



# STROKE PREDICTION

-

MARIANA MESA PÉREZ





# 1) EXPLORACIÓN

# SOBRE EL DATASET

Este set de datos es usado para predecir si un paciente tiene probabilidad de sufrir un derrame basado en parámetros de entrada como género, edad, enfermedades, tipo de trabajo, residencia, entre otros.

5110 FILAS  
12 COLUMNAS

COLUMNA	DESCRIPCIÓN
id	Identificador único
gender	Masculino, femino, otro
age	Edad del paciente
hypertension	0: paciente sin hipertensión. 1: paciente con hipertensión
heart_disease	0: paciente sin enfermedad cardiacas. 1: paciente con enfermedad cardiaca.
ever_married	No o Sí
work_type	Niños, trabajo gubernamental, privado, independiente, nunca trabajó
residence_type	Rural o Urbano
avg_glucose_level	promedio de nivel de glucosa en la sangre
bmi	Indice de masa corporal
smoking_status	Fumaba, nunca fumó, fuma, desconocido
stroke	1: paciente tuvo derrame. 0: paciente no tuvo derrame

## DUPLICADOS

Sin datos  
duplicados

## DATOS FALTANTES

201 en columna  
'bmi'

## COLUMNAS INNECESARIAS

Columna 'id', 5110  
datos únicos

## DATOS "INCONSISTENTES"

Name: smoking\_status

never smoked 1892

**Unknown 1544**

formerly smoked 885

smokes 789

En 'smoking\_status' al rededor del 30.41% del total de datos, es 'Unknown', que significa que no habían datos disponibles para 1544 pacientes.

Datos numéricos: presencia de edades decimales.

