




# INTROVERT OR EXTROVERT BEHAVIOR

**Vanessa Herrera Giraldo**



# PERSONALITY DATASET DESCRIPCION

Este dataset se enfoca en los rasgos de personalidad extrovertida e introvertida, con la finalidad de determinar a que grupo pertenece la persona, contiene información sobre comportamientos y hábitos sociales de las personas. Incluye variables como el tiempo que pasan solas, la frecuencia con la que asisten a eventos sociales y su interacción en redes sociales, es un problema de clasificación, donde la columna objetivo es Personality.



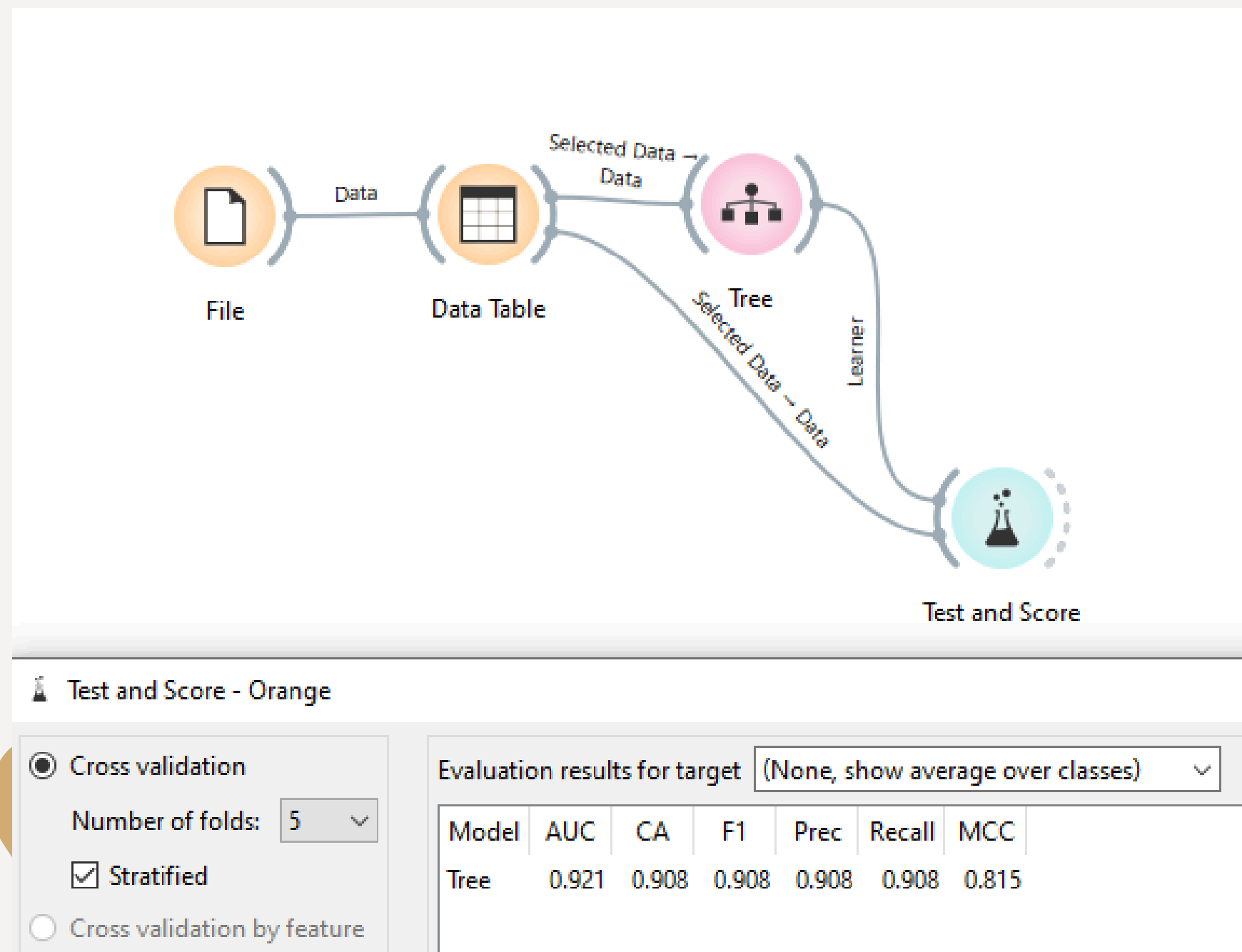
Es útil para psicólogos, sociólogos, empresas de marketing y científicos de datos que buscan entender o segmentar perfiles de personalidad para estudios o campañas.

link: <https://www.kaggle.com/datasets/rakeshkapilavai/extrovert-vs-introvert-behavior-data>



# PERSONALITY DATASET

## F1



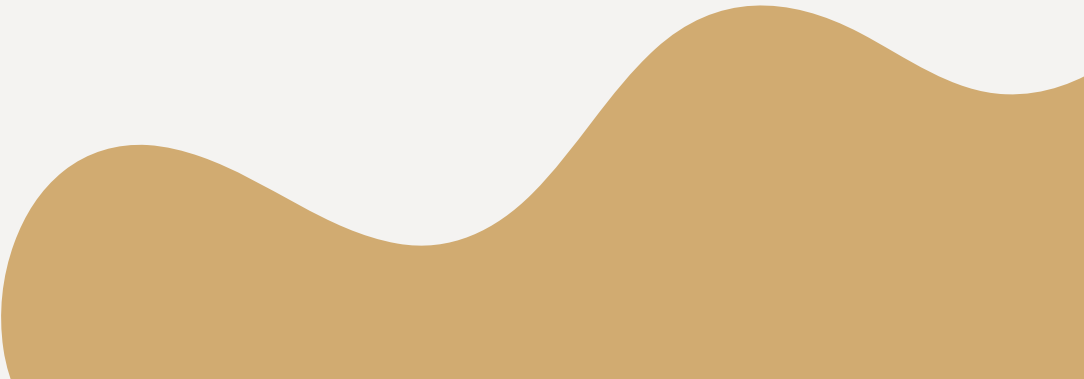

con base en el F1 del modelo hecho en orange se obtuvo un puntaje de 0.908, lo que indica que los datos podrían ser factibles para realizar las predicciones con el modelo (el f1 mide la calidad del rendimiento del modelo ).



# PERSONALITY DATASET

## ELIMINACION VARIABLES

No se eliminó ninguna variable, ya que considero que todas las columnas aportan información relevante para el objetivo de clasificación. Cada variable representa un aspecto del comportamiento social o personal que puede influir en el modelo, por lo tanto, no se identificaron columnas irrelevantes ni redundantes en esta etapa, además tienen relación con el target.



# PERSONALITY DATASET

## VARIABLES NUMERICAS Y CATEGORICAS

Data columns (total 8 columns):

| # | Column                    | Non-Null Count | Dtype   |
|---|---------------------------|----------------|---------|
| 0 | Time_spent_Alone          | 2900 non-null  | float64 |
| 1 | Stage_fear                | 2900 non-null  | object  |
| 2 | Social_event_attendance   | 2900 non-null  | float64 |
| 3 | Going_outside             | 2900 non-null  | float64 |
| 4 | Drained_after_socializing | 2900 non-null  | object  |
| 5 | Friends_circle_size       | 2900 non-null  | float64 |
| 6 | Post_frequency            | 2900 non-null  | float64 |
| 7 | Personality               | 2900 non-null  | object  |

dtypes: float64(5), object(3)  
memory usage: 181.4+ KB

después de cargar el dataset y su estructura general con ayuda de pandas, se logro identificar las variables numéricas y categóricas (stage\_fear, drained\_after\_socializing, personality), cabe resaltar que no tenia variables categóricas representadas como numéricas,.

# PERSONALITY DATASET

## VALORES NULOS

Número de valores nulos por columna antes de la eliminación:

|                           |   |
|---------------------------|---|
| Time_spent_Alone          | 0 |
| Stage_fear                | 0 |
| Social_event_attendance   | 0 |
| Going_outside             | 0 |
| Drained_after_socializing | 0 |
| Friends_circle_size       | 0 |
| Post_frequency            | 0 |
| Personality               | 0 |

Número de valores nulos por columna después de la eliminación:

|                           |   |
|---------------------------|---|
| Time_spent_Alone          | 0 |
| Stage_fear                | 0 |
| Social_event_attendance   | 0 |
| Going_outside             | 0 |
| Drained_after_socializing | 0 |
| Friends_circle_size       | 0 |
| Post_frequency            | 0 |
| Personality               | 0 |

como se logra ver en la imagen no habían valores nulos por lo tanto no se debía eliminar ninguno. el tratamiento de datos en este caso eliminacion se hace para evitar errores en el modelo y sesgos.

# PERSONALITY DATASET

## DESCRIPCION ESTADISTICA

|       | Time_spent_Alone | Stage_fear  | Social_event_attendance | Going_outside | Drained_after_socializing | Friends_circle_size | Post_frequency | Personality |
|-------|------------------|-------------|-------------------------|---------------|---------------------------|---------------------|----------------|-------------|
| count | 2900.000000      | 2900.000000 | 2900.000000             | 2900.000000   | 2900.000000               | 2900.000000         | 2900.000000    | 2900.000000 |
| mean  | 4.505816         | 0.486207    | 3.963354                | 3.000000      | 0.485172                  | 6.268863            | 3.564727       | 0.485862    |
| std   | 3.441180         | 0.499896    | 2.872608                | 2.221597      | 0.499866                  | 4.232340            | 2.893587       | 0.499886    |
| min   | 0.000000         | 0.000000    | 0.000000                | 0.000000      | 0.000000                  | 0.000000            | 0.000000       | 0.000000    |
| 25%   | 2.000000         | 0.000000    | 2.000000                | 1.000000      | 0.000000                  | 3.000000            | 1.000000       | 0.000000    |
| 50%   | 4.000000         | 0.000000    | 3.963354                | 3.000000      | 0.000000                  | 5.000000            | 3.000000       | 0.000000    |
| 75%   | 7.000000         | 1.000000    | 6.000000                | 5.000000      | 1.000000                  | 10.000000           | 6.000000       | 1.000000    |
| max   | 11.000000        | 1.000000    | 10.000000               | 7.000000      | 1.000000                  | 15.000000           | 10.000000      | 1.000000    |

Podemos deducir que las variables Social\_event\_attendance y Going\_outside presentan una posible distribución normal, ya que su media y mediana son prácticamente iguales, lo que facilitaría el ajuste de modelos predictivos y su rendimiento. Las variables Time\_spent\_Alone y Post\_frequency también muestran una cercanía considerable entre la media y la mediana, lo cual sugiere una posible distribución normal, aunque no tan clara como en los casos anteriores.

Por otro lado, la variable Friends\_circle\_size presenta una media notablemente mayor que la mediana, lo que indica un posible sesgo a la derecha.



# PERSONALITY DATASET

## DUPLICADOS

Número de registros duplicados antes de la eliminación: 402

DataFrame después de eliminar registros duplicados:

|   | Time_spent_Alone | Stage_fear | Social_event_attendance | Going_outside \ |  |
|---|------------------|------------|-------------------------|-----------------|--|
| 0 | 4.0              | 0          | 4.0                     | 6.0             |  |
| 1 | 9.0              | 1          | 0.0                     | 0.0             |  |
| 2 | 9.0              | 1          | 1.0                     | 2.0             |  |
| 3 | 0.0              | 0          | 6.0                     | 7.0             |  |
| 4 | 3.0              | 0          | 9.0                     | 4.0             |  |

|   | Drained_after_socializing | Friends_circle_size | Post_frequency | Personality |
|---|---------------------------|---------------------|----------------|-------------|
| 0 | 0                         | 13.0                | 5.0            | 0           |
| 1 | 1                         | 0.0                 | 3.0            | 1           |
| 2 | 1                         | 5.0                 | 2.0            | 1           |
| 3 | 0                         | 14.0                | 8.0            | 0           |
| 4 | 0                         | 8.0                 | 5.0            | 0           |

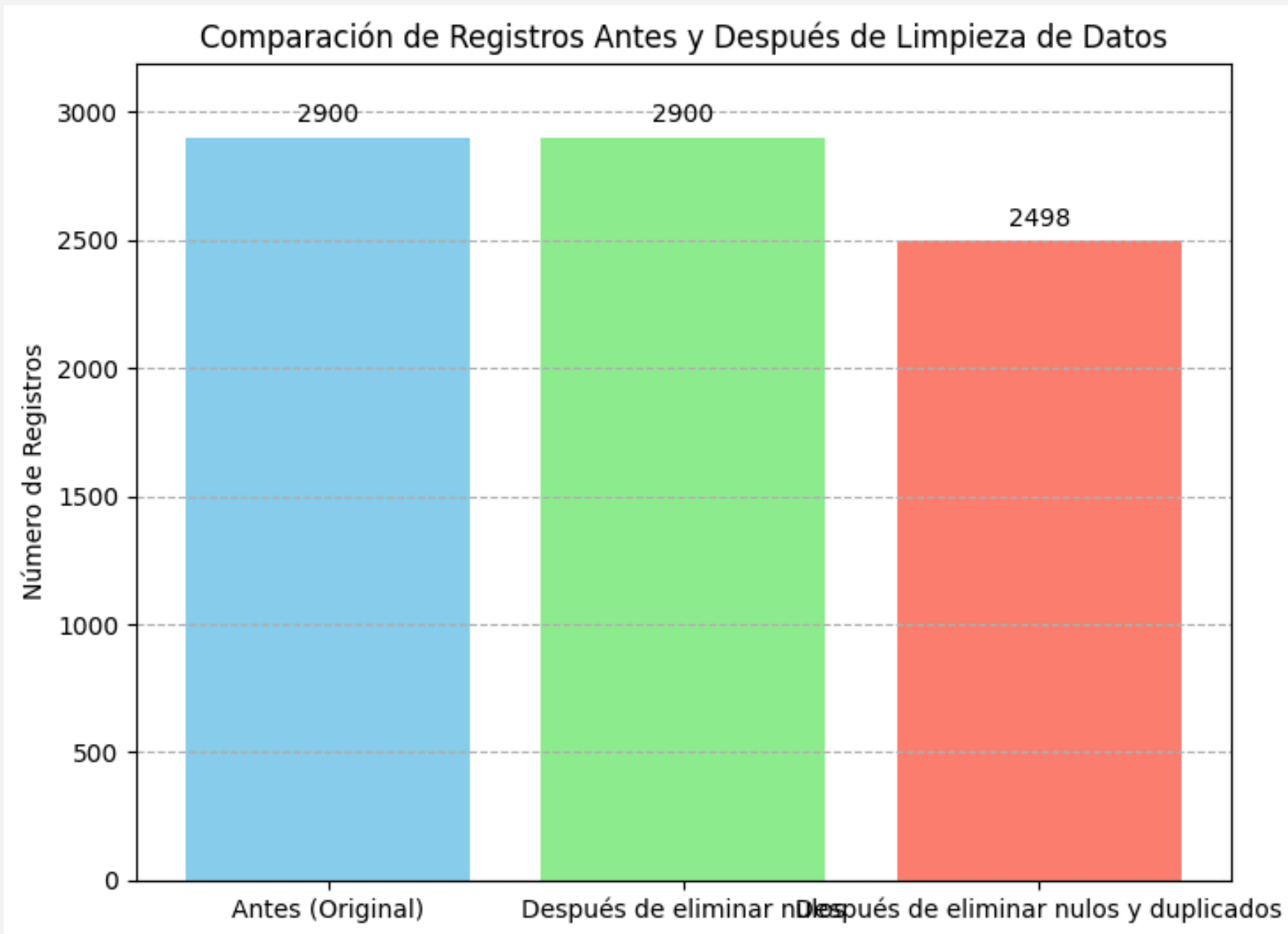
Número de registros después de eliminar duplicados: 2498

Duplicados restantes: 0

se puede observar 402 duplicados, que fueron eliminados como tratamiento de datos ya que estos pueden generar sesgos. se muestra que al final quedan 0 duplicados para confirmar el tratamiento.



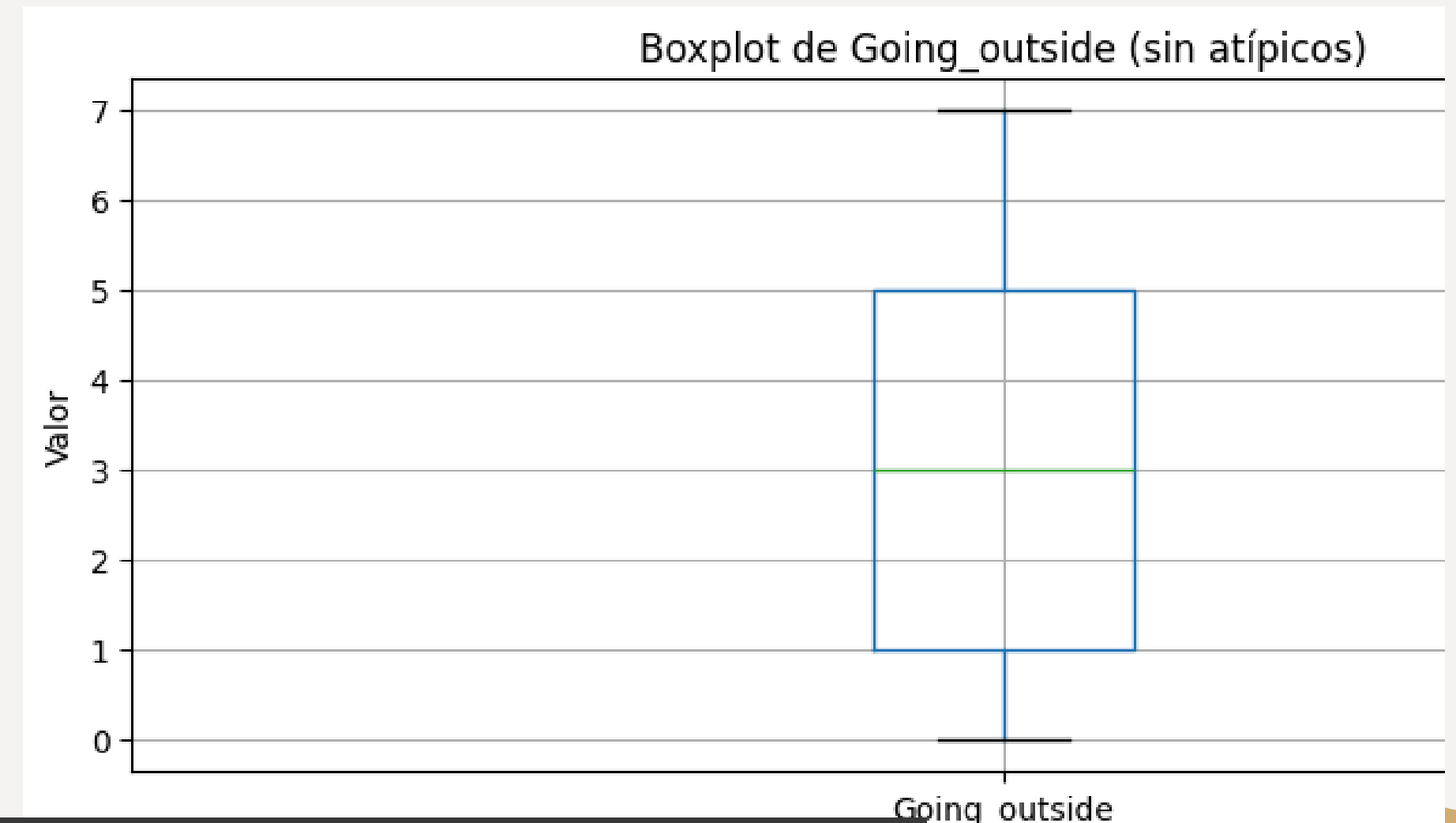
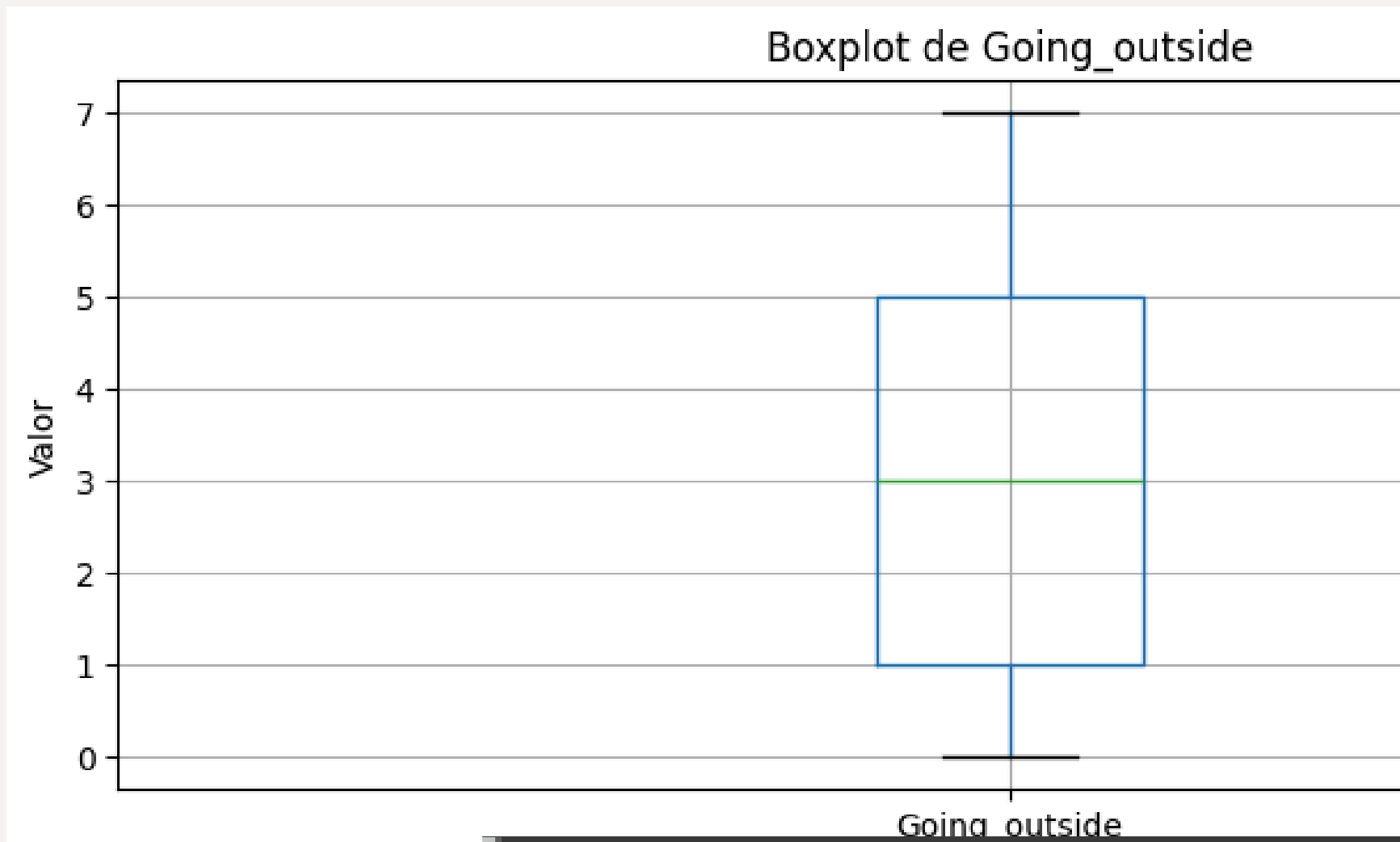
# PERSONALITY DATASET COMPARACION DATOS DESPUES DE LIMPIARLOS



