UNER Facultad de Ingeniería

Trabajo Practico de Modelado Estadístico
Tecnicatura en Procesamiento y Explotación de Datos.

Asignatura: Modelado Estadístico

DOCENTES:

- Blanco Mariana
- Aued Juan

INTEGRANTES:

- Busten Karen
- Durand Camila Ayelén

Fecha de entrega: 10/06/2025

Introducción

El presente informe describe el desarrollo y evaluación de un modelo de regresión logística con el objetivo de predecir la personalidad de los individuos, introvertida o extrovertida, a partir de variables relacionadas con su comportamiento social. Entre las variables consideradas se incluyen el tiempo que una persona pasa sola, la frecuencia de asistencia a eventos sociales, la sensación de agotamiento luego de socializar, el tamaño de su círculo de amistades y otros hábitos cotidianos. Estos factores fueron analizados para determinar su influencia en la categorización de la personalidad, partiendo de un conjunto de datos previamente estructurado.

Objetivos

En este proyecto nos enfocamos en crear un modelo de clasificación en el cual se clasifica a los individuos en extrovertidos o introvertidos aplicando los conocimientos adquiridos en la materia.

Descripción del dataset

Para realizar este trabajo utilizaremos un dataset publicado en Kaggle, el cual pueden encontrar en el siguiente link: https://www.kaggle.com/datasets/rakeshkapilavai/extrovert-vs-introvert-behavior-data

Dicho dataset contiene información sobre datos conductuales y sociales, capturando indicadores clave de extroversión e introversión. Cuenta con las siguientes características:

- Tamaño : El conjunto de datos contiene 2900 filas y 8 columnas.

- Variables que componen el dataset:
  - **Time_spent_Alone**: Número de horas (0–11) que una persona suele pasar sola diariamente
  - **Stage_fear**: Si la persona experimenta miedo escénico (Yes/No).
  - **Social_event_attendance**: Frecuencia (escala 0-10) de asistencia a eventos sociales.
  - **Going_outside**: Con qué frecuencia el individuo sale al exterior (escala 0-10).
  - **Drained_after_socializing**: Si el individuo se siente agotado después de socializar (Yes/No).
  - **Friends_circle_size**:Número de amigos cercanos (0–15).
  - **Post_frequency**: Frecuencia de publicación en redes sociales (0–10).
  - **Personality**: Variable objetivo, si el individuo es Introvertido o Extrovertido.

Exploración de los Datos

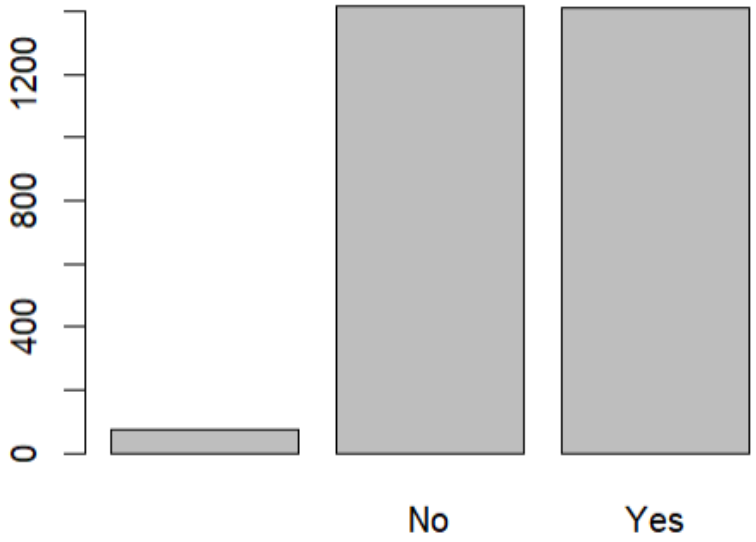
En esta etapa nos enfocamos en familiarizarnos con el dataset para su comprender la estructura y las características del conjunto de datos. Además identificar posibles problemas como valores nulos o atípicos, y preparar las variables para la aplicación del modelo de regresión logística.