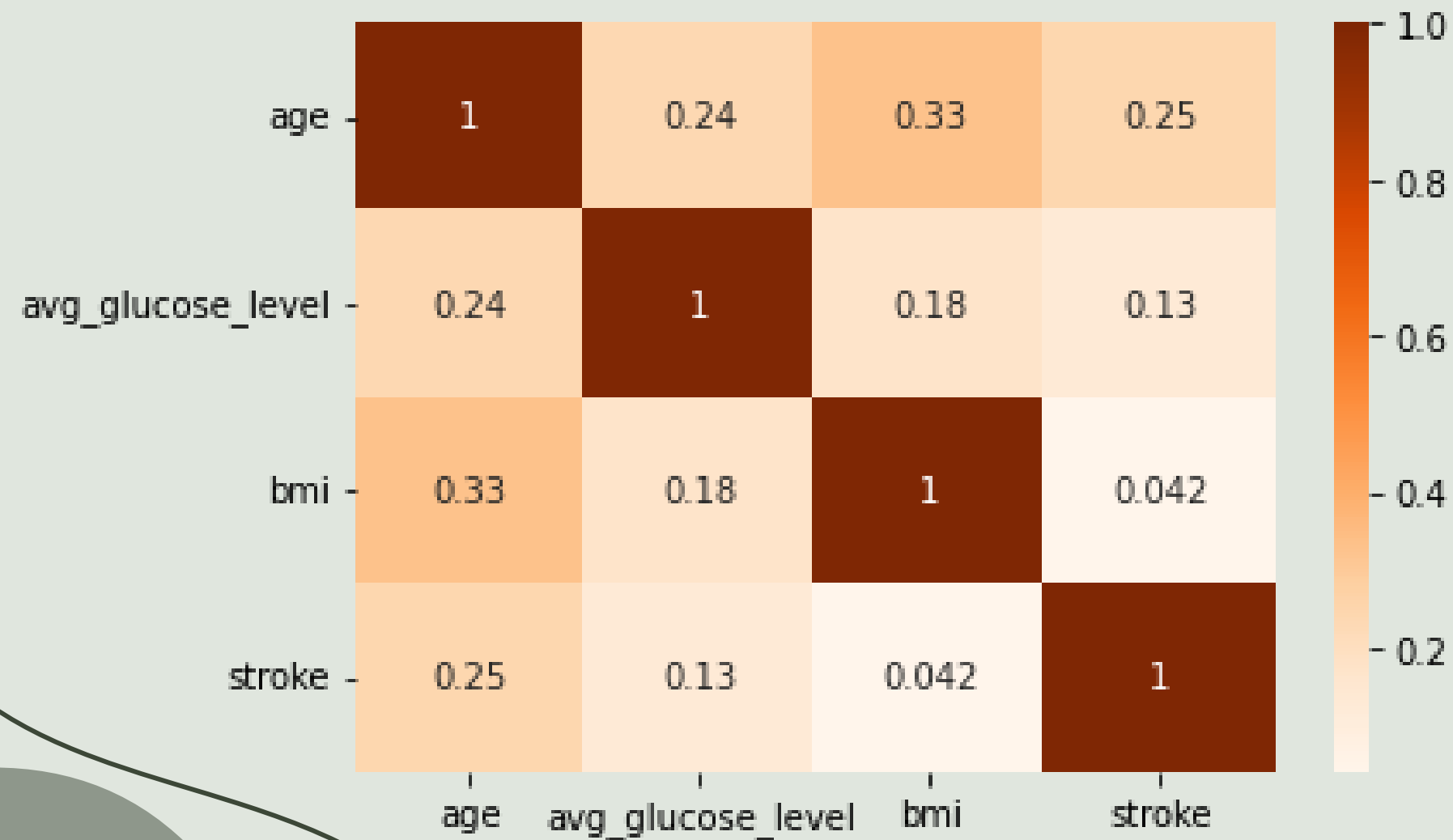


COLUMNA OBJETIVO: 'STROKE'

De los **5110** pacientes, **4861 (95.12%)** no sufrieron un derrame y **249 (4.87%)** sí.

Dataset desbalanceado.

# CORRELACIÓN




No hay relaciones significativas entre las características de estudio.

Más alta: 'bmi' - 'age'. 0.33



## 2) MANEJO DATOS FALTANTES E INCONSISTENTES

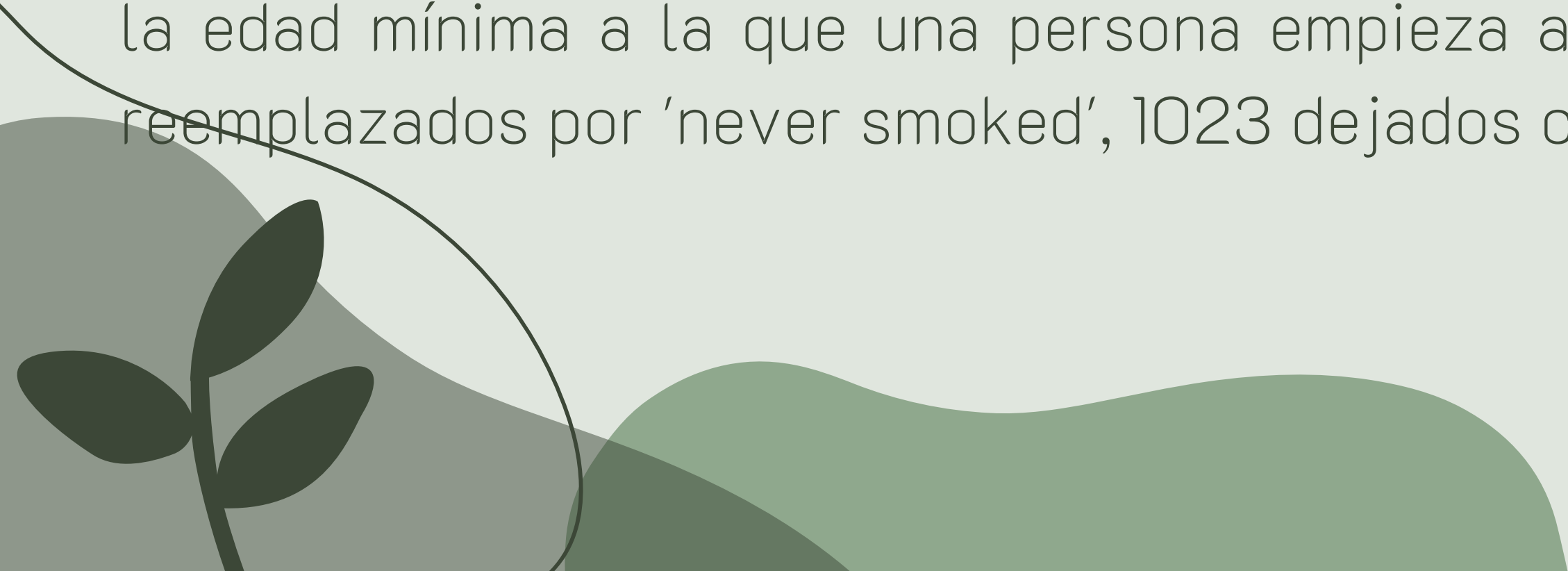




1) Datos faltantes columna 'bmi': 201 (3.93%). Se calculó la media de esta columna y se reemplazaron valores nulos por esta.

