APRENDIZAJE AUTOMÁTICO AVANZADO

INFORME TÉCNICO UNIDAD I – INGENIERÍA DE CARACTERÍSTICAS

PRESENTADO POR

Daniel José Rueda Lobato

Diego Andrés Valderrama Laverde

Catalina Piedrahita Jaramillo

PROFESOR:

José Lisandro Aguilar Castro

UNIVERSIDAD EAFIT

MEDELLÍN

SEPTIEMBRE 2020

1. OBJETIVO DE LA ITERACIÓN

Realizar la Ingeniería de Características sobre una base de datos, para posteriormente desarrollar modelos de clasificación y estudiar sus desempeños para escoger un subconjunto adecuado de características.

2. CONTEXTUALIZACIÓN DEL PROBLEMA.

Realizar la ingeniería de características sobre un conjunto de datos, para obtener los mejores descriptores para el diseño de un modelo supervisado. Para realizar este ejercicio se ha escogido una base de datos de la página de Kaggle con el que se presente clasificar cuáles individuos gana más de 50 mil dólares al año y cuáles no, utilizando como variables de entrada factores con relación a la edad, clase laboral, educación, estado civil, ocupación, raza, genero, horas laborales por semana, país de origen, ganancias y pérdidas de capital.

La base de datps tiene en total 15 variables, 6 cuantitativas y 9 cualitativas.

3. METODOLOGÍA

Con la ingeniería de características se busca obtener datos y descriptores apropiados para que los algoritmos de machine learning alcance un alto desempeño sin la necesidad de invertir grandes recursos, como por ejemplo tiempo de computación o el tiempo de recolección de información.

A continuación, se presentan los pasos desarrollados para esta actividad

3.1. Identificación de valores faltantes e imputación

Los valores faltantes es uno de los principales problemas al momento de realizar la preparación de datos, por ello como primer paso se debe realizar esta validación, revisando si existen valores nulos o caracteres extraños.

