Trabajo Final de Análisis de Datos sobre Empleados y Herramientas de ML.

Franco, Chiabo

UNRaf, Universidad Nacional de Rafaela, Santa Fe, Argentina.
Aprendizaje Automático y Grandes Datos (IC)

Autor Corresp.: fchiabo2017@gmail.com

Resumen:

Este estudio aplica técnicas de aprendizaje automático (ML) para desarrollar modelos predictivos sobre ingresos mensuales y satisfacción laboral en empleados. Se analizan datos de diversas variables laborales y demográficas para identificar factores clave que influyen en estas métricas. Se realizó un análisis exploratorio de datos utilizando herramientas como Pandas, Matplotlib y Seaborn. Se limpió y concatenó información de dos datasets, aplicando normalización y estandarización. Para predicción y clasificación, se usaron algoritmos como Random Forest, KNN y Regresión Lineal, además de técnicas de selección de características y ajuste de hiperparámetros. Random Forest mostró el mejor desempeño predictivo, con un coeficiente R^2 de 0.9011 y un MAE de 0.0316 en predicción de ingresos. En clasificación, alcanzó una precisión del 68% utilizando todas las variables.

El análisis detalló patrones significativos en deserción laboral, distribución salarial y factores demográficos. Los hallazgos confirman la relevancia de variables como nivel educativo y rol laboral en predicciones salariales, alineándose con estudios previos. La falta de mejora significativa al ajustar hiperparámetros sugiere un posible sobreajuste. Se recomienda explorar variables adicionales y enfoques integrados para mejorar la precisión y aplicabilidad de los modelos.

Palabras claves: Aprendizaje Automático, Ingreso Mensual, Satisfacción Laboral, Análisis de Datos, Algoritmos de Clasificación, Algoritmos de Predicción.

1. Introducción

En el panorama empresarial actual, la capacidad de predecir el ingreso mensual y la satisfacción laboral de los empleados es crucial para la toma de decisiones estratégicas, impactando directamente en la productividad, la retención del talento y el éxito organizacional. Este trabajo se enfoca en el análisis de un conjunto de datos de empleados mediante técnicas de aprendizaje automático, como la regresión y la clasificación, para desarrollar modelos predictivos que permitan estimar estas variables clave. Se explorarán características como el nivel de educación, el rol de trabajo, el tamaño de la empresa y otras variables relevantes, con el fin de identificar los factores más influyentes en la predicción y obtener modelos robustos y precisos que faciliten la toma de decisiones informadas en la gestión del talento. Este análisis se basa en el conocimiento actual sobre modelos predictivos, aprovechando la disponibilidad de grandes conjuntos de datos y algoritmos avanzados para mejorar la precisión y eficiencia en la estimación del ingreso mensual y la satisfacción laboral de los empleados.

Este análisis se basa en el conocimiento actual sobre modelos predictivos y utiliza grandes conjuntos de datos y algoritmos sofisticados para aumentar la precisión y eficiencia de la estimación de ingresos mensuales y la satisfacción laboral de los empleados. Numerosos estudios anteriores han abordado la predicción de estas variables utilizando diversos enfoques y estrategias. Por ejemplo, en State of the Art of Job Satisfaction Measures[1] se realiza una revisión sistemática de las medidas de satisfacción laboral, mientras que en Examination of the factors that predict job satisfaction[2] se examinan los factores que predicen la satisfacción laboral entre los empleados en el sector de tecnología de la información. Además, en The Efficient Measurement of Job Satisfaction[3] se propone un método eficaz para medir la satisfacción laboral, y en Job Satisfaction Indicators and Their Correlate[4] se examinan los indicadores de satisfacción laboral y sus correlatos. En conclusión, Income as a Predictor of Employe Job Satisfaction and Motivation[5] examina el ingreso como predictor de la satisfacción laboral y la motivación de los empleados.

Aunque estos estudios anteriores han aportado conocimientos valiosos sobre la predicción de los ingresos y la satisfacción laboral, existe una laguna en el uso de un enfoque integrado que combine numerosos algoritmos de aprendizaje automático y una amplia gama de factores predictivos. Este estudio tiene como objetivo llenar esta brecha comparando varios modelos de regresión y clasificación y evaluando qué tan bien predicen los ingresos mensuales y la

satisfacción laboral utilizando un conjunto completo de datos que incluye variables demográficas, laborales y de satisfacción. Además, se investigará el impacto de la selección de características en la precisión del modelo y se examinará la interpretabilidad de los hallazgos para facilitar la implementación práctica. El objetivo principal de esta investigación es desarrollar modelos predictivos precisos y robustos para el ingreso mensual y la satisfacción laboral de los empleados, utilizando algoritmos de aprendizaje automático. Para ello, se ha estructurado el trabajo en las siguientes secciones: Sección 2 - Metodología, donde se describe el conjunto de datos, el preprocesamiento de datos y los algoritmos utilizados; Sección 3 - Resultados de la predicción del ingreso mensual, donde se presentan los resultados de los modelos de regresión; Sección 4 - Resultados de la clasificación de la satisfacción laboral, donde se presentan los resultados de los modelos de clasificación; y Sección 5 - Conclusiones, donde se resumen los hallazgos principales y se discuten las implicaciones prácticas de la investigación.

2. Metodología

Análisis Exploratorio de Datos (EDA)

En la realización del Análisis Exploratorio de Datos (EDA) hemos determinado como interés estudiar la deserción laboral en las empresas y diversos factores que estén relacionados a esta o que influyan en la vida del trabajador. Para poder llevarla a cabo se seleccionó un dataset en la página de Kaggler, que contenía un total de 24 variables y 59598 registros de diferentes empleados. En el análisis que pretendemos hacer usaremos la librería Pandas en Python, esta nos permitirá manipular y analizar los datos, brindándonos estructuras y operaciones para su manipulación. También nos permite importar el conjunto de datos de diferentes medios, en nuestro caso desde un archivo CSV.

Detección de anomalías

Este conjunto de datos fue analizado en primer lugar en base a incongruencias presentes que pudimos detectar, tales como la edad de la persona con respecto a los años de antigüedad en la empresa, estos indicaban que los empleados iniciaron sus actividades teniendo menos de 16 años, lo cual es imposibles o al menos no es permitido legalmente. También se realizó una segunda depuración que involucra las variables anteriormente mencionadas, pero ahora con un límite de 18 años y el dato Healthcare (Cuidado de la Salud) de la variable Rol de Trabajo. Definimos en que un trabajador del área de salud no puede iniciar con menos de 18 años en el ámbito laboral ya que, desde nuestro entendimiento, no cuenta con los conocimientos básico requeridos que te puede brindar una carrera terciaria o universitaria específica del área, debido a que no es una rama que pueda ser estudiada de forma autodidacta. Por esa razón se decidió contemplarlos con el límite que se menciona para que se pueda comprender su incorporación como pasantes con al menos un año de estudios y con los saberes necesarias.

Limpieza de datos

Los filtros que establecimos fueron guardados en variables que luego, mediante el comando .drop(), procedimos a eliminar del dataset. Esta acción eliminó un total de 20145 registros, conformando un 33,8% de los datos que teníamos en nuestro poder.