se logra ver que antes de limpiar los datos se tenían 2900 registros, después de eliminar los nulos se tienen 2900, es decir no habían datos nulos, después de eliminar duplicados baja a 2498 registros, lo que indica que se eliminaron 402 registros, lo que coincide con las anteriores diapositivas.

# PERSONALITY DATASET OUTLIERS

Aunque el dataset no presentaba valores atípicos (outliers) en ninguna de sus variables, se aplicó igualmente el código para detectarlos y asegurarse de que no afectaran el análisis con datos que no representan el comportamiento general y evitar la distorsion en modelos mas sensibles.



Se eliminaron un total de 0 datos atípicos.

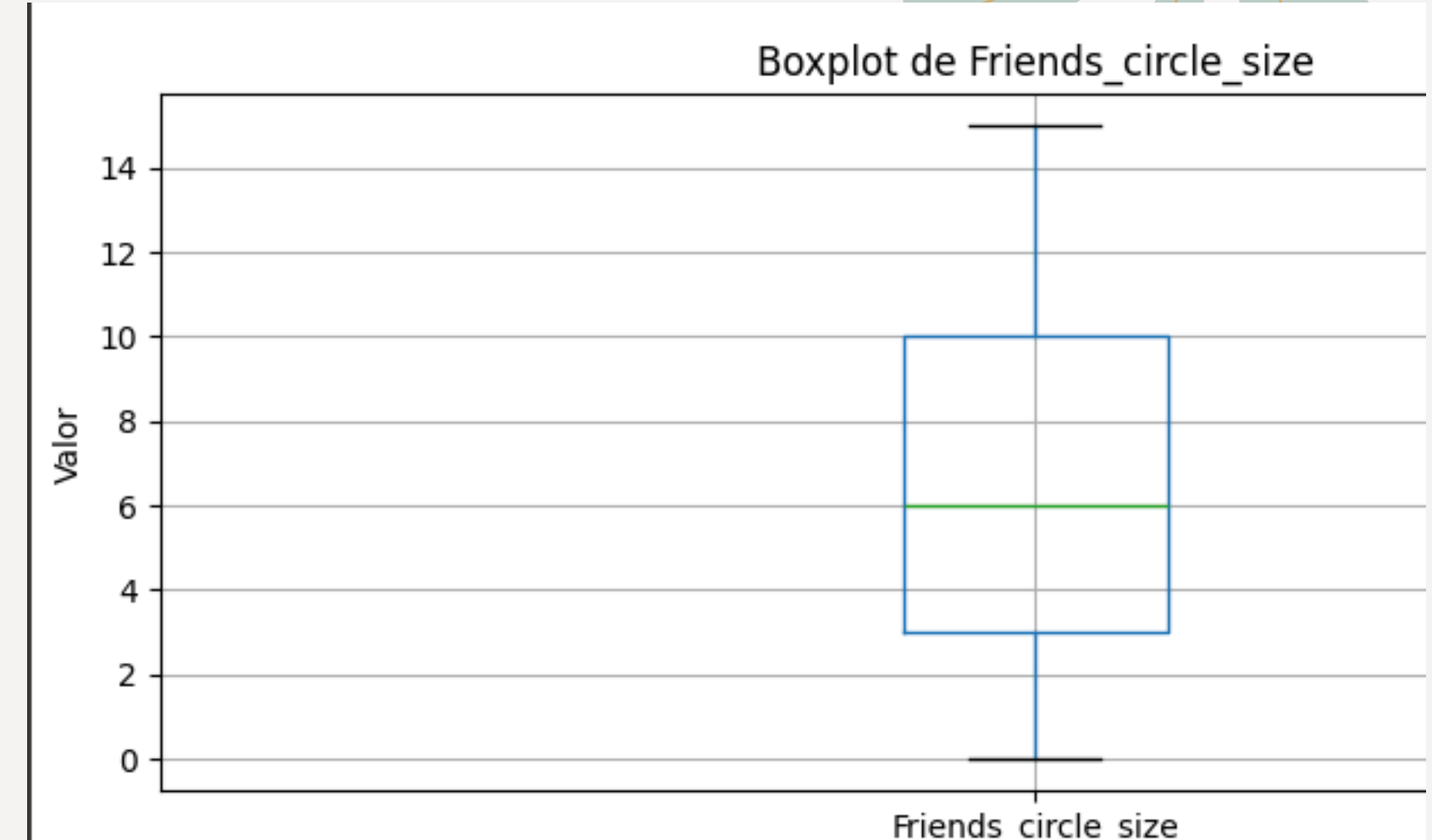
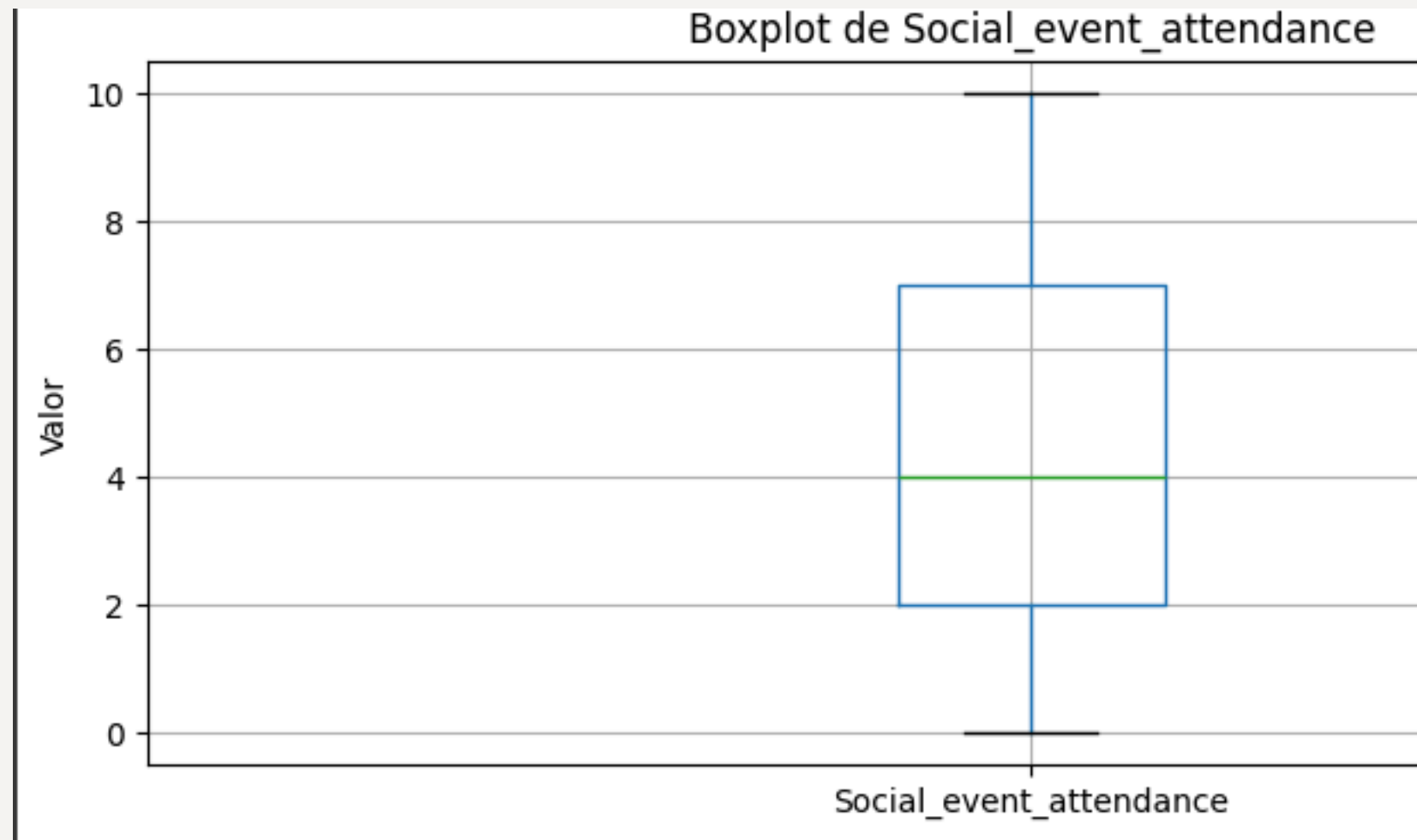
# PERSONALITY DATASET DESBALANCE DE CLASES



la clase 0 es extrovertido que serian 1403 personas y la clase 1 es introvertido que serian 1095 personas, por lo tanto hay mas personas extrovertidas que introvertidas pero sin ser una diferencia tan notable.

# PERSONALITY DATASET

## GRAFICOS UNIVARIADOS (BOX PLOT)

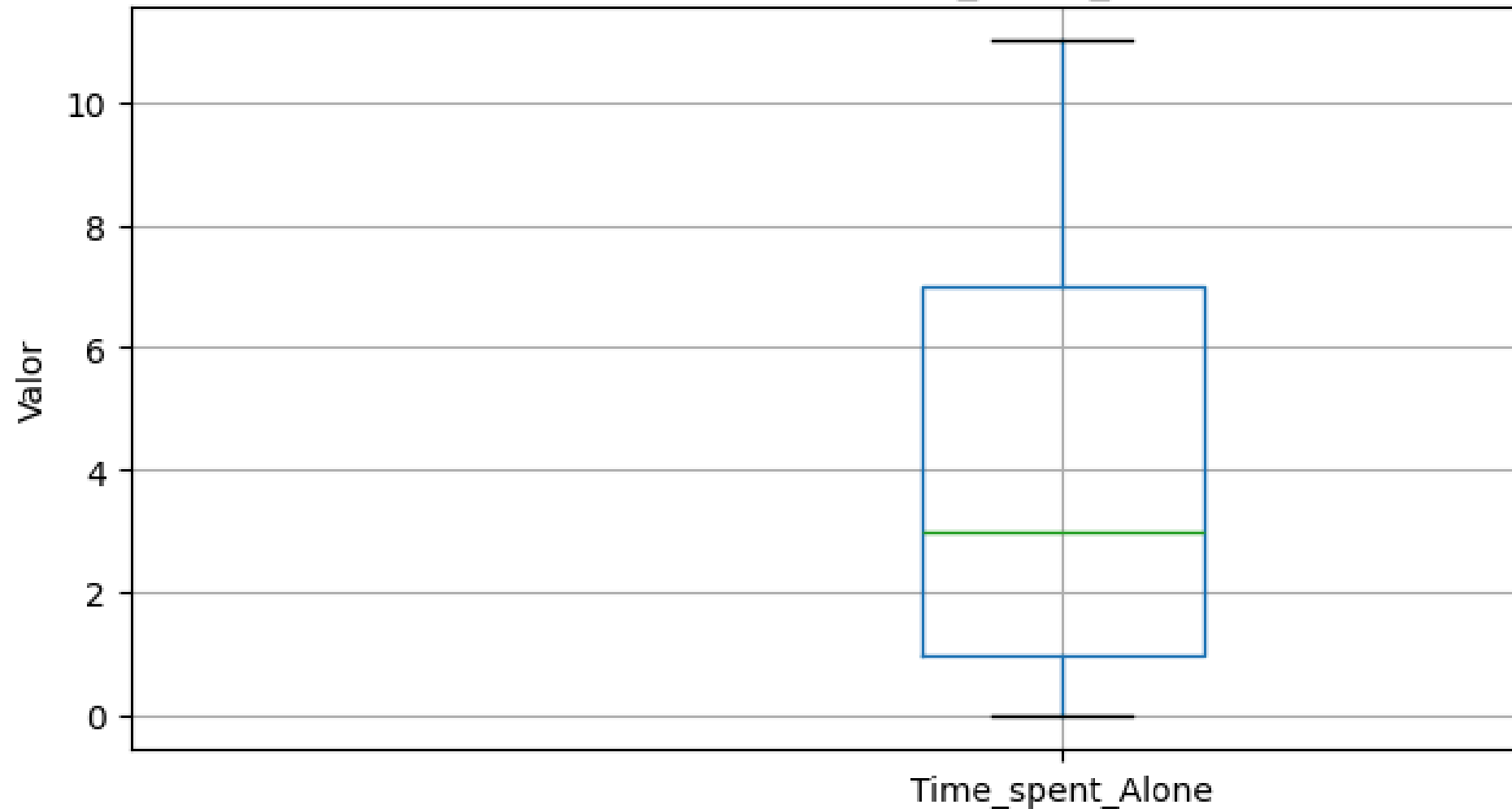


se podría decir que las variables social event attendance y friends circle size posiblemente siguen una distribución normal ya que vemos la mediana cerca del medio de la caja es decir para social en 4 y para friends en 6, la caja esta relativamente centrada, los bigotes no tienen una diferencia muy marcada y no había outliers. vemos que para social los datos estan centrados entre 2 y 7 eventos y para friends entre 3 y 10 amigos.

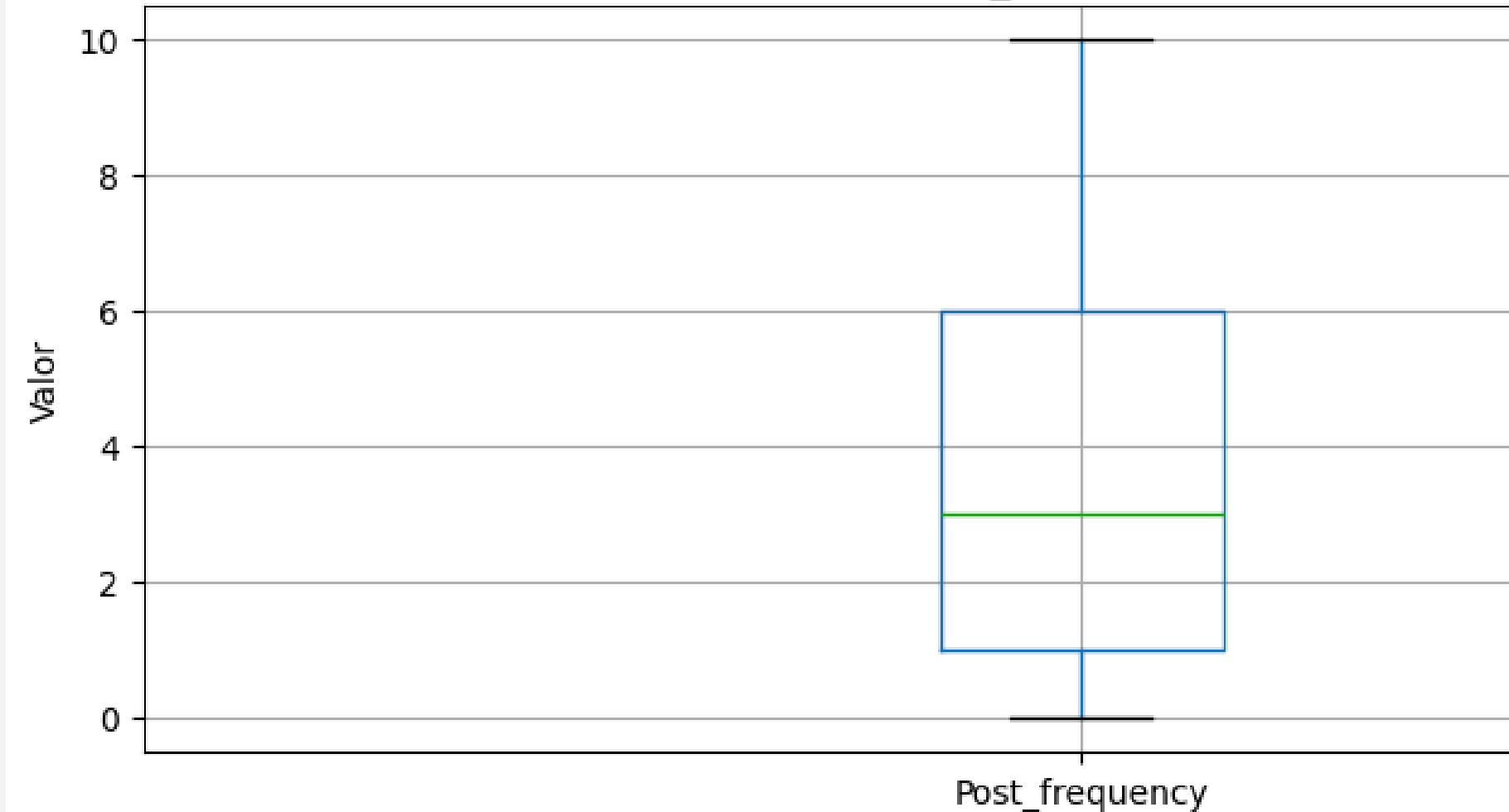
# PERSONALITY DATASET

## GRAFICOS UNIVARIADOS (BOX PLOT)

Boxplot de Time\_spent\_Alone (sin atípicos)



