preventing-employee-burnout-report) a 1.500 trabajadores de USA en el año 2020, un 52% de los encuestados reconoce presentar Burn Rate en período pre-pandemia Covid-19 y un 67% consideró que empeoró esta condición en el curso de la pandemia. Esta estadística permite inferir la importancia que tiene la prevención y estudio del burnout en todo tipo de organizaciones, más aún en períodos como el que nos encontramos actualmente (pandemia, inestabilidad política-económica, crisis inflacionaria, entre otros).

PRE PANDEMIA (2020)



PANDEMIA (2022)



OBJETIVOS

El objetivo de esta investigación es crear un modelo predictivo, a través de técnicas de Machine Learning, que permita identificar patrones y lograr pronosticar el nivel de burnout de un empleado, sobre la base de características individuales determinadas en el dataset.

La presente investigación busca responder las siguientes preguntas:

- *¿Cuáles son las características relevantes que pueden determinar el grado de burnout de un trabajador?*

- *¿Qué se puede concluir de los análisis estadísticos sobre el nivel de burnout?*

- *¿Qué modelo y algoritmo de Machine Learning permite obtener mejores resultados en la predicción del nivel de burnout de un empleado?*

DESCRIPCIÓN DE LOS DATOS

OBTENCIÓN Y COMPOSICIÓN

Estos dataset fueron extraídos de la página de kaggle.
Link: https://www.kaggle.com/datasets/blurredmachine/are-your-employees-burning-out?select=train.csv.

A continuación, se hace una presentación de los datos incluidos en el dataset de "Burn Rate".

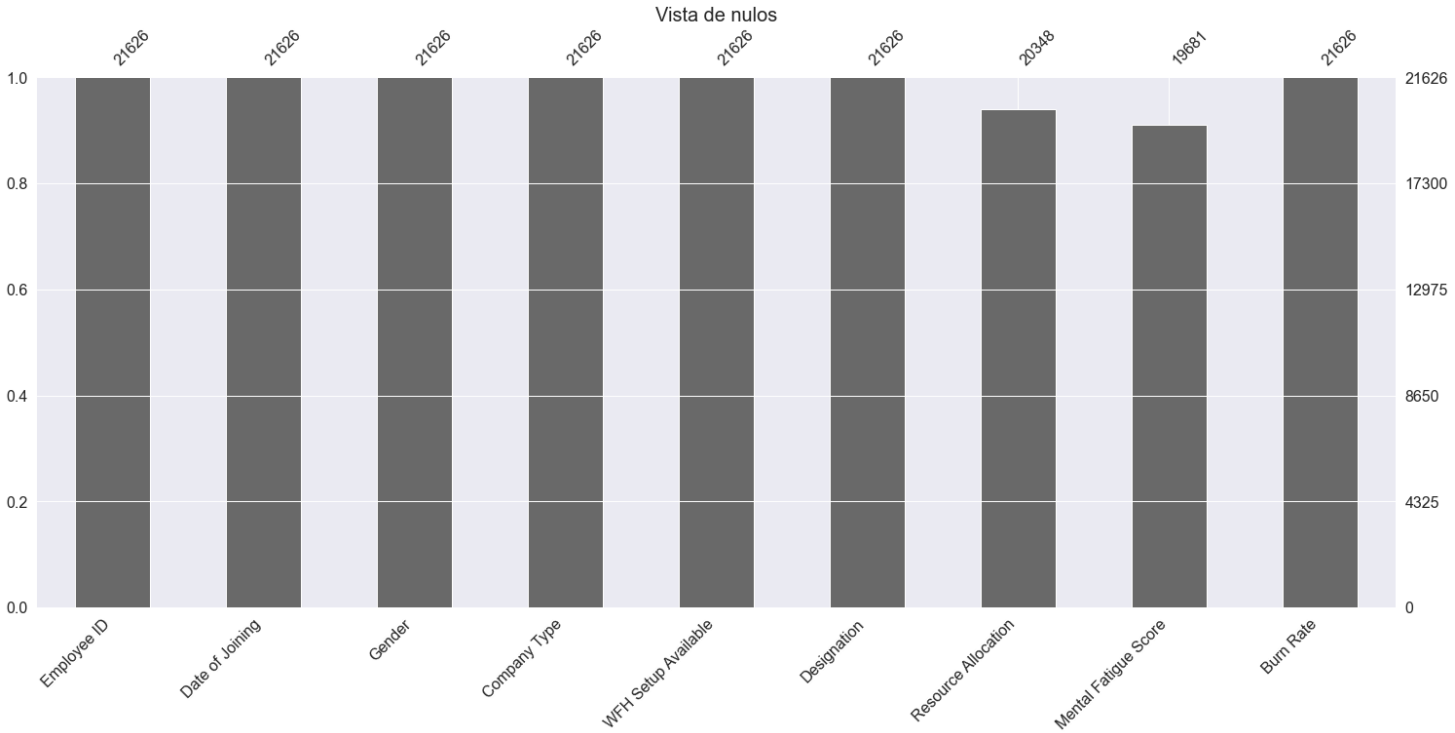
"Burn Rate"