Concatenación del segundo Dataset

En nuestra perspectiva, con la limpieza realizada hemos perdido una cantidad de información que no estábamos dispuestos a dejar de lado, por esa razón, se implantó el uso de un segundo dataset que luego de ser reestructurado y limpiado se lo desea concatenar en una unión con el original para así recuperar y ganar más registros que al inicio del estudio.

El nuevo conjunto de datos fue extraído nuevamente de la página de Kaggler y está compuesto de información sobre empleados de la empresa IBM y cuenta con 37 variables y 23436 registros. Como primer paso se eliminaron las variables que no son útiles y que no están presentes en el primer dataset, el comando utilizado es .drop() al igual que para la eliminación de registros. Luego se renombran las variables, comando .rename(), para que coincidan con las otras y se crean las variables faltantes para así poder ordenarlas al igual que el conjunto original. A continuación, se reemplazan los valores de las variables mediante el uso de .replace() y se eliminan los pocos registros con valores nulos para de esta forma tener un dataset completo y sin carencia de información. Terminando se realiza un cambio en los tipos de datos de cada columna ya que si difieren de las del original la concatenación no será posible. Para las variables de tipo cadena que queremos convertir en numéricas usamos la función pd.to_numeric() y para los valores float a int usamos .astype(int).

Finalmente ya se puede llevar a cabo la unión de los dos dataset y para eso usamos la sentencia dfconcatenado = pd.concat([df, df2], axis=0, ignore_index=True), y de esta forma nuestro dataset final es dfconcatenado que está formado por 24 variables y 62768 registros.

Gráficos

Parte del análisis que se contempla no solo es mediante número y porcentajes de las variables y sus datos, también se utilizan diferentes tipos de gráficos ya que de esta forma se puede apreciar mejor la información y genera una vista mucho más amigable para quienes desean leer los resultados del estudio.

Para esta instancia se utilizaron las librerías de matplotlib y seaborn, con la primera se puede generar una gran variedad de gráficos sin requerir muchas

líneas de código y seaborn apuesta por la visualización de datos estadísticos para entender mucho mejor los datos expuestos, sirve para hacer gráficos de datos estáticos de series temporales, elimina la superstición de gráficos y ayuda en su embellecimiento.

Mediante el uso de las librerías anteriores se confeccionaron gráficos de barras para las variables de tipo object e histogramas para las de tipo int además de diagrama de caja que nos permite ver si hay outliers presentes.

Normalización y Estandarización de datos

Para poder utilizar los datos que tenemos en el entrenamiento y prueba de un algoritmo se recomienda que los valores estén normalizados o estandarizados para mejor el rendimiento del modelo seleccionado, el tipo de tratado dependerá de cual dará mejor resultados, pero siempre es mejor probar con ambos métodos. La normalización consiste en reescalar los datos para que se encuentren en un rango específico, en nuestro caso entre -1 y 1. La estandarización reescala los datos para que tengan una media de 0 y una desviación estándar de 1, generando una distribución con una forma estándar. Estas técnicas de preprocesamiento se realizan en variables de tipo int64 por lo que primero hay que crear un diccionario donde se le asigne un valor numérico a cada clase de las variables de tipo object. Luego mediante la librería sklearn importamos MinMaxScaler para normalizar los datos y StandardScaler para el estandarizado de los mismos.

Algoritmos de predicción

Este análisis busca identificar cuál Algoritmo de Predicción es el más adecuado para nuestro modelo. Para ello realizamos una comparativa de métricas entre tres algoritmos que seleccionamos, Regresión Lineal, Regresión de Bosque Aleatorio y K Vecinos Más Cercanos (KNN). También se pensó en realizar el análisis con diferentes variables. El primer caso es mediante el Análisis de Componentes Principales (PCA), un método no paramétrico de reducción de dimensiones. PCA examina las variaciones en los datos, considerando a los atributos como portadores de información en forma de variaciones numéricas. Se seleccionan los componentes principales que explican la mayor parte de la varianza, y los datos se transforman a un nuevo espacio con menor dimensionalidad, conservando la mayor cantidad de información posible. Para hacerlo importamos PCA de la librería sklearn y definimos la cantidad de columnas, variables, que deseamos tener. Luego dividimos los datos en dos, el 80% se destina para el entrenamiento y el otro 20% es para la realización de las pruebas. Para lograr este fraccionamiento importamos train_test_split y definimos las variables de entrenamiento y prueba con la siguiente sentencia, X_trainPCA, X_testPCA, Y_trainPCA, Y_testPCA = train_test_split(PCA_df, y, test_size=0.2, random_state=2). Para el entrenamiento de los algoritmos se importan de sklearn el modelo a utilizar, LinearRegression, RandomForestRegressor y KNeighborsRegressor, luego con la función .fit(), se le pasan las variables de entrenamiento y luego con .predict() predice para nuestro conjunto de entrenamiento. Pero para saber si obtenemos buenos resultados tendremos que utilizar métricas que nos indicarán que tan acertada será la respuesta. Para eso importamos metrics en la librería que nos permitirá medir con R2, Mean Absolute Error (MAE) y Root Mean Squared Error (RMSE) si el algoritmo nos brindará buenos valores de salida. R2 indica qué tan bien se ajusta el modelo al dataset, mientras más será de 1 se encuentre mejor será, MAE mide el error promedio absoluto, es bajo o alto dependiendo del contexto y de la escala de la variable y RMSE es útil para evaluar la magnitud del error. Cuando las métricas de rendimiento del entrenamiento son muy diferentes a las de las pruebas, generalmente se interpreta como un síntoma de Sobreajuste (Overfitting) o Subajuste (Underfitting). Una vez obtenidas los valores que utilizaremos para comparar los algoritmos, también aplicamos los mismos pasos, pero ahora con todas las variables como clasificatorias y también se realiza una prueba con tres variables que nosotros hemos elegido, estas son 'Nivel de Educación, 'Rol de Trabajo' y 'Tamaño de Empresa'.

Hiperparámetros

Al completar con la instancia de entrenamiento y prueba de los modelos, se obtiene todas las métricas y se selecciona el mejor, pero antes, se le puede hacer un análisis de los hiperparámetros para poder determinar cuáles pueden mejor los resultados. Con la función .get_params() podremos visualizar los parámetros predeterminados, luego desde la librería sklearn importamos RandomizedSearchCV que analiza y muestras los mejores valores para los hiperparámetrso que nosotros queramos modificar. Para finalizar se les asigna dichos valores y se vuelve a entrenar al modelo para obtener las métricas actualizadas que nos indicaran si hay alguna mejora o no.

Algoritmos de clasificación

Cuando se desea entrenar un algoritmo para que sea capaz de clasificar en diferentes clases dependiendo los valores ingresados, el método suele ser parecido. Como algoritmos de clasificación nosotros utilizamos Clasificación por Bosques Aleatorio, Clasificación por K vecinos cercanos y SoftMax y al igual que en predicción se entrenaron y probaron con todas las variables del conjunto de datos, pero también con los mejores 8 que fueron seleccionadas mediante la función SelectKBest de sklearn. Su funcionamiento se basa en realizar un cálculo para identificar las mejores características para la clasificación, utiliza el estadístico F de análisis de

varianza (ANOVA). Este criterio calcula la relación estadística entre cada característica y la variable objetivo. Las métricas para estos modelos también cambian, utilizamos accuracy que nos indica la precisión, luego se analiza un promedio macro (macro avg) donde no se considera el peso de las clases, dándonos resultados de precision, recall y f1-score, por último, el promedio ponderado (weighted avg) considerando el número de instancias en cada clase (el support). También se nos brinda una Matriz de confusión que te muestra cómo el modelo clasifica las instancias de cada clase y dónde comete errores. Cada fila representa las instancias reales de una clase específica, mientras que cada columna representa las predicciones realizadas por el modelo para esas clases.

