Tecnológico de Monterrey

Actividad 7 Regresión logística

Nombre del Equipo

CAPT

Miembros

Danna Paola Arciniega Zúñiga | A01731987

Victoria Eugenia Téllez Castillo | A01732258

Mitzi Castelán Chávez | A01731147

Elizabeth Pérez García | A01366971

Fecha

18 de Octubre de 2023

Analítica de datos y herramientas de inteligencia artificial II

INTRODUCCIÓN

La regresión logística se le conoce como un tipo de análisis de regresión, utilizado para la producción de resultados de una variables categóricas en función de una o varias variables independientes. Dentro de este caso de estudio, se trabajara con una base de datos llamada Training Data Complete, con la finalidad de realizar diversos modelados y predicciones enfocados en encontrar el mejor análisis y comprensión entre variables dicotómicas y variables predictoras , tanto numéricas como categóricas.

PREPROCESAMIENTO

El preprocesamiento es la primera instancia por la que pasa la base de datos, en donde se conoce más a fondo sobre los datos que se tienen, además de ser una instancia importante ya que garantiza la calidad de los datos y la efectividad de análisis o modelos predictivos. Dicho lo anterior, después de realizar las respectivas y procesos para la visualización de los datos, no se encontraron valores nulos ni atípicos; a continuación se muestra el paso a paso el proceso de cada línea de código para la obtención de esta información.

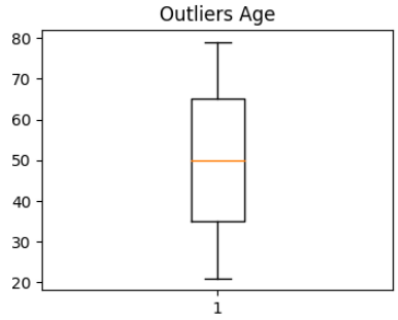
El primer paso fue conocer conocer la base de datos, por lo que se despues de leer el archivo, se solicitó imprimir los primeros cinco registros de la misma, posteriormente utilizando el comando .info() se obtuvo la información general sobre los datos, además de cómo están constituidos y que tipo de datos son de manera individual, en términos de columna; en ella se observa que hay 6 columnas con datos tipo enteros y las seis restantes con datos tipo objeto, que generalmente son cadenas. .

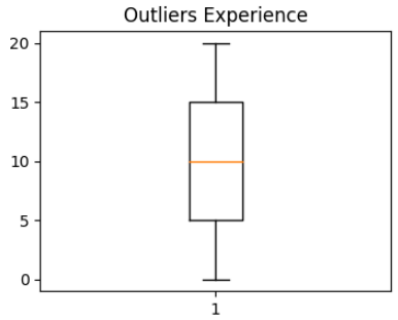
```
#OBTENER INFORMACION DEL DATA FRAME
df.info()
```