2) Datos decimales en columna 'age': conversión de datos flotantes a enteros.

3) 'smoking\_status' desconocido. Según una encuesta realizado en EEUU en 2020, la edad mínima a la que una persona empieza a fumar son los 12 años; 521 datos reemplazados por 'never smoked', 1023 dejados con esta como una categoría.





### 3) MODELOS DE PREDICCIÓN.

# 1. ÁRBOLES DE CLASIFICACIÓN

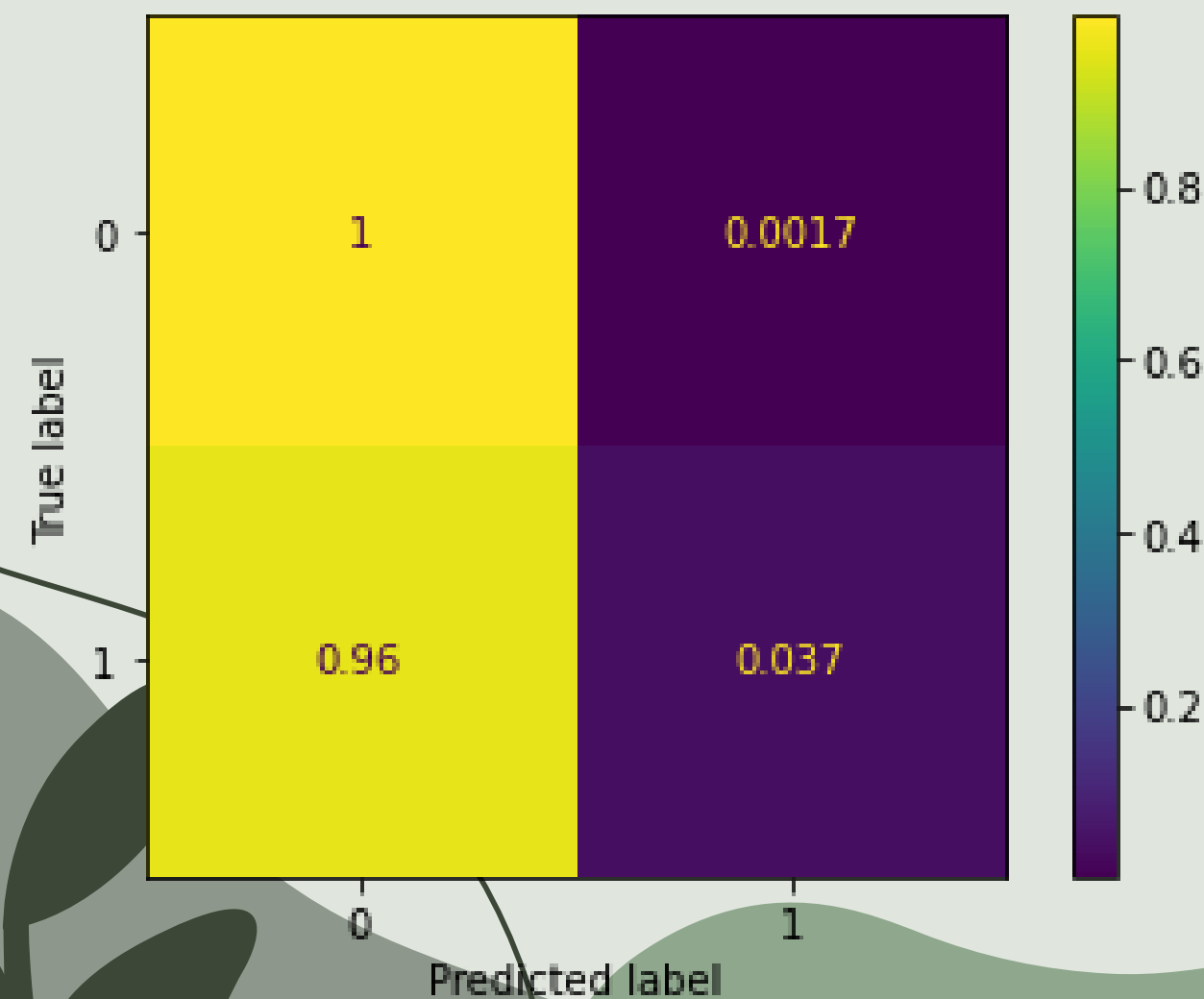
Profundidad máxima óptima: 3.