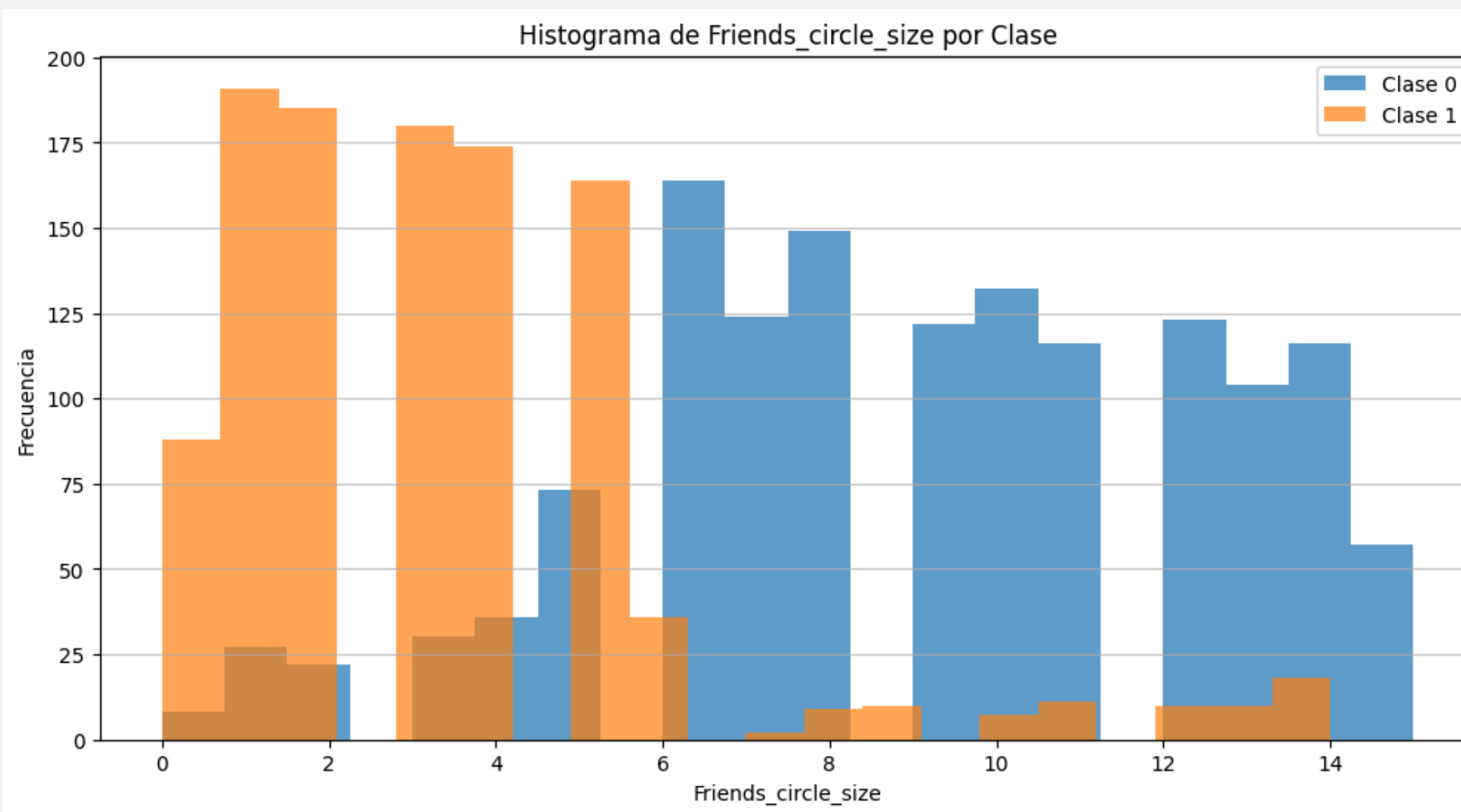
Boxplot de Post\_frequency (sin atípicos)



de los gráficos logro identificar que ambas variables o categorías están sesgadas hacia la derecha, vemos las cajas distribuidas hacia la derecha igual que la mediana, en el caso de time\_spent\_alone la mediana es 3, el máximo es 11 y el mínimo es 0 y en el caso de post\_frecuency la mediana es 3, el máximo es 10 y el mínimo es 0, en ninguno de los casos hay valores atípicos. **en general veo varias variables con posible distribución normal y otras con sesgos mas que todo a la derecha.**

# PERSONALITY DATASET

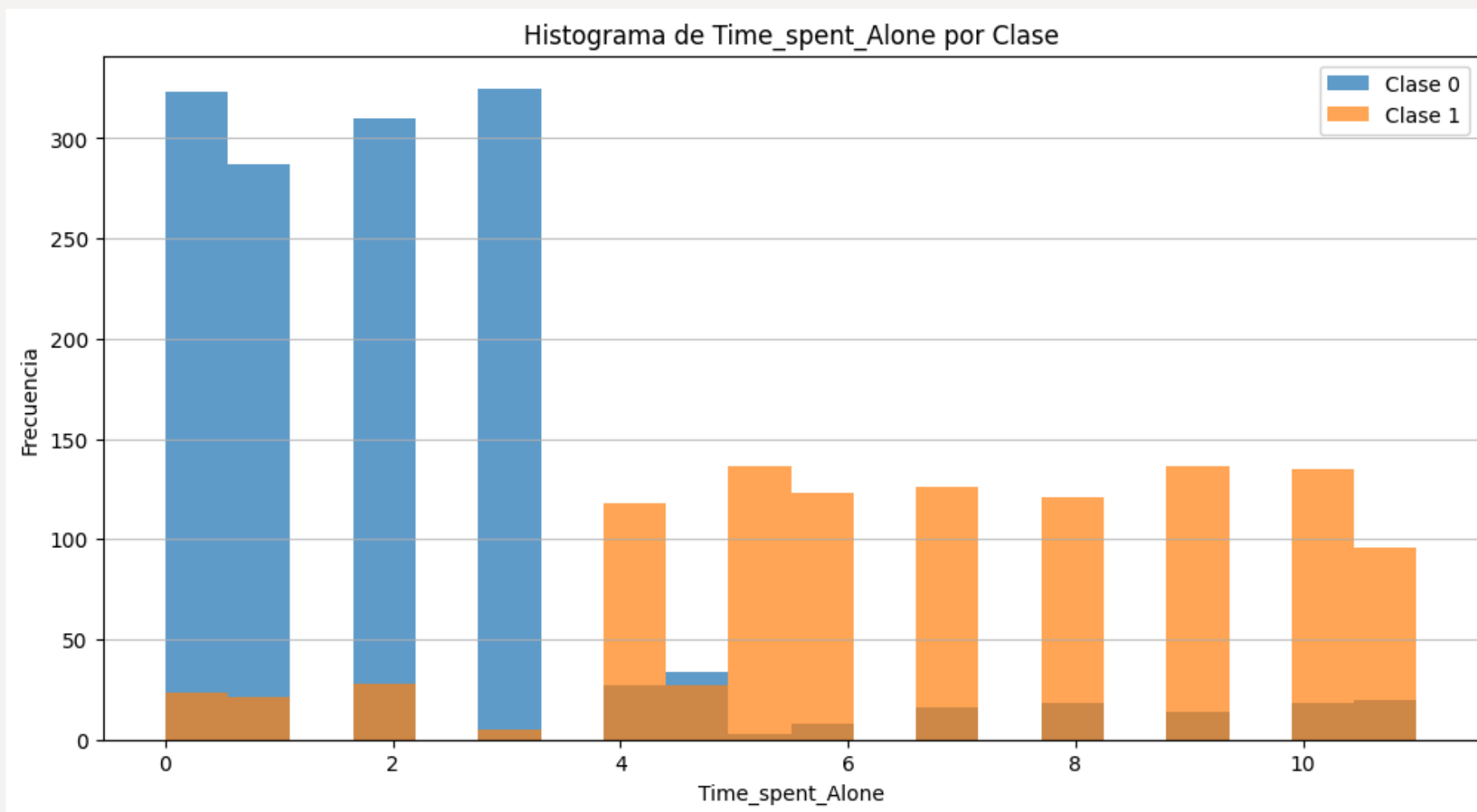
## GRAFICOS UNIVARIADOS (HISTOGRAMAS)



En la clase 1 (introvertidos) se logra evidenciar que es distribución sesgada a la izquierda y la clase 0 (extrovertidos) están sesgados hacia la derecha, los datos son distinguibles unos de otros donde el pico de la clase 1 es entre 0 y 5 y para la clase 0 es de 6 a 8, vemos mas dispersión en los extrovertidos y los datos mas concentrados en los introvertidos, donde se evidencia que los introvertidos suelen tener menos amigos.

# PERSONALITY DATASET

## GRAFICOS UNIVARIADOS (HISTOGRAMAS)



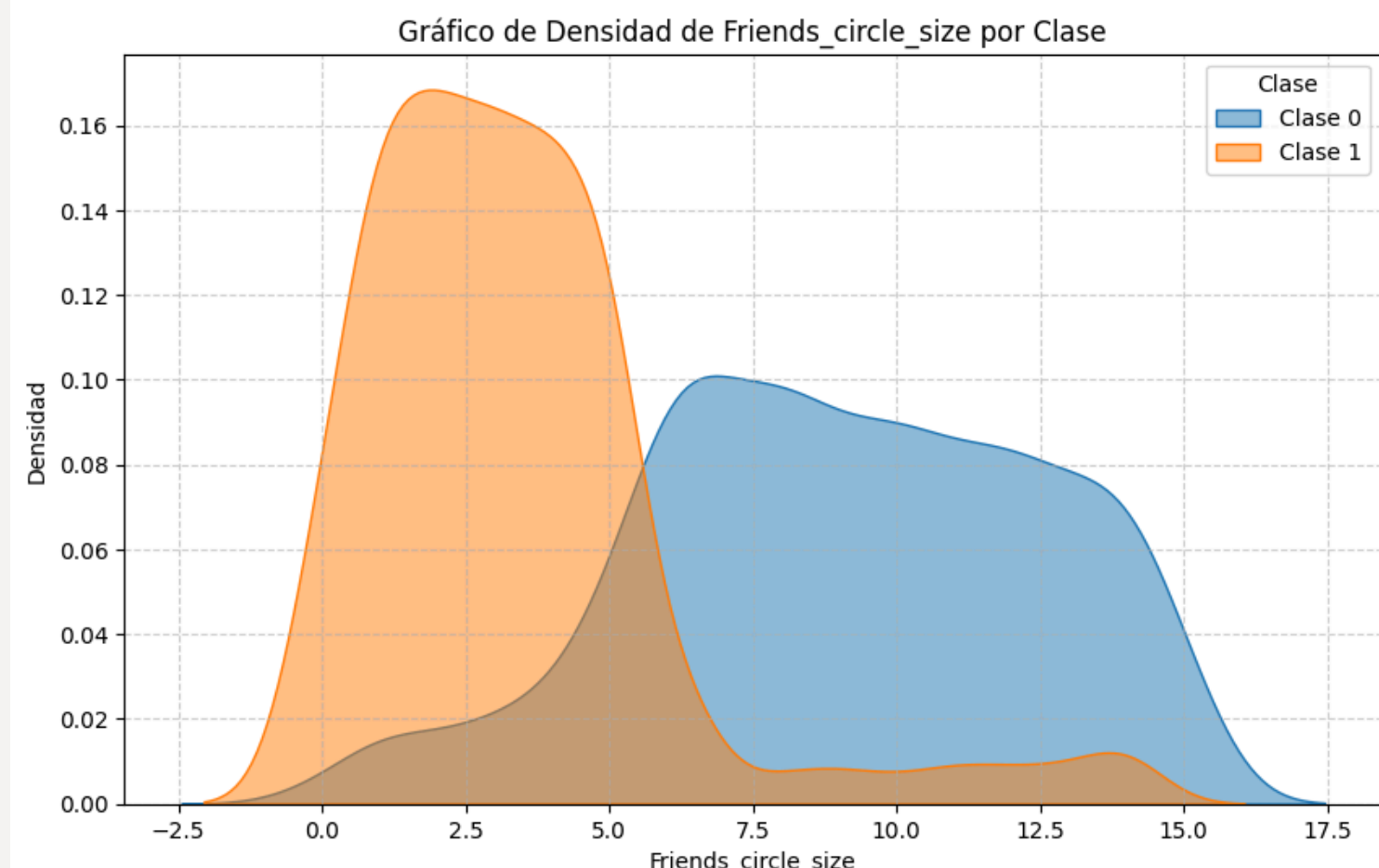
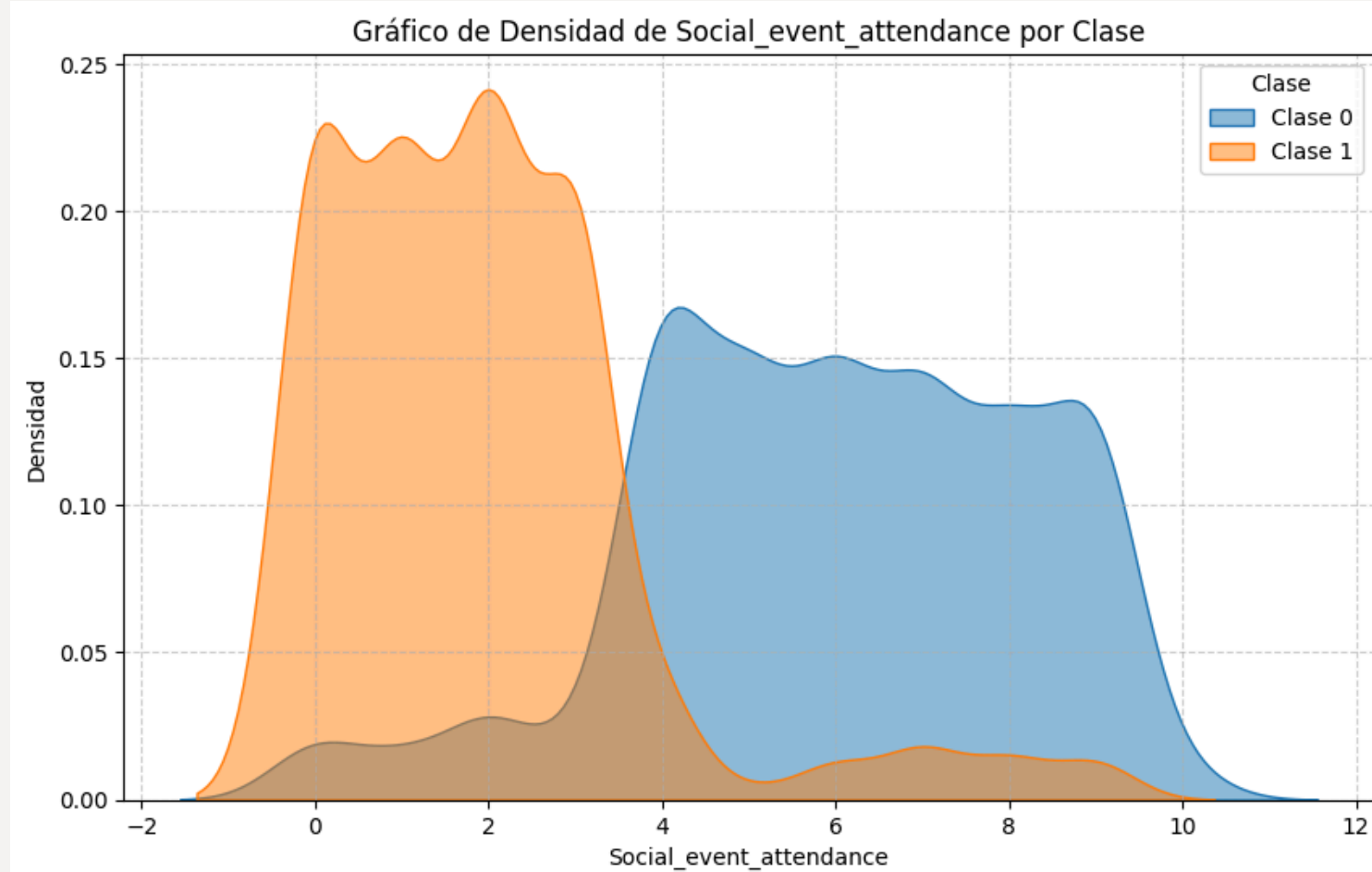
las dos clases en relacion con la variable `time_spent_alone` son bastante distinguibles, veo los datos para los dos bastante centrados en unos puntos, con una distribución de la clase 0 a la izquierda y de la clase 1 a la derecha, con picos para clase 0 entre 0 y 3 y para clase 1 entre 4 y 10, evidenciando que los extrovertidos pasan menos tiempo solos.

**en general los histogramas los veo con poca dispersión y poco sobre posicionados entre si, bastante distinguibles, lo que indica algo bueno para el modelo y la predicción.**



# PERSONALITY DATASET

## GRAFICOS UNIVARIADOS (DENSIDAD)



en las variables de social\_event\_attendance y friends\_circle\_size vemos una posible distribución normal en ambas clases por la forma de campana, aunque con un poco de dispersión, las clases son bastante separables o distinguibles una de otra, se ve un pico mas alto en los valores bajos en la clase

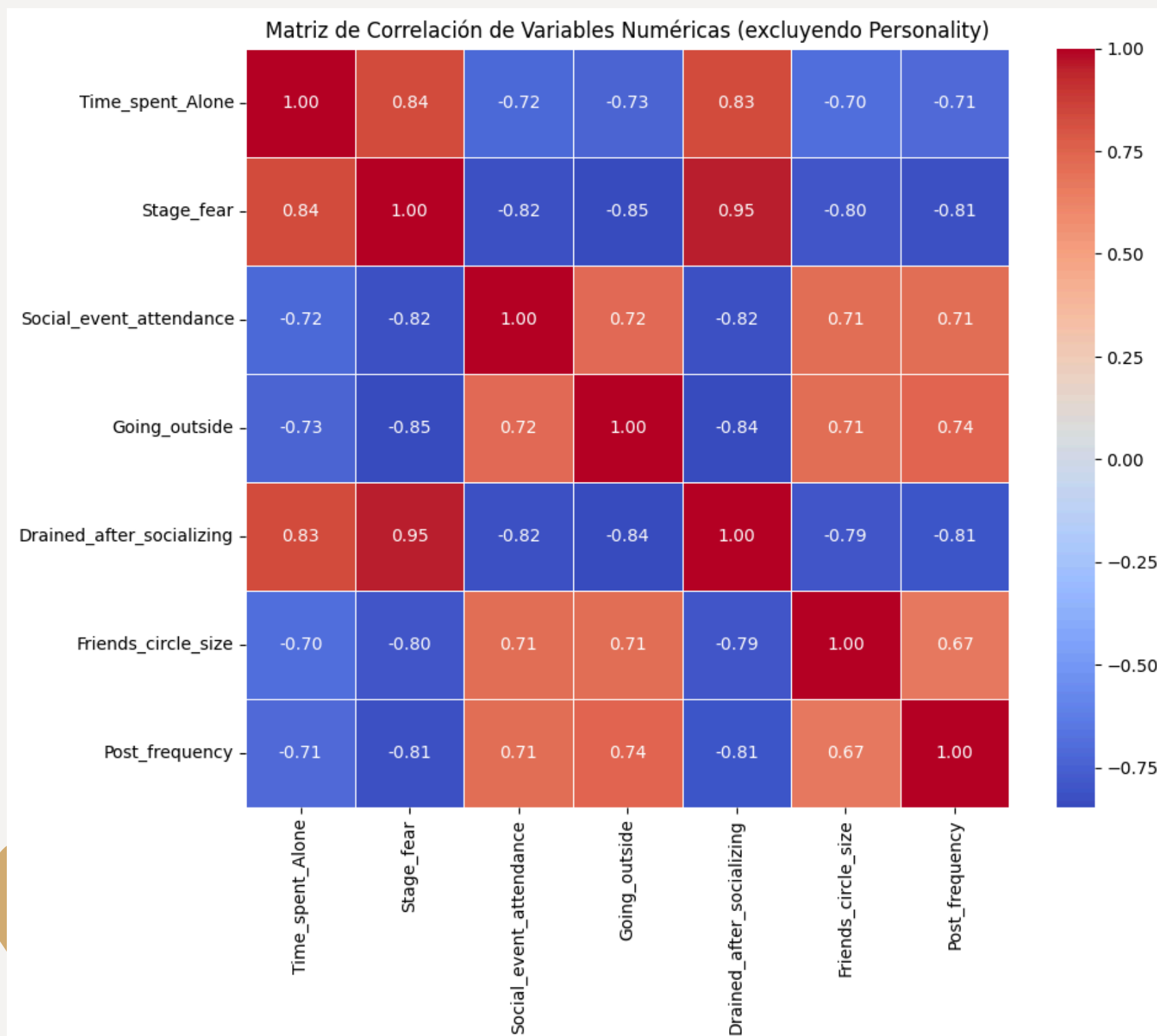
1 es decir en introvertidos, por lo tanto de las graficos podemos decir que los introvertidos suelen asistir a menos eventos y tienen menos amigos.

