**ANÁLISIS PREDICTIVO SOBRE LA DESERCIÓN DE EMPLEADOS**

Elaborado por:

DIANA CATALINA VELÁSQUEZ GAVIRIA

JUAN CAMILO CEBALLOS ARIAS

SANTIAGO ARISTIZABAL TORO

Materia: Estadística Multivariada Avanzada

Profesor: TOMÁS OLARTE HERNÁNDEZ

Maestría en ciencia de datos y analítica

Universidad Eafit

Medellín 2020

**ANTEPROYECTO**

* **Pregunta de investigación y objetivos**

La deserción de personal es un problema que enfrentan la mayoría de organizaciones por los costos que acarrea y los problemas de calidad y servicio que propicia, pues el personal nuevo que ingresa a la compañía no ha estado el tiempo suficiente para tener un desempeño óptimo. Por lo tanto, se hace necesario identificar los factores que causan la deserción y predecir la probabilidad de deserción de los empleados para establecer acciones que la mitiguen.

* *Objetivo General*

El objetivo de este análisis es aplicar diferentes modelos de aprendizaje sobre el problema seleccionado con el fin de predecir la probabilidad de deserción de empleados, utilizando las diferentes técnicas vistas en clase y comprendiendo los conceptos estudiados tales como error real de entrenamiento, error de aprendizaje, regularización, entre otros.

* *Objetivos específicos*
* Realizar un análisis descriptivo de los datos para entender el comportamiento de sus variables previo al modelado.
* Realizar transformación de variables para el proceso de modelado.
* Realizar selección de variables para disminuir la complejidad de los modelos.
* Aplicar modelos de aprendizaje automático para la predicción de la deserción de empleados, empezando por los modelos más simples.
* Escoger un modelo de todos los analizados de acuerdo con los mejores resultados en entrenamiento y test, luego de aplicar regularización para corregir el ajuste si es necesario.
* Hacer conclusiones sobre el análisis.

* **Metodología de Investigación**

Para llevar a cabo el análisis se utilizará la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), que consiste en:

* Comprensión del negocio.
* Estudio y comprensión de los datos.
* Composición de la base de datos (cantidad de variables, total de registros, tipos de variables).
* Análisis descriptivo (variables numéricas y categóricas).
* Selección de variables.
* Modelado.
* Evaluación de resultados.
* Despliegue.
* **Datos y análisis previo**

La base de datos utilizada para llevar a cabo el ejercicio es tomada de Kaggle y consiste en una base ficticia creada por los científicos de IBM. Este dataset consta de 1470 registros y 31 variables, entre las cuales se encuentra una con información referente a si el empleado dejó o no la compañía y las demás características hacen referencia a: rol del empleado en la compañía, información de satisfacción del empleado, información relacionada con el salario mensual, diario, etc. Cambios salariales de un año con respecto al siguiente, información de desempeño y frecuencia de viajes laborales.

Se empieza el análisis y como primer paso, se intenta entender los datos de manera general. Para ello se trabaja en una visualización inicial de los mismos y se resumen las variables:

Figura 1. Conociendo el dataset.

## Imagen que contiene captura de pantalla, pájaro Descripción generada automáticamente

Fuente: creación propia

También se describen las características/variables:

Figura 2. Resumen de las características.

Imagen que contiene captura de pantalla

Descripción generada automáticamente

Fuente: creación propia

Se define que la variable de interés, variable respuesta, es “Attrition” que toma valores de 0 para empleados retirados y 1 para empleados activos y está conformada así:

Figura 3. Visualización variable respuesta: “Attrition”.

Imagen que contiene captura de pantalla

Descripción generada automáticamente

Fuente: creación propia

El 16.12% de los empleados ha dejado la compañía, lo cual corresponde a 237 empleados y el 83.88% permanece en ella equivalente a 1233 empleados.

Posteriormente se realiza un análisis de cada una de las variables de la base con respecto a la variable de interés (attrition), dividiendo el conjunto de datos original en dos subsets, uno para variables numéricas (14 variables) y otro para variables categóricas (16 variables):

* *Análisis de variables numéricas:*

En el conjunto de variables numéricas se analiza el promedio de cada variable para las dos clases de la variable Attrition así:

Figura 4. Promedio de cada variable para las clases de “Attrition”.

Imagen que contiene captura de pantalla

Descripción generada automáticamente

Imagen que contiene captura de pantalla

Descripción generada automáticamente

Fuente: creación propia

Con este análisis se busca entender cómo varía el promedio tanto para empleados retirados como para los activos e identificar las variables donde se presenta la mayor diferencia a simple vista, representadas con un círculo naranja.

Además, para complementar el entendimiento se realizan boxplots comparativos que permiten observar la distribución y variabilidad de los datos en cada categoría de Attrition.