Durante la exploración inicial, se identificó la presencia de valores faltantes (NA) en 5 columnas. La función colSums(is.na(datos)) nos permitió cuantificar el número de NAs por variable, revelando un total de 333 valores nulos distribuidos en las siguientes columnas: Time_spent_Alone, Social_event_attendance, Going_outside, Friends_circle_size y Post_frequency.

```
> colSums(is.na(datos))
        Time_spent_Alone               Stage_fear   Social_event_attendance
                      63                        0                        62
           Going_outside Drained_after_socializing      Friends_circle_size
                      66                        0                        77
         Post_frequency               Personality
                      65                        0
> sum(is.na(datos))
[1] 333
```

Para abordar estos valores faltantes en las variables numéricas, se optó por imputar la mediana en lugar de la media. La mediana es una medida de tendencia central más robusta frente a la presencia de valores atípicos, lo que la hace adecuada para rellenar datos sin introducir distorsiones significativas en la distribución de las variables.

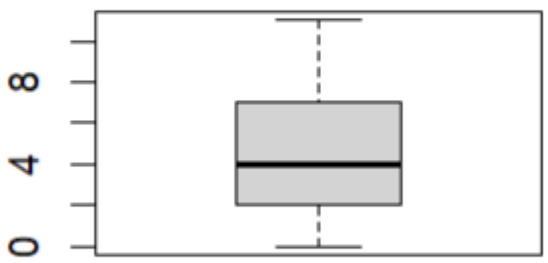
Adicionalmente, se detectaron celdas vacías ("") en las variables cualitativas, las cuales se trataron como valores faltantes y se eliminaron las filas correspondientes para asegurar la integridad de los datos categóricos. Esto se realizó mediante el reemplazo de cadenas vacías por NA y posterior eliminación de las filas completas con na.omit().



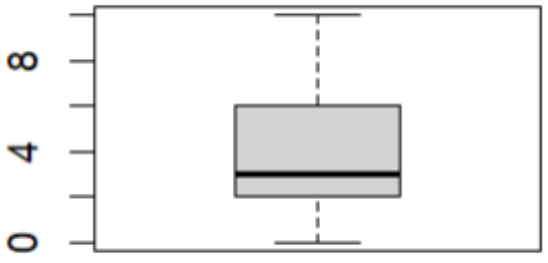
Esta acción resultó en la reducción del número de observaciones a 2776, garantizando un conjunto de datos limpio para el análisis.

Se realizó un análisis de la dispersión de las variables numéricas mediante diagramas de caja (boxplot()). Estos gráficos visuales permitieron identificar la presencia de posibles valores atípicos (outliers).

```
#Analizamos la disperción de los datos, obsevamos que no presentan outliers
boxplot(datos$Time_spent_Alone)
boxplot(datos$Social_event_attendance)
boxplot(datos$Going_outside)
boxplot(datos$Friends_circle_size)
boxplot(datos$Post_frequency)
```



> boxplot(datos$Time_spent_Alone)



> boxplot(datos$Social_event_attendance)

La observación de los boxplots sugirió que las variables numéricas no presentaban outliers significativos que pudieran distorsionar el análisis.

Las variables categóricas Stage_fear, Drained_after_socializing y Personality fueron analizadas mediante gráficos de barras (barplot(table(as.factor())) para visualizar la distribución de sus categorías.

```
#Analizamos las variables cualitativas
barplot(table(as.factor(datos$Stage_fear)))
barplot(table(as.factor(datos$Drained_after_socializing)))
barplot(table(as.factor(datos$Personality)))
```

Posteriormente, las variables fueron convertidas a factores (factor()), que es el tipo de dato adecuado para variables categóricas en R. Para la variable objetivo Personality, que originalmente contenía las categorías "Extrovert" e "Introvert", se realizó una transformación a un formato binario (0 y 1). Se asignó '1' a "Extrovert" y '0' a "Introvert".