Streamlit

Como parte final de este estudio, se planteó la idea de llevar al modelo entrenado de forma funcional a internet para que demás personas puedan acceder y probar tanto la predicción del ingreso mensual como también la clasificación de la satisfacción laboral. Para poder llevarlo a cabo nos ayudamos de Streamlit que es una biblioteca de Python que nos permitió crear la aplicación web de forma sencilla. Primero mediante el uso de la librería pickle creamos un archivo .pkl que contendrá nuestro modelo entrenado, luego se crea la aplicación en Python donde se llama a dicho archivo y definen detalles de visualización y funcionalidad. Para correrla de forma local se ejecuta en la terminal la sentencia *streamlit run nombre_aaplicacion.py* y para poder cargarlo a la web creamos un repositorio en GitHub donde contenga los archivos .pkl y la aplicación, la página de Streamlit tiene una opción donde mediante este método lo puedo subir sin ningún problema. Una vez en internet, cualquier persona mediante el enlace podrá probar la aplicación.

3. Resultados

3.1 Visualización de datos

En el Análisis Exploratorio de Datos se estudió la variable de Deserción, que para nosotros es la más importante debido a que los motivos que causan dicha acción son la razón por la fue elegido este Dataset al principio.

Como se puede observar en la gráfica (Figura 1), el 35,9% de los empleados abandonan sus empresas. Esto no remarca una fuerte crisis de deserción, pero si da a entender que poco menos de la mitad de los empleados se van a otras empresas o a trabajar por su propia cuenta.

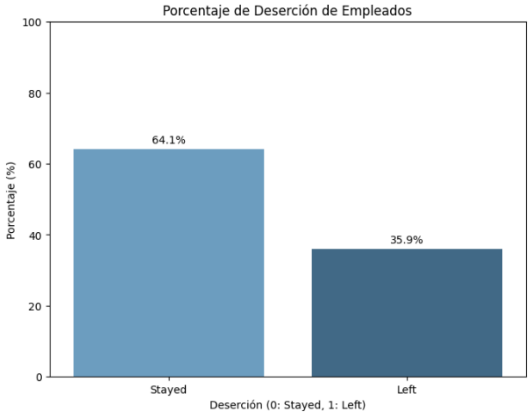

Figura 1: Gráfico de barras porcentual.

Se analizan las variables por separadas ya nos interesa, en primera instancia, cual es la frecuencia de los datos. Para las variables de tipo categóricas empleamos gráficos de barras, esto nos permitirá tener un vistazo previo de los datos y cuales nos serán de utilidad en un futuro.

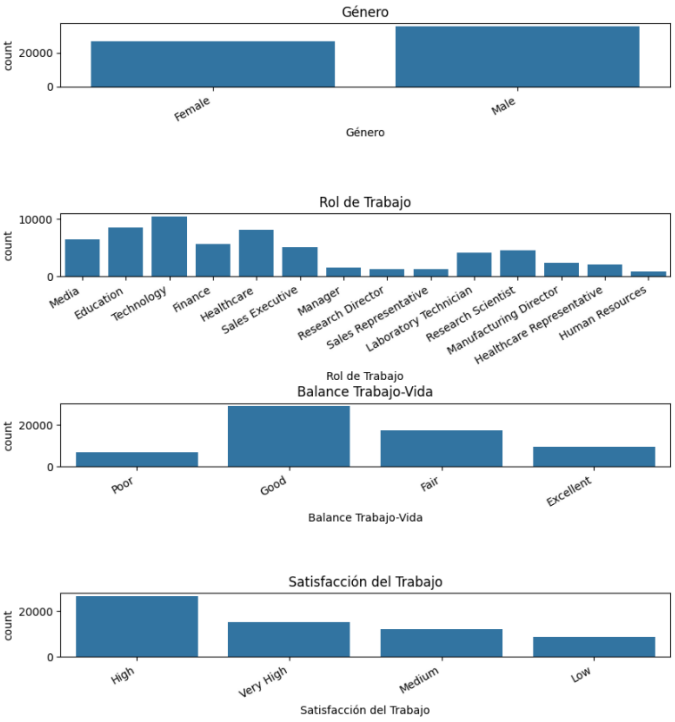

Figura 2: Gráficos de barras de las variables categóricas Género, Rol de Trabajo, Balance Trabajo-Vida y Satisfacción Laboral.

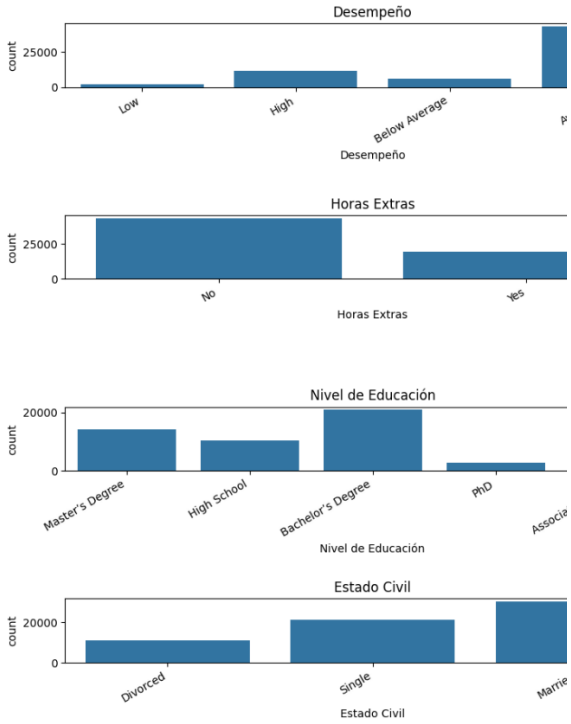

Figura 3: Gráficos de barras de las variables categóricas Desempeño, Horas Extras, Nivel de Educación y Estado Civil.

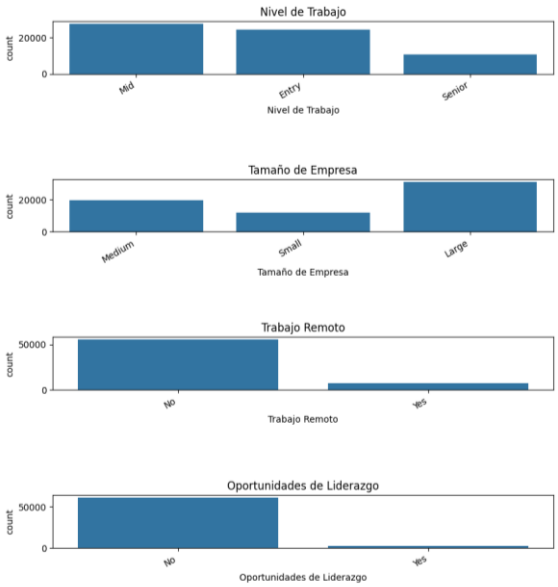

Figura 4: Gráficos de barras de las variables categóricas Nivel de Trabajo, Tamaño de Empresa, Trabajo Remoto y Oportunidades de Liderazgo.