**en general los graficos de densidad los veo con algo de dispersión, las clases distinguibles entre si y todas con posible distribución normal.**

# PERSONALITY DATASET

## ANALISIS MATRIZ

### CORRELACION (BIVARIADO)



Hay varias variables fuertemente correlacionadas. En la correlación positiva resalta la relación entre Stage\_fear y Drained\_after\_socializing (0.95), indicando que quienes tienen más miedo escénico se agotan más al socializar. También, Stage\_fear y Time\_spent\_Alone (0.84) muestran que a mayor miedo, más tiempo pasan solos.

En la correlación negativa, se destaca que a mayor miedo escénico (-0.85) o agotamiento social (-0.84), menos suelen salir las personas. En general, veo fuertes relaciones entre varias variables que podrían reflejar comportamientos similares.

# PERSONALITY DATASET

## ANALISIS CHI-CUADRADO (BIVARIADO)

Realizando Test de Chi-cuadrado para Variables Categóricas vs. Result:

-----  
Analizando la relación entre 'Stage\_fear' y 'Personality'...

Estadístico Chi-cuadrado: 1700.7751

Valor p: 0.0000

Grados de libertad: 1

Conclusión: 'Stage\_fear' es SIGNIFICATIVA ( $p < 0.05$ ) - Hay una asociación estadísticamente significativa con 'Personality'.

-----  
Analizando la relación entre 'Drained\_after\_socializing' y 'Personality'...

Estadístico Chi-cuadrado: 1693.9527

Valor p: 0.0000

Grados de libertad: 1

Conclusión: 'Drained\_after\_socializing' es SIGNIFICATIVA ( $p < 0.05$ ) - Hay una asociación estadísticamente significativa con 'Personality'.

-----  
Reporte de Significancia (Nivel alpha = 0.05):

-----  
Variables Categóricas Significativas (asociación con 'Personality'):

- Stage\_fear (Valor p: 0.0000)

- Drained\_after\_socializing (Valor p: 0.0000)

Variables Categóricas NO Significativas (sin asociación con 'Personality'):

Todas las variables categóricas testeadas fueron estadísticamente significativas.

-----

Las variables que resultaron más importantes y con mayor relación con el tipo de personalidad fueron:

Stage\_fear, Drained\_after\_socializing

recordemos que chi-cuadrado se aplica a las variables categóricas, yo solo tengo 3.

Esto significa que estas dos características están muy ligadas a si una persona es más introvertida o extrovertida, por lo tanto, son claves para predecir la personalidad. esto se define ya que su  $p = 0.0000$  en ambos casos

# PERSONALITY DATASET

## ANALISIS ANOVA (BIVARIADO)

Reporte de Significancia de Variables Numéricas (ANOVA con Nivel  $\alpha = 0.05$ ):

-----  
Variables Numéricas Significativas (media difiere entre grupos de 'Personality'):

- Time\_spent\_Alone (Valor p: 0.0000)
- Social\_event\_attendance (Valor p: 0.0000)
- Going\_outside (Valor p: 0.0000)
- Friends\_circle\_size (Valor p: 0.0000)
- Post\_frequency (Valor p: 0.0000)

Variables Numéricas NO Significativas (media no difiere significativamente entre grupos de 'Personality'):

Todas las variables numéricas testeadas fueron estadísticamente significativas.  
-----

Las cinco variables numéricas evaluadas mediante Anova (time\_spent\_alone, social\_event\_attendance, going\_outside, friends\_circle\_size y post\_frequency) resultaron estadísticamente significativas ( $p = 0.0000$ ), lo que indica que sus medias difieren entre personas extrovertidas e introvertidas.

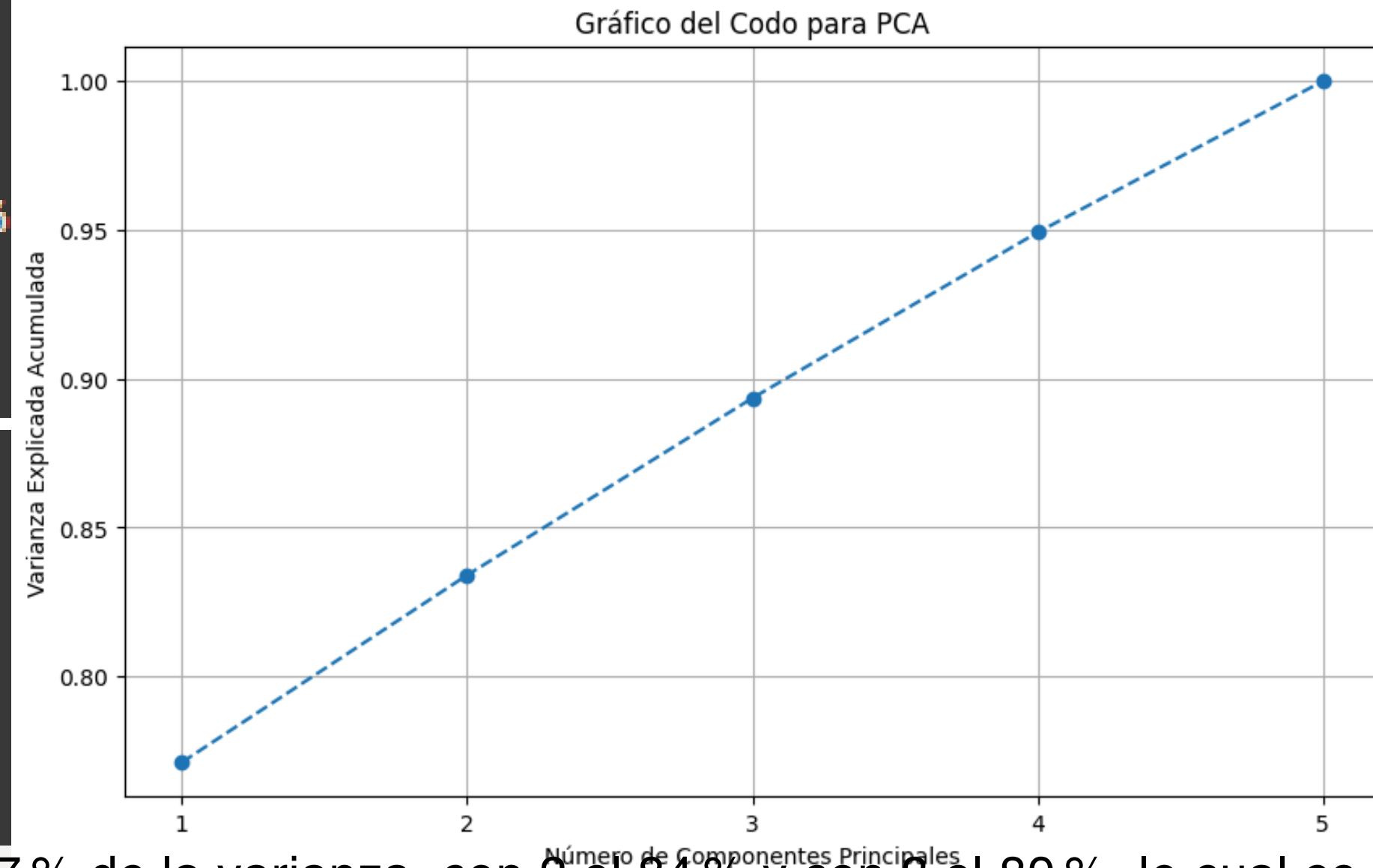
Esto significa que estas características están fuertemente relacionadas con el tipo de personalidad y son clave para predecirla.



## Matriz de Loadings:

|                         | PC1       | PC2       | PC3      | PC4       | PC5       |
|-------------------------|-----------|-----------|----------|-----------|-----------|
| Time_spent_Alone        | -0.468556 | -0.201383 | 0.852196 | -0.082593 | 0.082709  |
| Social_event_attendance | 0.428628  | 0.237643  | 0.174695 | -0.603417 | 0.604296  |
| Going_outside           | 0.486763  | -0.449545 | 0.182886 | 0.612531  | 0.390295  |
| Friends_circle_size     | 0.414229  | 0.630909  | 0.441836 | 0.261955  | -0.408078 |
| Post_frequency          | 0.433832  | -0.550303 | 0.120735 | -0.430408 | -0.555993 |

|   | PC1       | PC2       | PC3      | Personality |
|---|-----------|-----------|----------|-------------|
| 0 | 0.425156  | 0.024034  | 0.255252 | 0.0         |
| 1 | -0.822265 | -0.213964 | 0.008904 | 1.0         |
| 2 | -0.545634 | -0.053308 | 0.213832 | 1.0         |
| 3 | 0.908568  | -0.042458 | 0.072104 | 0.0         |
| 4 | 0.404915  | 0.079302  | 0.065595 | 0.0         |



En el gráfico se observa que con 1 componente se explica el 77 % de la varianza, con 2 el 84 % y con 3 el 89 %, lo cual es aceptable para reducir dimensiones sin perder mucha información.

En la matriz de loadings:

- PC1 representa una dimensión de sociabilidad: valores altos indican que la persona sale, publica y tiene más amigos; valores bajos, más aislamiento.
- PC2 está influido negativamente por post\_frequency y going\_outside, lo que podría asociarse con menor interacción social.
- PC3 refleja, en menor medida, diferencias en el nivel de interacción directa o posibles rasgos de aislamiento.

# PERSONALITY DATASET COMPARACION MODELOS

## LIMPIO

|            | Model                        | Accuracy | AUC    | Recall | Prec.  | F1     | Kappa  | MCC    | TT (Sec) |
|------------|------------------------------|----------|--------|--------|--------|--------|--------|--------|----------|
| <b>gbc</b> | Gradient Boosting Classifier | 0.9325   | 0.9631 | 0.9373 | 0.9128 | 0.9242 | 0.8634 | 0.8649 | 0.1750   |
| <b>nb</b>  | Naive Bayes                  | 0.9308   | 0.9042 | 0.9321 | 0.9136 | 0.9220 | 0.8598 | 0.8611 | 0.0330   |
| <b>knn</b> | K Neighbors Classifier       | 0.9251   | 0.9476 | 0.9242 | 0.9084 | 0.9155 | 0.8482 | 0.8496 | 0.0490   |

## NORMALIZADO

|            | Model                        | Accuracy | AUC    | Recall | Prec.  | F1     | Kappa  | MCC    | TT (Sec) |
|------------|------------------------------|----------|--------|--------|--------|--------|--------|--------|----------|
| <b>gbc</b> | Gradient Boosting Classifier | 0.9325   | 0.9631 | 0.9373 | 0.9128 | 0.9242 | 0.8634 | 0.8649 | 0.1800   |
| <b>nb</b>  | Naive Bayes                  | 0.9308   | 0.9042 | 0.9321 | 0.9136 | 0.9220 | 0.8598 | 0.8611 | 0.0260   |
| <b>knn</b> | K Neighbors Classifier       | 0.9268   | 0.9465 | 0.9295 | 0.9072 | 0.9176 | 0.8517 | 0.8530 | 0.0470   |

## P C A

|            | Model               | Accuracy | AUC    | Recall | Prec.  | F1     | Kappa  | MCC    | TT (Sec) |
|------------|---------------------|----------|--------|--------|--------|--------|--------|--------|----------|
| <b>lr</b>  | Logistic Regression | 0.9314   | 0.9055 | 0.9334 | 0.9137 | 0.9227 | 0.8610 | 0.8623 | 0.0270   |
| <b>nb</b>  | Naive Bayes         | 0.9314   | 0.9086 | 0.9334 | 0.9137 | 0.9227 | 0.8610 | 0.8623 | 0.0230   |
| <b>svm</b> | SVM - Linear Kernel | 0.9314   | 0.9045 | 0.9334 | 0.9137 | 0.9227 | 0.8610 | 0.8623 | 0.0410   |

En los datos normalizados, el modelo con mejor desempeño fue Gradient Boosting Classifier, destacando en todas las métricas. Al aplicar PCA, el mejor modelo fue Regresión Logística, pero con un rendimiento ligeramente menor al GBC, ya que PCA reduce la varianza para simplificar los datos, lo que puede perder información relevante para modelos más complejos.

en conclusión El mejor modelo general es Gradient Boosting Classifier sobre datos limpios o normalizados, ya que conserva toda la información y obtiene las mejores métricas globales.

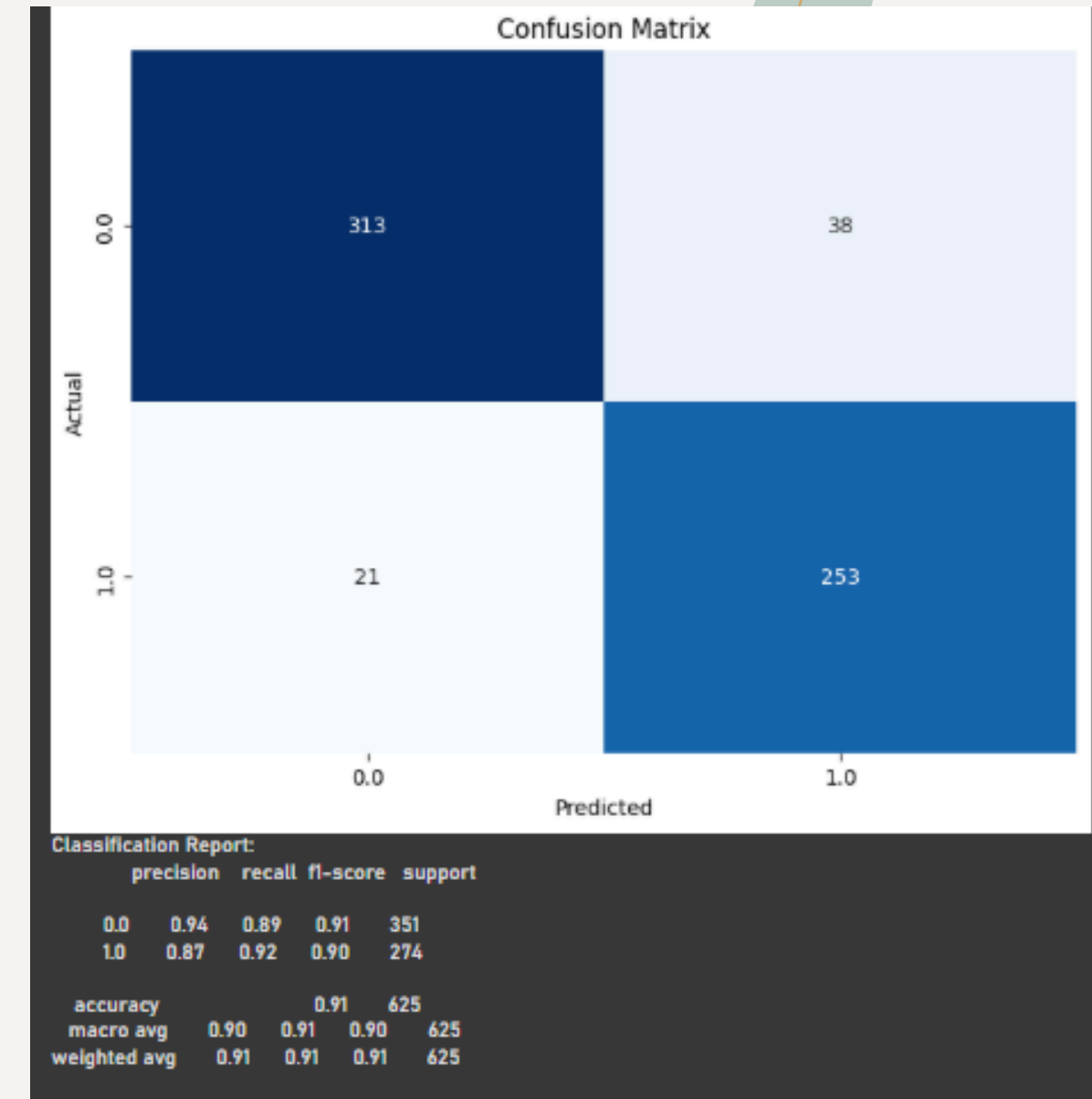
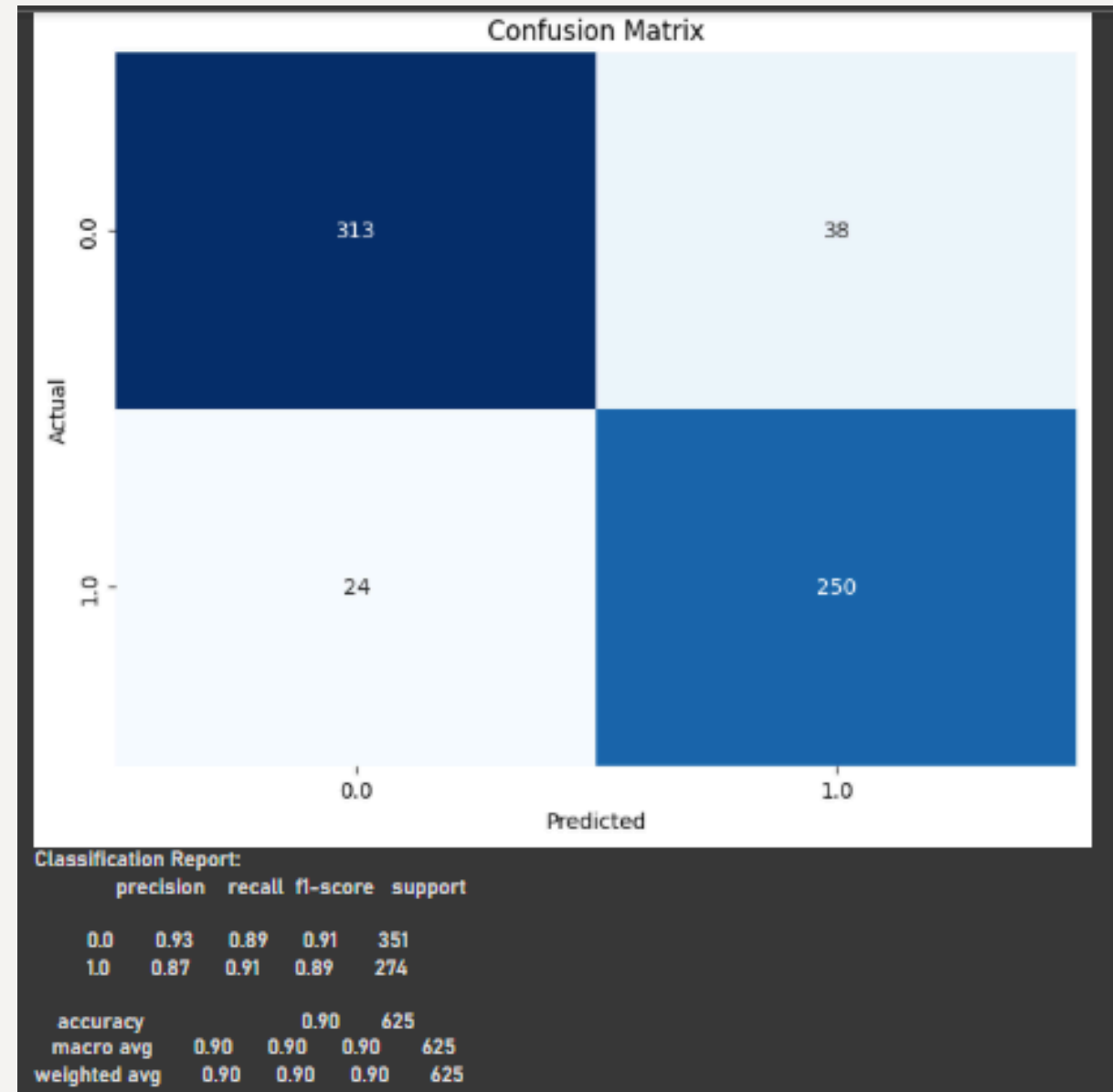
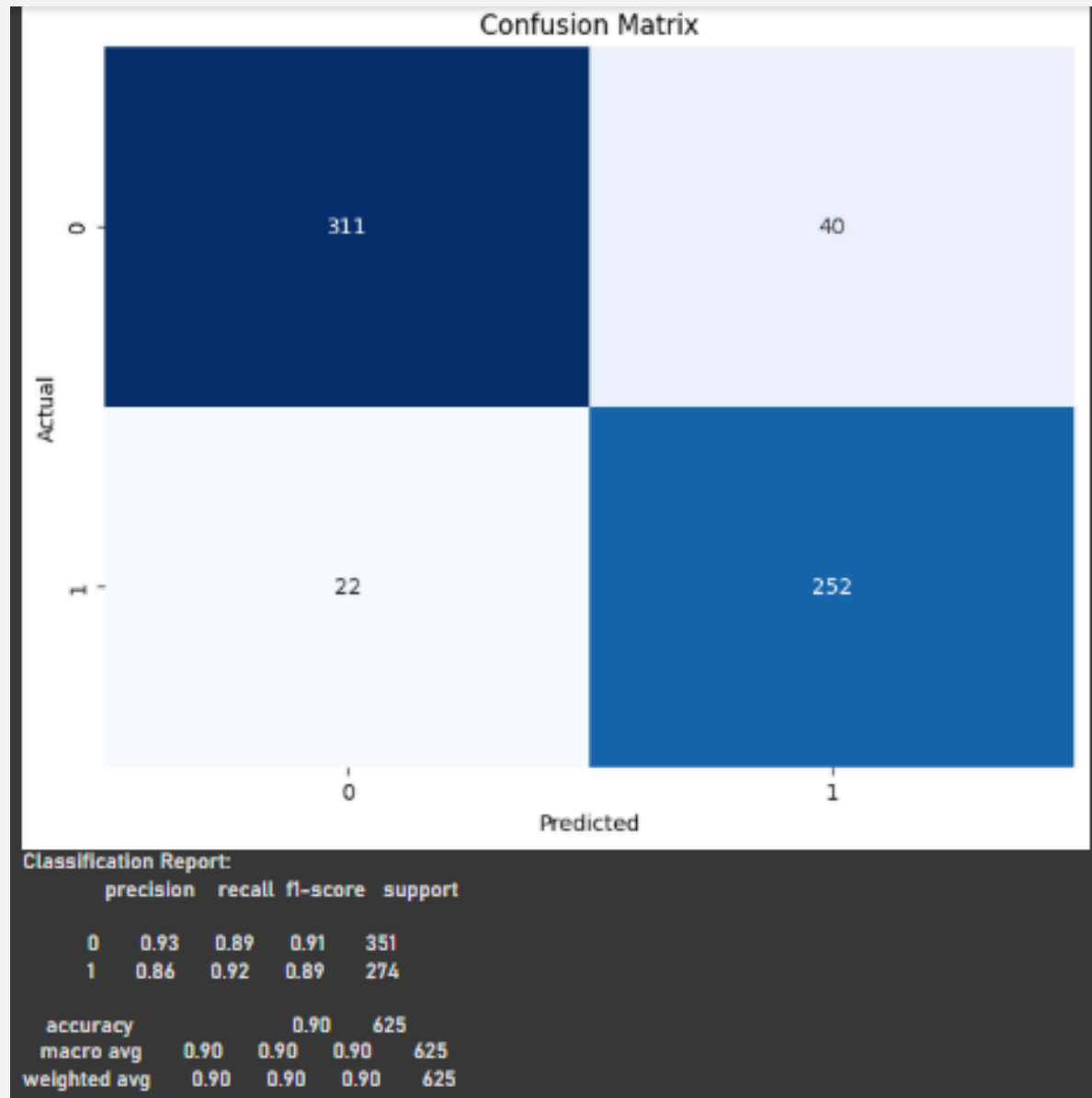
# PERSONALITY DATASET