Finalmente, se verificó la estructura del conjunto de datos (str(datos)) y se generó un nuevo resumen (summary(datos)) para confirmar que todas las transformaciones se habían aplicado correctamente y que el dataset estaba listo para la fase de modelado.

```
> str(datos)
Classes 'data.table' and 'data.frame':  2776 obs. of  8 variables:
 $ Time_spent_Alone         : num  4 9 9 0 3 1 4 2 10 0 ...
 $ Stage_fear               : Factor w/ 2 levels "No","Yes": 1 2 2 1 1 1 1 1 2 1 ...
 $ Social_event_attendance  : num  4 0 1 6 9 7 9 8 1 8 ...
 $ Going_outside            : num  6 0 2 7 4 5 3 4 3 6 ...
 $ Drained_after_socializing: Factor w/ 2 levels "No","Yes": 1 2 2 1 1 1 1 1 2 1 ...
 $ Friends_circle_size      : num  13 0 5 14 8 6 7 7 0 13 ...
 $ Post_frequency           : num  5 3 2 8 5 6 7 8 3 8 ...
 $ Personality              : Factor w/ 2 levels "0","1": 2 1 1 2 2 2 2 2 1 2 ...
```

```
> summary(datos)
 Time_spent_Alone Stage_fear Social_event_attendance Going_outside
 Min.   : 0.000   No :1399   Min.   : 0.000          Min.   :0.000
 1st Qu.: 2.000   Yes:1377   1st Qu.: 2.000          1st Qu.:1.000
 Median : 4.000              Median : 3.000          Median :3.000
 Mean   : 4.497              Mean   : 3.947          Mean   :2.999
 3rd Qu.: 7.000              3rd Qu.: 6.000          3rd Qu.:5.000
 Max.   :11.000              Max.   :10.000          Max.   :7.000
 Drained_after_socializing Friends_circle_size Post_frequency   Personality
 No :1399                  Min.   : 0.000      Min.   : 0.000   0:1345
 Yes:1377                  1st Qu.: 3.000      1st Qu.: 1.000   1:1431
                           Median : 5.000      Median : 3.000
                           Mean   : 6.228      Mean   : 3.554
                           3rd Qu.:10.000      3rd Qu.: 6.000
                           Max.   :15.000      Max.   :10.000
>
```

Hipótesis

El presente estudio parte de la hipótesis principal de que la personalidad de un individuo, categorizada como introvertida o extrovertida, puede ser predicha significativamente a

través de la observación de sus hábitos sociales y personales. Esta hipótesis será evaluada mediante la construcción de un modelo de regresión logística, buscando identificar cuáles de las variables consideradas son los predictores más influyentes en la determinación de la personalidad introvertida o extrovertida.

Pregunta: *¿Cuáles de estás variables están más relacionadas con la personalidad del individuo?*

Modelado

Antes de construir el modelo de regresión logística, se realizó la codificación "dummy" de las variables categóricas nominales Drained_after_socializing y Stage_fear utilizando la función dummy_cols(). Esto convierte cada categoría de una variable cualitativa en una nueva columna binaria (0 o 1), lo que permite que sean interpretadas por el modelo de regresión. La columna original seleccionada se eliminó después de la transformación para evitar redundancias. Es importante destacar que la variable objetivo Personality no fue transformada en esta instancia, ya que ya había sido codificada como 0 y 1 en la fase de exploración.

```
#Aplicamos Dummies sobre las variables tipo factor menos la objetivo
datos <- dummy_cols(datos,
                    select_columns = c("Drained_after_socializing","Stage_fear"),
                    remove_selected_columns = T)
```

Para evaluar la capacidad de generalización del modelo, el conjunto de datos se dividió en dos subconjuntos: entrenamiento y prueba. Se asignó el 70% de las observaciones al conjunto de entrenamiento (datos.train) y el 30% restante al conjunto de prueba (datos.test), se eligió dicho porcentaje para asegurar una cantidad considerable de datos de testeo y así obtener mejores resultados.
Luego se construyó un modelo de regresión logística (glm con family = "binomial") utilizando todas las variables predictoras disponibles. El resumen del modelo (summary(modelo.datos)) proporcionó información sobre la significancia estadística de cada predictor y la calidad general del ajuste.