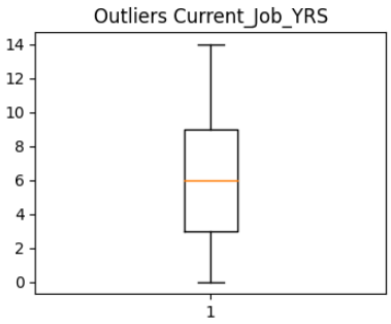
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 252000 entries, 0 to 251999
Data columns (total 13 columns):
 #   Column             Non-Null Count   Dtype 
---  ------             --------------   ----- 
 0   Id                 252000 non-null  int64 
 1   Income             252000 non-null  int64 
 2   Age                252000 non-null  int64 
 3   Experience         252000 non-null  int64 
 4   Married/Single     252000 non-null  object
 5   House_Ownership    252000 non-null  object
 6   Car_Ownership      252000 non-null  object
 7   Profession         252000 non-null  object
 8   CITY               252000 non-null  object
 9   STATE              252000 non-null  object
 10  CURRENT_JOB_YRS    252000 non-null  int64 
 11  CURRENT_HOUSE_YRS  252000 non-null  int64 
 12  Risk_Flag          252000 non-null  int64 
dtypes: int64(7), object(6)
memory usage: 25.0+ MB
```

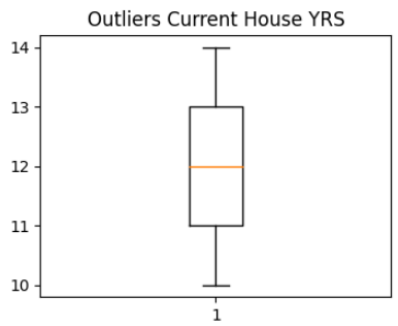
Para la continuación del preprocesamiento, se identificaron los valores tanto nulos como atípicos para así analizar las columnas y reemplazarlos de la mejor manera para comenzar con la manipulación de datos, sin embargo, no fue necesario ningún tipo de limpieza de valores, ya que dentro del dataframe no se contaba con valores nulos ni con valores atípicos. Para tener una representación visual de la falta de valores atípicos de cada una de las columnas principales del conjunto de datos, se realizaron histogramas de cajas y de barras, las cuales se muestran a continuación.

Outliers 'Caja' | Income, Age, Experience, current house, Current Job

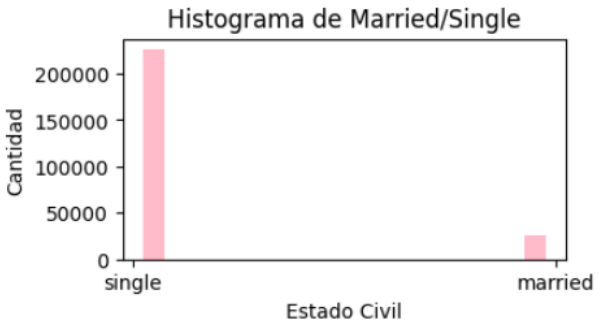


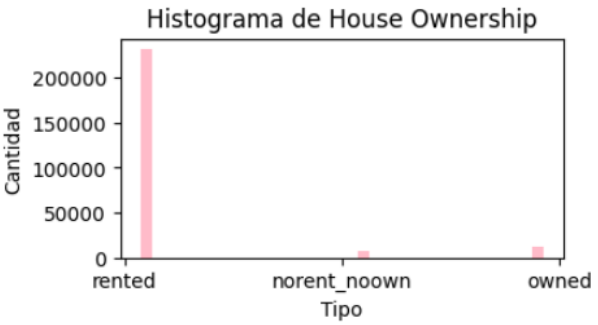


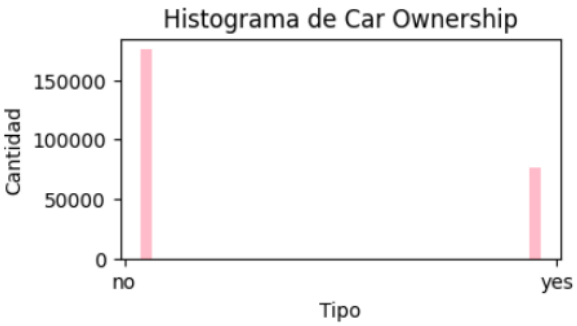


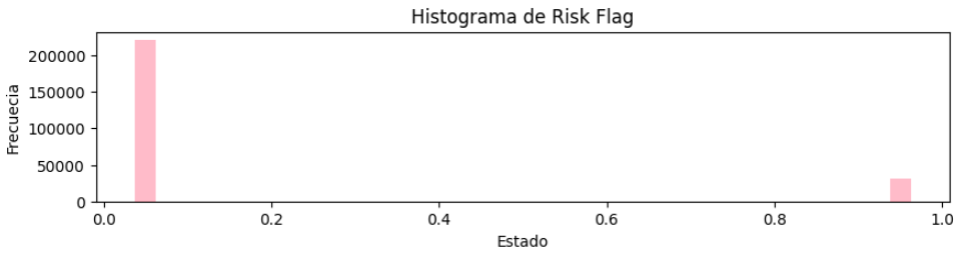


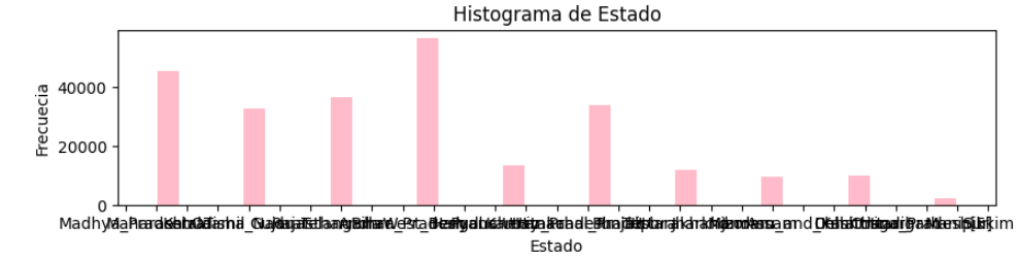
Histograma | Married/Single, Howse Ownership, Car Ownership, Profession,City, State, Risk Flag

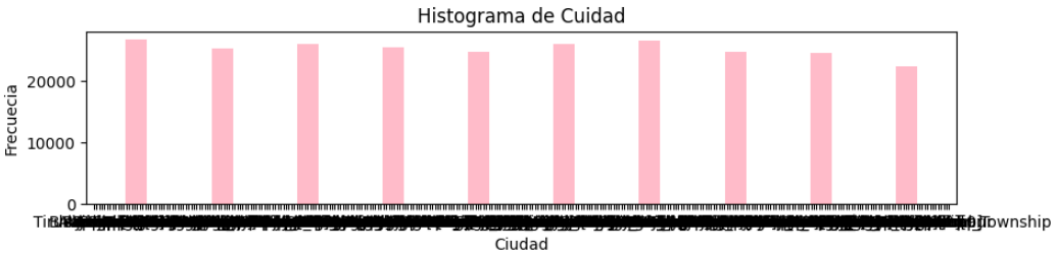












La creación de histogramas para las previas variables "Married/Single," "House Ownership," "Car Ownership," "Profession," "City," "State," y "Risk Flag" proporciona una representación visual de cómo se distribuyen las categorías en cada variable categórica, estos histogramas nos permite visualizar la frecuencia de cada categoría, identificar desequilibrios o tendencias en la distribución y destacar categorías dominantes. Además, la comparación de las distribuciones entre diferentes variables categóricas puede revelar patrones y relaciones potenciales, es importante crearlas y observarlas para la toma de decisiones en futuros análisis.

REGRESIÓN LOGÍSTICA

Ya teniendo una base de datos limpia, en este caso sin modificaciones, se comenzó a realizar una serie de regresiones logísticas, tomando en cuenta ciertas características de la base de datos,.

Debido a la magnitud de los datos, fue necesario realizar una serie de filtros, donde únicamente se seleccionaron ciertas variables,para que el uso de la regresión logística fuera más precisa y arrojará datos, permitiendo analizar los mismos.

De la misma manera, y para una ejecución de código de la manera más eficiente, se realizó una función, tomando en cuenta cada una de las instancias de la regresión logística, además de las métricas de evaluación, dicho lo anterior, la función mencionada anteriormente es la siguiente:

```
def entrenar_evaluar_modelo(X, y, pos_label, test_size=0.3, random_state=None):
# Dividimos el conjunto de datos en entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=test_size, random_state=random_state)

# Escalamos los datos
escalar = StandardScaler()
X_train = escalar.fit_transform(X_train)
X_test = escalar.transform(X_test)

# Definimos el algoritmo de regresión logística
algoritmo = LogisticRegression()

# Entrenamos el modelo
algoritmo.fit(X_train, y_train)

# Realizamos una predicción
y_pred = algoritmo.predict(X_test)

# Calculamos las métricas de evaluación
matriz = confusion_matrix(y_test, y_pred)
precision = precision_score(y_test, y_pred, average='binary', pos_label=pos_label)
exactitud = accuracy_score(y_test, y_pred)
sensibilidad = recall_score(y_test, y_pred, average='binary', pos_label=pos_label)
f1 = f1_score(y_test, y_pred, average='binary', pos_label=pos_label)

# Imprimimos los resultados
print('Matriz de confusión:')
print(matriz)
print('\nPrecisión del modelo:')
print(precision)
print('\nExactitud del modelo:')
print(exactitud)
print('\nSensibilidad del modelo:')
print(sensibilidad)
print('\nPuntuación F1 del modelo:')
print(f1)