## KNN

### LIMPIO

### NORMALIZADO

### PCA



Con KNN, todos fallan más en falsos positivos (predicen introvertido cuando es extrovertido), siendo el modelo limpio el que más falla con 40 casos. En falsos negativos (personas introvertidas clasificadas como extrovertidas), el peor es el normalizado con 24 errores. Aunque los tres tienen el mismo F1 en macro avg, el modelo con PCA tiene mejor recall, por eso lo considero el mejor enfoque.



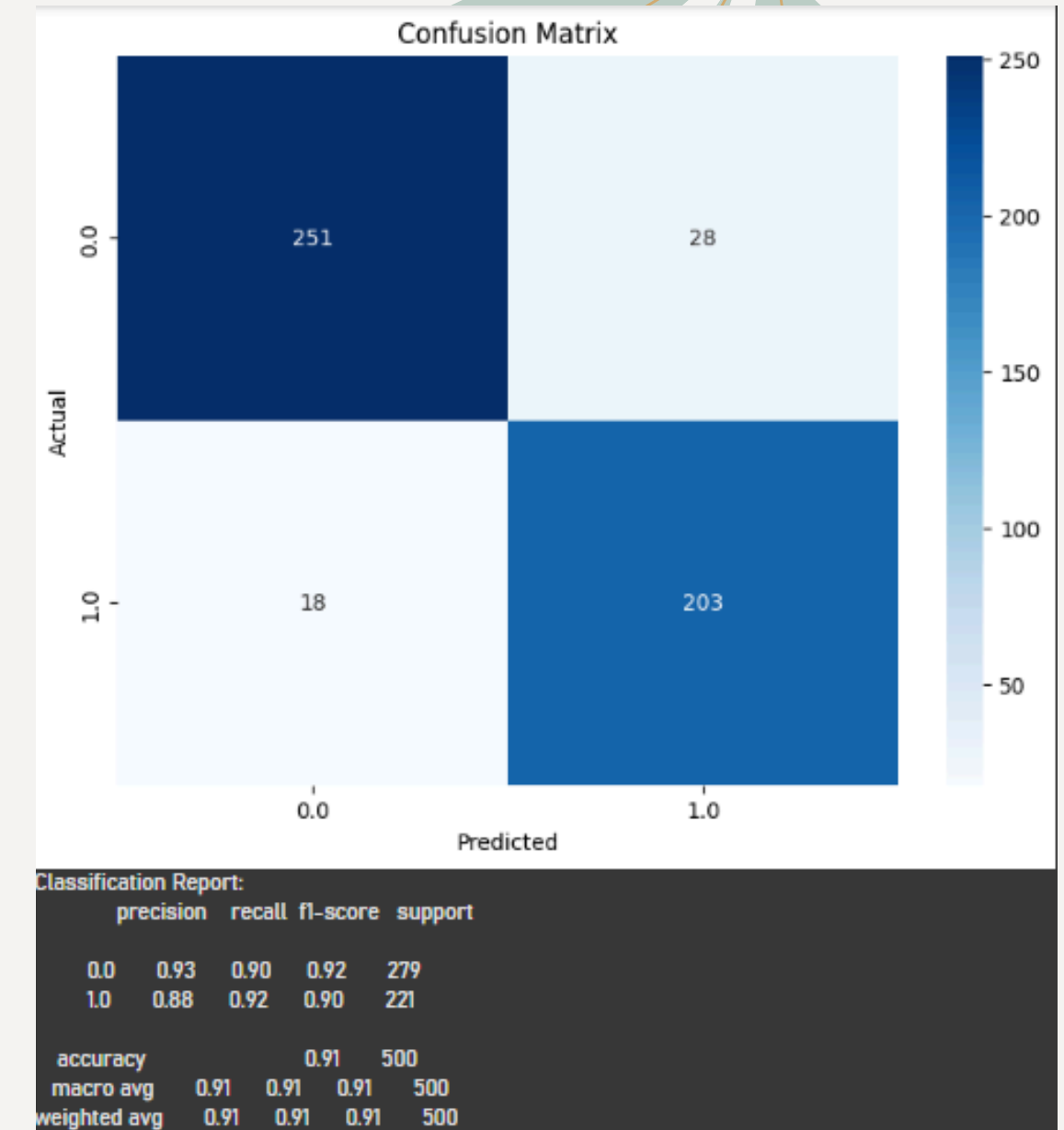
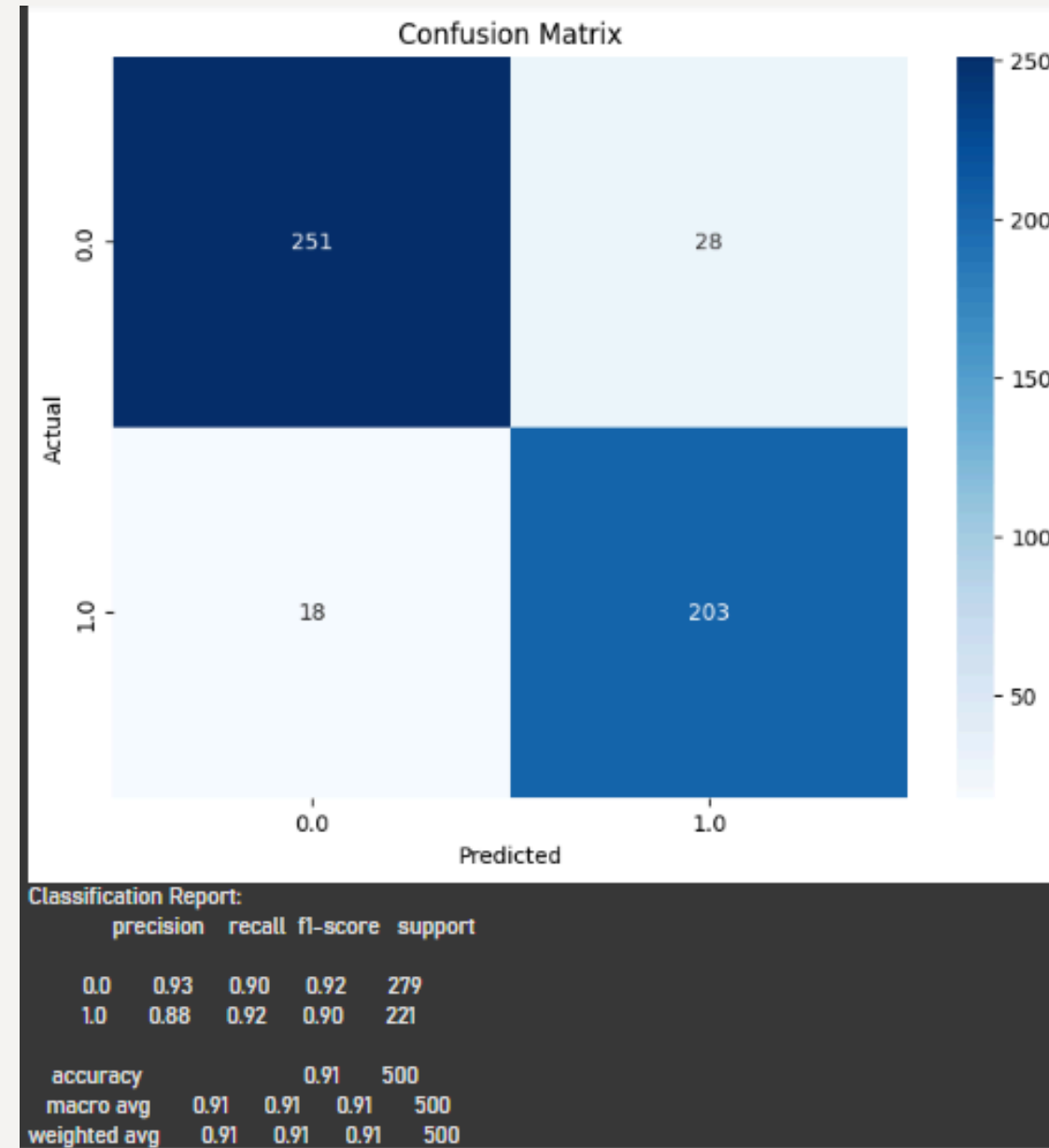
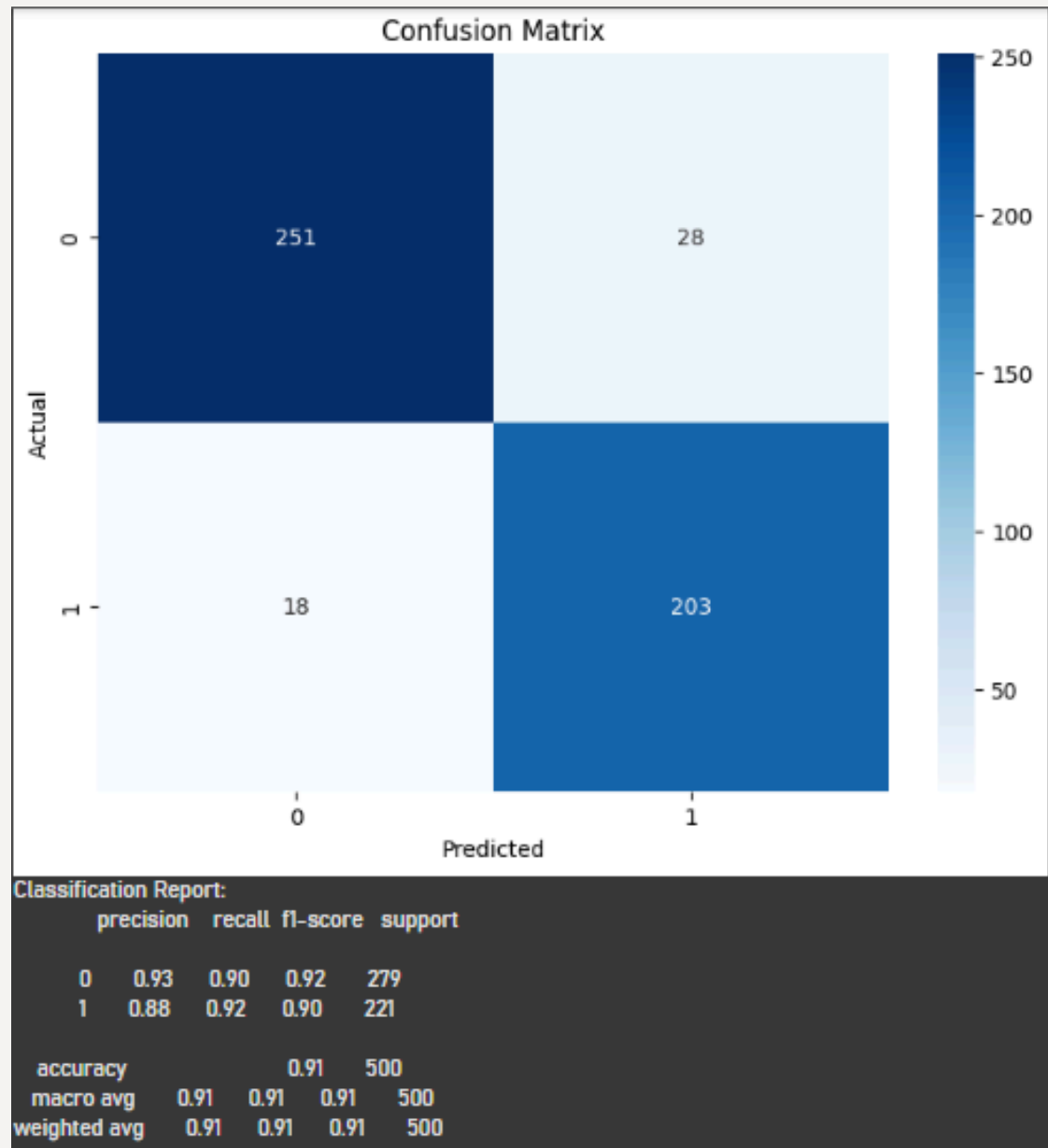
# PERSONALITY DATASET

## NAIVE BAYES

### LIMPIO

### NORMALIZADO

### PCA

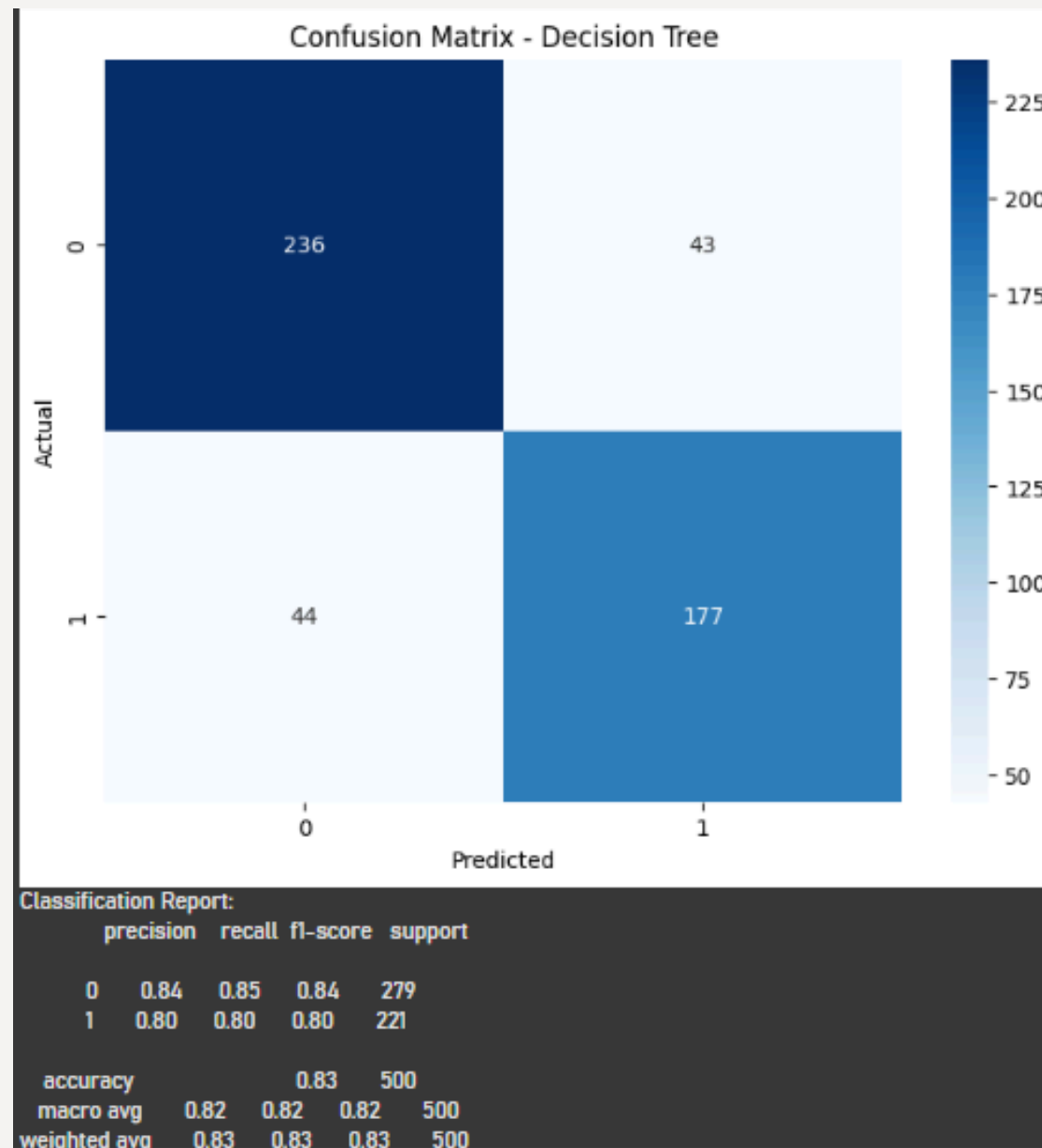


con naive bayes todos los enfoque tanto limpio como normalizado como PCA tienen la misma métrica de macro avg, mismo recall y misma precisión, también se equivocan en la misma proporción en los falsos positivos y los falsos negativos, así que pienso que podríamos tomar PCA ya que tendríamos la misma calidad de predicción pero con menos datos, es decir un sistema mas liviano. vemos que sube un punto en f1 en comparación con KNN.