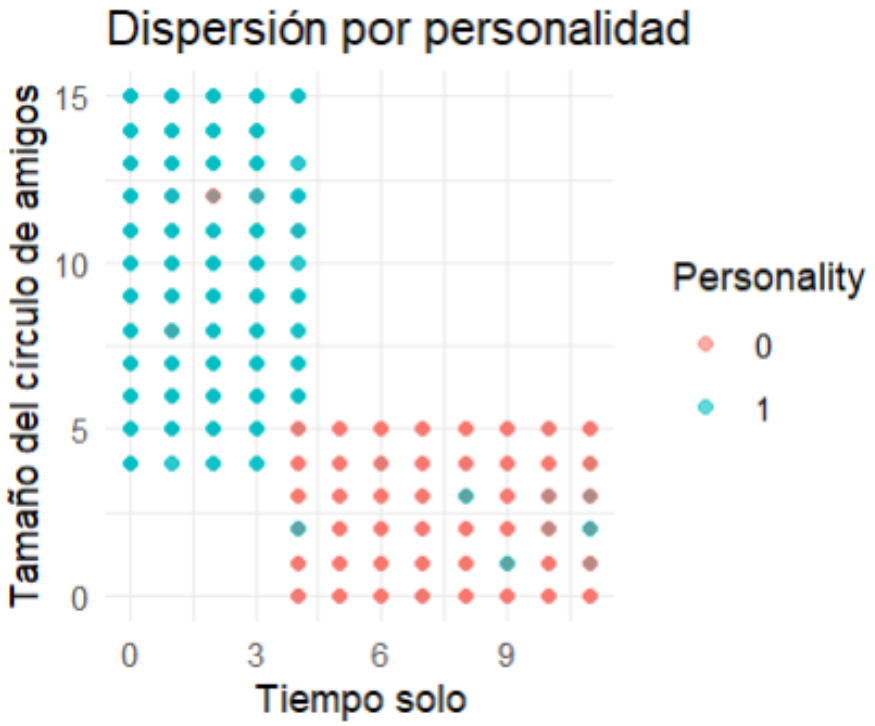
```
Call:
glm(formula = Personality ~ ., family = "binomial", data = datos.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8919  -0.3897   0.1619   0.3255   2.8245

Coefficients: (3 not defined because of singularities)
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -3.17438    0.46715  -6.795 1.08e-11 ***
Time_spent_Alone             0.22686    0.05116   4.434 9.23e-06 ***
Social_event_attendance     -0.14817    0.06253  -2.370  0.01780 *
Going_outside               -0.23282    0.08589  -2.711  0.00672 **
Friends_circle_size         -0.18824    0.04130  -4.558 5.17e-06 ***
Post_frequency              -0.09242    0.06133  -1.507  0.13181
Drained_after_socializing_No 10.30500    0.80001  12.881  < 2e-16 ***
Drained_after_socializing_Yes       NA         NA      NA       NA
Stage_fear_No                       NA         NA      NA       NA
Stage_fear_Yes                      NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Los coeficientes del modelo inicial revelaron las siguientes relaciones significativas (p-value < 0.05):
Time_spent_Alone: presenta un coeficiente positivo (0.22686, $p < 0.001$), lo cual indica que, contra intuitivamente, a mayor tiempo que una persona pasa sola, mayor es la probabilidad de que sea extrovertida.

Drained_after_socializing_No: tiene un coeficiente alto (10.30500, $p < 0.001$), lo que indica que las personas que no se sienten agotadas después de socializar tienen una probabilidad significativamente mayor de ser extrovertidas.

Social_event_attendance (−0.14817, $p \approx 0.017$), **Going_outside** (−0.23822, $p \approx 0.0067$), **Friends_circle_size** (−0.18824, $p < 0.001$), y **Post_frequency** (−0.09242, $p \approx 0.13$, no significativa) tienen coeficientes negativos, lo que sugiere que un menor número de amigos, menor frecuencia de salidas o de participación en eventos sociales se asocia con una mayor probabilidad de ser introvertido. De estos, los coeficientes de asistencia a eventos sociales, salir afuera y tamaño del círculo social son estadísticamente significativos, reforzando su relevancia como predictores del tipo de personalidad.

Se observó que las variables Drained_after_socializing_Yes, Stage_fear_No y Stage_fear_Yes fueron automáticamente excluidas del modelo debido a colinealidad perfecta. Esto significa que estas variables estaban perfectamente correlacionadas con otras variables en el modelo o con la constante, lo que impide su inclusión independiente ya que no aportan información adicional.



Se construyó un gráfico de dispersión para explorar visualmente la relación entre el tiempo pasado solo (Time_spent_Alone) y el tamaño del círculo de amigos (Friends_circle_size), diferenciando por tipo de personalidad (introvertido = 0, extrovertido = 1).

A primera vista, el gráfico sugiere que los individuos con mayor tiempo en soledad tienden a tener un círculo de amigos más pequeño, especialmente entre los introvertidos. Esta observación podría llevar a pensar que el tiempo en soledad está negativamente asociado con la extroversión.

Sin embargo, el modelo de regresión logística muestra un coeficiente positivo y significativo para la variable Time_spent_Alone (Estimate = +0.2289), lo que indica que, controlando por otras variables, un mayor tiempo en soledad incrementa la probabilidad de que un individuo sea clasificado como extrovertido.

Evaluación

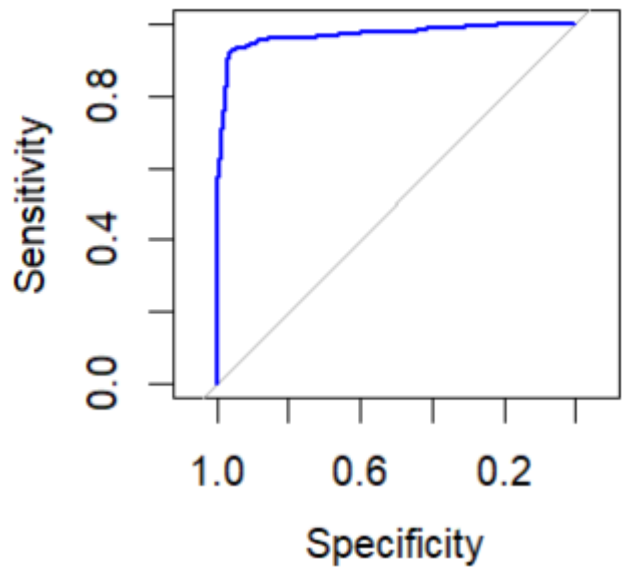
El primer modelo de regresión logística fue entrenado para predecir la personalidad de los individuos (introvertido = 0, extrovertido = 1) a partir de variables relacionadas con el comportamiento social. El rendimiento del modelo fue evaluado mediante una matriz de confusión, métricas de clasificación y la curva ROC.