Accuracy= 0.9381.

Precision= 0.6

Recall = 0.03.

f1=0.07.



# 2. K VECINOS MÁS CERCANOS

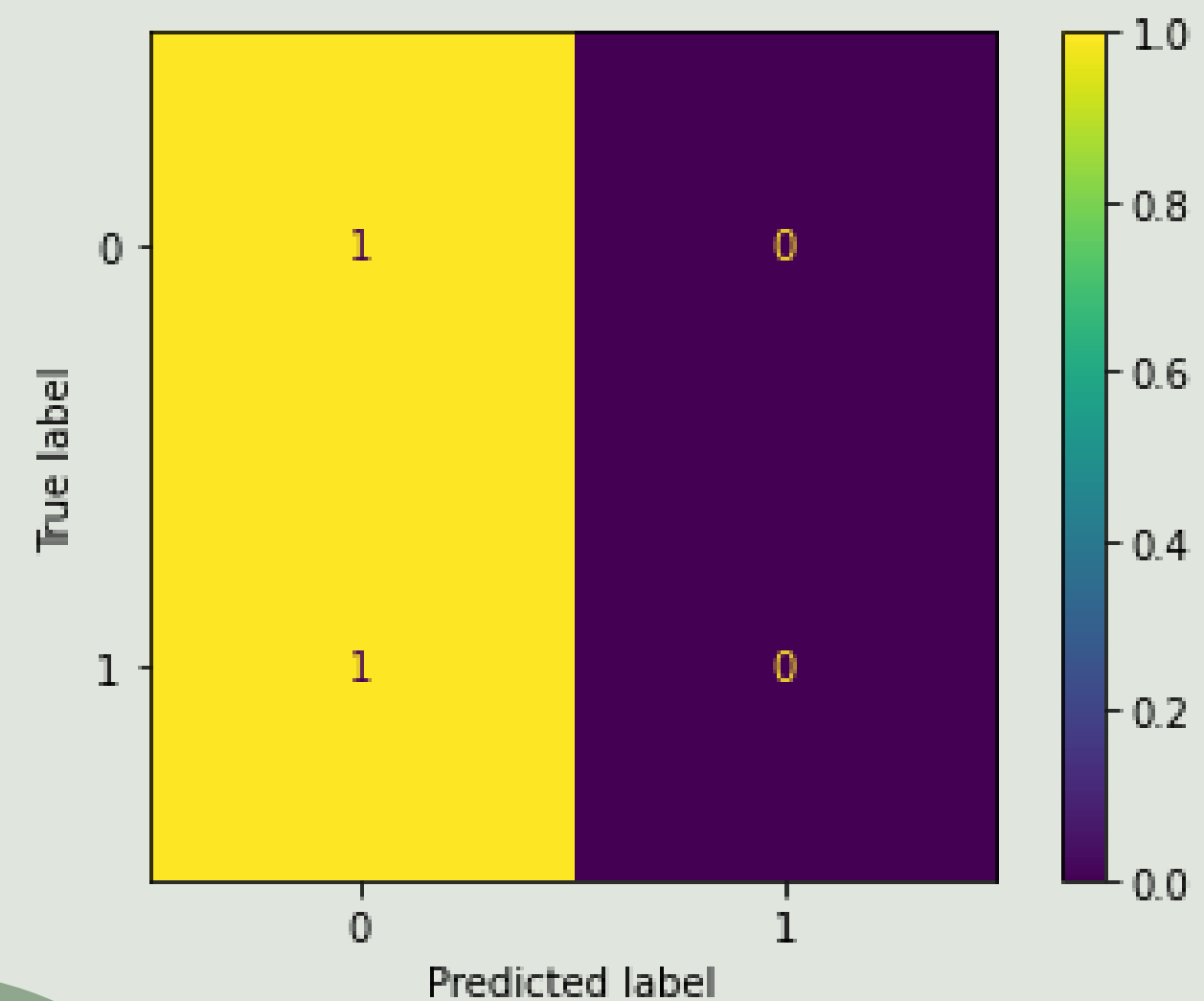
# optimo de vecinos: 4.