return X_train, X_test, y_train, y_test, y_pred
```

Esta función nos permite que únicamente al momento de realizar las regresiones, se llame a la función, donde las variables dependientes e independientes sean las que se modifican.

FILTRO UNO

El primer filtro se enfoca en la columna profesión, específicamente en aquellas con las frecuencias más altas. Psychologist & Computer Hardware Engineer, son las profesiones que representan estos valores, por lo que el dataframe a utilizar en los primeros tres modelos contiene la información únicamente de las personas con estas dos profesiones.

Modelo UNO

Tomando como variable dependiente profesión y como variable independiente income, considerando como etiqueta positiva " Computer Hardware Engineer", los resultados que arrojó fueron los siguientes:

<< Matriz de Confusión >>

True Positive	False Positive
815	838
False Negative	True Negative
608	968

<< Evaluación >>

Precisión	0.5727
Exactitud	0.5522
Sensibilidad	0.4930
F1 Score	0.5299

Esta regresión predice la profesión basada en los ingresos. La matriz de confusión indica que hay 815 verdaderos positivos, que a su vez indica el total de predicciones correctas de la clase Computer Hardware Engineer, el valor de los verdaderos negativos, nos dice que el modelo fue capaz de predecir correctamente una clase que no es la elegida, mientras que el modelo seiscientas ocho veces predijo la clase objetivo, cuando no era el caso, además de que el modelo no pudo predecir la clase objetivo "Computer Hardware Engineer" ochocientos treinta y ocho veces la clase cuando en realidad lo era.

La precisión indica la proporción de predicciones correctas, incluidas las TP y TN en comparación con todas las predicciones realizadas, en este caso la precisión es de un 57%. La exactitud, como su nombre lo dice, representa la proporción de predicciones correctas en general, qué tan exacto es al momento de predecir, donde en este caso en particular representa un 55%.

La sensibilidad, también conocida el true positive rate o la tasa de verdaderos positivos, mide la proporción de casos positivos que son correctamente identificados por el modelo, donde la sensibilidad del modelo es menor al 50%, indicando que el modelo en un 51% de los casos son clasificados erróneamente como negativos.

Finalmente, el F1 score, es una métrica que indica una medida de la precisión general del modelo, tomando en cuenta tanto los falsos positivos como los falsos negativos, es este caso, el puntaje o la medida obtenida fue de un 53%. En general los resultados se encuentran balanceados ya que sus

métricas se encuentran arriba de un 50%, sin embargo esto no indica una predicción realmente buena.

Modelo DOS

Tomando como variable dependiente profesión y como variables independiente income y age, considerando como etiqueta positiva " Psychologist", los resultados que arrojó fueron los siguientes:

<< Matriz de Confusión >>

True Positive	False Positive
773	857
False Negative	True Negative
667	932

<< Evaluación >>

Precisión	0.5210
Exactitud	0.5280
Sensibilidad	0.5829
F1 Score	0.5502

Esta regresión predice la profesión basada en el ingreso y la edad. Tomando en cuenta los resultados de la matriz de confusión, se puede concluir que el modelo predice correctamente la profesión objetivo 773 veces, si bien, analizando los datos de la matriz las proporciones entre los verdaderos negativos y verdaderos negativos, no es muy diferente, lo que a primera vista sugiere que hay un cierto equilibrio en términos de identificar correctamente ambas clases. Sin embargo, los valores falsos, negativos y positivos, tienen una diferencia en proporción más grande, además de ser valores elevados, indicando que el modelo, tiene dificultades para predecir con precisión ambas clases.

La evaluación indica que aproximadamente el 52% de las predicciones realizadas por el modelo son correctas, un el 52% de las predicciones son precisas en general, el 58% de los verdaderos casos positivos fueron identificados correctamente por el modelo, y finalmente en términos de rendimiento, el modelo tiene un rendimiento moderado en términos de equilibrio entre la precisión y la sensibilidad, debido a su puntaje de 55% en la métrica de F1 score.

Modelo TRES

Tomando como variable dependiente profesión y como variables independiente income, age y experience, considerando como etiqueta positiva " Computer Hardware Engineer". Este modelo predice la profesión basada en tres variables, donde los resultados que arrojó fueron los siguientes:

<< Matriz de Confusión >>

True Positive	False Positive
873	712
False Negative	True Negative
802	842

<< Evaluación >>

Precisión	0.5212
Exactitud	0.5311
Sensibilidad	0.5508
F1 Score	0.5356

La matriz de confusión a primera instancia, se puede considera como una matriz semi balanceada, simplemente al observar los valores, sin embargo, la realidad es que no es considerada como una matriz balanceado, ya que aunque hay un equilibrio en la capacidad del modelo para identificar correctamente ambas clases, los resultados de los falsos positivos y negativos, son altos, lo que a su vez sugiere que el modelo tiene cierta dificultad para predecir con precisión ambas clases.

El resultado de la precisión, indica que en un 52% de todas las predicciones son correctas, con una exactitud del 53%, donde las predicciones realizadas por el modelo son correctas,sugiriendo que el modelo es moderadamente confiable en general. El porcentaje que arrojo la métrica de sensibilidad, indica que el modelo en un 55% es capaz de identificar los casos verdaderos positivos, y finalmente con el puntaje de F1 score del 54%, se infiere que el modelo es capaz de lograr un equilibrio entre la precisión y la capacidad de identificación correcta de los casos positivos.

En resumen, el modelo genera una matriz una balanceada, por los mismo las predicciones logran tener resultados altos en términos de positivos, pero al mismo tiempo altos en términos negativos;además al observar los comportamiento de las métricas de evaluación, se puede ver que los porcentaje aunque pasan de un 50%, estos son relativamente bajos. Se puede decir que este modelo logra predecir en gran medida la profesión basada en el ingreso, la edad y la experiencia.

FILTRO DOS