- ***Employee ID:*** *El ID único asignado a cada empleado (ejemplo: fffe390032003000)*
- ***Date of Joining:*** *La fecha y hora en que el empleado se unió a la organización (ejemplo: 2008-12-30)*
- ***Gender:*** *El género del empleado (Hombre/Mujer)*
- ***Company Type:*** *El tipo de empresa donde trabaja el empleado (Servicio/Producto)*
- ***WFH Setup Available:*** *¿Está disponible el trabajo desde casa para el empleado? (Sí/No)*
- ***Designation****: La designación del empleado de trabajo en la organización. En el rango de [0.0, 5.0] mayor es la designación mayor.*
- ***Resource Allocation:*** *La cantidad de recursos asignados al empleado para trabajar, es decir. número de horas de trabajo. En el rango de [1.0, 10.0] (más alto significa más recursos)*
- ***Mental Fatigue Score:*** *El nivel de fatiga mental al que se enfrenta el empleado. En el rango de [0.0, 10.0] donde 0.0 significa sin fatiga y 10.0 significa fatiga total.*
- ***Burn Rate:*** *El valor que necesitamos predecir para cada empleado indicando la tasa de Burn out mientras trabaja. En el rango de [0.0, 1.0] donde cuanto más alto es el valor, más se cansa.*