Accuracy= 0.9374.

Precision=0

Recall = 0.

f1=0



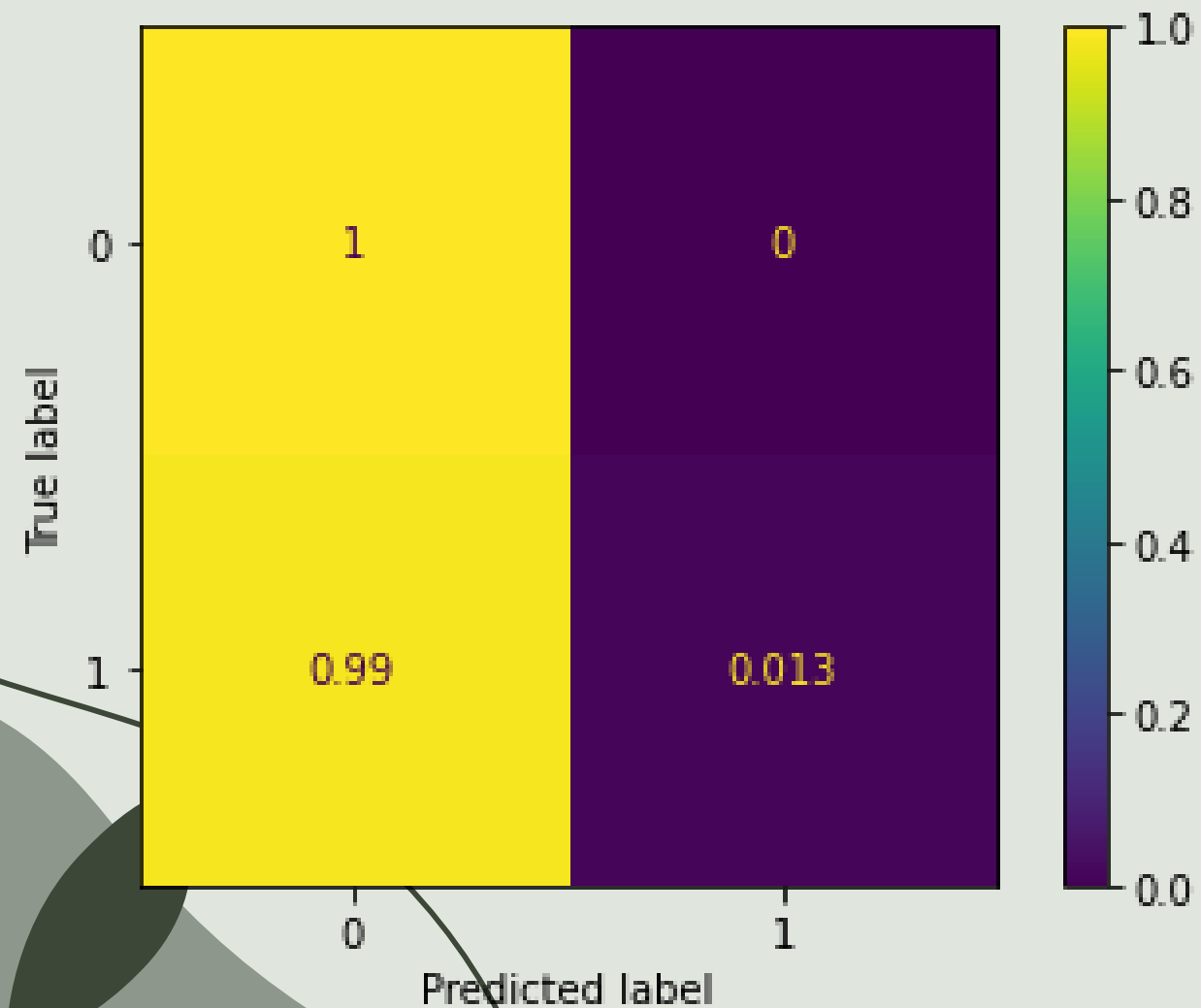
## 3. REGRESIÓN LOGÍSTICA

Accuracy= 0.9381.