Figura 5. Boxplot de cada variable para las clases de “Attrition”.

Imagen que contiene interior, pared

Descripción generada automáticamente

Fuente: creación propia

En los boxplots se pueden visualizar de mejor manera aquellas variables donde pareciera haber diferencia entre ambas categorías de Attrition, lo cual da una idea inicial sobre qué variables pueden ser de interés para explicar la deserción de los empleados.

Continuando con el análisis numérico se realiza un histograma de dichas características de forma univariada para comprender sus distribuciones y se encuentra que algunas variables como DailyRate, HourlyRate, MontlyRate presentan distribuciones muy uniformes.

Finalmente se realiza una matriz de correlación para estas variables y se encuentra que las que tienen mayores coeficientes de correlación son las que se presentan a continuación resaltadas con color naranja, lo cual tiene mucho sentido. Por ejemplo, mientras más años en la compañía, existen más ingresos, más años en el rol actual, más años desde el último ascenso o con el jefe actual:

Figura 5. Matriz de correlación – Variables numéricas.Imagen que contiene captura de pantalla

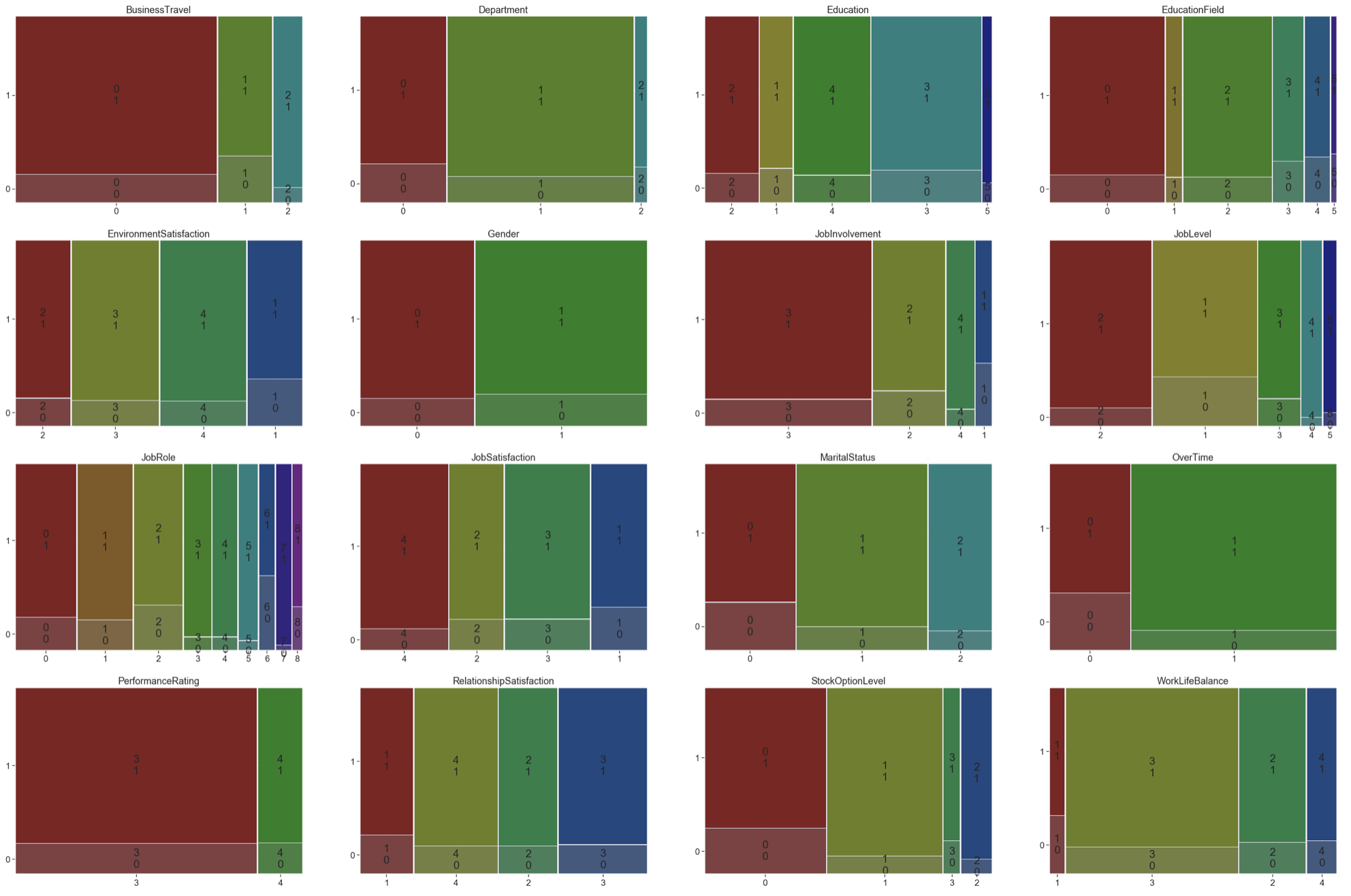
Descripción generada automáticamente

Fuente: creación propia

* *Análisis de variables categóricas*

De la misma manera se realiza un análisis sobre las variables categóricas con respecto a la variable “Attrition” para entender como varían las proporciones de cada categoría entre los niveles de la variable respuesta. Para llevar a cabo este análisis se construyen mosaic plots y tablas de contingencia, para identificar aquellas características donde dichas proporciones parecen diferentes en empleados activos y retirados:

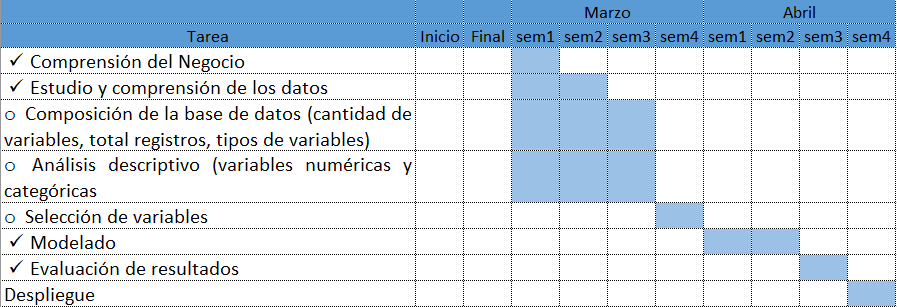
Figura 6. Mosaic plots – Variables categóricas.

Señal de pulgar hacia arriba Señal de pulgar hacia arriba Señal de pulgar hacia arriba Señal de pulgar hacia arriba Señal de pulgar hacia arriba Señal de pulgar hacia arriba Señal de pulgar hacia arriba Señal de pulgar hacia arriba Señal de pulgar hacia arriba Señal de pulgar hacia arriba Señal de pulgar hacia arriba Señal de pulgar hacia arriba 

Fuente: creación propia

Se resaltan con una mano aquellas variables donde parece tenerse diferencia en los niveles. Las tablas de contingencia presentan resultados similares.

* **Plan (diagrama de gantt)**



* **Implicaciones éticas**

El mayor problema que puede afrontar esta clase de estudios está relacionado con la información del empleado. No es claro qué tan ético es que una empresa tenga acceso a cierta información personal del mismo. No es claro si es ético si a la hora de contratar un candidato con características personales con un riesgo alto de deserción laboral como vivir lejos del trabajo o estar casado, deba ser motivo para descartar a un candidato o no. Por ejemplo, si determinado género o determinado grupo de edad resultan en una mayor probabilidad de rotación, ¿es ético descartar al candidato que pertenezca al grupo de alta probabilidad, aunque aparentemente sea bueno y cumpla con los requisitos para trabajar? Las empresas deben analizarlo y tomar la decisión.

Cuando se tenga implementado el modelo, es posible que haya que hacer algunos ajustes en la empresa como reducir o aumentar el número de horas trabajadas, cambiar el cargo, poner más o menos supervisores, ajustar salarios. Una gran implicación podría ser que, empleados con un nivel educativo mayor o personas con más años en la empresa, tengan mayor probabilidad de rotar. En este caso, para reducir la deserción, la empresa puede optar por cambiar los empleados con determinados años de trabajo o con un alto nivel educativo, lo cuál podría hacer incluso que el personal cuando cumpla cierta cantidad de tiempo en la compañía, rote debido a esta política. Dependiendo de las variables significativas del modelo y de las políticas de la empresa con estas variables, podría haber muchos o pocos cambios.

* **Aspectos legales y comerciales**

Es posible que estos modelos se limiten en alguna información personal dependiendo del marco jurídico del país o la región. Por ejemplo, en algunas regiones puede ser ilegal saber dónde vive el empleado, su estado civil o el número de hijos que tiene, lo que podría limitar un estudio de este tipo. En cuanto a la exposición de los resultados, las compañías implementando esta clase de modelos deben ser cautelosas, pues intervienen en el recurso humano, un tema sensible, pues existe la posibilidad de que, si un empleado se entera de que es catalogado como alta probabilidad de rotación, es posible que su rendimiento cambie, que deserte sin haberlo pensado o incluso en algunos casos podría demandar a la compañía. El potencial del modelo radica en mantener empleados estables a lo largo del tiempo, esto como se mencionó anteriormente, puede aumentar la productividad y rentabilidad de la empresa.

* **Bibliografía**

Kaggle. (2018). Data files used for models (Versión 2) [Archivo de datos]. Recuperado de https://www.kaggle.com/carmelgafa/data-files-used-for-models