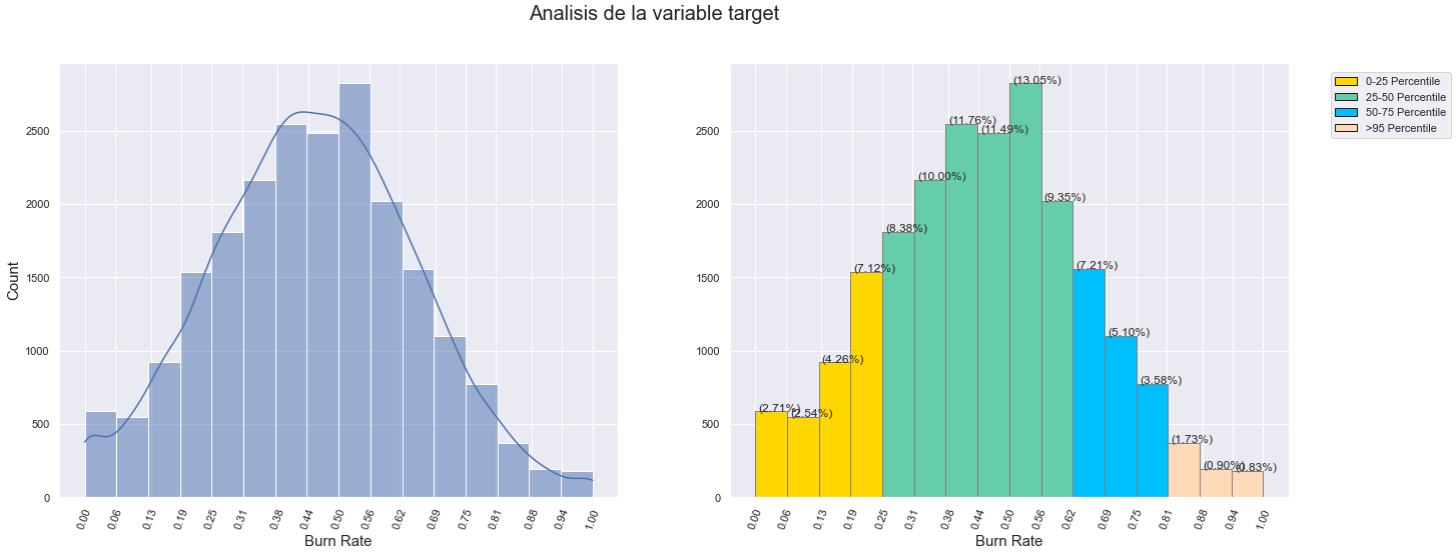
HALLAZGOS POR EL EDA

Realizamos un análisis exploratorio y obtuvimos los siguientes resultados:

Contamos con un dataset de 9 columnas por 22.750 registros, en la cual nuestra variable target ("Burn Rate") cuenta con 1.124 registros vacíos, la misma es eliminada ya que no se podría generar el modelo predictivo. A su vez, las variables "Date of Joining" y "Employee ID" también son eliminadas ya que no son relevantes para nuestra predicción.



Como podemos ver en el grafico 2 variables relevantes presentan nulos, "Mental Fatigue Score" y "Resource Allocation", lo cual corresponden a un %9 (1.945) y %6 (1.278) de las observaciones totales respectivamente.



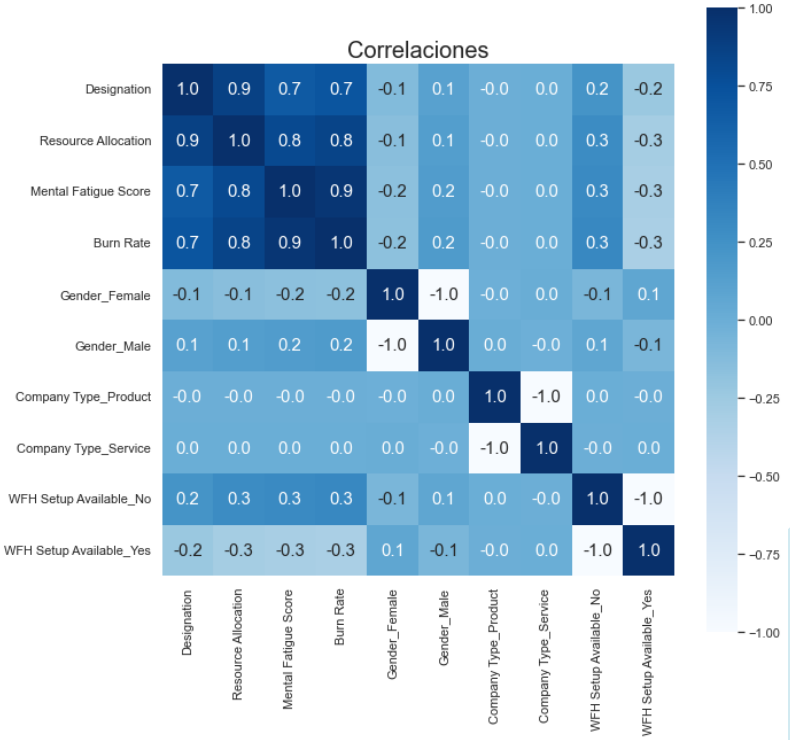
Se generaron 2 gráficas de histogramas para nuestra variable objetivo con el fin de comprender su distribución, podemos observar que se acerca a una distribución normal, los cuales sus valores se concentran en el intervalo [0.25,0.81].

Planteamos la hipótesis de que podrían existir fuertes correlaciones de las features con respecto a la variable target. A continuación, representamos las mismas en un heatmap.