Precision= 1.0

Recall = 0.012.

f1=0.024.



## 4. BAGGED TREES

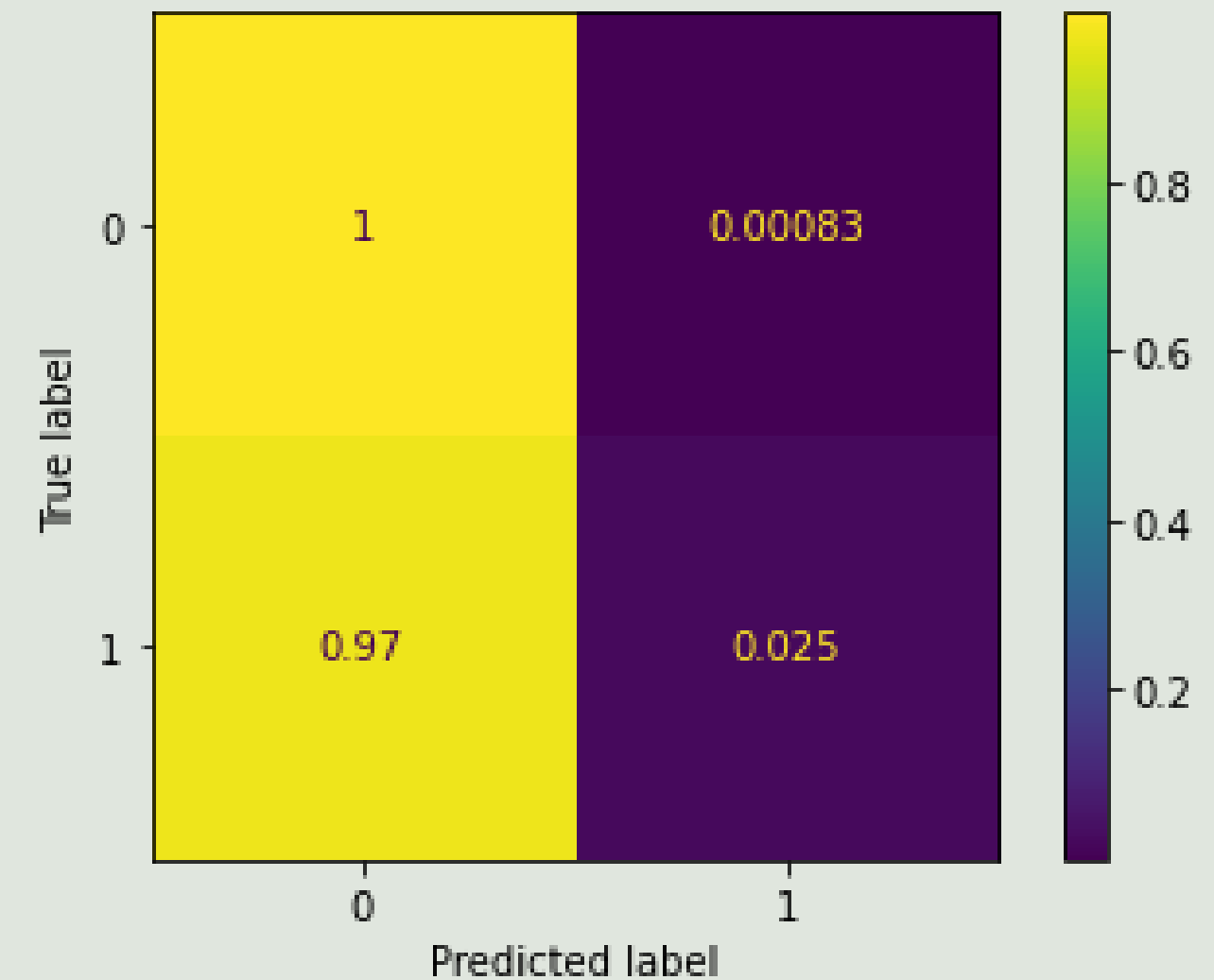
# optimo de estimadores: 50.

Accuracy= 0.9381.

