INFORME CASO DE NEGOCIO

ELABORADO POR:

ANDRÉS FELIPE PENNA HERNÁNDEZ

JHONATAN VALENCIA OCAMPO

STEVEN ZAPATA ZULETA

PROFESOR:

JUAN CAMILO ESPAÑA LOPERA

ANALÍTICA



UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE INGENIERÍA

DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

1. **Diseño de solución propuesto.**

Problema de negocio

De acuerdo al contexto de la empresa, se logra identificar una situación indeseada en una organización, en donde, existe una alta tasa de deserción (15% anual) de empleados. Lo que puede desencadenar en una serie de efectos adversos tales como: Bajo rendimiento y altos costos.

Con base a lo anterior, el grupo traza el objetivo de disminuir la tasa de deserción por parte de los trabajadores anualmente.

Problema analítico

Se establece tres focos desde el apartado de analítica, los cuales son:

1. Definir las variables que son relevantes para diseñar un modelo que permita predecir el perfil de riesgo de deserción de un aspirante en el proceso de selección.
2. Comprender qué factores pueden hacer que un aspirante presente tendencias a la deserción a futuro.
3. Realizar modelo que prediga si el perfil del aspirante presenta riesgos de deserción con el fin de apoyar la toma de decisiones en el proceso de selección.

Marco

De acuerdo al contexto de la empresa, el equipo de analítica evaluó la situación desde múltiples puntos de vista con el objetivo de determinar cómo abordar la problemática, entre las posibilidades estudiadas se opta por plantear un modelo que permita predecir en el proceso de selección si una persona que aspira a ingresar a la empresa presenta un perfil con alto riesgo de desertar.

La propuesta busca apoyar al departamento de RRHH, brindándoles una herramienta que agilice las primeras etapas de la selección de personal y asegure la confiabilidad de la permanencia del personal con base en la identificación de factores que pueden influir en la deserción. Se focalizan los esfuerzos en dicho proceso ya que el modelo se basa principalmente de datos precisos reduciendo la incertidumbre y evitando la subjetividad de cierta información, además, de poder identificar posteriores etapas que ayudan a reducir la problemática por medio de la generación de otras estrategias.

Adicionalmente, según Portafolio en su artículo “El costo de la alta rotación de personal para las empresas” enfatiza en la importancia de una buena selección de personal, puesto que desde un inicio se puede ahorrar en costos que se incurran como lo es la capacitación, asimismo es común que las personas deserten a corto plazo. Por lo anterior la solución que se plantea puede ser vital para la empresa en su objetivo de reducir la alta rotación.

En el apartado de anexos se muestra el diagrama de procesos donde se plantea la interacción del modelo con el departamento de recursos humanos.

**b. Limpieza y transformación**

En el proceso de limpieza y transformación de datos, se logra que los datos contenidos sean consistentes, faciliten su interpretación y predicciones. En este proceso se eliminan datos nulos, duplicados o incoherencias que puedan afectar los resultados de estudio. Los datos que se presentan como nulos son reemplazados por el promedio del total de cada característica, se rellenan con 0 *fillna* (0) o con el número previo más próximo *ffill*, todo esto se realiza según algunas características propias y lógica aplicada para cada variable.

De acuerdo con el enfoque del caso de estudio basado en la deserción, se evalúa la importancia de algunos datos presentes en la base que no tienen relación con el objetivo final, por lo tanto, se eliminan.

**c. Análisis exploratorio**

En la exploración de las bases de datos, se analiza la composición de cada una para determinar que variables en primera instancia podrían ser de valor para el modelo, se tiene en cuenta el análisis estadístico y las categorías de las que consta cada una de estas variables.

Se presenta un análisis gráfico que facilite la comprensión de la relación entre las variables explicativas y la variable respuesta (*Target*), debido al planteamiento del diseño de la solución la exploración se limita a variables que puedan captarse en el proceso de selección de personal. Entre las variables se dispone de características de cada aspirante como el género, edad, estado civil y formación; por otra parte también se analizan variables que captan aspiraciones y trayectoria laboral.