No se encontraron valores nulos cuando se realizó la primera inspección

Tabla 1 Identificación de valores nulos sobre la base de datos original

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48842 entries, 0 to 48841
Data columns (total 15 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             48842 non-null  int64
 1   capital-gain    48842 non-null  int64
 2   capital-loss    48842 non-null  int64
 3   hours-per-week  48842 non-null  int64
 4   educational-num 48842 non-null  int64
 5   workclass       48842 non-null  object
 6   fnlwgt          48842 non-null  int64
 7   education       48842 non-null  object
 8   marital-status  48842 non-null  object
 9   occupation      48842 non-null  object
 10  relationship    48842 non-null  object
 11  race            48842 non-null  object
 12  gender          48842 non-null  object
 13  native-country  48842 non-null  object
 14  income          48842 non-null  object
dtypes: int64(6), object(9)
memory usage: 5.6+ MB
```

En la segunda inspección validando ahora caracteres extraños se puedo identificar el símbolo de pregunta “?” para 3 de las 9 variables cualitativas, pesando muy poco en cada una

- Workclass 5.7%
- Occupation 5.7%
- Native-country 1.7%

Los datos con este carácter se consideran valores faltantes, y dado que se presenta en pocas categorías con poco peso en cada una de ellas no se procederá a la eliminación de los registros, se ha realizado una imputación de datos, remplazando los faltantes con la moda.

Tabla 2 Total de caracteres extraños por categoría

workclass	2799
education	0
marital-status	0
occupation	2809
relationship	0
race	0
gender	0
native-country	857
income	0
dtype: int64	

3.2. Estadística descriptiva.

Con la estadística descriptiva se busca entender aún más los datos con los que se cuenta, identificando por ejemplo que información es cuantitativa o cualitativa, que información hay por categorías, que tanta variabilidad presentan, entre otras.

Estadística descriptiva variables cuantitativas

Gráfica 1 Histogramas variables cuantitativas

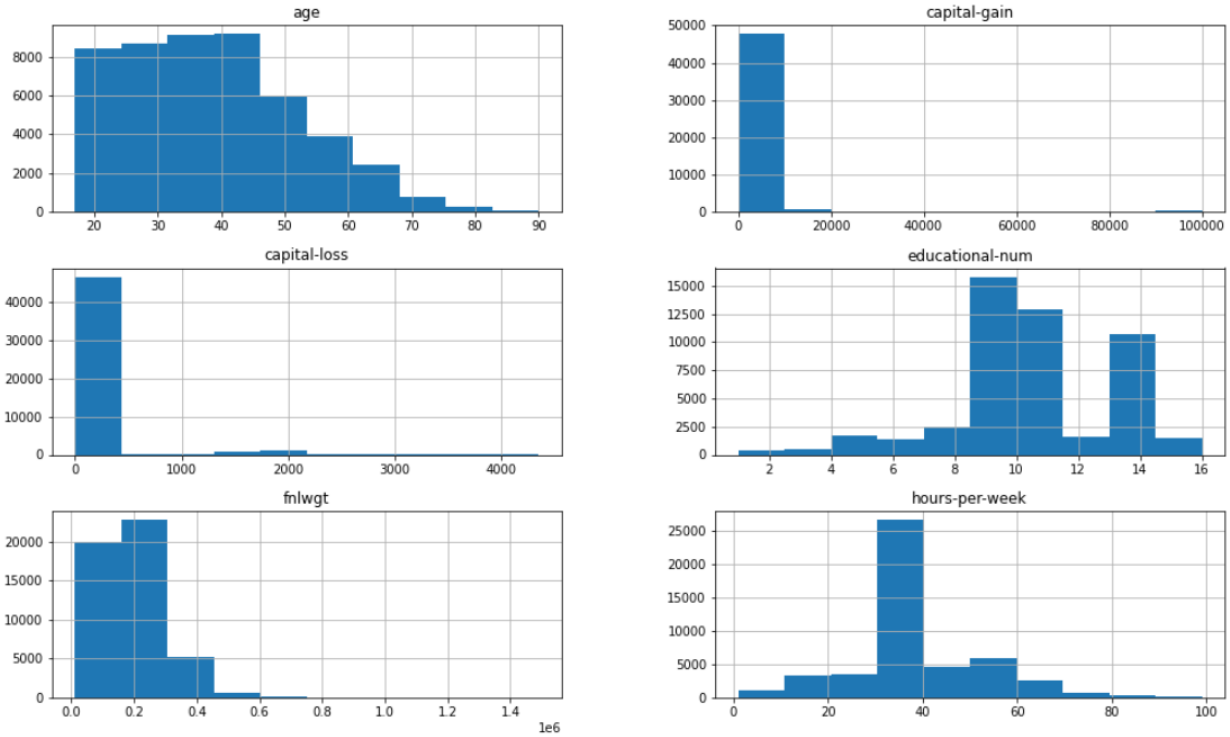


Tabla 3 Estadísticas descriptivas de las variables cuantitativas

	age	capital-gain	capital-loss	hours-per-week	educational-num
count	48842.000000	48842.000000	48842.000000	48842.000000	48842.000000
mean	38.643585	1079.067626	87.502314	40.422382	10.078089
std	13.710510	7452.019058	403.004552	12.391444	2.570973
min	17.000000	0.000000	0.000000	1.000000	1.000000