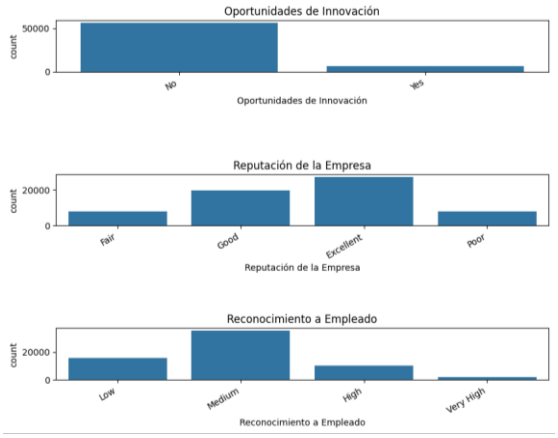

Figura 5: Gráficos de barras de las variables categóricas Oportunidades de Innovación, Reputación de la Empresa y Reconocimiento a Empleado.

Haciendo una observación acerca de estos gráficos podemos deducir lo siguiente:

Género: Se visualiza que el valor "Masculino" tiene una mínima diferencia positiva por lo que determina una mayor población masculina en el ámbito laboral.

Rol del trabajo: Hay 14 variables para determinar el trabajo de los empleados, la mayoría se encuentran en Tecnología seguido de Educación. Posteriormente se realizará un gráfico específico para su estudio.

Balance Trabajo-Vida: En el gráfico se percibe claramente que los empleados pueden mantener una relación Buena entre su vida personal y el trabajo. Es notoria la diferencia por lo que se podría indagar más en detalle cuales son los empleados en específico que gozan de este balance.

Satisfacción del Trabajo: Amplia diferencia entre el valor de satisfacción Alta con los demás. A simple vista, los empleados están conformes con sus trabajos y el ambiente en el que se encuentran.

Desempeño: El desempeño de los empleados en lo general es Promedio, no suelen destacarse una su mayoría, pero tampoco son ineficientes.

Horas Extras: La mayoría no cuentan con horas extras, al parecer, solo el 33% aproximadamente pueden acceder a realizarlas.

Nivel de educación: En esta variable, los valores tienden a estar más parejos, excepto por PhD. Destaca por sobre los demás el Bachiller, que para nosotros sería un título Medio-Superior.

Estado Civil: Esta variable puede relacionarse con otras como con Balance Trabajo-Vida. Principalmente se indica que gran parte de los trabajadores se encuentran Casados.

Nivel de Trabajo: En esta variable, se define el nivel y dificultad de obligaciones que tienen los empleados debido a su estatus dentro de la empresa. Aproximadamente el 80% de los trabajadores se encuentran en nivel de Ingresante y Medio, por una pequeña diferencia, mayormente Medio.

Tamaño de Empresa: Los empleados manifestaron cual es el tamaño de la empresa en la que se encuentran. Al rededor de la mitad de ellos la definieron como un Gran Empresa.

Trabajo Remoto: Esta variable es interesante de estudiar ya que al ser una modalidad actualmente en constante crecimiento, la mayoría de los empleados indicaron que no la realizan. Se desempeñarán futuros gráficos y combinaciones de variables para encontrar una razón a esta situación.
Oportunidades de Liderazgo: Se puede visualizar que son muy pocos los empleados que pueden aspirar a liderar un sector o grupo de trabajo.
Oportunidades de Innovación: Ocurre un evento idéntico al de la variable Oportunidad de Liderazgo, es escasa la cantidad de empleados que tienen esta oportunidad.
Reputación de la Empresa: La mayoría de los trabajadores opinan que trabajan en empresas con una Excelente reputación, también hay otra gran parte que la denominan una Buena empresa. Son pocos los casos de empleados con una visión negativa de la empresa.
Reconocimiento a Empleado: Tal como ocurre en la variable de Desempeño, los empleados determinan en su mayoría, que el reconocimiento a ellos es Moderado, no se les premia o recompensan sus acciones tan a menudo, pero tampoco se los desestima o ignoran.

En el caso de las variables numéricas se optó por dos tipos de gráficos, un histograma por cada variable que contiene una línea continua que suaviza la distribución de los datos para visualizar mejor la tendencia general, mostrando cómo se distribuyen los valores a lo largo del rango, y también se utilizó el diagrama de cajas que nos es útil para encontrar outliers, estos son valores atípicos que desvían significativamente la tendencia general de un conjunto de datos. Estos últimos tiene la peculiaridad de que están representando a las variables y su relación con Deserción, de esta forma podemos presenciar si esto produce alguna diferencia entre ellas.

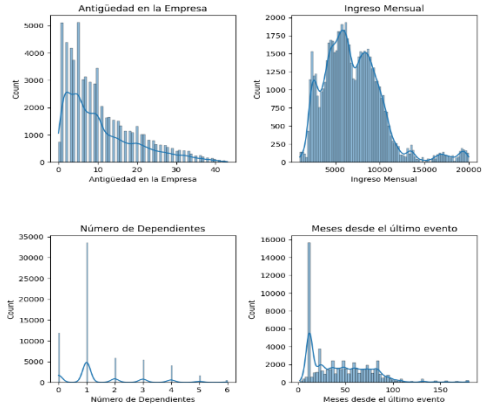

Figura 6: Histogramas de las variables ID Empleado, Edad, Número de Promociones y Distancia a Casa.

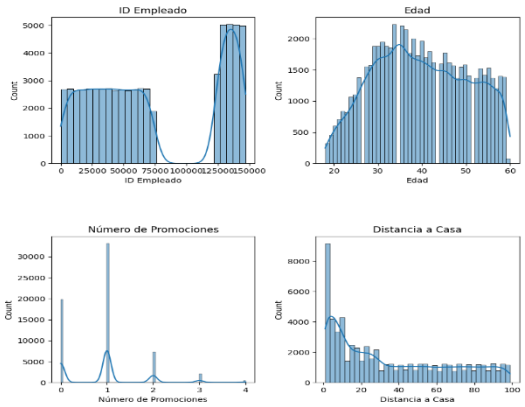

Figura 7: Histogramas de las variables Antigüedad en la Empresa, Ingreso Mensual, Número de Dependientes y Meses desde el último evento.

Estos gráficos nos permiten presenciar la información de mejor manera tal que nos facilita la obtención de conclusiones como que en el gráfico de **ID Empleado** se sugiere que los empleados podrían haber sido asignados con IDs en diferentes momentos o bajo diferentes esquemas de numeración. Sin embargo, el ID no tiene importancia en el análisis en sí.

En **Edad**, la mayoría de los empleados se encuentran en un rango de edad joven a mediana. Hay una disminución en la cantidad de empleados mayores de 45 años, lo que podría indicar una tendencia de deserción o jubilación en edades más avanzadas.

Antigüedad en la Empresa, la mayoría de los empleados son relativamente nuevos en la empresa, con pocos alcanzando más de 10 años de antigüedad. Esto podría indicar una alta rotación de personal o una estructura joven de la empresa.

Ingreso Mensual, esto sugiere que la mayoría de los empleados tienen ingresos en un rango 2000 a 12500, con una minoría que recibe sueldos más altos, posiblemente debido a diferencias en roles o niveles de experiencia.

Número de Promociones, esto podría indicar una estructura de promoción limitada, donde pocos empleados son promovidos. La falta de promoción podría estar relacionada con la deserción.

