UANL
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Maestría en Ciencia de Datos

Aprendizaje Automático

Alvaro Pequeño Mondragón
1726520

Reporte #4

21 de julio de 2024

Para el reporte 4 se requiere realizar un modelo de clustering en el cual se usarán los datos que se utilizaron en el reporte anterior el cual es el reporte 3. En el reporte anterior se utilizó la información correspondiente a las preparatorias y las facultades donde se toma en cuenta el numero de empleados por dependencia y el sueldo diario promedio que paga cada dependencia.

Primero se visualizan los datos originales y se entrena el modelo:

```
#Se mantienen los datos usados en la practica anterior:
model3 = sueldo_dependencia.loc[(sueldo_dependencia['tipo'] == 'PREPARATORIA') | (sueldo_dependencia['tipo'] == 'FACULTAD')]
model3_x = model3.drop(['dependencia', 'tipo', 'sueldo_neto'], axis = 1, inplace = False)
model3_y = model3.drop(['dependencia', 'sueldo_neto', 'num_empleados', 'pago_diario_promedio'], axis = 1, inplace = False)
```

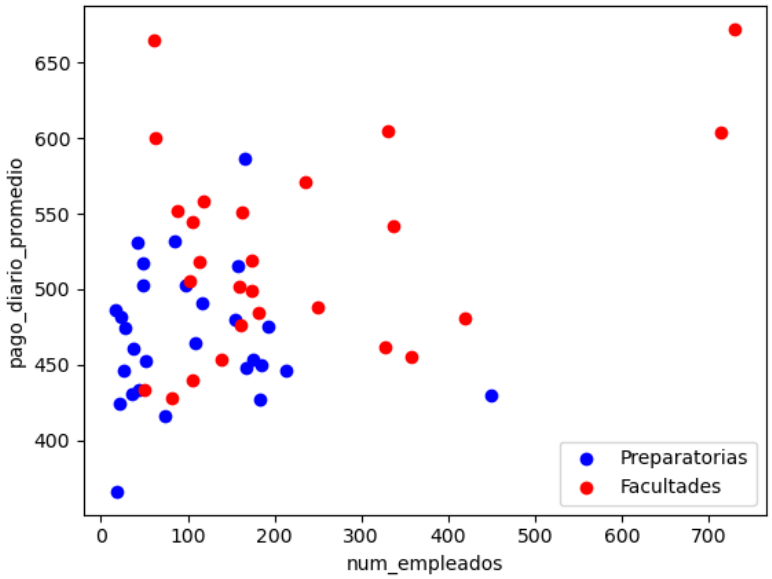
```
#Se separan los datos para observar gráficamente la separación actual de los datos:
num_prepas = model3['tipo'].value_counts()['PREPARATORIA']
model3_prepas = model3_x.iloc[:num_prepas]
model3_facultades = model3_x.iloc[num_prepas:]
```

```
#Se grafican los valores para observar visualmente a cada grupo:
plt.scatter(model3_prepas['num_empleados'], model3_prepas['pago_diario_promedio'], color='blue', label = 'Preparatorias')
plt.scatter(model3_facultades['num_empleados'], model3_facultades['pago_diario_promedio'], color='red', label = 'Facultades')
plt.xlabel("num_empleados")
plt.ylabel("pago_diario_promedio")
plt.legend()
```

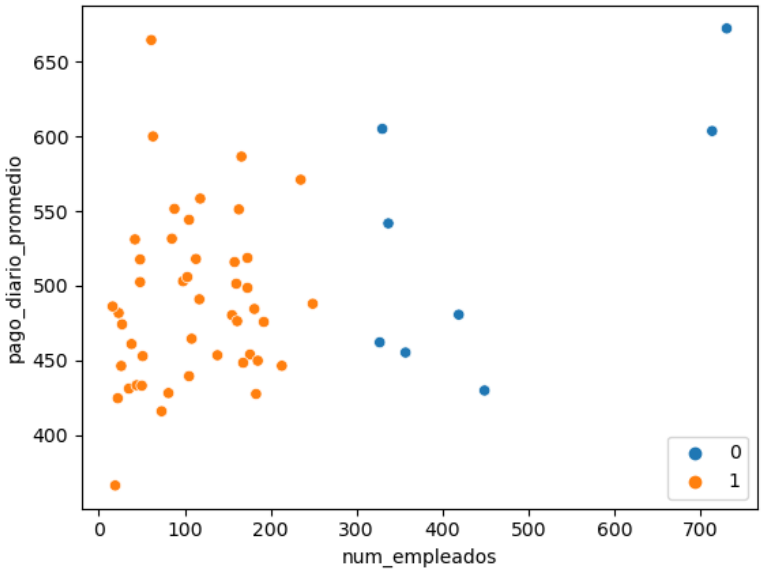
```
Outputs are collapsed ...
```

```
#Se realiza el modelo de agrupación:
model_kmeans = KMeans(n_clusters = 2, random_state = 0, n_init='auto')
model_kmeans.fit(model3_x)
```

A continuación, se muestran los datos originales y su correcta clasificación:



Una vez entrenado el modelo podemos extraer la información sobre los grupos que identificó el modelo, como en este caso estamos hablando de 2 clases que son preparatoria y facultad en los parámetros del modelo se especificó que el modelo nos identificara 2 agrupaciones, a continuación, se muestra de manera gráfica las 2 agrupaciones que fueron identificadas:



En la imagen anterior se puede observar que el modelo si pudo identificar 2 agrupaciones pero al comparar esta grafica con la de los datos correctamente clasificados podemos ver que los grupos del modelo no separa de manera totalmente correcta las facultades de las preparatorias, si se observa prácticamente todos los puntos por debajo de 300 empleados los clasifica como un mismo grupo cuando en realidad en esa sección hay puntos que corresponden a las 2 clases disponibles.