25%	28.000000	0.000000	0.000000	40.000000	9.000000
50%	37.000000	0.000000	0.000000	40.000000	10.000000
75%	48.000000	0.000000	0.000000	45.000000	12.000000
max	90.000000	99999.000000	4356.000000	99.000000	16.000000

Algunas observaciones sobre las variables cuantitativas.

- La variable edad tiene un rango amplio con un mínimo de 17 años y un máximo de 90. Aunque se puede apreciar que la mayoría está entre los 20 y los 60. La edad promedio es de 38 años
- Para el capital ganado y perdido se puede destacar grandes desviaciones estándar, la medía de capital-gain es 1079 y de capital-loss 87, pero para ambos la mediana es cero, lo que indicaría alta simetría. En el histograma puede apreciarse que para ambas variables en la mayoría de los casos el valor es cero. Dado esto pueden ser buenas candidatas para convertirse en una sola variable, reemplazándolas por el total (total=gain-loss) ya que podría brindar la misma información.
- Para la variable de horas trabajadas por semana se destaca que el 75% de las personas trabaja 45 horas por semana o menos.
- Para el caso de años estudiados puede apreciarse que la media y la mediana son iguales 10 años, al ver la desviación de 2.5 años y el histograma podemos apreciar que en efecto la mayoría de las personas parecen tener entre 7.5 y 12.5 años de estudio

Estadística descriptiva variables cualitativas

Tabla 4 Estadísticas descriptivas variables cualitativas

	workclass	education	marital-status	occupation	relationship	race	gender	native-country	income
count	48842	48842	48842	48842	48842	48842	48842	48842	48842
unique	8	16	7	14	6	5	2	41	2
top	Private	HS-grad	Married-civ-spouse	Prof-specialty	Husband	White	Male	United-States	<=50K
freq	36705	15784	22379	8981	19716	41762	32650	44689	37155

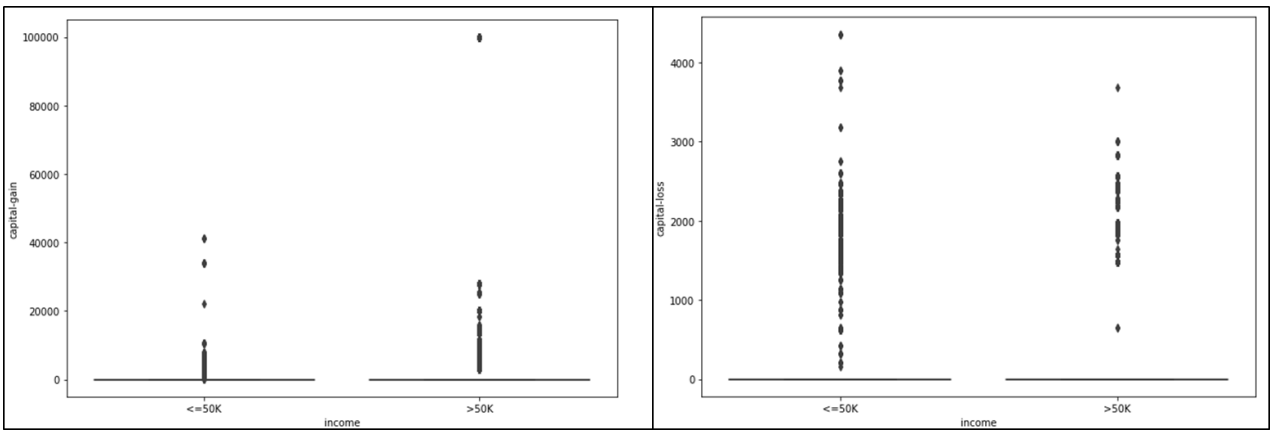
Algunas observaciones sobre las variables cualitativas

- Puede apreciarse que para algunas variables existen muchas categorías, y con los diagramas de barras del notebook pudo identificarse que algunas cuentan con muy pocos datos.
- Con respecto a la clase de trabajo, se puede destacar que más de la mitad de las personas encuestadas en esta base de datos pertenecen al sector privado

- Hay 16 tipos de niveles educativos en la base de datos y la moda representa el 32.3%
- Del estado marital el 46% de las personas indican estar casadas
- Hay 14 categorías para ocupación, la más representativa Pro-speciality equivale al 20% del total, lo que implica que no hay un peso absoluto de la moda