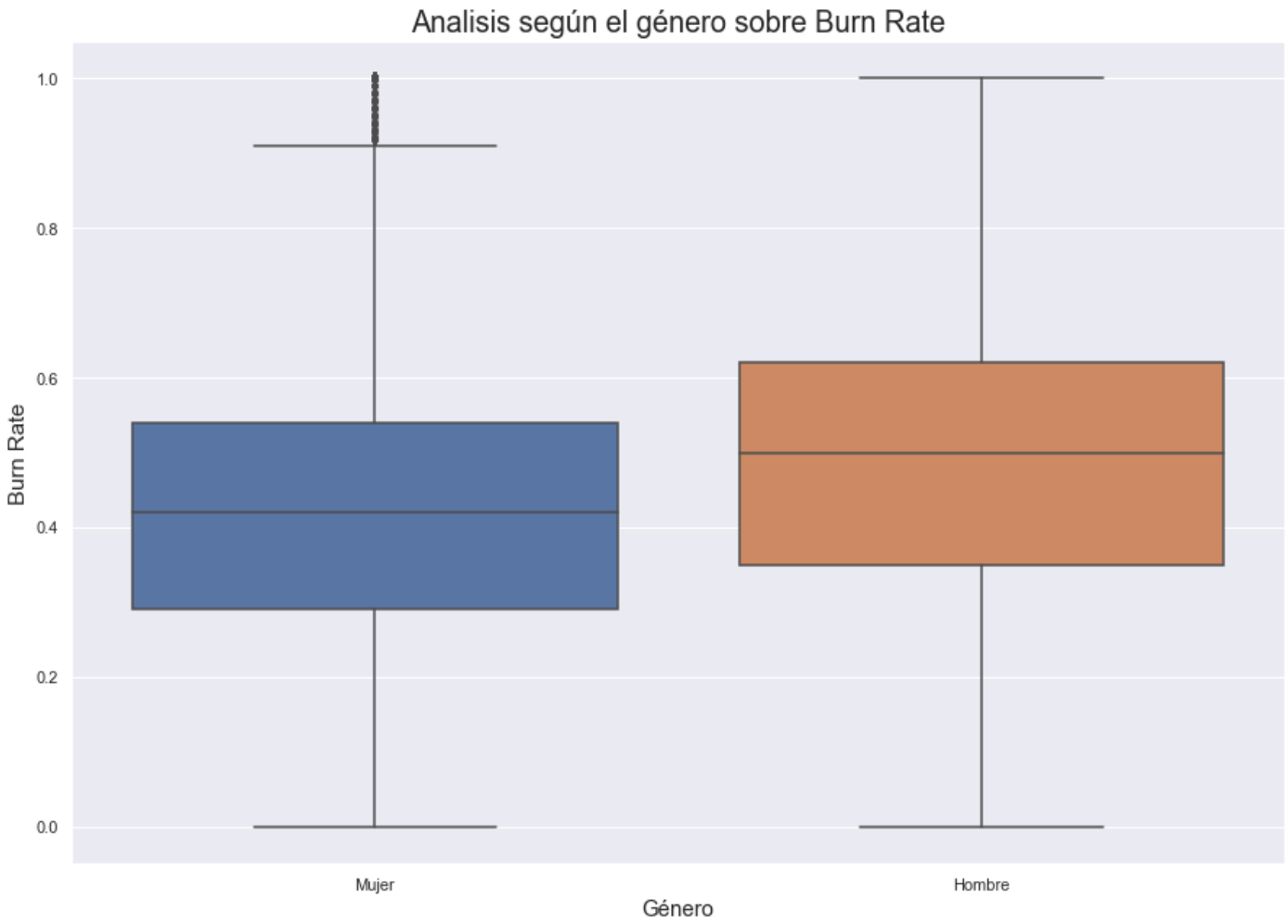
*Las features "**Designation**", "**Resource Allocation**" y "**Mental Fatigue Score**" tiene una alta correlación positiva con el target "**Burn Rate**", superior al 0.7.*

Teniendo en cuenta esta conclusión, posteriormente aplicaremos una estandarización/normalización de las features mencionadas.

Por otro lado, las demás features no muestran una correlación relevante.