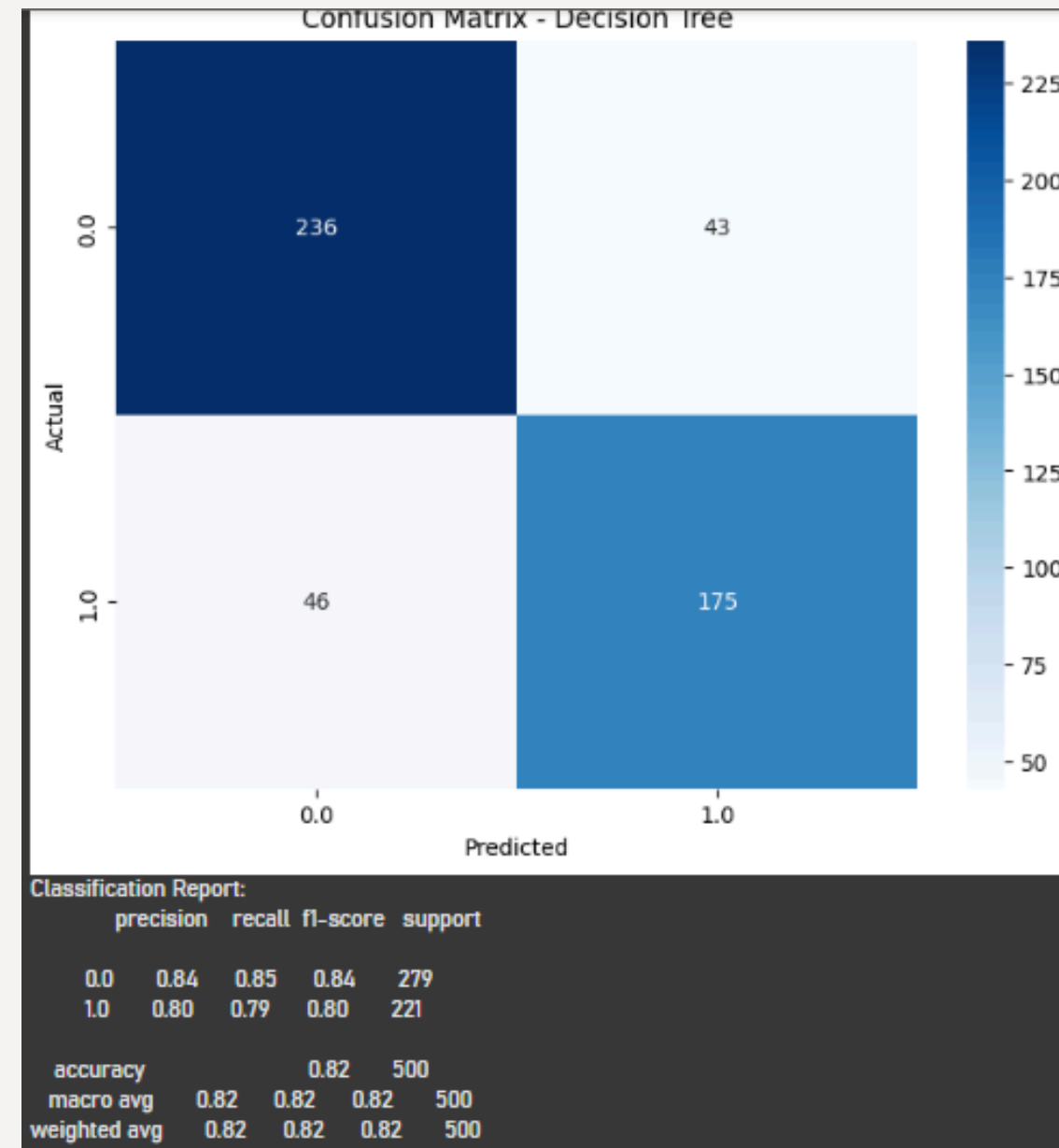
# PERSONALITY DATASET

## ARBOL DE DECISIONES

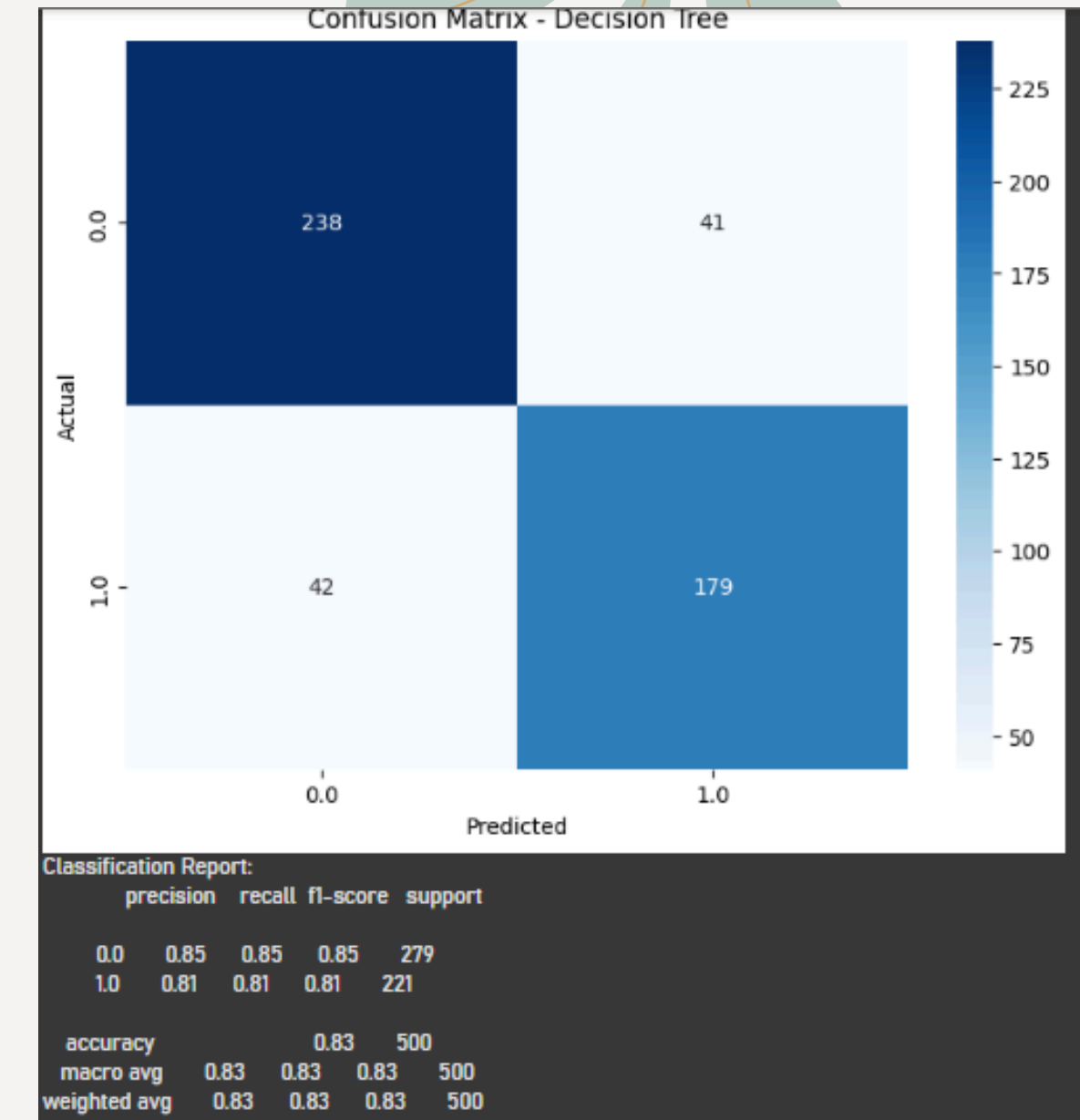
### LIMPIO



### NORMALIZADO



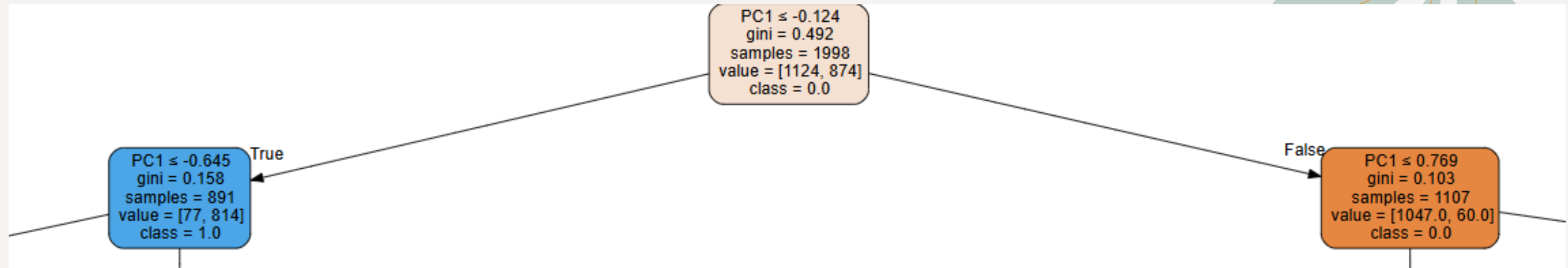
### PCA



el mejor enfoque es el PCA con un macro avg de 0.83, teniendo 41 falsos positivos, 42 falsos negativos, 179 negativos reales y 238 positivos reales, se evidencia una reducción en el f1 score en comparación con naive bayes y KNN.

# PERSONALITY DATASET

## ARBOL DE DECISIONES (PCA)

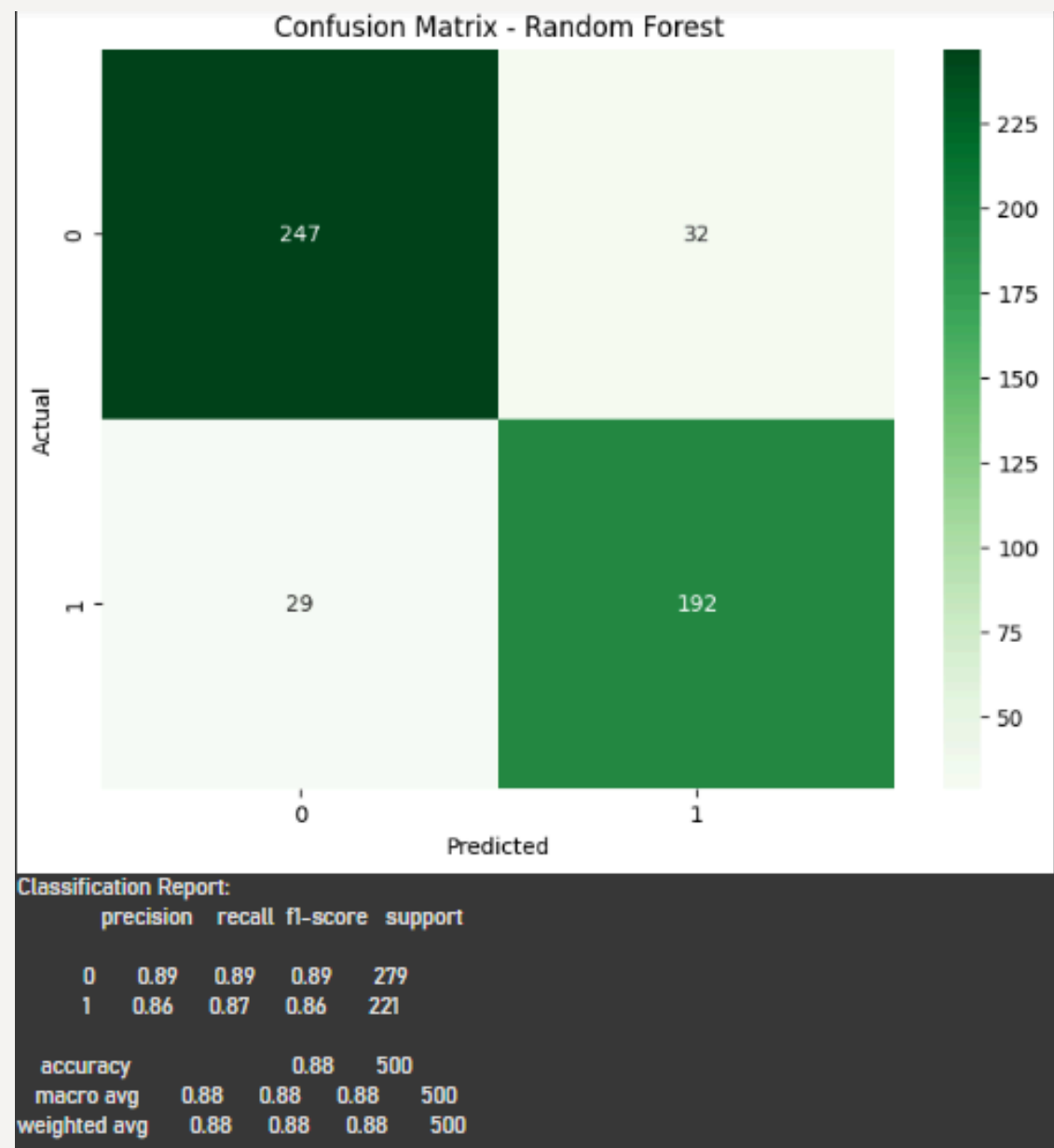


vemos que el árbol utiliza como raíz el componente 1, lo que indica que PC1 es la más importante para predecir la variable "Personality".

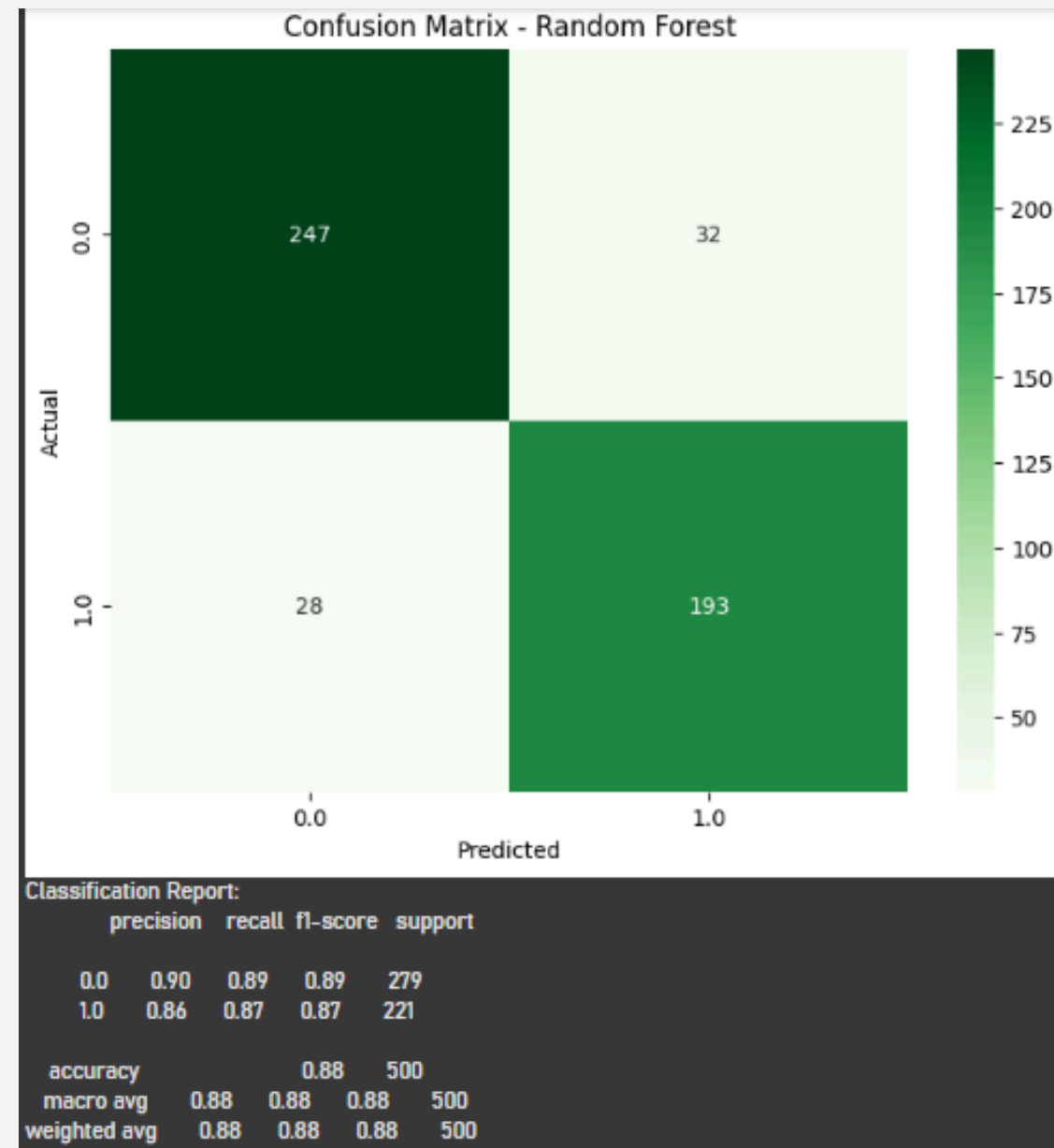
en nodo raíz se tiene que si el valor es menor o igual a -0.124 entonces va a la rama izquierda, si no va a la derecha. en la izquierda se agrupan las personas con valores bajos de PC1 (es decir, con comportamientos más introvertidos como lo mencione en la diapositiva de PCA). en la derecha se agrupan los de valores más altos en PC1, es decir, personas con más amigos, más actividad social y más publicaciones. entonces se evidencia que PC1 tiene la mayor parte de la información y que además las clases son distinguibles.