- De raza y genero puede destacarse que más de la mitad son hombres blancos
- El 92% de las personas de este base de datos tiene como país de origen Estados Unidos, lo que no es de extrañar dado que este censo es en ese país.

De las conclusiones obtenidas después de realizar la estadística descriptiva es convertir el capital gain and loss en una variable de capital total. Esto dado que para el 75% de los datos en ambos casos el valor es cero; esto también puede apreciarse en los boxplot donde en ambas categorías de la variable respuesta el valor acumulado es cero.

Gráfica 2 Boxplot Capital gain vs income y Capital loss vs income



3.3. Codificación One-hot y Estandarización

La base de datos tiene 9 variables cualitativas, pero para poder realizar un análisis de puntos atípicos y el entrenamiento de modelos es necesario convertirlas en variables numéricas, para ello se utilizó el one-hot encoding, el cual consiste en distribuir los valores de una columna en varias variables indicadoras asignando 1 si el individuo pertenece a la categoría y 0 de lo contrario, expresando así la relación entre la columna y el individuo.

Después de realizar esta transformación la base de datos pasa a tener una dimensión de 48842 filas por 104 columnas.

También se han estandarizado las variables numéricas restándole su media y dividiendo por su desviación estándar. Esto es aconsejable hacerlo ya que de esta manera se pueden hacer mejores comparaciones entre variables numéricas eliminando por ejemplo influencias de unidades de medida, al tener grandes valores en una y escalas más bajas para otras.

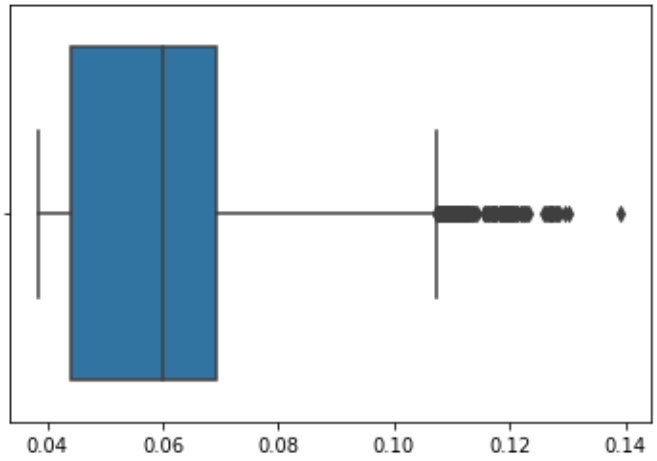
3.4. Identificación de outliers

Realizar la identificación de valores atípicos es otro importante paso de la preparación de los datos, porque su presencia puede afectar los modelos de machine learning generando que tengan una baja precisión.

Para la identificación de puntos atípicos dado que la base de datos es mixta, se ha utilizado la distancia de Gower, y para medir esta distancia de ha estimado un vector centro tomando la mediana de las variables cuantitativas y la moda de las cualitativas. El procedimiento ha sido entonces medir la distancia de Gower de cada individuo al vector centro.

Utilizando los bigotes del boxplot se han identificado los outliers. Los que fueron equivalentes a 294 puntos, 0.6% del total.

Gráfica 3 Boxplot distancias de Gower



3.5. Selección de características

3.5.1. Análisis estadístico univariante para selección de características (Método Filtro)