Para algunas relaciones interpretadas se estandarizan los datos con el objetivo de evitar sesgos por la composición de la población. A partir de los hallazgos se infiere que los datos poseen valor, ya que existen patrones en las variables que pueden explicar la deserción, por ejemplo, las personas con menor edad son más inestables esto se puede atribuir a sus aspiraciones salariales o profesionales, otro claro ejemplo es la relación entre el nivel de responsabilidad al interior de la empresa y los salarios que dejan ver que estas son claves para los trabajadores. Al finalizar esta etapa se puede dar continuidad con una orientación más clara hacia donde dirigir los esfuerzos.

**d. Selección de algoritmos y técnicas de modelado**

La transformación de variables categóricas a numéricas binarias es necesaria para realizar un modelo de aprendizaje automático, por tal razón, una vez se hallan las variables de interés que son categóricas son transformadas por medio de la función *get\_dummies* a excepción de la variable *Target* y el identificador del trabajador (Id). Además, es importante mejorar el comportamiento y precisión en la predicción del modelo, con ayuda del escalamiento de features, dado que convierte los datos en términos comparables evitando sesgos, se escalan características con el fin de estandarizarlas mediante el método *StandardScaler*.

**e. Selección de variables**

En la selección de variables se utilizó un método llamado *SelectFromModel* para determinar las variables más importantes del conjunto de datos, este método es aplicado a los modelos *Decision Tree*, *Random Forest* y *Gradient Boosting Classifier*, los cuales se evaluarán más adelante para elegir el de mejor desempeño. Posteriormente en el *SelectFromModel* se define como umbral el valor de 1, lo que significa que se van a seleccionar todas las variables que tengan una importancia igual o por encima de 1, luego se definen como variables finales del modelo aquellas que fueron arrojadas por el método en cada evaluación de los 3 modelos. Estas variables permiten al modelo reducir el ruido, el costo de procesamiento y aumentar la capacidad de predicción.

**f. Comparación y selección de técnicas**

Se inicia la selección de algoritmos de acuerdo a el esquema de *Sklearn*, pero debido al desbalance de los datos, modelos como *SVC* y *K-Neighbors* presentan problemas en las predicciones por lo que se optó por modelos que ofrecieran mayor robustez, mejorando la precisión en casos de desbalance aunque esto también implica complejidad en la interpretación. En este caso se evalúan los modelos *Decision Tree*, *Random Forest* y *Gradient Boosting Classifier.*

Se realiza la partición de los datos en entrenamiento y prueba, posteriormente se entrena cada modelo con sus respectivos parámetros para finalmente realizar la predicción del modelo. Así luego, poder seleccionar el modelos a trabajar, donde se requiere una evaluación de desempeño que defina la precisión del modelo por medio de las pruebas *Cross Validation* y *K Fold*, la matriz de confusión y el reporte de clasificación. Métricas que al ser comparadas soportan la decisión del modelo a seleccionar.

Los resultados arrojados en cada uno de los modelos, usando los features seleccionados anteriormente para predecir los aspirantes que pueden desertar, se presentan a continuación:

| **Modelo** | **Accuracy score** | **Matriz de confusión** |
| --- | --- | --- |
| ***Decision Tree*** | 0.9505 |  |
| ***Random Forest*** | 0.9618 |  |
| ***Gradient Boosting Classifier*** | 0.8900 |  |

Con base en los resultados, el modelo de *Random Forest* es el que obtiene un mejor desempeño, sin embargo, el modelo *Decision Tree* tiene métricas cercanas en precisión y la matriz de confusión permite identificar que este modelo predice mejor los aspirante que pueden desertar, por lo que se selecciona este segundo modelo. Además de que es un modelo de aprendizaje individual que representa mayor sencillez en interpretación y menor costo computacional a diferencia de los modelos de ensamble como *Random Forest.*

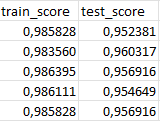
**g. Afinamiento de hiper-parámetros**

Después de seleccionar el modelo, se somete a un afinamiento de hiper-parámetros con la intención de mejorar el desempeño. Se analizan los posibles parámetros a evaluar, seleccionado así: *max depth, min samples split, min samples leaf, max features y criterion*.

Luego se usa la técnica búsqueda rejilla, después de obtener la mejor combinación de hiper-parámetros se guarda el modelo con este ajuste para realizar nuevamente predicciones.