# PERSONALITY DATASET RANDOM FOREST

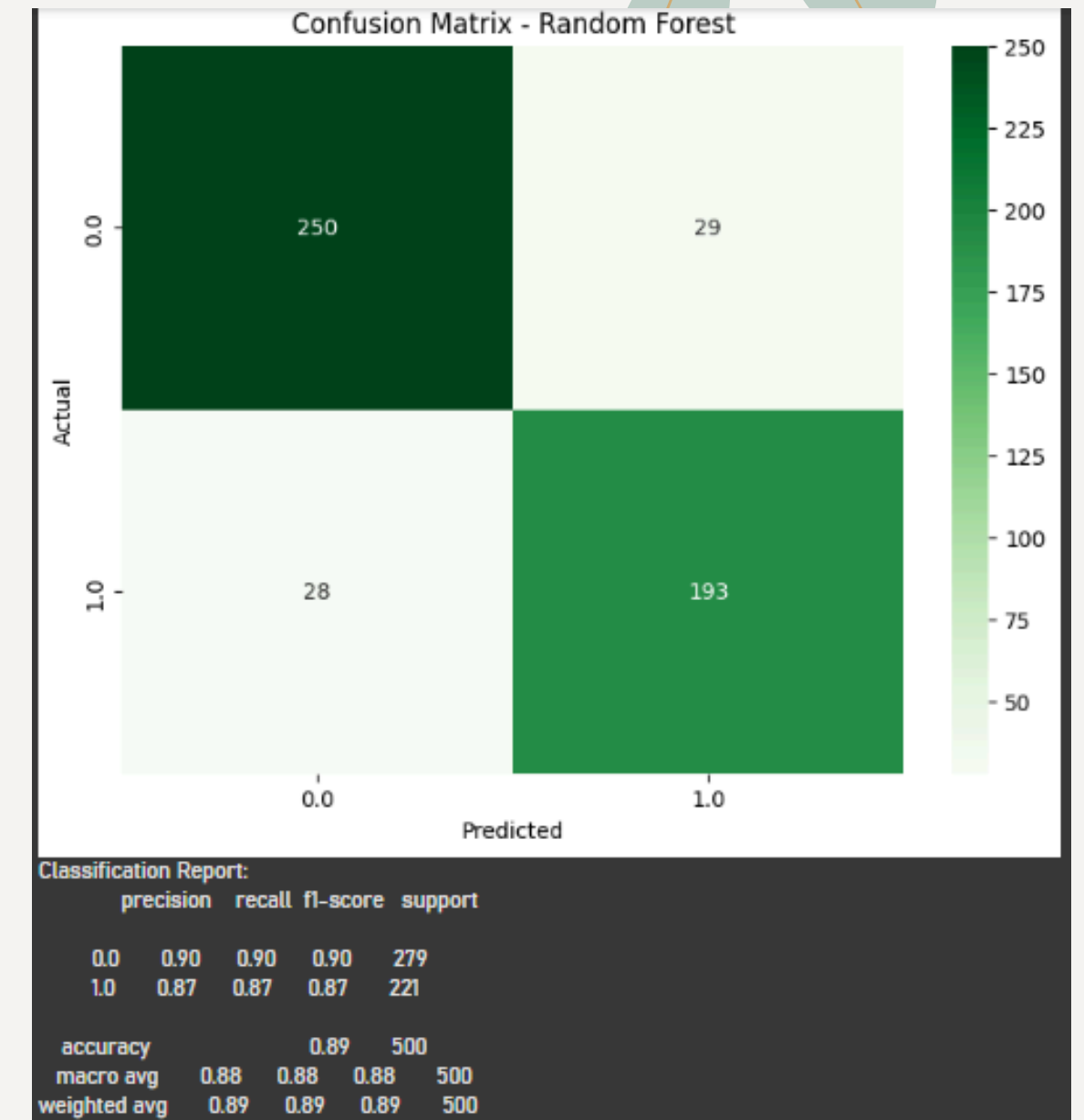
## LIMPIO



## NORMALIZADO

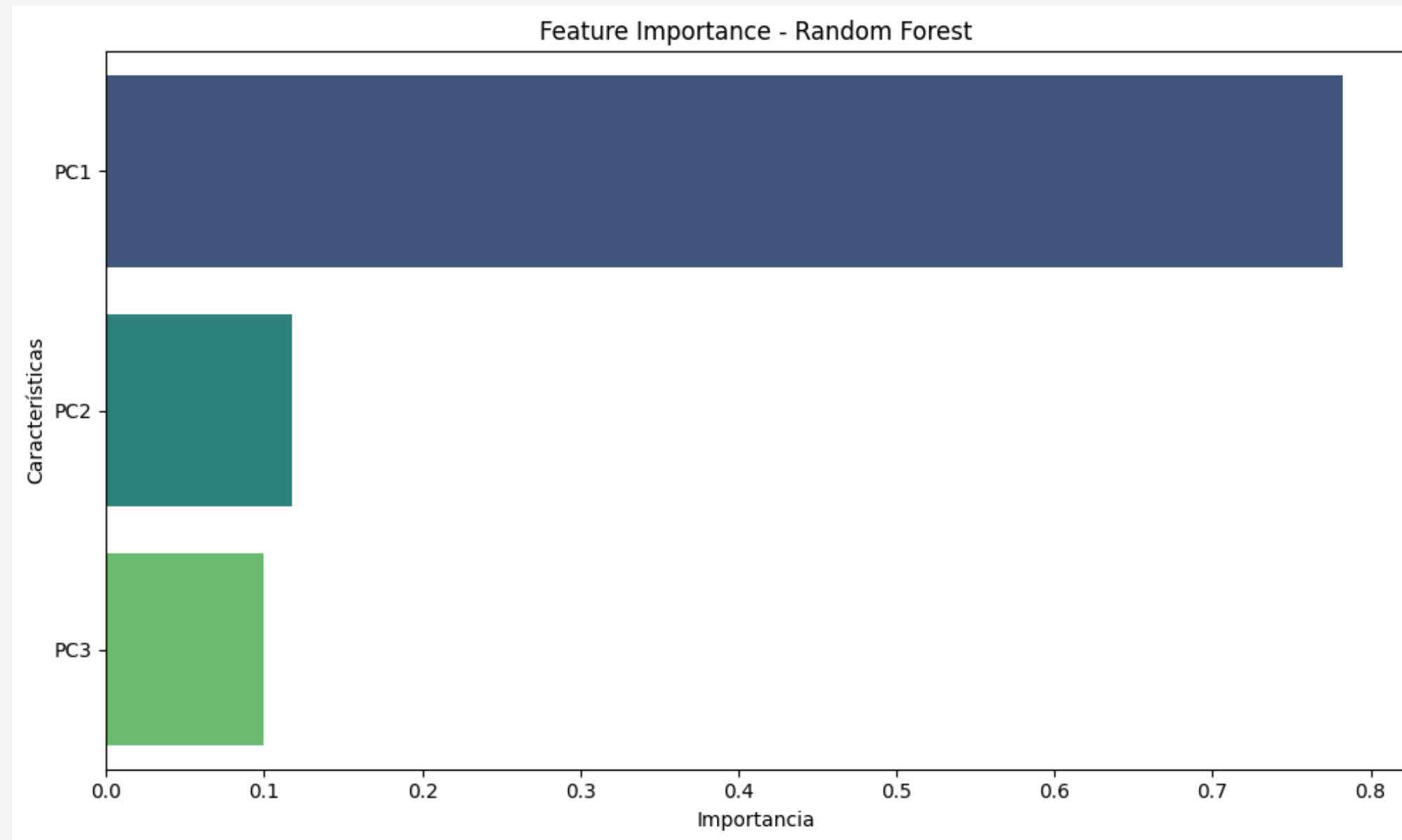


## PCA



se evidencia que todos los enfoques tienen el mismo puntaje en macro avg pero el PCA tiene menos errores en los falsos positivos y falsos negativos, por lo tanto decidí que PCA seria el mejor enfoque. en comparación con KNN y naive bayes el puntaje es inferior y con árbol de decisiones es mejor este modelo.

# PERSONALITY DATASET RANDOM FOREST (PCA)



PC1 tiene una importancia altísima, lo que indica que es la que más influye en la predicción del modelo.

PC2 y PC3 tienen muy poca importancia relativa comparado con PC1, Esto sugiere que la mayoría de la información útil está concentrada en PC1 (como ya se había mencionado en la diapositiva de PCA).

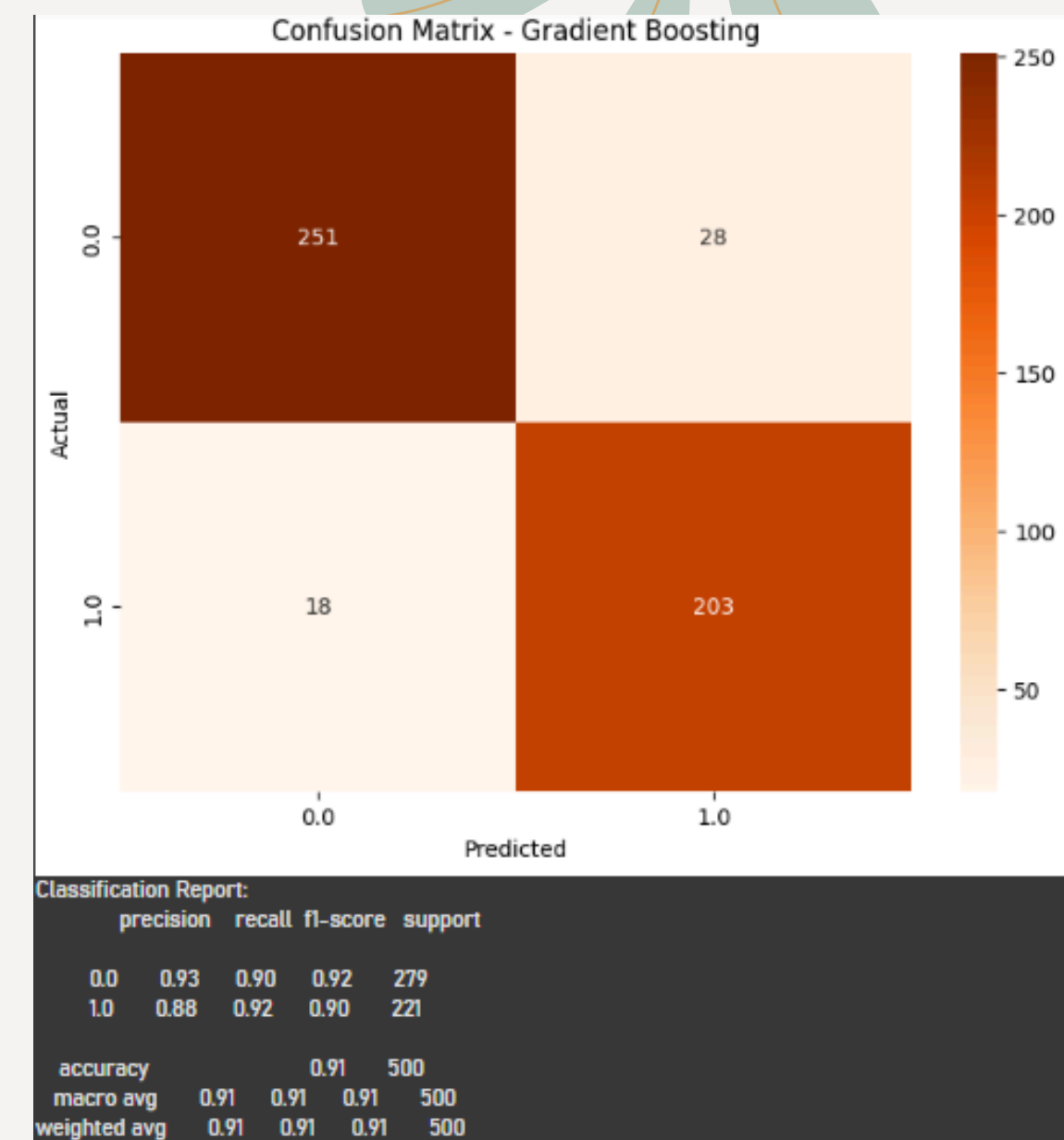
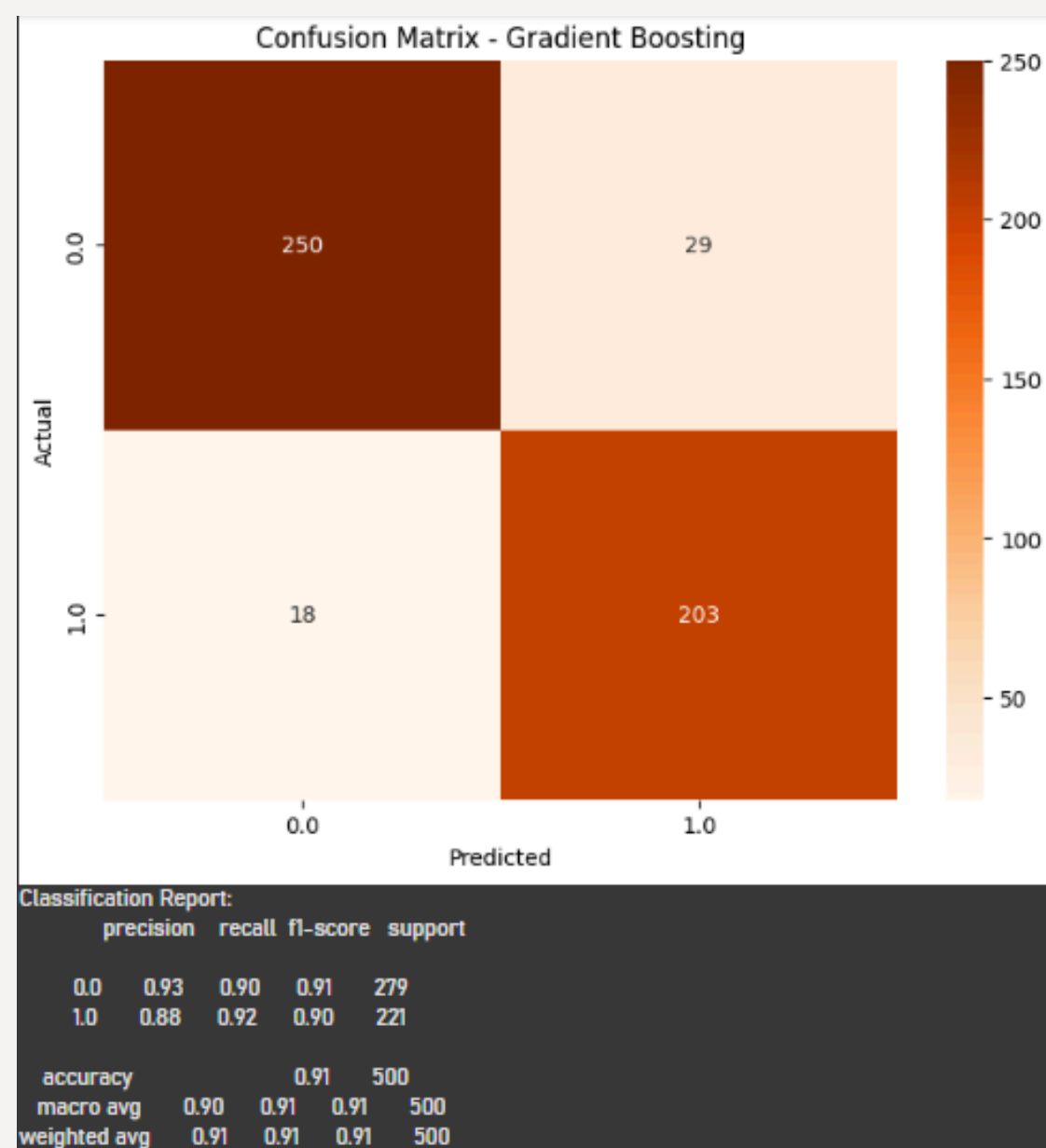
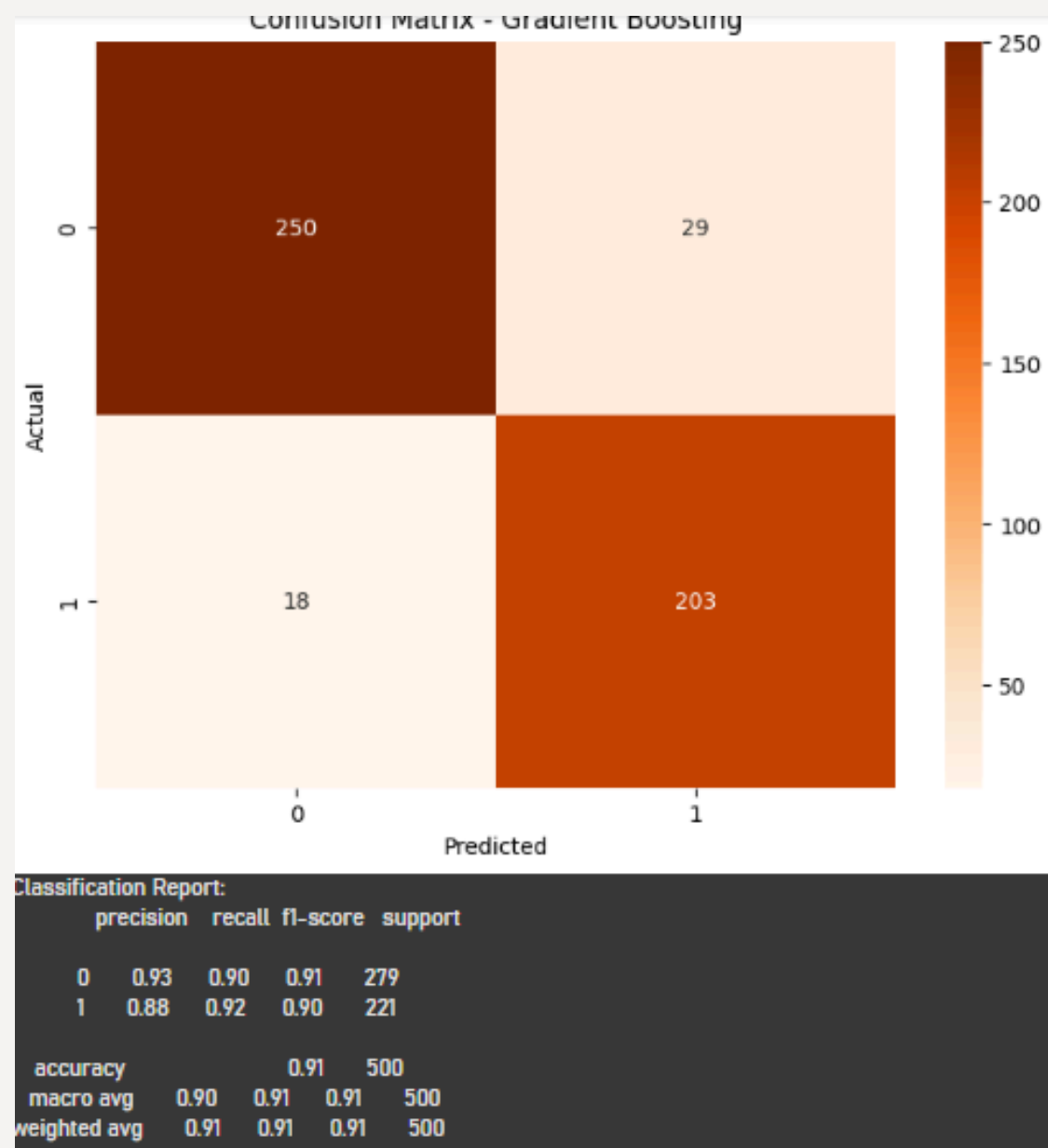
# PERSONALITY DATASET

## GRADIENT BOOSTING

### LIMPIO

### NORMALIZADO

## PCA

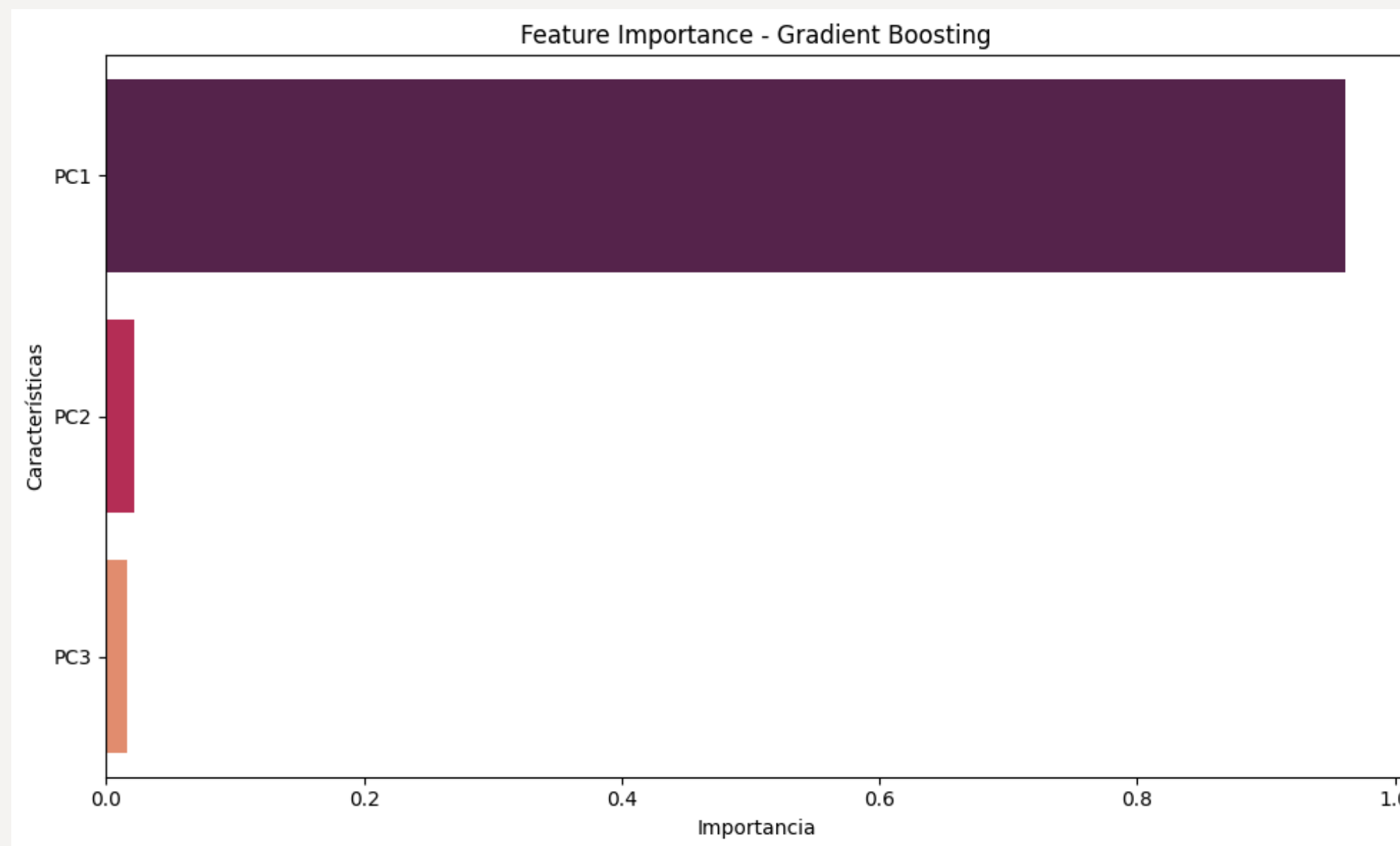


podemos ver que todos los enfoques tiene en mismo puntaje de macro avg, sin embargo pienso que el mejor seria PCA ya que tuvo menos errores en los falsos positivos en comparación con los demas enfoques. en comparación con los otros modelos esta al mismo nivel que el naive bayes y tiene mejor puntaje que KNN, arboles y random forest.