Con este proceso se da una mirada individual a cada variable y su relación con la variable respuesta, esto con el objetivo de identificar si agrega o no valor y entonces concluir si la variable será parte del subconjunto de características para el modelo.

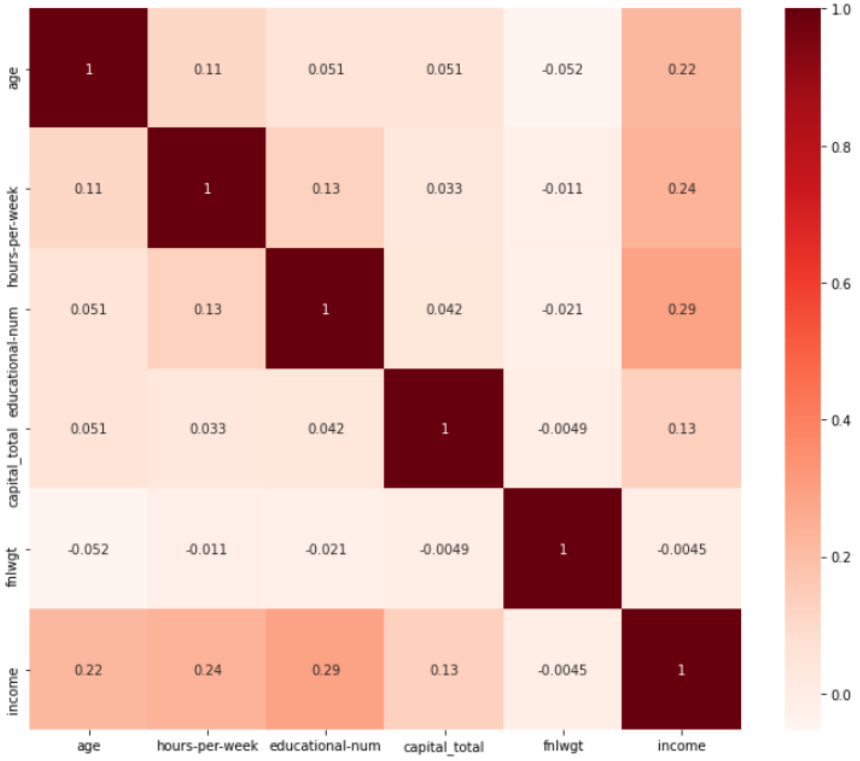
Dado que la base de datos cuenta con variables cuantitativas y cualitativas y la variable respuesta es cualitativa, usaremos la correlación de Kendall para evaluar la correlación de las variables numéricas con income y chi-cuadrado para evaluarlo con las cualitativas.

Análisis variables cuantitativas

Tabla 5 Correlación de Kendall y Valor P de las variables

	Variable	Tau	p_value
0	age	0.222680	0.000000e+00
1	hours-per-week	0.237639	0.000000e+00
2	educational-num	0.289631	0.000000e+00
3	capital_total	0.133651	5.709543e-203
4	fnlwgt	-0.004497	2.249291e-01

Gráfica 4 Correlación de Kendall de las variables cuantitativas

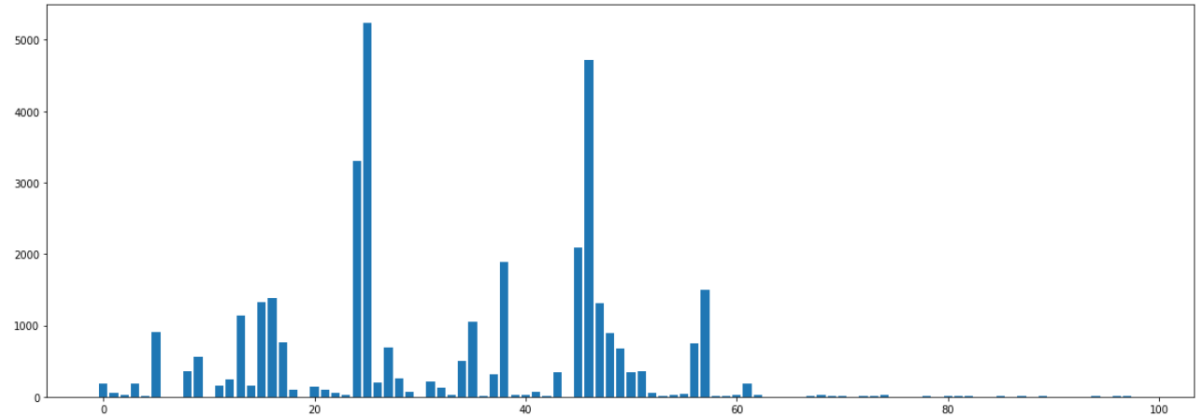


Para las variables cuantitativas mediante este análisis se concluye que 4 ellas son significativas y deben permanecer en el modelo

Análisis variables cualitativas

La función utilizada para esta operación fue SelectKBest, mediante esta función también se obtiene un puntaje que resalta la significancia de una variable, a continuación, el grafico para nuestro caso en el que se puede apreciar como algunas resaltan de manera muy significativa. Las 3 primeras fueron marital-status_2=Married-civ-spouse, relationship_2=Husband y marital-status_1=Never-married

Gráfica 5 Score de las variables cualitativas



Para seleccionar las variables cualitativas mediante la prueba de Chi-cuadrado se utilizó el valor p filtrando aquellas con un valor menor a 0.05, esto nos dio como resultado 76 variables significativas.

En total mediante este método se obtiene 80 características seleccionadas

3.5.2. Recursive Feature Elimination (Método de Envoltorio)

La eliminación de características mediante Recursive Feature Elimination (RFE) es una estrategia de envoltura en la que se usa un algoritmo de clasificación predeterminado para eliminar características que no se consideran relevantes, el punto de partida es todo el conjunto, el modelo clasifica las características por importancia, descartando las menos importantes y reajustando el modelo. Es un proceso repetitivo hasta que quede un número especifico de características.

Gráfica 6 Código RFE con Random Forest. Librería Sklearn