**h. Evaluación y análisis del modelo**

Una vez realizados los modelos con afinamiento de hiper-parámetros, se aplica el método de *Classification Report* para efectuar una comparativa respecto al modelo sin el afinamiento, se obtuvo un incipiente mejor desempeño con el afinamiento de hiper-parámetros. La proporción de datos clasificados de forma correcta como verdaderos es mayor en el dato de interés (desertor), con un 0,79 ( *Recall* en hiper-parámetros) y 0,77 (*Recall* sin hiper-parámetros). En resumen, la comparativa de estos reportes de clasificación, establecen en concreto los mejores desempeños y elementos para identificar problemas de sesgo en las predicciones.

En relación con el rendimiento del modelo con afinamiento de hiper-parámetros, se divide el modelo en subconjuntos mediante la técnica de *Cross-Validation* donde se evalúa y entrena el modelo en diferentes combinaciones de conjuntos, estos resultados o puntuaciones de entrenamiento *Train\_score* y de prueba *Test\_score* son bastante similares como se observa la siguiente imágen: 

Por medio de esto se puede deducir que el modelo no está sobre-ajustado, además las puntuaciones son altas, alrededor de 95% , por dicha razón el modelo funciona correctamente.

**i. Despliegue del modelo**

El objetivo del proyecto de analítica comprende el predecir a los aspirantes que tienen riesgo de deserción e identificar las variables que más influyen en esa deserción. Con esto en mente, se genera un ejemplo en donde se utiliza una base de datos que representa los nuevos aspirantes y predecimos el riesgo de deserción a partir de estos datos. Después de estas predicciones se filtran los aspirantes que resultaron con riesgo de deserción y finalmente se entregan estos resultados a los encargados de evaluar a los aspirantes en el proceso de selección, esto mediante una tabla de Excel. Los resultados se guardan en una tabla en excel, esta información se suministra con el objetivo de que apoye la toma de decisiones sobre los aspirantes.

Por otro lado, se identifican las variables más importantes que arroja el modelo de árbol de decisión que indica los aspectos que influyen en mayor proporción en la deserción de un aspirante, esta información se suministra al departamento de recursos humanos para enfocar los esfuerzos en disminuir la influencia de aquellos factores que representan más riesgo.

Es importante destacar que se automatiza el proceso de predicción, facilitando la manipulación de los nuevos datos y simplificando la ejecución del modelo para generar la predicción cada año.

**j. Conclusiones**

1. Es importante partir del diseño de la solución para fijar un objetivo y comprender qué elementos son importantes a la hora de realizar un modelo predictivo, además, define implícitamente el por qué y para qué es necesario llevar este caso a la analítica.

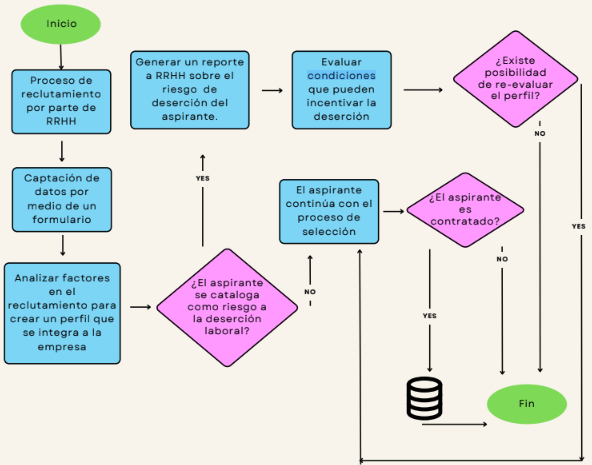
1. El perfilamiento de personal a la hora de realizar reclutamiento, es importante ya que parte desde las necesidades y características necesarias para una labor en la empresa. Por esto, se le da gran valor de importancia al obtener estos datos en este modelo.
2. Se logra identificar los factores más influyentes que aumentan el riesgo de deserción en un aspirante y construir un modelo que obtiene un desempeño efectivo al predecir la deserción de los mismos.

**Bibliografía:**

Portafolio. (2016). *El costo de la alta rotación de personal para las empresas*. Extraído de:

<https://www.portafolio.co/economia/empleo/costos-de-la-alta-rotacion-de-personal-en-las-empresas-502333>

**Anexo:**

****