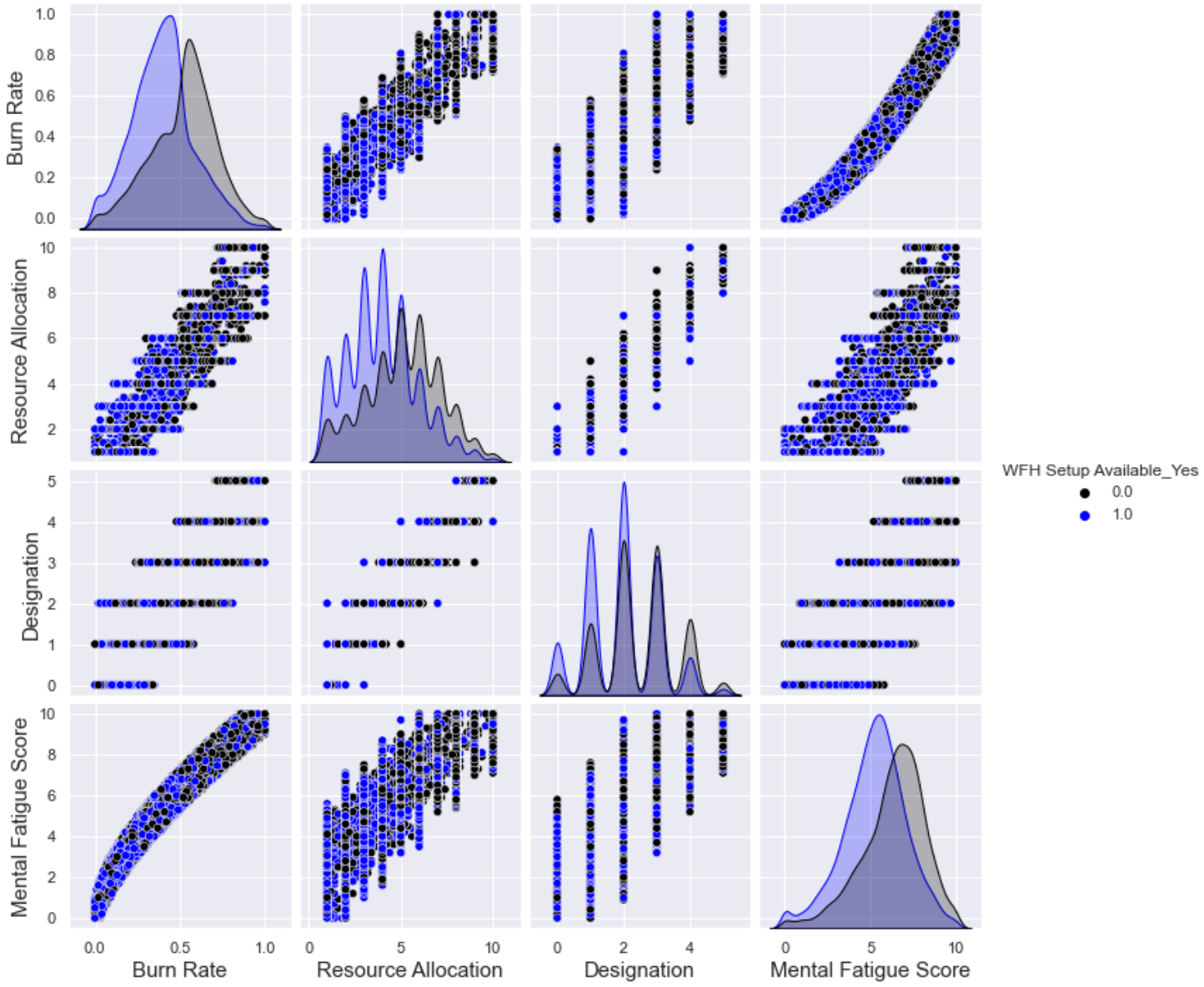
En la siguiente gráfica, realizamos un análisis de comportamiento de género con respecto al “Burn Rate”.



Podemos observar que los hombres tienen mayor nivel de Burn Rate que las mujeres.

Además, detectamos outliers en las mujeres, los cuales serán tenidos en cuenta y tratados en la estandarización.