```
from sklearn.feature_selection import RFE
from sklearn.ensemble import RandomForestRegressor
```

```
df_X_res = None
df_X_res = pd.DataFrame(X)
df_y_res = None
df_y_res = pd.DataFrame(y)
#Usando RFE para elegir caracteristicas
rfe_selector = RFE(RandomForestRegressor(n_estimators=300), n_features_to_select=None, step=10, verbose=5)
rfe_selector.fit(X, y)
```

Para este ejercicio académico se ha escogido Random Forest como el algoritmo de clasificación y establecido obtener la mitad de las características. Teniendo como resultado las 5 variables cuantitativas y 47 variables cualitativas. Total 52 características seleccionadas

3.5.3. Support Vector Classification lsvc (Método Integrado).

Este es un método integrado ya que realizar una estimación para determinar cuáles características contribuyen para realizar la clasificación, para este ejercicio se le ha incluido una penalización L1 o regularización de Lasso la cual se aplica sobre los coeficientes que multiplican cada una de las características reduciendo a cero aquellos que se consideren menos relevantes.

Gráfica 7 Código Support Vector Machine con penalización L1. Librería Sklearn

```
#  Configurar modelo svm
lsvc = LinearSVC(C=1,                              # Parámetro de regularización por defecto de la libreria
                 penalty="l1",                     # Regularizacion L1 para seleccion de caracteristicas
                 max_iter = 20000,
                 dual=False).fit(X, y)