El segundo filtro también se enfoca en la columna de profesión, ahora enfocados en los que tiene las frecuencias más bajas, donde Statistician & Drafter, son las respectivas. Este nuevo filtro fue utilizado para los siguientes tres modelos.

Modelo CUATRO

Tomando como variable dependiente profesión y como variables independiente income, age y experience, considerando como etiqueta positiva " Statistician". Este modelo predice la profesión basada en tres variables, donde los resultados que arrojó fueron los siguientes:

<< Matriz de Confusión >>	
True Positive	False Positive
638	976
False Negative	True Negative
584	1152

<< Evaluación >>	
Precisión	0.5414
Exactitud	0.5343
Sensibilidad	0.6636
F1 Score	0.5963

La matriz de confusión arroja que hay 638 verdaderos positivos, prediciendo correctamente la profesión; 1152 verdaderos negativos, indicando que se identificaron correctamente mil ciento cincuenta y dos casos como negativos; hay 976 casos donde se identificó erróneamente como positivos cuando en realidad era negativa o en términos del escenario cuando se idéntico de manera incorrecta la posición cuando en realidad no lo era, y finalmente se presentan 584 casos como negativos cuando son positivos. Asimismo se puede observar que los resultados de cada instancia de la matriz no se encuentran relativamente bien relacionadas, lo que representa una matriz no equilibrada.

Pasando con los resultados de la evaluación, estos indican que la precisión del modelo es del 54%, lo que indica que alrededor de ese porcentaje las predicciones positivas realizadas son correctas. La exactitud por su parte indica que alrededor del del 53% del total de las predicciones son correctas, independientemente de si son positivas o negativas. Por parte de la sensibilidad del modelo, su porcentaje del 66, indica que el modelo fue capaz en un 66% de identificar correctamente los casos verdaderamente positivos. Finalmente con la última métrica, su porcentaje fue del 60%, representando un equilibrio razonable entre las métricas de sensibilidad y precisión al momento de considerar un caso como falso positivo o como falso negativo.

Modelo CINCO

Tomando como variable dependiente profesión y como única variable independiente la experiencia, además de considerar como etiqueta positiva la profesión Statistician. Una vez utilizado el modelo de regresión logística, los resultados fueron los siguientes:

<< Matriz de Confusión >>	
True Positive	False Positive
257	1357
False Negative	True Negative
305	1431

<< Evaluación >>	
Precisión	0.5133
Exactitud	0.5039
Sensibilidad	0.8243
F1 Score	0.6326

En esta nueva matriz, los datos nos indican que el modelo fue capaz de predecir la profesión correctamente, cuando el caso lo era 257 veces, a su vez puede predecir 1431 veces la otra profesión cuando realmente lo era; basándonos en los resultados falsos, estos indican que la profesión con el modelo tiene la capacidad de indicar todos los casos positivos cuando en realidad era negativos, lo que en este espacio indica que se identificaron 1357 casos donde la profesión era Statistician y la realidad era la profesión Drafter, y viceversa, sin embargo aquí la diferencia es en número de casos.

Ahora bien, la sensibilidad de este modelo es de un 82%, siendo en modelo hasta el momento con la mejor puntuación, indicando que tiene una gran capacidad para identificar correctamente los verdaderos positivos en relación con el total de casos positivos presentes en el conjunto de datos. Su precisión pasa del 50%, lo que quiere decir que solamente puede realizar correctamente las predicciones de la mitad del conjunto de datos, además, tomando en cuenta el porcentaje de exactitud este nos dice, que independientemente de si es verdadero a falso, de igual manera solo el 50% de las predicciones del conjunto de datos son exactas. Mientras que el porcentaje de F1 score, comenta que en relación de la precisión y la sensibilidad, el modelo es capaz de proporcionar una medida de la precisión general en un 63%.

Modelo SEIS

Para la creación del sexto modelo, se tomaron en cuenta las variables de Experience, Age e Income como las independientes, y como la variable dependiente, Profession, la cuál, como se mencionó anteriormente, hace referencia únicamente a los registros de profesiones de Estadística (Statistician) y la profesión de Diseñador técnico (Drafter) .
Gracias a la función previamente realizada y mencionada, se obtuvieron los siguientes resultados:

<< Matriz de Confusión >>

True Positive	False Positive
615	995
False Negative	True Negative
525	1215

<< Evaluación >>

Precisión	0.5395
Exactitud	0.5463
Sensibilidad	0.3820
F1 Score	0.4473

Como se muestra en la primera tabla a la izquierda, contamos con una tabla donde se indican las clasificaciones de las muestras, en donde se indica si estas fueron clasificadas de la manera correcta o no. Dentro de la tabla contamos con 615 valores clasificados como verdaderos positivos, haciendo referencia a la cantidad de instancias que el modelo es capaz de clasificar de manera correcta. EL resultado de 995 nos indica que el modelo hacalasificado esta cantidad como positivos cuando en

realidad eran negativos; el siguiente escenario, tomando en cuenta los 525 hacen referencia a los falsos negativos, es decir casos que son positivos pero fueron clasificados como negativos, para finalizar, nuestro modelo ha clasificado de manera correcta, al decir que son negativos el resultado de 1215.

Pasando a la evaluación, la precisión nos arroja un resultado de 0.5395, indica la proporción de predicciones que son verdaderamente positivas, es decir, aproximadamente el 53% de las predicciones son certeras. La exactitud es capaz de medir la proporción de predicciones correctas de todas las clases que conforman el modelo, clave valiosa para poder examinar de manera profunda una matriz, el resultado de esta métrica fue del 55%, siendo un poco más de la mitad de lo esperado. Pasando a la sensibilidad o recall, este modelo demuestra tener un 54%, haciendo referencia a que alrededor del 54% de los casos positivos que son verdaderos, fueron detectados.
La última métrica que se encuentra en la tabla F1 Score es de las más importantes ya que demuestra el balance que se tiene entre las métricas, el resultado de 44% es bajo y demuestra se pueden realizar mejoras en el rendimiento del modelo.