Buscamos comprender si un empleado realiza home office o no frente a la variable target y las features de mayor correlación, para ello lo representamos en un pairplot.



Podemos ver una mayor concentración y aumento de “Burn Rate” en los trabajadores que no realizan su actividad laboral de manera presencial, sucede lo mismo con la fatiga mental ("Mental Fatigue Score") y cantidad de horas trabajadas ("Resource Allocation").

A mayor jerarquía (“Designation”) menor trabajo presencial y viceversa.

ELECCIÓN DEL ALGORITMO

A partir de nuestros hallazgos en el EDA, se seleccionan los siguientes procesos y elecciones de algoritmos.

1. *Algoritmos y técnicas de aplicación.*

Al tener features importantes con nulos se tomó la decisión de imputarle estos mediante el algoritmo KNNImputer, este algoritmo es una clase de scikit-learn que se utiliza para completar o predecir los valores faltantes en un conjunto de datos. Es un método más útil que funciona con el enfoque básico del algoritmo KNN en lugar del enfoque de completar todos los valores con la media o la mediana. En este enfoque, especificamos una distancia de los valores perdidos que también se conoce como el parámetro K. El valor faltante se predecirá en referencia a la media de los vecinos más cercanos.

Para este algoritmo se utilizaron los siguientes parámetros:

- *n_neighbors=5,*
- *weights='uniform',*
- *metric="nan_euclidean"*

Para transformar en variables numéricas aplicamos a las features categóricas binarias ("Company Type", "Gender", "WFH Setup Available") la técnica de One Hot Encoding y descartamos una de las categóricas por ser perfectamente colineales.

Al estar transformadas a numéricas aplicamos la técnica de "Standard Scaler" de los datos, para darle un peso similar a todas las observaciones.

2. Algoritmos regresión.

En los modelos de regresión es difícil predecir el valor exacto, por esto se busca estar lo más cerca posible del valor real. La mayoría de las métricas van a centrarse en medir: lo cerca (o lejos) que están las predicciones de los valores reales.

En estos modelos hay que considerar los valores anómalos/outliers y si queremos penalizar errores grandes o no. A su vez al tener una variable target en un intervalo de [0,1] algunas métricas pueden no tener un comportamiento correcto, como el MAPE.

Por ello decidimos utilizar la medida RMSE (Raíz cuadrada del error cuadrático medio), ya que cuanto mayor sea el RMSE mayor será la diferencia entre los valores predichos y observados, lo que significara que tan bien nuestro modelo se ajusta a un conjunto de datos.

Por consiguiente, para predecir nuestra variable objetivo "Burn Rate", se definieron los algoritmos LinearRegression, RandomForestRegressor, KNeighborsRegressor, GradientBoostingRegressor, DecisionTreeRegressor, XGBRegressor, LGBMRegressor con los parámetros detallados debajo para comparar luego su métrica de RMSE.

- ***LinearRegression***
  - *n_jobs=-1*

- ***RandomForestRegressor***
  - *max_depth=10,*
  - *n_jobs=-1,*
  - *random_state=45*