# Configurar modelo de selección de variables

model = SelectFromModel(lsvc, prefit=True)        # Meta-transformador para seleccionar características basadas en pesos de importancia.
X_new = model.transform(X)
```

En total con este método se escogieron 91 características.

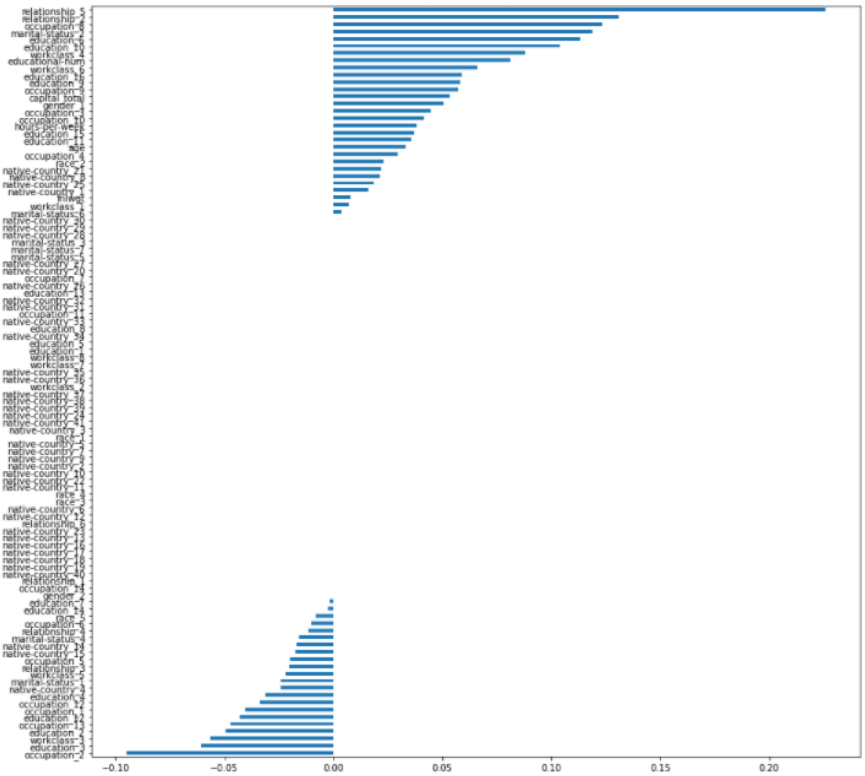
3.5.4. Estimación de regresión Lasso con penalización. (Integrado)

Realizar una regresión de Lasso con una penalización significa que se ha incluido un término regulador Alpha, con el que se busca dar una sanción a los coeficientes para lograr un modelo más simple, esto provoca que algunos coeficientes relacionados a características que se consideren poco relevantes para estimar la variable respuesta terminen siendo cero. Por ello este método también se considera integrado, ya que realiza la selección automática de características durante la estimación.

Cuando se realiza este proceso parte del reto está en definir el Alpha optimo, por ello se decidió usar la función LassoCV de la librería Sklearn que ayuda encontrar el Alpha óptimo para la regularización.

En total con este método se escogieron 51 características. A continuación, se muestra en la gráfica que indica el peso de las variables seleccionadas. Las 2 más importantes fueron relationship_5=Wife y occupation_2=Farming-fishing

Gráfica 8 Diagrama de barras importancia de características en Lasso

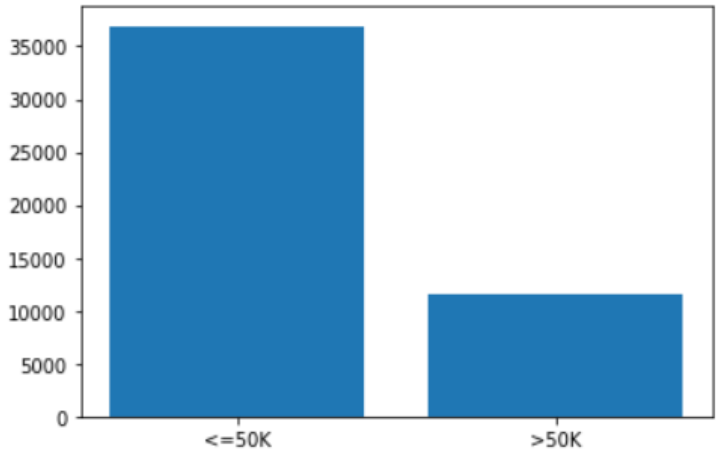


3.6. Balance de categorías

Ahora se procede a dividir la base de datos en entrenamiento y testeo, lo cual se hizo 70-30, y se revisa cual es la representación total de datos que se tiene en el conjunto de entrenamiento por cada categoría de la variable respuesta.

Al revisar la representación de los datos en las 2 categorías de la variable respuesta se pudo apreciar un desbalance, donde la categoría de quienes gana más de 50 mil dólares al año cuenta con menos datos que la categoría de quienes ganan menos

Gráfica 9 Diagrama de barras. Total, de datos por categoría

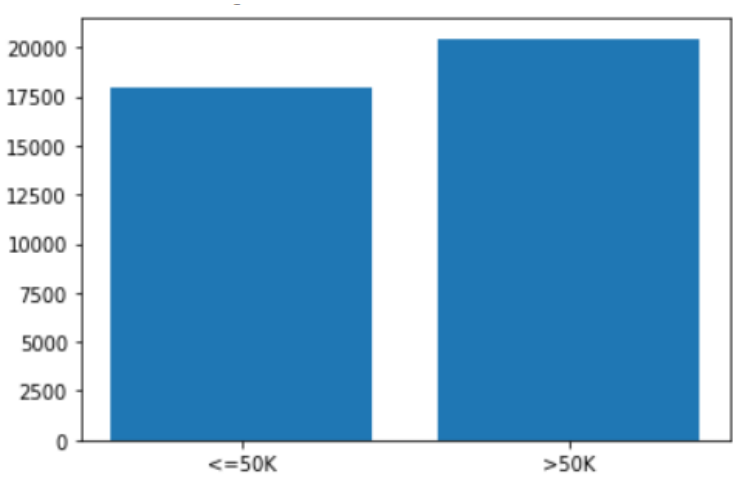


Es importante realizar un balance de categorías ya que la poca representación de una clase puede implicar dificultades futuras para el entrenamiento del modelo de machine learning, como por ejemplo que la subrepresentación de una clase provoque que el algoritmo solo aprenda la dominante, entre otras posibles dificultades.

En el gráfico de barras se puede apreciar la gran diferencia de datos para cada categoría, por ello decidimos no solo sobremuestrear sino también submuestrear, es decir, combinar ambas técnicas para obtener una base de datos de entrenamiento más balanceado, el procedimiento utilizado fue SMOTEENN, que usa la técnica de SMOTE para sobremuestrar y ENN para submuestrear.

Es de resaltar que estos resultados son variables para cada corrida y para cada uno de los subconjuntos resultantes en la selección de características, pero en general se puede apreciar cómo se obtiene un base de datos balanceada, a continuación, una gráfica ejemplo de uno de los resultados

Gráfica 10 Diagrama de barras. Total, de datos por categoría después de realizar SMOTEEN a set de entrenamiento



Cuando se realiza la ingeniería de características es importante tener claro el objetivo para definir las estrategias a realizar, en este caso nosotros consideramos el de reducir dimensionalidad teniendo una buena exactitud, por ello después de haber realizado este ejercicio de limpieza, análisis, selección de características y balanceo de categorías procederemos en la iteración II de supervisado a evaluar cada uno de los resultados para identificar con que combinación de características y modelo se logra la mejor precisión.