Distancia a Casa, la mayoría de los empleados viven cerca de su lugar de trabajo, lo cual podría ser un factor positivo en la retención de empleados. La dispersión de valores sugiere que algunos empleados están dispuestos a desplazarse más lejos.

Número de Dependientes, esto podría indicar que los empleados tienen pocas responsabilidades laborales, lo cual puede influir en su estabilidad en el trabajo y decisiones de deserción.

Meses desde el último evento, la presencia de muchos empleados con eventos recientes podría reflejar una empresa dinámica, pero la presencia de outliers podría sugerir una falta de reconocimiento o oportunidades para ciertos empleados, lo cual podría llevar a la deserción.

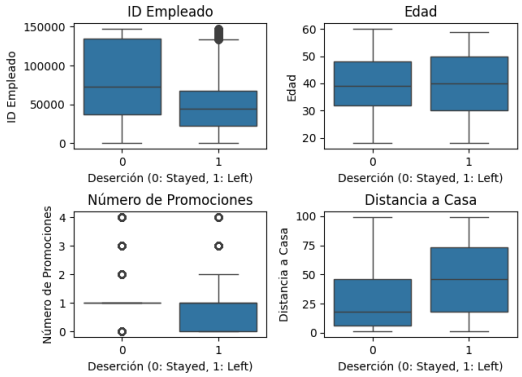

Figura 8: Diagrama de Cajas de las variables ID Empleado, Edad, Número de Promociones y Distancia a Casa.

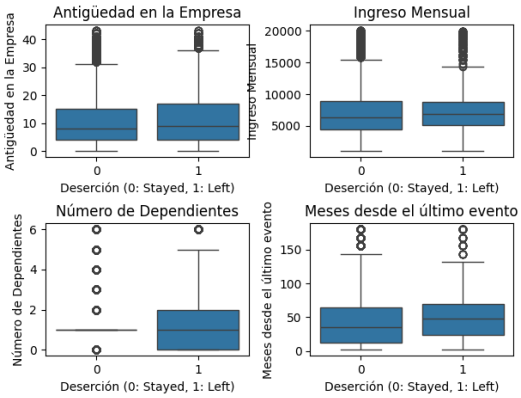

Figura 9: Diagrama de Cajas de las variables Antigüedad en la Empresa, Ingreso Mensual, Número de Dependientes y Meses desde el último evento.

Las observaciones que podemos apreciar en estos casos son las siguientes:

Edad: Se observa que la mediana de edad para los empleados que se quedaron (0) es más alta en comparación con la de los que desertaron (1). Esto sugiere que los empleados más jóvenes tienden a desertar más.

Antigüedad en la Empresa: La mediana de antigüedad en la empresa es menor para los empleados que desertaron (1). Además, se observan valores atípicos (outliers) en ambos grupos, lo que indica que hay empleados con muchos años en la empresa que también pueden desertar.

Ingreso Mensual: No parece haber una diferencia significativa en los ingresos mensuales entre los empleados que se quedaron (0) y los que se fueron (1). Los outliers indican empleados con ingresos excepcionalmente altos en ambos grupos.

Número de Promociones: Los empleados que se quedaron (0) tienen un rango de promociones más amplio, mientras que la mayoría de los que desertaron (1) no recibieron promociones o recibieron muy pocas. Esto sugiere que la falta de crecimiento profesional podría estar relacionada con la deserción.

Distancia a Casa: Los empleados que se quedaron (0) tienen una menor distancia promedio a casa en comparación con los que desertaron (1). Esto podría indicar que un largo desplazamiento es un factor que contribuye a la deserción.

Número de Dependientes: Los empleados que desertaron (1) tienden a tener más dependientes en comparación con los que se quedaron (0). Hay muchos outliers en el grupo que se quedó, lo que indica que hay empleados con un número excepcionalmente alto de dependientes que no desertan.

Meses desde el último evento: La mediana es similar en ambos grupos, pero hay más outliers en el grupo que desertó (1). Esto sugiere que la falta de eventos significativos durante un tiempo prolongado podría estar relacionada con la deserción, aunque no es concluyente.

Como resultado de estos primeros análisis surgió el interés de indagar más en profundidad en algunas variables y su relación con otras, por ejemplo ¿Cuál es el rango de edades más propenso a desertar? Esta incógnita nos hizo desagregar la variable Edad en diferentes rangos, de 18 a 25, de 26 a 35, de 36 a 45, de 46 a 55 y de 56 a 60 que es la edad máxima registrada. Al realizar el gráfico (Figura 10) podemos apreciar el 49,8% de las personas de entre 18 a 25 años son propensos a desertar, lo cual, nos da a entender que los jóvenes son el grupo de empleados con más rotaciones en el mercado laboral. Los siguen los trabajadores de 56 a 60 años con un 41,6%, esto se puede ocasionar por retiros voluntarios previos a jubilaciones, ya sea para mujeres de 60 años en adelante o 55 años para trabajos insalubres. Luego la tendencia de los demás grupos de edad se mantiene entre el 30% al 38%, estas edades corresponden a aquellos que tienen un trabajo estable o que no consideran prudente renunciar para cambiar por otra empresa, buscan la estabilidad y la adquisición de años o estatus en una empresa.

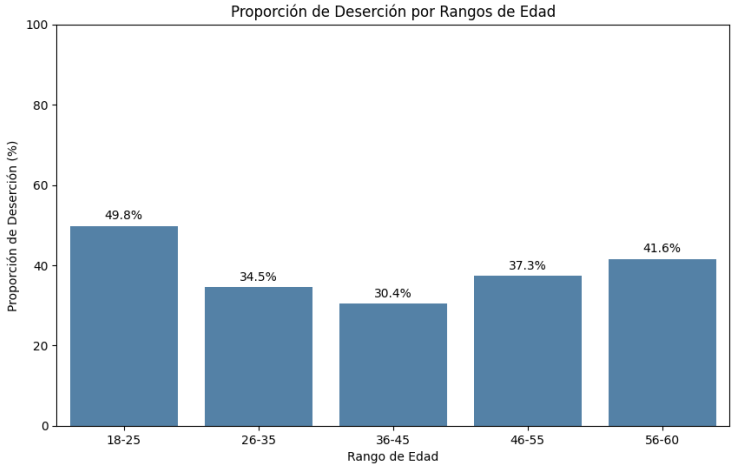

Figura 10: Gráfico de barras con rango de edades y su porcentaje de deserción.

Otra variable de gran interés es Rol de Trabajo ya que para nosotros es importante saber cuáles son los puestos que más empleados abarca, por lo que

mediante un gráfico de Pareto podemos observar que las primeras 6 categorías (Technology, Education, Healthcare, Media, Finance, Sales Executive) parecen cubrir aproximadamente el 80% de la frecuencia total. Estas son las categorías más relevantes en el análisis y podrían considerarse el "punto de enfoque" si se busca priorizar esfuerzos o recursos. Las categorías restantes (a la derecha de la línea roja) tienen menos impacto en términos de frecuencia y podrían no ser tan prioritarias para el análisis o la acción.