FILTRO TRES

El tercer filtro se enfoca en la columna de cuidad, donde se enfoca únicamente en las ciudades con los registros más altos, Indore, segundo de Bhopal, esto para construir una conjunto de datos más específico y observar si el modelo tiene mejores resultados. Este nuevo filtro fue utilizado para el siguiente modelos.

Modelo SIETE

Para el séptimo filtro se necesitó realizar un filtro, siendo este el tercero del código, en donde se busca analizar a las ciudades de Indore y Bhopal, estas ahora unidas para ser nuestra variable dependiente, y Age, Income y Current_house_years como nuestras variables independientes, los resultados de clasificación se muestran y son interpretados a continuación:

<< Matriz de Confusión >>

True Positive	False Positive
232	153
False Negative	True Negative
110	207

<< Evaluación >>

Precisión	0.6784
Exactitud	0.6254
Sensibilidad	0.6026
F1 Score	0.6382

El modelo fue capaz de clasificar de manera correcta y positiva 232 valores, fue capaz de identificar 207 valores negativos de manera correcta, tuvo un total de 263 clasificaciones erróneas, distribuidas de la siguiente manera, 110 fueron instancias positivas verdaderas clasificadas de manera incorrecta

como negativas, y 153 fueron negativas verdaderas, clasificadas como positivas. Esta matriz de confusión a pesar de haber clasificado algunas instancias de manera errónea, se destaca ya que se cuentan con más predicciones correctas que incorrectas; enfocándonos ahora en la tabla referente a su evaluación, observamos que contamos con una precisión del 67.8%, indicando la proporción de predicciones correctas, el resultado de la exactitud se interpreta como el 62.5% de las predicciones de todo el modelo fueron correctas, el 60.2% del modelo pudo identificar los verdaderos positivos y el resultado de las métricas de manera general fue del 63.8%, siendo un resultado bueno, sin embargo mejorable.

FILTRO CUATRO

El tercer filtro también se enfoca en la columna de ciudad, sin embargo, en lugar de utilizar las frecuencias más altas, ahora utiliza el las ciudades con las frecuencias más bajas para construir este nuevo conjunto de datos, donde las ciudades que arrojó un análisis previo como las dos con registros más bajos fueron: Karaikudi y Katni. Este nuevo filtro fue utilizado para el siguiente modelos.

Modelo OCHO

Continuando con nuestro enfoque en las ciudades, como se mencionó en la parte anterior, tomando en cuenta las ciudades de Karaikudi y Katni, representando nuestra variable dependiente, y las mismas columnas del modelo anterior, para continuar con el mismo análisis, pero diferentes ciudades.

<< Matriz de Confusión >>

True Positive	False Positive
69	64
False Negative	True Negative
53	78

<< Evaluación >>

Precisión	0.5656
Exactitud	0.5568
Sensibilidad	0.5188
F1 Score	0.5412

Se nos presentan los siguientes resultados, el modelo tiene la capacidad suficiente para predecir correctamente las clases objetivo un total de 69 veces, y de igual manera, logra predecir 78 negativos verdaderos. Al analizar de manera completa la tabla, observamos que no contamos con una diferencia significativa entre las predicciones correctas e incorrectas, sin embargo, las correctas son mayores, pero por una ligera diferencia de 30. Este hallazgo es considerado como positivo, ya que se deduce que se encuentra un balance entre las predicciones correctas e incorrectas.

La segunda tabla, la de evaluación, nos arroja los resultados en métricas de la matriz de confusión, como se observa, contamos con una precisión del 56%, referente a las predicciones correctas de todo el modelo en general; la exactitud nos indica que el 55.7% de las predicciones positivas fueron

clasificadas con éxito; el rendimiento usando la métrica de recall es de 51.9%, proporción de los casos positivos bien clasificados por el modelo; de manera general entre las métricas se demuestra que el modelo cuenta con un balance del 54%, se deduce que el modelo presenta ciertas dificultades en mantener bajo equilibrio la precisión y sensibilidad.

FILTRO CINCO

Este nuevo filtro, se enfoca en los resultados de la columna de experiencia, específicamente en los registros con experiencia de 18 años y 16 años.

Modelo NUEVE

El siguiente modelo al igual que anteriores, hizo uso de un filtro, ahora nuestra variable objetivo toma en cuenta la experiencia entre los 16 y 18 años y tomamos en cuenta para las variables independientes la columna de Age y Current_Job_Years, los resultados son los siguientes:

<< Matriz de Confusión >>

True Positive	False Positive
953	2691
False Negative	True Negative
895	2874

<< Evaluación >>

Precisión	0.5157
Exactitud	0.5163
Sensibilidad	0.2615
F1 Score	0.3471

La matriz de confusión, sin previo análisis no demuestra tener en balance sus clasificaciones, no se encuentra un correcto equilibrio entre la identificación de sus clases ya que la primera columna demuestra tener menos de la mitad de valores que la primera; contamos con un total de 953 valores positivos predecidos de manera correcta, con 2,874 valores negativos correctos, sin embargo el modelo no fue capaz de predecir 2691 correctos, ya que estas clases fueron incorporadas a la columna de falsos positivos, y contamos con 895 clases en la columna de falsos negativos, es decir, negativos que en realidad debían ser positivos. Nuestro análisis es corroborado con la ayuda de la tabla de evaluación, donde nos encontramos con una precisión del 51%, la mitad de lo esperado, indicando la proporción de predicciones correctas; la exactitud del 51.6% indica las predicciones realizadas por el modelo que son correctas; sin embargo, siendo el balance entre las de 34%, de los más bajos en todo nuestro análisis, este modelo no puede ser interpretado como eficiente a la hora de realizar una clasificación correcta de casos positivos.

FILTRO SEIS

El nuevo filtro, de igual manera se enfoca en la columna de experiencia, donde ahora se utilizan los registros con la experiencia de cero y de trece años.

Modelo DÍEZ

Tomamos los registros de nuestro filtro que conforman la columna de Experience con los valores que deseamos, y lo usamos como la variable objetivo, la independiente. Las variables dependientes continúan siendo las columnas de Age y Current_Job_Years. Los resultados se muestran y analizan a continuación:

<< Matriz de Confusión >>

True Positive	False Positive
3294	0
False Negative	True Negative
0	3617

<< Evaluación >>

Precisión	1.0
Exactitud	1.0
Sensibilidad	1.0
F1 Score	1.0

Como primera instancia observamos que contamos con una matriz de confusión casi perfecta, siendo que todos los clasificadores estuvieron correctos; este resultado se debe a dos posibles opciones, a que hay una falla dentro del modelo, o que en efecto, los resultados si están correctos y clasificados adecuadamente. Las métricas de evaluación comprueban, a todas ser igual a uno, que el modelo realiza de manera correcta todas las predicciones, tanto generales como enfocadas en las clases positivas correctas, el principal y más importante hallazgo fue que nos percatamos de que la mayoría de personas que tienen de 0 a 10 años de experiencia eran en su actual trabajo. Para tener una matriz de confusión más confiable, se realizó una función y modelo referente a la experiencia, pero ahora de manera dicotómica logrando así una mejor y más segura confiabilidad en los modelos.

FILTRO SIETE

Para este nuevo filtro, fue necesario el uso de una función, donde el objetivo de la misma era identificar y reemplazar los valores en dos categorías, mayor a la media y menor a la media, específicamente en la columna de experiencia, transformando los datos diversos de esa categoría, únicamente en dos datos, dependiendo de la característica predefinida, obtenido así que la columna de experiencia tuviera solamente datos numéricos 0 y 1.

Modelo ONCE

Para la creación del onceavo modelo, se utilizó, como se menciona en el filtro número siete, la columna de new_experience como variable objetivo y la columna de Current Job Years como la variable independiente. Los resultados son:

<< Matriz de Confusión >>	
True Positive	False Positive
24062	12332
False Negative	True Negative
9497	29709

<< Evaluación >>	
Precisión	0.7170
Exactitud	0.7113
Sensibilidad	0.6612
F1 Score	0.6879

Contamos con una matriz de confusión con grandes números de predicciones, nuestro modelo cuenta con 24,062 predicciones positivas correctas, y 29,709 predicciones negativas correctas; contamos con 12,332 predicciones clasificadas como positivos cuando en realidad eran negativos, y 9,497 clasificados como negativos cuando en realidad eran positivos. Nuestra matriz de confusión parece tener un buen balance y predicción de resultados, esto se comprueba gracias a la tabla de evaluación, donde observamos que la precisión de las predicciones correctas fue del 72%, un porcentaje elevado en comparación con otros modelos, con una exactitud de manera general de las clasificaciones del 71%, de igual manera elevado. La sensibilidad muestra un decremento en comparación con las métricas de este mismo modelo, pero de igual manera elevado en comparación con otros, siendo este del 66%. De manera general, el balance de las métricas es del 69%, un rendimiento en su mayoría eficiente.

FILTRO OCHO

De la misma manera que en filtro pasado, se utilizó una función para identificar y reemplazar datos, la diferencia es que en este nuevo filtro se utiliza la columna de edad, sin embargo la finalidad se queda igual.

Modelo DOCE

Las variables utilizadas como independientes fueron las mismas que usamos en el filtro, enfocadas en las edades, y como variables dependientes Current_House_Years y Experience. Los resultados fueron los siguientes:

<< Matriz de Confusión >>	
True Positive	False Positive
34084	4247
False Negative	True Negative
32692	4577

<< Evaluación >>	
Precisión	0.5104
Exactitud	0.5114
Sensibilidad	0.8892
F1 Score	0.6486

Al igual que en modelos anteriores, no se tiene un balance dentro de la matriz, contamos con un total de 34,084 predicciones positivas correctas, 4,577 predicciones negativas correctas, y el restante fueron errores dentro de la clasificación de resultados. Las predicciones con valores más fuertes se observan dentro de la primera columna, de valores positivos correctos y falsos negativos. La evaluación consta de las métricas de precisión, donde nos indica cuantas predicciones correctas positivas logró sacar el modelo, su porcentaje fue del 51%, poco más de la mitad, con una exactitud general de acierto del modelo del 51%, mismo porcentaje; la sensibilidad demuestra ser superior, siendo del 89% del modelo un resultado eficiente, y aún más, al juntar la métrica de F1, la cuál es el balance, al ser este resultado del 65%, de igual manera, este modelo prueba ser de los mejores dentro de todo nuestro análisis e interpretación.

Modelo TRECE

Para el último modelo, debido a los resultados del modelo pasado, hicimos uso del mismo filtro, es decir, de igual manera nuestra variable objetivo fueron las edades, y las variables independientes fueron Current_House_Years, Current_Job_Years e Income, los resultados son:

<< Matriz de Confusión >>

True Positive	False Positive
33441	4760
False Negative	True Negative
32295	5104

<< Evaluación >>

Precisión	0.5087
Exactitud	0.5099
Sensibilidad	0.8754
F1 Score	0.6435

El modelo fue capaz de predecir de manera correcta un total de 38,545 instancias, divididas 33,441 como verdaderas positivas y 5,104 como verdaderas negativas; el modelo demuestra tener un buen balance entre predicciones correctas e incorrectas; esto se comprueba en la tabla de evaluación, donde contamos con un porcentaje del 0.5057 indicando que 51% de las predicciones positivas fueron correctas, en total, de manera general, contamos con una exactitud de las predicciones del 51%, y un recall de casi el 90%, en general un rendimiento del 64%, inferior en modelos anteriores, pero demostrando tener un mejor desempeño de manera general.

MODELO	PRECISIÓN
Profession = Psychologist + Computer_hardware_engineer	
Var. Independiente → Income Var. Dependiente → Profession	0.573
Var. Independiente → Income , Age Var. Dependiente → Profession	0.521
Var. Independiente → Income, Age, Experience Var. Dependiente → Profession	0.521
Profession = Statistician + Drafter	
Var. Independiente → Income, Age, Experience Var. Dependiente → Profession	0.541
Var. Independiente → Experience Var. Dependiente → Profession	0.513