Precision= 0.66

Recall = 0.025.

f1= 0.048



# 5. RANDOM FOREST

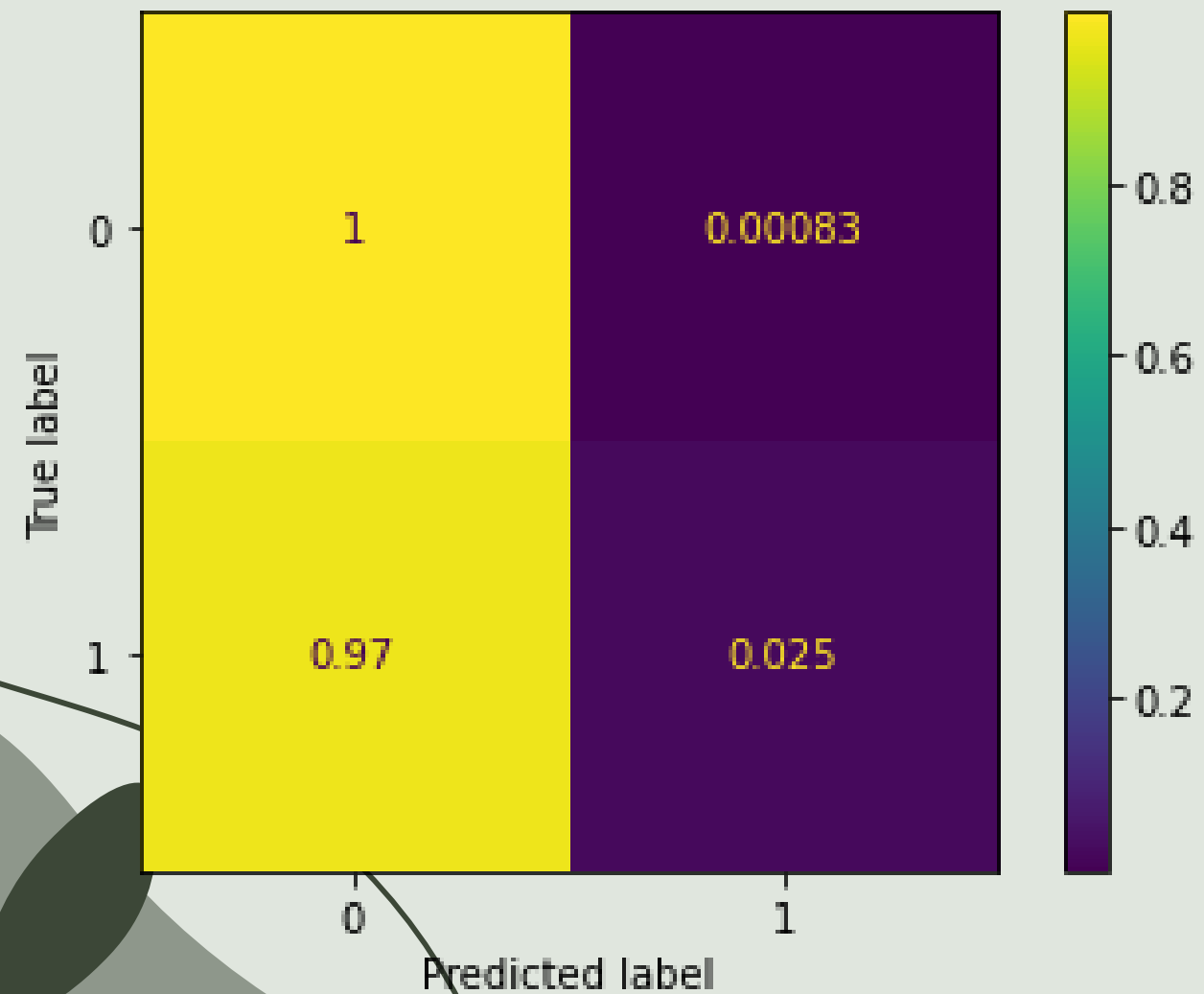
# optimo de estimadores: 120.

Accuracy= 0.9381.

Recall = 0.025.

Precision= 0.66

f1=0.048.