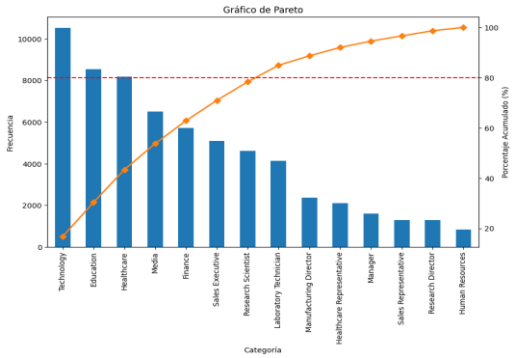

Figura 11: Gráfico de Pareto de la variable Rol de Trabajo.

En cuanto a relacionar variables hay dos situaciones que nos parecían más relevantes. La primera tiene como variable principal el Número de Promociones y su relación directa con las variables de Edad y Antigüedad en la Empresa, pero como no se empleó un gráfico 3D se realizaron por separados en dos gráficos de dispersión en 2D y contiene una línea roja que muestra la tendencia para ver visualmente si hay alguna correlación entre las variables. En el primer gráfico (Figura 12) se muestra la relación con la edad, parece no haber una correlación clara entre la edad y el número de promociones, ya que la línea de tendencia es plana. En el siguiente (Figura 13) hay una relación ligeramente negativa, lo que podría indicar que los empleados con más años en la empresa tienden a recibir menos promociones. Sin embargo, esta relación no parece ser muy pronunciada.

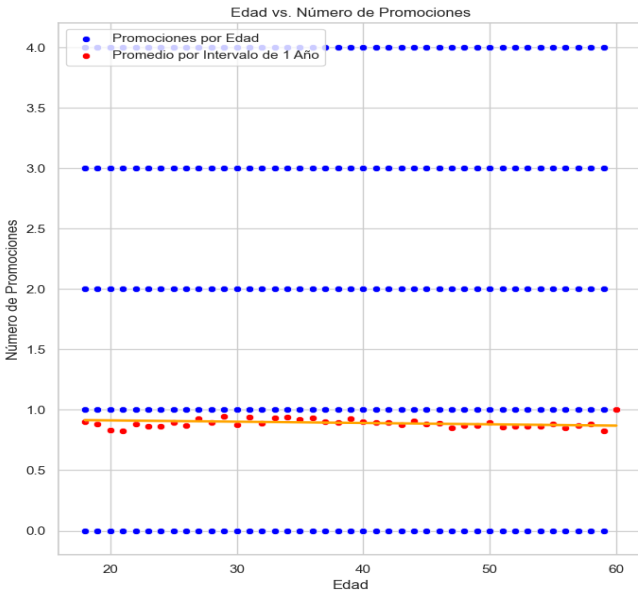

Figura 12: Diagrama de Dispersión Edad vs Número de Promociones.

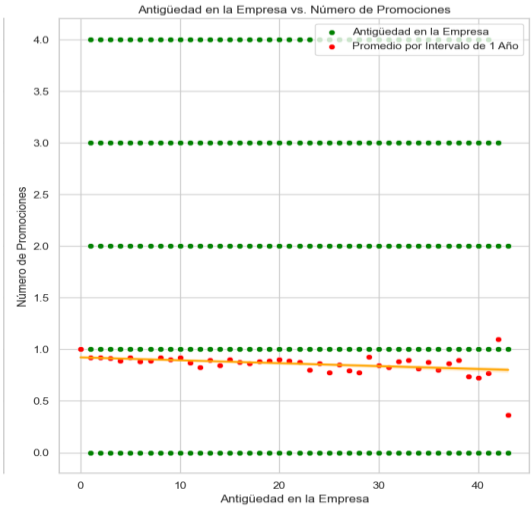

Figura 13: Diagrama de Dispersión Antigüedad en la Empresa vs Número de Promociones.

La segunda relación importante que quisimos analizar fue entre la variable Trabajo Remoto y Distancia a Casa, esto se debe a que últimamente esta modalidad de trabajo está cada vez más en tendencia. Para su visualización utilizamos un gráfico de violín el cual es muy similar al diagrama de cajas. En la Figura 14 se puede apreciar que los empleados que trabajan de forma remota parecen vivir más lejos de su lugar de trabajo en promedio, ya que su mediana está más alta que la del grupo que no trabaja de forma remota. Los empleados que no trabajan de forma remota tienden a vivir más cerca, con la mayoría concentrados en distancias de hasta 20 km. Y, por último, la distribución de la distancia a casa es más variada para aquellos que trabajan de forma remota, mientras que para los que no trabajan de forma remota, la mayoría de los empleados vive relativamente cerca.

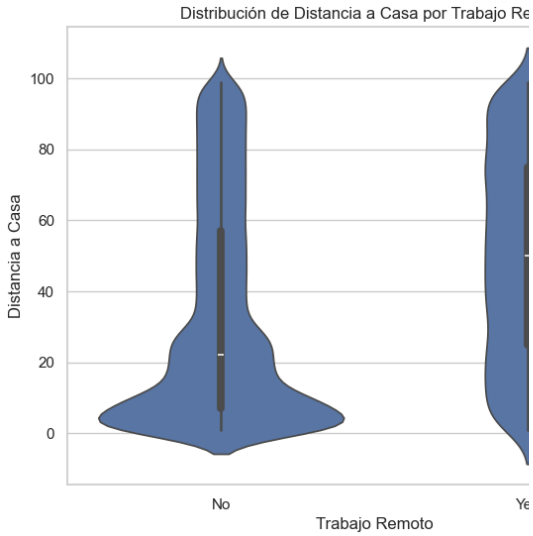

Figura 14: Gráfico de Violín que relaciona Distancia a Casa con Trabajo Remoto.

3.2 Resultados de los algoritmos

Para poder obtener el mejor algoritmo, tanto de predicción como de clasificación, tuvimos que realizar una serie de pruebas y comparativas para llegar a la conclusión de cual se adapta mejor a nuestro dataset.

En el caso de los algoritmos de predicción se utilizaron datos normalizados ya que los estandarizados nos generaban una baja relación, siempre cerca de 0, y un valor de MAE muy alto para lo que es el contexto de nuestro conjunto de datos, aproximadamente 0.98.

Para poder dar una mejor perspectiva de las métricas y su comparación se colocarán los resultados en modo de resumen para cada instancia. La primera comparará los algoritmos mencionados anteriormente en la sección de Metodología con las variables de PCA, luego lo hará con todas las variables y por último con las variables seleccionadas por nosotros las cuales también fueron indicadas con anterioridad.

Cabe aclarar que los valores de entrada para las pruebas pertenecen a un registro ya existente el cual su Ingreso Mensual es de 5534 por lo que además de las métricas se busca aquel que este más cerca de ese valor.