Var. Independiente → Experience, Age, Income Var. Dependiente → Profession	0.540
City = Indore + Bhopal	
Var. Independiente → Age, Income, Current House YRS Var. Dependiente → City	0.678
City = Karaikudi + Katni	
Var. Independiente →Age, Income, Current House YRS Var. Dependiente → City	0.566
Experience = 18 + 16	
Var. Independiente → Age, Current Job YRS Var. Dependiente → Experience	0.516
Experience = 0 + 13	
Var. Independiente → Age, Current Job YRS Var. Dependiente → Experience	1.00
Experience = 0 (menor a la media) + 1 (mayor a la media)	
Var. Independiente → Current Job YRS Var. Dependiente → Experience	0.717
Age = 0 (menor a la media) + 1 (mayor a la media)	
Var. Independiente → Current House YRS, Experience Var. Dependiente → Age	0.510

MODELO	EXACTITUD
Profession = Psychologist + Computer_hardware_engineer	
Var. Independiente → Income Var. Dependiente → Profession	0.552
Var. Independiente → Income , Age Var. Dependiente → Profession	0.528
Var. Independiente → Income, Age, Experience Var. Dependiente → Profession	0.531
Profession = Statistician + Drafter	
Var. Independiente → Income, Age, Experience Var. Dependiente → Profession	0.534
Var. Independiente → Experience Var. Dependiente → Profession	0.504
Var. Independiente → Experience, Age, Income Var. Dependiente → Profession	0.546
City = Indore + Bhopal	
Var. Independiente → Age, Income, Current House YRS Var. Dependiente → City	0.625
City = Karaikudi + Katni	
Var. Independiente →Age, Income, Current House YRS Var. Dependiente → City	0.557
Experience = 18 + 16	
Var. Independiente → Age, Current Job YRS Var. Dependiente → Experience	0.516
Experience = 0 + 13	
Var. Independiente → Age, Current Job YRS Var. Dependiente → Experience	1.00
Experience = 0 (menor a la media) + 1 (mayor a la media)	
Var. Independiente → Current Job YRS Var. Dependiente → Experience	0.711
Age = 0 (menor a la media) + 1 (mayor a la media)	
Var. Independiente → Current House YRS, Experience Var. Dependiente → Age	0.511

MODELO	SENSIBILIDAD
Profession = Psychologist + Computer_hardware_engineer	
Var. Independiente → Income Var. Dependiente → Profession	0.493
Var. Independiente → Income , Age Var. Dependiente → Profession	0.583
Var. Independiente → Income, Age, Experience Var. Dependiente → Profession	0.551
Profession = Statistician + Drafter	
Var. Independiente → Income, Age, Experience Var. Dependiente → Profession	0.664
Var. Independiente → Experience Var. Dependiente → Profession	0.824
Var. Independiente → Experience, Age, Income Var. Dependiente → Profession	0.382
City = Indore + Bhopal	
Var. Independiente → Age, Income, Current House YRS Var. Dependiente → City	0.603
City = Karaikudi + Katni	
Var. Independiente →Age, Income, Current House YRS Var. Dependiente → City	0.519
Experience = 18 + 16	
Var. Independiente → Age, Current Job YRS Var. Dependiente → Experience	0.262
Experience = 0 + 13	
Var. Independiente → Age, Current Job YRS Var. Dependiente → Experience	1.00
Experience = 0 (menor a la media) + 1 (mayor a la media)	
Var. Independiente → Current Job YRS Var. Dependiente → Experience	0.661
Age = 0 (menor a la media) + 1 (mayor a la media)	
Var. Independiente → Current House YRS, Experience Var. Dependiente → Age	0.875

MODELO	F1 - SCORE
Profession = Psychologist + Computer_hardware_engineer	
Var. Independiente → Income Var. Dependiente → Profession	0.530
Var. Independiente → Income , Age Var. Dependiente → Profession	0.550
Var. Independiente → Income, Age, Experience Var. Dependiente → Profession	0.536
Profession = Statistician + Drafter	
Var. Independiente → Income, Age, Experience Var. Dependiente → Profession	0.596
Var. Independiente → Experience Var. Dependiente → Profession	0.633
Var. Independiente → Experience, Age, Income Var. Dependiente → Profession	0.447
City = Indore + Bhopal	
Var. Independiente → Age, Income, Current House YRS Var. Dependiente → City	0.638
City = Karaikudi + Katni	
Var. Independiente →Age, Income, Current House YRS Var. Dependiente → City	0.541
Experience = 18 + 16	
Var. Independiente → Age, Current Job YRS Var. Dependiente → Experience	0.347
Experience = 0 + 13	
Var. Independiente → Age, Current Job YRS Var. Dependiente → Experience	1.00
Experience = 0 (menor a la media) + 1 (mayor a la media)	
Var. Independiente → Current Job YRS Var. Dependiente → Experience	0.688
Age = 0 (menor a la media) + 1 (mayor a la media)	
Var. Independiente → Current House YRS, Experience Var. Dependiente → Age	0.643

CONCLUSIÓN

Con las previas tablas pudimos obtener los resultados de las diferentes métricas, como lo fue la precisión, exactitud, sensibilidad y f1-score, la regresión logística nos ayudó a obtener resultados de diferentes variables categóricas en función de una o diversas variables independientes, el modelo con los resultados más elevados se observa cuando se relacionan la variable 'Current Job YRS' y la variable dependiente de "Experience",en donde en tres de las cuatro métricas obtuvo el mayor resultado. Donde la métrica de sensibilidad , con la combinación "Current House YRS, Experience" como variables independientes y "Age" como variable dependiente tuvo un mejor rendimiento. Es importante mencionar que entre más cercano a uno es mejor, por lo que esta es considerada nuestro fundamento para identificar los mejores modelos. Los mejores resultados en términos de métrica fueron los siguientes: F1-Score (0.688), Precisión (0.717), Exactitud(0.711) y Sensibilidad (0.875)

Para todos los casos se decidió estar intercalando las variables independientes en relación con la dependiente, en ciertos casos los resultados arrojados eran 0 puesto que no se tenía ninguna relación entre las mismas, se decidió manipular con ellas para que se pudiera obtener una matriz de confusión y en ella observar los resultados, si existía el mismo número de datos para una y otra variable, esta nos da un panorama, con lo que podemos valorar que tan bueno es el modelo de clasificación basado en aprendizaje automático, en todos los resultados se observó que entre mayor relación tuvieran las variables independientes con las dependientes los resultados arrojados eran mayores en cuanto a todas las métricas generadas, así como tomar en cuenta el número de instancias en relación a cada clase a analizar..