4. CONCLUSIONES

- Es importante cuando se va a realizar la ingeniería de características realizar la estadística descriptiva porque esto permite tener un mejor detalle de las variables y sus interacciones con la variable respuesta, en nuestro caso nos permitió identificar variables que tenían potencial de ser combinadas.

- Antes de realizar la selección de características es importante realizar los pasos de identificación de valores faltantes y estadística descriptiva para tomar medidas con relación a eliminar o imputar información faltante, ya que información equivocada o nula puede tener un impacto negativo en las técnicas de selección.

- Es esencial identificar el tipo de datos que contiene la base de datos, si las variables son cuantitativas, cualitativas o si es mixto, ya que esto determina el tipo de procedimiento y cálculos que se pueden realizar, no todas las métricas o metodologías aplican para datos cualitativos, como por ejemplo aquellas relacionadas con las distancias para identificar valores atípicos.

- Cuando se trabaja con una base de datos que contiene variables categorías cada una de las cuales, con múltiples categorías, es importante validar cuando se realiza la selección si cada una de ellas ha sido significativa ya que de concluirse que no, se podría simplificar la recolección de la información o de las encuestas para ocasiones futuras. Por ejemplo, en nuestro caso en varios de los métodos de selección utilizados fueron eliminados varias de las categorías relacionadas a país.

- Cuando se tiene variables categorías es importante definir estratégicamente las categorías que tendrá cada una de ellas, ya que es necesario crear variables indicadoras para que la información cualitativa sea

entendida por los algoritmos de machine learning y un gran número de categorías en múltiples variables categorías incrementara significativamente la matriz de datos.

- Es importante graficar el estado de las categorías en el set de entrenamiento para identificar que tanta representatividad tiene cada grupo, pues esto puede determinar la decisión de qué tipo de estrategia utilizar para balancear si sobremuestrear, submuestrear o mezclar.

5. REFERENCIAS

A continuación, algunas referencias con relación a información de librerías y metodologías utilizadas

Brownlee, J. (2020, August). Retrieved from Recursive Feature Elimination (RFE) for Feature Selection in Python: https://machinelearningmastery.com/rfe-feature-selection-in-python/

Brownlee, J. (2020, Agosto). Retrieved from How to Choose a Feature Selection Method For Machine Learning: https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/

Brownlee, J. (2020, Agosto). *How to Combine Oversampling and Undersampling for Imbalanced Classification*. Retrieved from https://machinelearningmastery.com/combine-oversampling-and-undersampling-for-imbalanced-classification/

Kes, S. (2019, Mayo). *Medium*. Retrieved from How to calculate Gower's Distance using Python: https://medium.com/analytics-vidhya/concept-of-gowers-distance-and-it-s-application-using-python-b08cf6139ac2

Nogueria, F., Lemaitre, G., Victor, D., & Aridas, C. (n.d.). *imbalanced-learn*. Retrieved from https://imbalanced-learn.readthedocs.io/en/stable/generated/imblearn.combine.SMOTEENN.html

pypi. (n.d.). *Gower*. Retrieved from https://pypi.org/project/gower/