PCA:

- **Regresión Lineal:**
  - R2: 0.2497
  - RMSE: 0.1498
  - MAE: 0.1175
- **Regresión Árbol Aleatorio:**
  - R2: 0.6771
  - RMSE: 0.0982
  - MAE: 0.0655
- **K Vecinos más Cercanos (KNN):**
  - R2: 0.5920
  - RMSE: 0.1104
  - MAE: 0.0768

Todas las variables(normalizadas):

- **Regresión Lineal:**
  - R2: 0.0920
  - RMSE: 0.1648
  - MAE: 0.1291
  - Predicción: 6998.13
- **Regresión Árbol Aleatorio:**
  - R2: 0.9011
  - RMSE: 0.0543
  - MAE: 0.0316
  - Predicción: 5956.72
- **K Vecinos más Cercanos (KNN):**
  - R2: 0.6503
  - RMSE: 0.1022
  - MAE: 0.0678
  - Predicción: 7472.40

Variables elegidas (sesgada):

- **Regresión Lineal:**
  - R2: 0.0202
  - RMSE: 0.1711
  - MAE: 0.1321
  - Predicción: 5886.89
- **Regresión Árbol Aleatorio:**
  - R2: 0.7563
  - RMSE: 0.0853
  - MAE: 0.0598
  - Predicción: 4461.82
- **K Vecinos más Cercanos (KNN):**
  - R2: 0.7018
  - RMSE: 0.0944
  - MAE: 0.0659
  - Predicción: 7860.00

Con estos resultados llegamos a la conclusión de que el algoritmo que mejor se adapta a nuestro dataset es Regresión Árbol Aleatorio con todas las variables. Esto se debe a que no solo mostros las mejores métricas, sino que también se aproxima mucho al valor correcto. Por eso, con el afán de querer mejor dichos resultados, se les ajusto los hiperparámetros con los que serían los mejores valores: {'max_depth': 48, 'max_features': 'sqrt', 'min_samples_leaf': 3, 'min_samples_split': 2, 'n_estimators': 276}. Esto nos da un R2 de 0.8904 y un MAE de 0.0397, lo cual es un poco peor que si dejamos los valores predefinidos del modelo, esto se puede comprobar ya que el resultado de la prueba realizada es 5998.13, eso es una diferencia de 41,41 en el ingreso mensual.

Para los algoritmos de clasificación a diferencia de los de predicción, se utilizaron valores numéricos sin normalizar ni estandarizar ya que los resultados no muestran ninguna diferencia en sí. En la sección de Metodología se mencionaron los algoritmos y métricas que se utilizaron, se aclara este punto para no ser redundante en la explicación. Y como dato a tener en cuenta, se utilizó Regresión Logística para poder

clasificar una variable que contiene 4 clases, esto fue un error debido a que este modelo sería adecuado para clasificaciones binarias y no multivariadas, por eso se incorporó el algoritmo de SoftMax, ambos sin darnos buenos resultados pero que se mostraran a continuación.

K vecinos cercanos con todas las variables:

- accuracy = 0.61
- macro avg: precision = 0.62, recall = 0.58 y f1-score = 0.59
- weighted avg: precision = 0.62, recall = 0.61 y f1-score = 0.61

K vecinos cercanos con las mejores características:

- accuracy = 0.52
- macro avg: precision = 0.51, recall = 0.47 y f1-score = 0.48
- weighted avg: precision = 0.52, recall = 0.52 y f1-score = 0.51

Bosques Aleatorios con todas las variables:

- accuracy = 0.68
- macro avg: precision = 0.87, recall = 0.60 y f1-score = 0.65
- weighted avg: precision = 0.80, recall = 0.68 y f1-score = 0.66

Bosques Aleatorios con las mejores características:

- accuracy = 0.56
- macro avg: precision = 0.59, recall = 0.49 y f1-score = 0.51
- weighted avg: precision = 0.57, recall = 0.56 y f1-score = 0.54

Regresión Logística con todas las variables:

- accuracy = 0.42
- macro avg: precision = 0.19, recall = 0.27 y f1-score = 0.20
- weighted avg: precision = 0.26, recall = 0.42 y f1-score = 0.30

Regresión Logística con las mejores características:

- accuracy = 0.42
- macro avg: precision = 0.19, recall = 0.26 y f1-score = 0.19
- weighted avg: precision = 0.26, recall = 0.42 y f1-score = 0.29

SoftMax con todas las variables:

- accuracy = 0.42
- macro avg: precision = 0.44, recall = 0.26 y f1-score = 0.19
- weighted avg: precision = 0.39, recall = 0.42 y f1-score = 0.29

SoftMax con las mejores características:

- accuracy = 0.42
- macro avg: precision = 0.18, recall = 0.26 y f1-score = 0.19
- weighted avg: precision = 0.26, recall = 0.42 y f1-score = 0.29

Como conclusión podemos presenciar que cuando se realizan las pruebas con las variables seleccionadas que mantienen un sesgo, los resultados son significativamente peores que cuando se realiza con todas las variables. En el caso de la regresión logística y softmax, en ambos las pruebas nos dan un resultado similar y a su vez muy malo, por lo que son los primeros en ser descartados. Luego entre KNN y Bosques Aleatorios el mejor con diferencia es Bosques Aleatorios con todas las variables, si bien sus métricas son aceptables se podrían mejorar ajustando los hiperparámetros.

El resultado final de este estudio es la elaboración de la web funcional que pueda predecir el Ingreso Mensual y clasificar la Satisfacción Laboral de los empleados. Anteriormente se mencionó la herramienta Streamlit, una librería de Python que nos ayudará a crear y subir a internet nuestros modelos en formato de cuestionario para que otras personas la puedan probar.

En la primera imagen (Figura 15) se puede observar el inicio de la página web donde se encuentra su nombre “equipoPI”, un subtitulo donde nos da a entender de que trata o cual es el tema principal, en este caso “Análisis del Empleado”, y también las pestañas que contiene “Visualización EDA”, “Modelo de Predicción” y “Modelo de Clasificación”. La primera sección contiene datos del dataset utilizado para entrenar los modelos y análisis previos que se mencionaron a lo largo de este documento, contiene otra variedad de gráficos (Figura 16) pero mantiene la misma información. Cabe aclarar que en el sitio web hay más gráficos, pero no se adjuntaran al escrito para no prolongarlo más de lo debido con imágenes.

Figura 15: Inicio de página web.

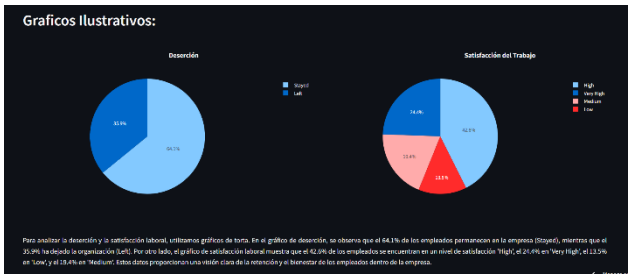

Figura 16: Gráficos que se encuentran en la página web.

En la segunda pestaña se encuentra el modelo de predicción, este esta presente como un formulario con las diferentes variables como casillas para que el usuario pueda completarlo. En las variables de tipo objeto se les añadió una lista desplegable y limitada con los valores que se encuentran en el dataset, esto con el fin de que pueda encontrar valores que le permitan comparar. A su vez, las variables numéricas si pueden ser cargadas por el usuario, estos valores tienen están limitados para enteros y con un rango mínimo y máximo. Para finalizar y poder visualizar el resultado de deberá de presionar el botón con la leyenda de “Realizar Predicción”.

Figura 17: Modelo de Predicción.

La tercera y última pestaña se trata del modelo de clasificación que está conformado y funciona de la misma forma que el modelo anteriormente mencionado.

Figura 18: Modelo de Clasificación.

Como última observación de esta página web y posibles mejoras en un futuro, sería conveniente incorporar gráficos con los cuales interactuar para la visualización de las relaciones entre diferentes variables y valores. Otra alternativa muy útil sería limitar las variables para los modelos o que el usuario decida cuales variables le interesa tener para la predicción o clasificación.