- ***KNeighborsRegressor***
  - *n_jobs=-1*

- ***GradientBoostingRegressor***
  - *max_depth=10*
  - *random_state=45*

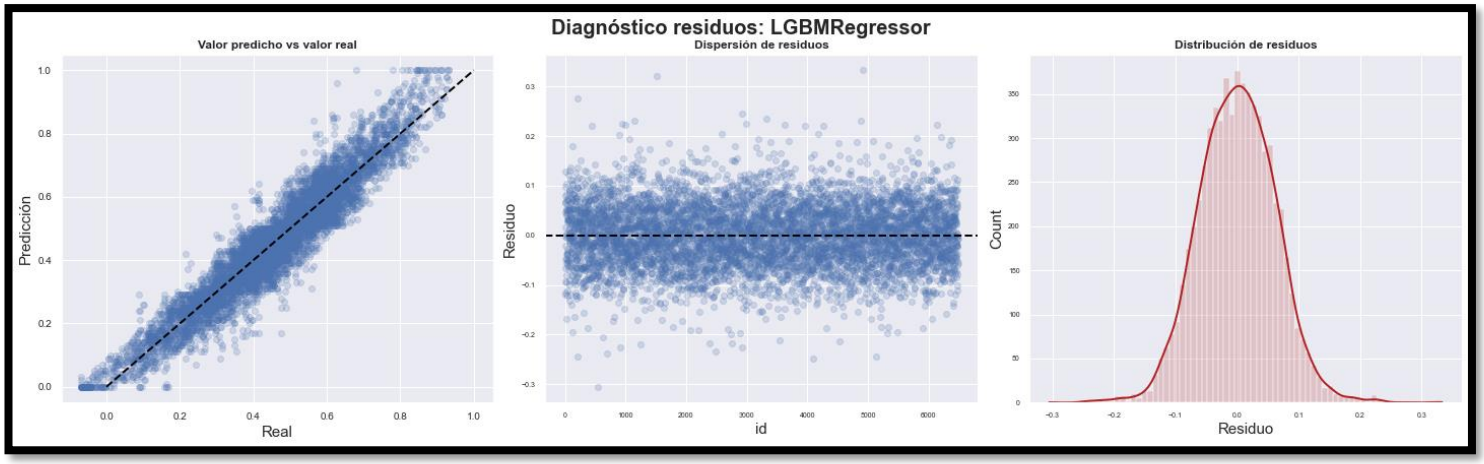
- ***DecisionTreeRegressor***
  - *max_depth=10*

- *random_state=45*

- ***XGBRegressor***
  - *max_depth=6*
  - *learning_rate=0.300000012*
  - *n_estimators=100*
  - *random_state=45*
  - *n_jobs=-1*

- ***LGBMRegressor***
  - *random_state=45*

Gráficos de residuos del mejor modelo según la métrica RMSE.



	Nombre_modelo	R2	RMSE	MAE
6	LGBMRegressor	0.917457	0.059065	0.039289
1	RandomForestRegressor	0.923246	0.059299	0.038805
5	XGBRegressor	0.924773	0.059651	0.039444
4	DecisionTreeRegressor	0.919891	0.060854	0.040000
0	LinearRegression	0.902022	0.061949	0.041659
3	GradientBoostingRegressor	0.935826	0.062826	0.040964
2	KNeighborsRegressor	0.922354	0.064303	0.040000

Notamos que, por cada modelo, las métricas utilizadas tienen baja variabilidad entre ellas.
Como podemos visualizar en la comparación de las métricas el modelo LGBMRegressor tiene el RMSE más bajo, por lo que usaremos este algoritmo para nuestro proyecto.

Para continuar nuestra mejora del modelo aplicamos la técnica de Hipertunnig de parámetros con RandomizedSearchCV a nuestro algoritmo elegido, utilizando los siguientes parámetros globales:

- *num_estimators*
- *max_depth*
- *Learning_rate*
- *random_state=45*

Realizando una validación cruzada de 10 pliegues por 50 iteraciones, un total de 500 entrenamientos. Obtuvimos la mejor RMSE (0,059015) a través de los siguientes parámetros:

- *n_estimators: 566,*
- *max_depth: 10*
- *learning_rate: 0.011*

RMSE = 0,059015.
Best_model_RMSE = 0.058966.

FUTUROS PASOS

- Se desarrollará una aplicación open source para que cualquier persona pueda cargar sus datos para conocer su “Burn Rate” laboral.
- Se ofrecerá la aplicación a empresas o servicios de medicina ocupacional, consultoras de recursos humanos para que pueda evaluar el “Burn Rate” de sus empleados.

CONCLUSIONES

- *Se desarrollo un modelo para predecir el estrés laboral de los empleados con el fin de mejorar la productividad laboral. Creemos que con este modelo cada empresa podrá comparar de manera eficiente sus métricas de productividad laboral y a partir de allí mejorarla, conocer el estado de estrés que tienen sus empleados, y establecer un indicador preventivo para evitar o retardar el desarrollo de Síndrome del Burn Out.*
- *Hubiese sido interesante contar con más variables para enriquecer nuestro modelo.*