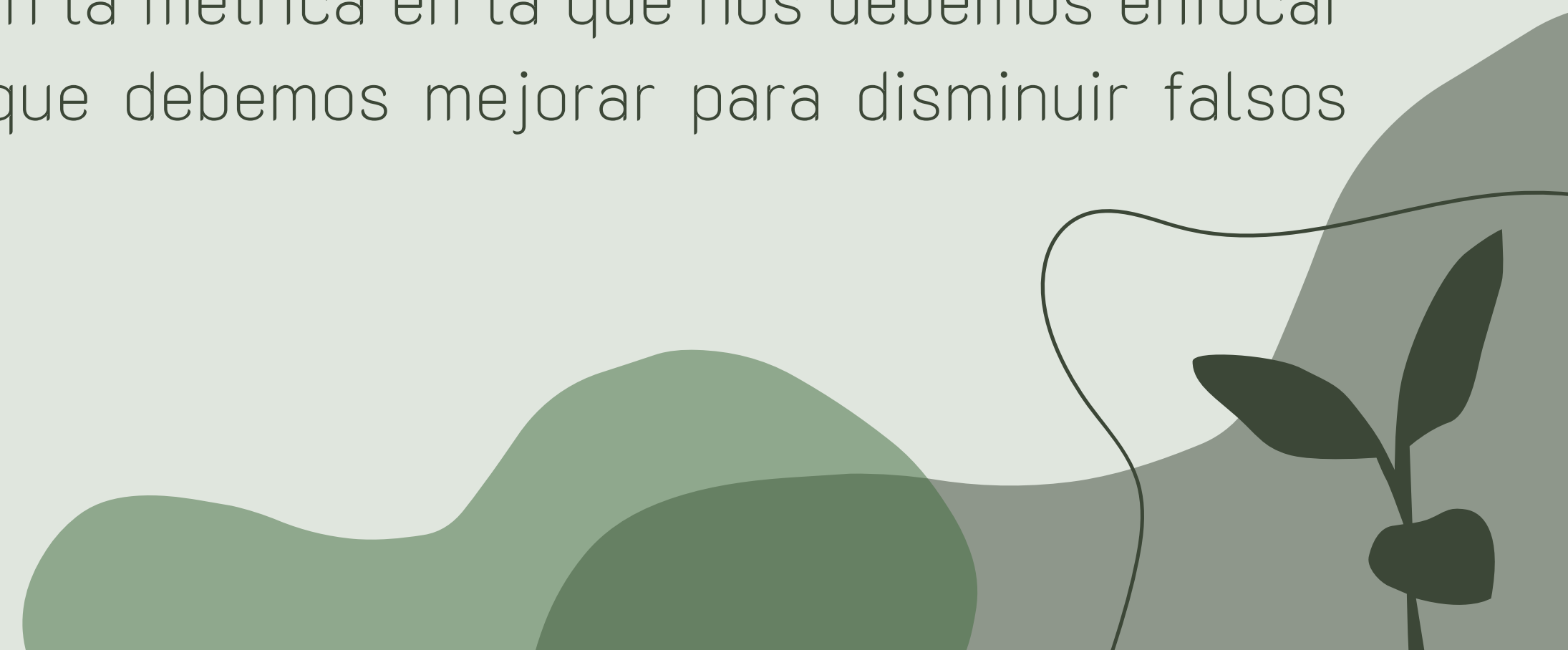




# ERRORES MÁS CRÍTICOS

En este caso específico, son más críticos los errores tipo II (falsos negativos). Ya que no se tomarían los controles necesarios, a tiempo, para evitar un derrame.

Según esto las métricas en la métrica en la que nos debemos enfocar es recall. Es la métrica que debemos mejorar para disminuir falsos negativos.



# RESUMEN DE MÉTRICAS.

Models	Accuracy	Recall	Precision	f1
Decision tree	0.938185	0.0375	0.600000	0.070588
KNN	0.937402	0.0000	0.000000	0.000000
Logistic regression	0.938185	0.0125	1.000000	0.024691
Bagged trees	0.938185	0.0250	0.666667	0.048193
Random Forest	0.938185	0.0250	0.666667	0.048193



## CONCLUSIÓN

En todos los modelos la métrica 'recall' es bastante ineficiente, lo que quiere decir que todos los modelos tienen problemas para predecir los verdaderos positivos (pacientes que efectivamente tuvieron un derrame), esto debido a que el dataset está bastante desbalanceado, lo que hace que la métrica 'accuracy' sea engañosa.

'Mejor modelo considerando lo anterior': árbol de decisión de clasificación.