En la sección de Referencias se incluirá el link [6] de acceso a esta página para una mejor visualización y análisis.

4. Discusión

1.Antecedentes.

Diversas investigaciones han demostrado la importancia del nivel de educación, el rol de trabajo y el tamaño de la empresa como factores predictivos del ingreso mensual [2, 4, 5]. En nuestro análisis, estas variables también resultaron ser relevantes para la predicción, lo que confirma los hallazgos de estudios previos. Además, se ha evidenciado que el uso de técnicas de selección de características, como SelectKBest, puede mejorar el rendimiento de los modelos de predicción [1], aunque en nuestro caso, el uso de todas las variables disponibles proporcionó mejores resultados.

2.Declaración de resultados.

El objetivo principal de este análisis fue predecir el ingreso mensual esperado para un empleado nuevo en función de su nivel de educación, rol de trabajo y tamaño de la empresa. Para ello, se evaluaron tres algoritmos de regresión: Regresión Lineal, Random Forest y KNN, utilizando tanto todas las variables del dataset como un subconjunto seleccionado mediante SelectKBest. Los resultados mostraron que el modelo Random Forest, con hiperparámetros predeterminados y utilizando todas las variables disponibles, obtuvo el mejor rendimiento en términos de R^2, RMSE y MAE.

3.Resultado(s) (in)esperado(s).

Si bien se esperaba que el Random Forest tuviera un buen rendimiento en esta tarea de predicción, resultó sorprendente que la optimización de hiperparámetros no mejorara significativamente las métricas de evaluación. Esto podría deberse a un sobreajuste del modelo con hiperparámetros más complejos o a la naturaleza del dataset.

4.Referencia a investigaciones anteriores (comparación).

La precisión obtenida por el modelo Random Forest es comparable a la reportada en estudios previos que utilizaron algoritmos similares para predecir variables relacionadas con el trabajo, como la satisfacción laboral [3]. Sin embargo, es importante destacar que la precisión puede variar dependiendo de las características específicas del dataset y de las variables predictoras utilizadas.

5. Explicación del resultado(s) insatisfactorio(s).

La falta de mejora significativa en el rendimiento del modelo Random Forest tras la optimización de hiperparámetros podría deberse a un sobreajuste, como lo sugieren [Amani Shoman, 2009] [2] en su análisis de los factores que predicen la satisfacción

laboral. También es posible que la configuración predeterminada de los hiperparámetros sea ya adecuada para este dataset en particular.

6. Ejemplificación.

Como modo de ejemplo para poder saber que tan buenos resultados nos entrega el modelo con valores predefinido, se replicó el mismo método de predicción hacia otras variables como objetivo. Uno de ella fue Antigüedad en la Empresa que sus métricas de entrenamiento fueron muy buenas, R2: 0.9441 y MAE: 0.0300, pero al momento de las pruebas esos valores bajaron consideradamente, R2: 0.5974 y MAE: 0.0823. Al intentar mejorarlo con los hiperparámetros tampoco se vio un incremento favorable.

7. Deducción e hipótesis.

Los resultados obtenidos sugieren que el Random Forest es un algoritmo efectivo para predecir el ingreso mensual de un empleado, utilizando variables como el nivel de educación, el rol de trabajo y el tamaño de la empresa. Se plantea la hipótesis de que la inclusión de otras variables relevantes, como la experiencia laboral o el sector industrial, podría mejorar aún más la precisión del modelo.

8. Recomendaciones.

Estos estudios requieren de muchas pruebas e intentos para poder encontrar valores adecuados y perfeccionar los modelos cada vez más. Con esta apreciación podemos decir que el tiempo y el compromiso son esenciales, ya que se puede probar con diferentes variables, algoritmos, conjuntos de datos, hiperparámetros, relaciones y gráficos para la visualización de los resultados.

9. Justificación.

Este análisis contribuye al área de la predicción de ingresos al proporcionar un modelo efectivo y fácil de implementar, similar a lo propuesto por [Joseph Ryan Owens, Pankti Patel; 2019] [5] en su estudio sobre el ingreso como predictor de la satisfacción laboral. Además, se ha demostrado la importancia de utilizar todas las variables disponibles en este caso particular, lo que podría tener implicaciones para futuras investigaciones en el área.

5. Conclusiones

Este estudio demuestra el potencial del aprendizaje automático para predecir ingresos mensuales y clasificar niveles de satisfacción laboral en empleados, utilizando variables clave como nivel educativo, rol de trabajo y tamaño de la empresa. Los principales resultados muestran que el algoritmo Random Forest ofreció el mejor rendimiento en predicción, con un coeficiente R^2 de 0.9011 y un MAE de 0.0316. En tareas de clasificación, también fue el más efectivo, alcanzando una precisión del 68%. Además, el análisis exploratorio reveló patrones significativos, como la alta deserción en empleados jóvenes y la relación entre distancia al trabajo y modalidad remota.

Desde una perspectiva teórica, los hallazgos confirman estudios previos sobre la relevancia de las características demográficas y laborales en modelos predictivos. Desde el ámbito práctico, este trabajo proporciona una herramienta robusta para apoyar decisiones en gestión de recursos humanos, como la retención de talento y planificación salarial.

Para mejorar este estudio, se sugiere explorar modelos más avanzados como redes neuronales. Además, el ajuste de hiperparámetros podría ser optimizado utilizando técnicas más exhaustivas, como búsqueda en malla, para mejorar aún más la precisión de los resultados. También, una validación en datasets de mayor diversidad garantizaría su aplicabilidad en diferentes contextos organizacionales. Finalmente, desarrollar aplicaciones en tiempo real y evaluar su impacto en contextos empresariales diversos representaría un paso clave hacia una implementación más amplia.

- **Referencias**

[1] N. Van Saane, J. K. Sluiter, J. H. A. M. Verbeek, and M. H. W. Frings-Dresen, "State of the Art of Job Satisfaction Measures: A Systematic Review", ResearchGate, abril 2018. Disponible: https://www.scielo.br/j/tpsy/a/xbTN7gyT3zdVRVJDBrN7Pgf/?lang=en

[2] Amani Shoman, "Examination of the factors that predict job satisfaction", SJSU ScholarWorks, pp. 55, diciembre 2009. Disponible: https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=4990&context=etd_theses

[3] Angelika Lepold, Norbert Tanzer, Anita Bregenzer, Paulino Jiménez, "The Efficient Measurement of Job Satisfaction: Facet-Items versus Facet Scales", PMC, pp. 19, junio 2018. Disponible: https://www.mdpi.com/1660-4601/15/7/1362

[4] S. E. Seashore and T. D. Tobor, "Job Satisfaction Indicators and Their Correlates", Deep Blue Repositories, vol. 18, no. 3, pp. 36, febrero 1975. Disponible: https://deepblue.lib.umich.edu/bitstream/handle/2027.42/67361/10.1177_000276427501800303.pdf

[5] J. R. Owens and P. Pate, "Income as a Predictor of Employee Job Satisfaction and Motivation", UTC Scholar, pp. 4, 2019. Disponible: https://scholar.utc.edu/cgi/viewcontent.cgi?article=1184&context=rcio

[6] F. Chiabo, noviembre 2024. equipoPI. Disponible: https://mlwebpage-jrkm24wmd9utewn5k34psx.streamlit.app/