```
> datos.conf
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 400  30
         1  19 384
```

```
> datos.conf$byClass
         Sensitivity          Specificity       Pos Pred Value       Neg Pred Value
           0.9546539           0.9275362           0.9302326           0.9528536
           Precision              Recall                  F1          Prevalence
           0.9302326           0.9546539           0.9422850           0.5030012
Detection Rate Detection Prevalence    Balanced Accuracy
           0.4801921           0.5162065           0.9410951
```

La **precisión global** del modelo fue del **94,12%**, con un intervalo de confianza del 95% entre 92,3% y 95,62%. La **sensibilidad** (capacidad para detectar correctamente a los introvertidos) fue de **95,47%**, mientras que la **especificidad** (capacidad para detectar correctamente a los extrovertidos) fue de **92,75%**. Además, el modelo obtuvo un valor F1 de **0,94**, lo que indica un buen equilibrio entre precisión y recall.

La curva ROC mostró un rendimiento sobresaliente, con un área bajo la curva (AUC) cercana a 1, lo cual confirma la buena capacidad discriminativa del modelo.



Optimización del Modelo

Con el objetivo de mejorar la interpretación y eficiencia del modelo, se ajustó un segundo modelo de regresión logística utilizando únicamente las variables que resultaron estadísticamente significativas en el modelo inicial. Las variables seleccionadas fueron:

- Time_spent_Alone
- Drained_after_socializing_No
- Social_event_attendance
- Going_outside
- Friends_circle_size

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8032  -0.3905   0.1667   0.3256   2.8653

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                -3.26931    0.46223  -7.073 1.52e-12 ***
Time_spent_Alone            0.22889    0.05112   4.478 7.55e-06 ***
Drained_after_socializing_No  9.91504    0.75222  13.181  < 2e-16 ***
Social_event_attendance    -0.15159    0.06239  -2.430  0.01512 *
Going_outside              -0.24666    0.08541  -2.888  0.00388 **
Friends_circle_size        -0.18764    0.04130  -4.544 5.53e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

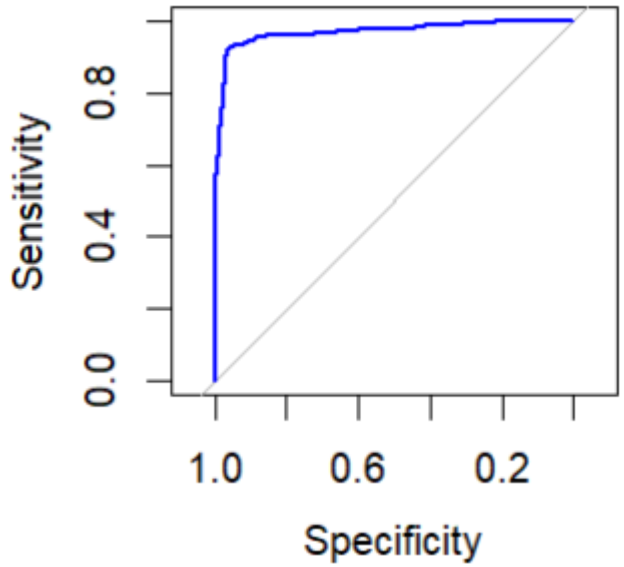
Time_spent_Alone presentó un coeficiente positivo (0.23), lo cual indica que a mayor tiempo que una persona pasa sola, mayor es la probabilidad de que sea extrovertida. La variable *Drained_after_socializing_No* tuvo un coeficiente altamente positivo (9.91), lo que sugiere que las personas que **no** se sienten agotadas después de socializar tienen muchas más probabilidades de ser extrovertidas.
Por otro lado, variables como *Social_event_attendance*, *Going_outside* y *Friends_circle_size* presentaron coeficientes negativos lo que implica que un mayor tamaño del círculo de amistades se asocia con una mayor probabilidad de ser extrovertido. En conjunto, estos coeficientes reflejan relaciones lógicas y esperadas según el comportamiento social relacionado con la personalidad.

El modelo optimizado mostró un excelente rendimiento predictivo. La precisión global (Accuracy) alcanzó el 94,12%, mientras que la sensibilidad (capacidad para identificar correctamente a los introvertidos) fue del 95,47% y la especificidad (detección correcta de extrovertidos) fue del 92,75%. El valor F1 fue de 0,94, lo que evidencia un equilibrio adecuado entre precisión y exhaustividad.