# PERSONALITY DATASET

## GRADIENT BOOSTING (PCA)



como vemos nuevamente PC1 tiene una importancia altísima, lo que indica que es la que más influye en la predicción del modelo, incluso es mas importante en este modelo que en el de random forest. asi mismo PC2 y PC3 tienen muy poca importancia relativa comparado con PC1, incluso menos que en random forest.



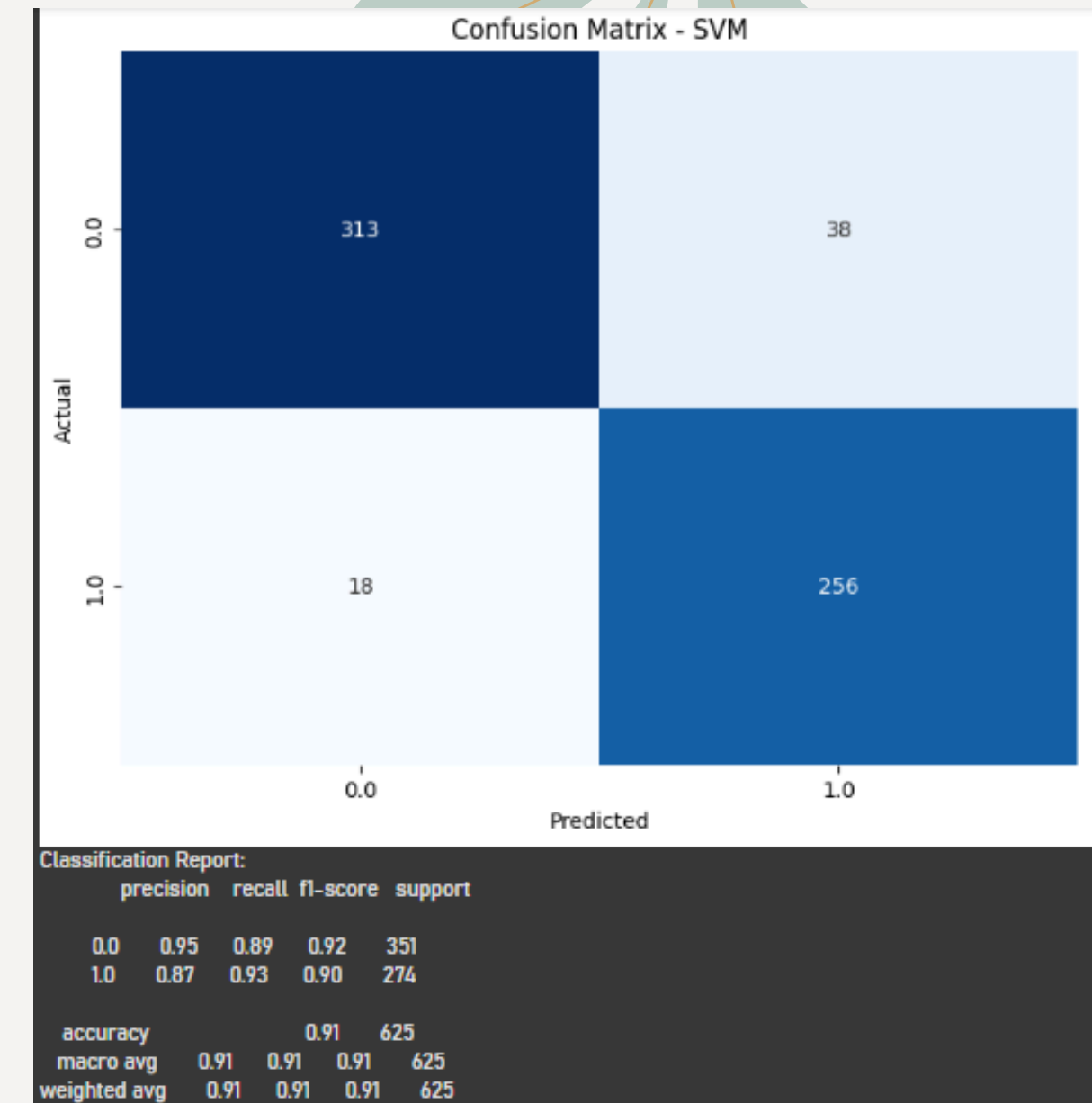
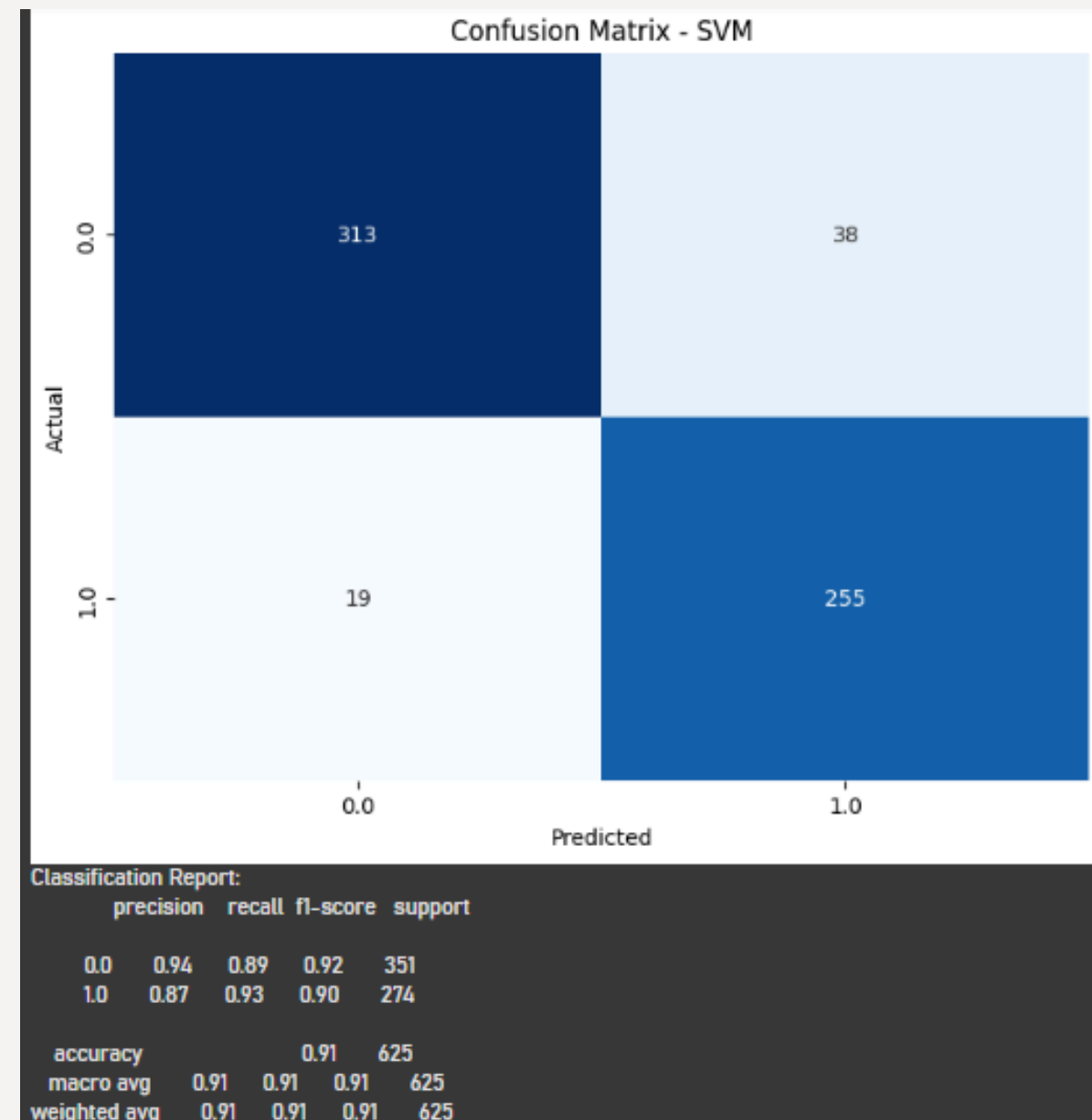
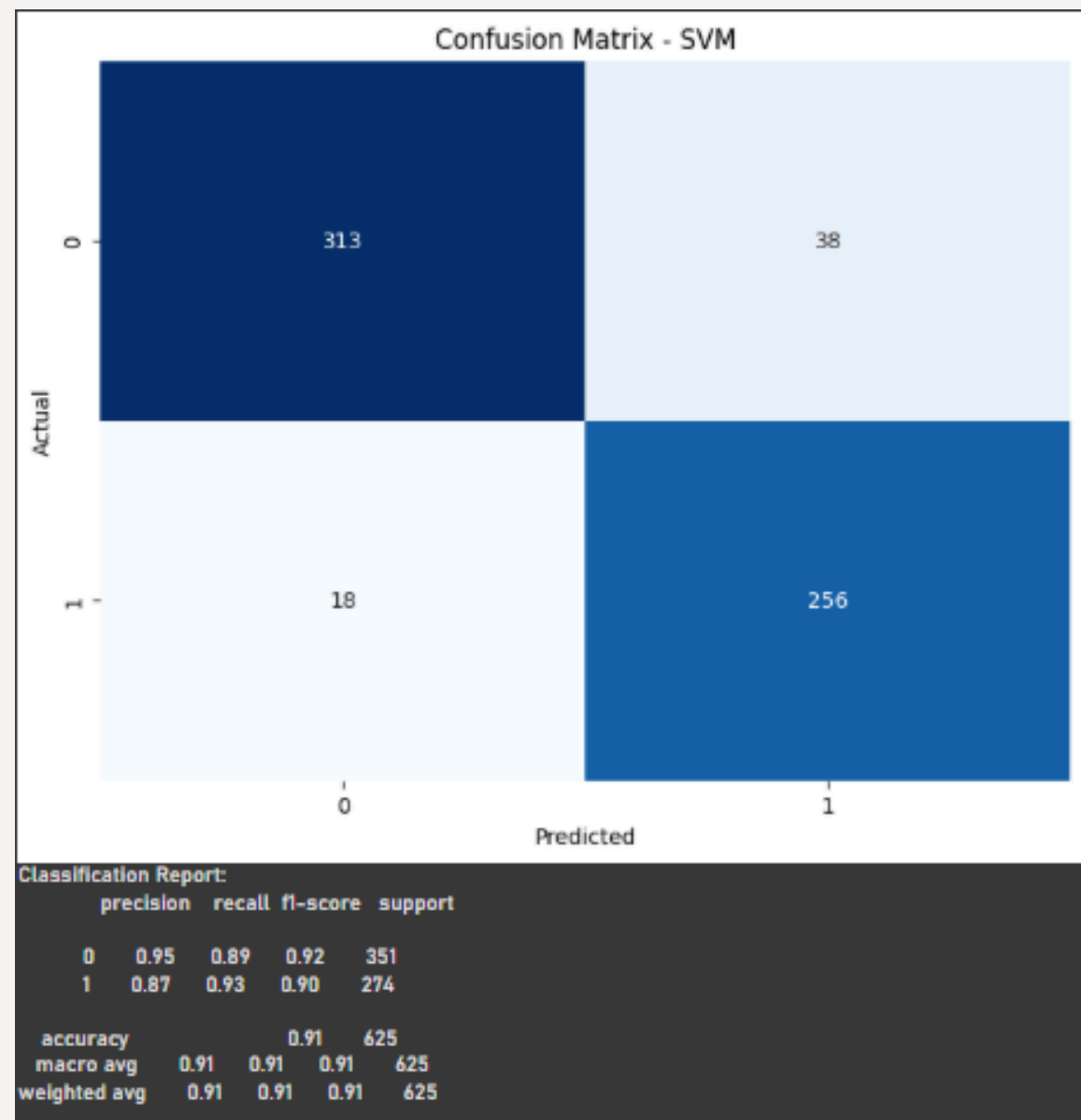
# PERSONALITY DATASET

## SVM

### LIMPIO

### NORMALIZADO

### PCA



veo que respecto a macro avg todos los enfoques tienen el mismo puntaje, y el limpio y PCA tienen la misma cantidad de fallas y el mismo puntaje, entonces cualquiera de los dos seria bueno para el modelo, por conveniencia decidí elegir el PCA. En comparación con los otros modelos esta tiene el mismo macro avg en f1 que naive bayes y que gradient boosting, pero tiene mas fallos que los otros dos en falsos positivos.

# PERSONALITY DATASET

## SVM (PCA)

Training SVM with kernel: linear  
Macro Avg F1-score for linear kernel: 0.9096485068770394

Training SVM with kernel: poly  
Macro Avg F1-score for poly kernel: 0.8981318348687938

Training SVM with kernel: rbf  
Macro Avg F1-score for rbf kernel: 0.9096485068770394

Macro Avg F1-scores for different kernels:  
linear: 0.9096485068770394  
poly: 0.8981318348687938  
rbf: 0.9096485068770394

Los kernels lineal y rbf obtuvieron el mejor desempeño, con un macro avg de f1 de 0.90, lo que indica que ambos modelos clasifican de forma equilibrada a introvertidos y extrovertidos. El kernel polinómico tuvo un desempeño un poco menor, por lo tanto no sería la mejor opción para este problema.

# PERSONALITY DATASET

## SELECCION DE CARACTERISTICAS

El mejor enfoque fue PCA, ya que mantuvo el rendimiento con menos variables, haciendo el modelo más simple y entendible. Naive Bayes y Gradient Boosting fueron los mejores modelos, con el mismo macro F1 (0.91), métricas iguales y errores similares.

Ambos son buenas opciones: Naive Bayes por su simplicidad y Gradient Boosting por su robustez. en este caso decidí elegir gradient boosting.

|   | Num_Features | Selected_Features | F1_Score |
|---|--------------|-------------------|----------|
| 0 | 1            | [PC1]             | 0.899070 |
| 1 | 2            | [PC1, PC2]        | 0.901126 |

Entonces con gradient boosting y PCA obtuve las características, donde se evidencia que: el mejor F1 Score macro promedio fue 0.9011, al utilizar dos componentes principales (PC1 y PC2), esto indica que agregar PC2 a PC1 mejora ligeramente el desempeño del modelo. vemos también que no aparece PC3, lo que indica que no aportaba un valor muy significativo al modelo.

# PERSONALITY DATASET VALIDACION CRUZADA

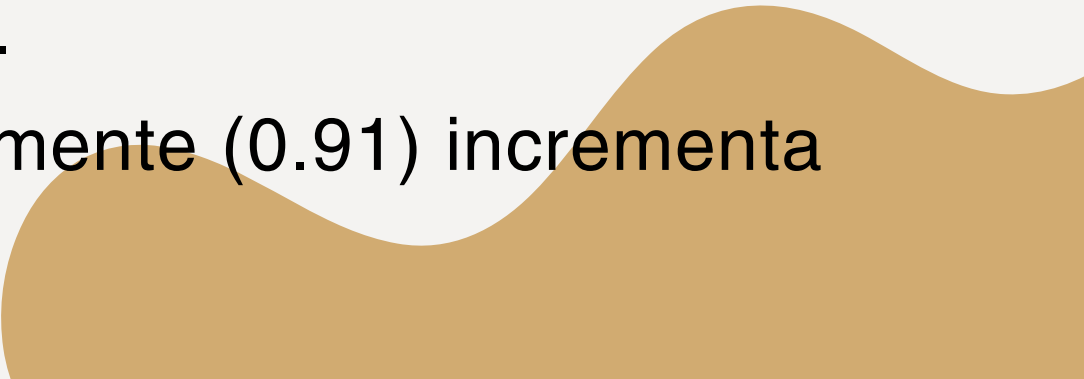


Resultados Finales por Fold y Promedio:

|   | Fold     | Accuracy | Precision | Recall   | F1 Score |
|---|----------|----------|-----------|----------|----------|
| 0 | 1        | 0.922000 | 0.922046  | 0.922000 | 0.922019 |
| 1 | 2        | 0.928000 | 0.928465  | 0.928000 | 0.928098 |
| 2 | 3        | 0.932000 | 0.932248  | 0.932000 | 0.932064 |
| 3 | 4        | 0.903808 | 0.904329  | 0.903808 | 0.903937 |
| 4 | 5        | 0.921844 | 0.921812  | 0.921844 | 0.921824 |
| 5 | Promedio | 0.921530 | 0.921780  | 0.921530 | 0.921588 |

la validación cruzada se hizo con el enfoque de PCA y el modelo de gradient boosting, donde vemos que hay 5 folds con métricas altas y muy similares, que indica que el modelo es robusto y confiable, ya que mantiene un rendimiento alto y parejo en cada iteración. el mejor desempeño fue en el fold 3 con 0.9320, el promedio fue de 0.9215.

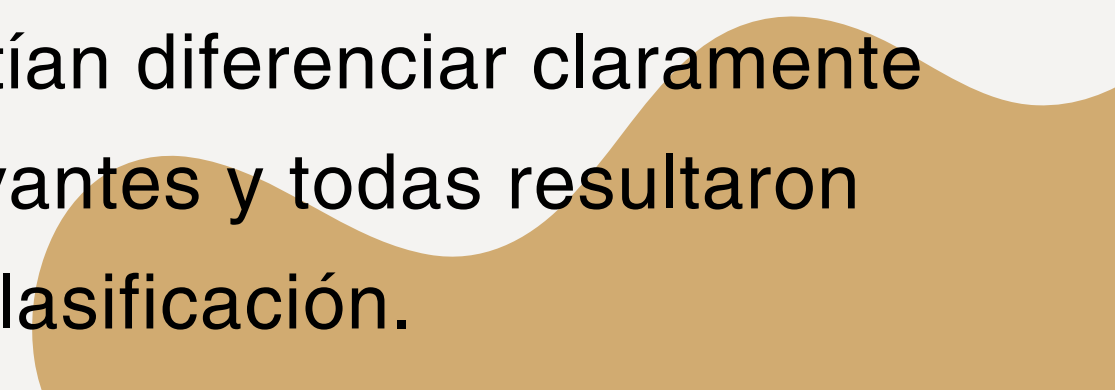
comparado con el macro avg f1 que obtuvimos al aplicar el modelo inicialmente (0.91) incrementa ligeramente al aplicar validación cruzada.



# PERSONALITY DATASET

## CONCLUSIONES



- El mejor enfoque fue PCA, ya que logró reducir las dimensiones de 5 a 3 componentes sin perder rendimiento ni precisión, manteniendo el 89% de la información y facilitando el análisis.
  - En cuanto a los modelos, Naive Bayes y Gradient Boosting Classifier fueron los más destacados, obteniendo métricas idénticas. se eligió Gradient Boosting por ser un modelo más robusto y flexible, capaz de capturar relaciones no lineales y complejas entre variables, con este modelo se alcanzo un 0.91 en f1 score y aumenta ligeramente al aplicar validación cruzada.
  - En el análisis gráfico univariado y bivariado, se observó que la mayoría de las variables seguían una distribución aproximadamente normal (por la forma de campana y la cercanía entre media y mediana), presentaban una ligera dispersión y permitían diferenciar claramente entre clases. Además, las variables mostraron correlaciones relevantes y todas resultaron estadísticamente significativas para el problema de clasificación.
- 



# GRACIAS POR SU ATENCIÓN

---

[www.unsitiogenial.es](http://www.unsitiogenial.es)

---