Brahayan Stiven Gil Henao

Jhon Alexander Bedoya Carvajal

Sebastián Salamanca Mendez

Ana María Ospina Arredondo.

INFORME RESULTADOS CASO ESTUDIO PREDICCIÓN DE ABANDONO.

1.

a. Diseño de solución propuesto:

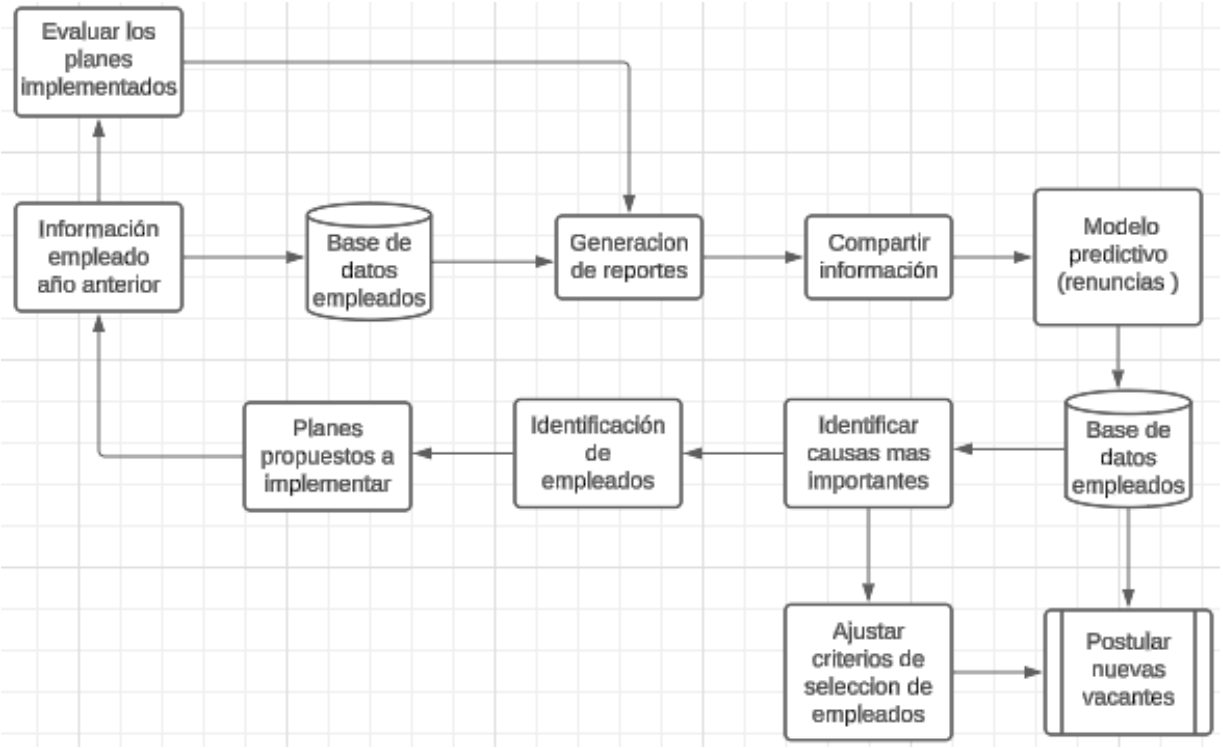

Imagen 1

El modelo presentado en la imagen anterior se ha diseñado con el propósito específico de prever la probabilidad de deserción de empleados en una organización. Basado en datos históricos de empleados, este modelo utiliza técnicas de aprendizaje supervisado automático para identificar patrones y relaciones entre las variables, el modelo funcionará de la siguiente, manera:

- Se entrena un modelo supervisado de clasificación que predice si un empleado va a renunciar o no a su puesto de trabajo el año siguiente. Esto es modificable según las evaluaciones que se vean con las acciones tomadas.

- Las predicciones generadas por el modelo se usarán para determinar los departamentos en los que se prevé que habrá más renuncias y tomar acciones preventivas.

- Las predicciones de renuncias se dejan en una base de datos a la que tienen acceso las personas de recursos humanos para utilizar dicha intervención con las áreas o directamente con los empleados que de prevé van a renunciar.

- Se genera una lista de las variables que más influyeron en las renuncias de los empleados y se presentan a recursos humanos para que generen estrategias que permitan mejorar estas variables y así disminuir la tasa de renuncias. Además, estas variables serán tomadas en cuenta como aspectos importantes en los criterios de selección de personal.

- La información utilizada para la predicción es la que se tiene disponible al cierre del año anterior de los empleados y última encuesta de satisfacción y desempeño.

- El modelo se re-entrena cada año, en cuanto se tengan los nuevos resultados de las encuestas de satisfacción y desempeño. Predicción anual

b. Limpieza y transformación:

Para implementar la solución de la problemática se contaba con 2 bases de datos las cuales fueron unidas en una sola base usando la herramienta SQL donde se exploran las variables categóricas y numéricas para tener conocimiento de su información, se pasa a una limpieza breve donde se realizó el manejo de valores nulos y asignación de categorías descriptivas para algunas variables, al final usando la herramienta de SQL se decide separar los datos excluyendo a las personas que renunciaron en 2015 ya que no son relevantes para el modelo. Se exporta el dataset y se lleva a visual studio el cual permitirá realizar un análisis más exhaustivo a la base de datos, allí se analiza la correlación entre variables y se decide eliminar 3 variables, tartar valores atípicos y transformación de las variables.

c. Análisis exploratorio:

Para este análisis se realizaron gráficos de torta, barra, boxplot, y gráficos de frecuencia para cada variable donde se pueda observar donde la variable objetivo tiene mayor predominancia.

d. Selección de algoritmos y técnicas de modelado:

Para abordar este problema de clasificación binaria de retiro de empleados, se seleccionaron algoritmos con alta interpretabilidad para comprender el proceso de toma de decisiones y la importancia de las variables. para ello se utilizarán:

- Árboles de decisión: Representan decisiones tomadas en forma de árbol. Fáciles para visualizar y entender. Brindan información sobre la importancia relativa de las características.
- Regresión logística:Ofrece los coeficientes de regresión para cada variable predictora. Esto significa que nos indica qué tanta magnitud de relación hay y la dirección de

cada relación con la probabilidad de la que se retire. Un coeficiente positivo aumenta la probabilidad de retiro y un coeficiente negativo la disminuye.

Además, se evaluarán algoritmos más complejos como Random Forest y Extreme Gradient Boosting. Aunque menos interpretables, pueden tener mejor rendimiento. Se comparará su desempeño con los modelos más simples. Si la diferencia no es significativa, se preferirán los modelos simples por su interpretación. En caso contrario, se continuará con los modelos más complejos.

e. Selección de variables: Se usan modelos de selección de variables y técnicas de selección con modelos para tener una mayor comprensión de las variables y priorizar las características más relevantes para el análisis.

f. Comparación y selección de técnicas: Después de haber realizado la selección ganadora en el punto anterior, se prueban los 4 modelos planteados y se realiza un análisis que arroja como resultado que el modelo ideal es el árbol de decisión tanto por sus métricas como por su simplicidad a la hora de la interpretación de los datos.

g. Afinamiento de hiper parámetros: Se realizaron múltiples interacciones en la búsqueda del modelo más sencillo y de mejor desempeño, ajustando principalmente la profundidad del árbol.

h. Evaluación y análisis del modelo: Para la evaluación del modelo se usa la curva roc, demostrando que el modelo predice mucho mejor los resultados que un clasificador aleatorio. El análisis del modelo se hace siguiendo un caso específico durante su trayecto en el árbol de decisión, en el árbol cuando se cumple con el criterio se va hacia la izquierda y si no se cumple se va hacia la derecha hasta llegar a un nodo final con la decisión de abandonar la empresa o no .

- ★ En el caso específico analizado, la primera decisión se centra en determinar si la edad de la persona es inferior a 33.5 años. Dado que la persona tenía 29 años, esta condición se cumple, llevándola al siguiente nodo hacia la izquierda.
- ★ En la siguiente evaluación, se analiza si la persona ha trabajado menos de 3.5 años en la compañía. Dado que ha trabajado 4 años, esta condición no se cumple, lo que la dirige al siguiente nodo hacia la derecha.Si hubiera cumplido la condición avanzaría al siguiente módulo donde se evalúa si ha sido entrenado menos de 1.5 veces y si la cumplia se quedaba en la compañía, como no se puede trabajar menos años se puede suponer cual es la diferencia con una persona que haya trabajado menos años, la cual podría ser que la persona ya está cansada en la compañía, por lo tanto se recomienda hacerle un acompañamiento psicológico.
- ★ Prosiguiendo, se verifica si la persona no está soltera. Al no cumplirse esta condición, se avanza hacia el nodo siguiente en la dirección derecha.
- ★ Luego, se examina si la persona recibió menos de 3.5 entrenamientos el año pasado. Al cumplirse esta condición, se procede hacia la izquierda. En caso de que el

trabajador haya sido entrenado 4 veces en total, avanzaría al siguiente nodo, que evalúa si su edad es mayor a 28.5 años, y al cumplirse, permanecería en la compañía.

- ★ La condición siguiente implica que la satisfacción laboral no esté clasificada como "muy alto". Al no cumplirse esta condición, se continúa hacia el nodo de la derecha.
- ★ Posteriormente, se considera si el aumento porcentual del salario fue menor al 14.5%. Al no cumplirse esta condición, se sigue hacia el nodo de la derecha.

- ★ La siguiente evaluación es si la persona tiene menos de 7.5 años en la compañía. Al cumplirse esta condición, se procede con el nodo hacia la izquierda.
- ★ Luego, se analiza si el salario es inferior a 24740. Al cumplirse esta condición, la persona decide abandonar la empresa.

  Sin embargo, si la persona aumenta su salario a más de 24740, avanzaría al siguiente nodo que evalúa si la distancia desde su casa es menor a 11km. Al cumplirse esta condición, optaría por quedarse en la compañía.

  Este análisis se puede aplicar a cada uno de los trabajadores pronosticados para retirarse de la compañía, permitiendo la implementación de acciones correctivas sugeridas anteriormente para modificar los resultados.

i. **Despliegue del modelo:** Se seleccionan las columnas para crear un nuevo dataframe

Se realiza el preprocesamiento de los datos para tener una mayor facilidad y entendimiento de lectura, asignando las categorías a cada una de las variables categóricas y tratando los datos faltantes con la moda. Se definieron los límites de los outliers de las variables numéricas para realizar correctamente el tratamiento de los datos faltantes. Finalmente obtenemos las dummies, se hacen las predicciones con los nuevos datos y las variables seleccionadas, al final se entregan las predicciones en el orden de la probabilidad de que un trabajador abandone junto con la ruta que sigue en el árbol de decisión y los datos correspondientes a las variables seleccionadas para el análisis.

Arbol:https://drive.google.com/file/d/1o1AwrMFHaQKAtaXJzLzPPwQ8yRqJGeQA/view?usp=sharing