```
> datos.conf_opt
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 400  30
         1  19 384
```

```
> #
> datos.conf_opt$byClass
         Sensitivity          Specificity       Pos Pred Value       Neg Pred Value
           0.9546539           0.9275362           0.9302326           0.9528536
           Precision              Recall                  F1          Prevalence
           0.9302326           0.9546539           0.9422850           0.5030012
      Detection Rate Detection Prevalence    Balanced Accuracy
           0.4801921           0.5162065           0.9410951
```

La curva ROC también mostró un excelente comportamiento con un AUC de 0,967, lo cual confirma una capacidad discriminativa muy alta para diferenciar entre personalidades introvertidas y extrovertidas.



Discusión

Uno de los hallazgos más interesantes del análisis fue la aparente contradicción entre el gráfico de dispersión y el coeficiente estimado para la variable Time_spent_Alone. Visualmente, el gráfico sugiere que los individuos que pasan más tiempo solos tienden a tener un círculo de amigos más pequeño, lo cual podría asociarse intuitivamente con una personalidad más introvertida. Sin embargo, el modelo de regresión logística mostró un coeficiente positivo y estadísticamente significativo para esta variable, indicando que, al controlar por otras variables, un mayor tiempo en soledad incrementa la probabilidad de ser clasificado como extrovertido.

Este resultado muestra la importancia de los modelos multivariados, que permiten aislar el efecto de cada variable en presencia de otras. Es posible que el tiempo en soledad, en combinación con no sentirse cansado tras socializar o con un círculo social más reducido, represente un tipo de extroversión más introspectiva o selectiva, lo cual no se capta fácilmente en un análisis bivariado.

Además, se compararon dos modelos: uno completo con todas las variables disponibles y otro optimizado que incluyó únicamente aquellas que resultaron significativas. Ambos modelos arrojan exactamente el mismo rendimiento predictivo, con una precisión del 94.12% y un AUC de 0.9678. Esto sugiere que el modelo optimizado es preferible, ya que ofrece la misma capacidad de predicción con menor complejidad y mayor interpretabilidad.

Conclusión

El modelo de regresión logística desarrollado logró clasificar con alta precisión a los individuos como extrovertidos o introvertidos, utilizando variables relacionadas con sus hábitos sociales y preferencias personales. La variable Time_spent_Alone, inicialmente contraintuitiva, resultó ser un predictor positivo de extroversión cuando se consideraron otras variables en conjunto.

La comparación entre el modelo completo y el modelo optimizado demostró que es posible reducir la cantidad de variables sin sacrificar rendimiento, lo cual mejora la eficiencia del modelo y facilita su interpretación. Este enfoque puede ser útil en contextos donde se requiere una clasificación precisa de la personalidad con un conjunto limitado de indicadores.