Reporte de Clasificación: Predicción de Tratamiento de Salud Mental en el Trabajo

1. Descripción de la Base de Datos

La base de datos utilizada corresponde a una encuesta aplicada en el año 2014 relacionada con la salud mental en el ámbito laboral del sector tecnológico. Contiene respuestas de personas trabajadoras que fueron consultadas respecto a: situaciones personales (edad, género, historia familiar), ambiente laboral (beneficios, políticas de ayuda, percepción del empleador) y tratamiento recibido o no por salud mental.

Variable objetivo (target): ***treatment*** – indica si la persona ha recibido tratamiento por un problema de salud mental (Yes o No).

2. Preprocesamiento de los Datos

Para asegurar la calidad del análisis y el entrenamiento de los modelos, se realizaron los siguientes pasos de preprocesamiento:

- **2.1 Eliminación de columnas irrelevantes:**
  - *Timestamp, state, Country, comments*
- **2.2 Limpieza de datos inconsistentes:**
  - Se eliminaron registros con edades menores a 16 o mayores a 100 años
- **2.3 Normalización del género:**
  - Se unificaron múltiples formas de identificar *'Male', 'Female'* y otros como 'Other'
- **2.4 Imputación de valores faltantes:**
  - Se rellenaron valores nulos de *'self_employed*' con la moda y *'work_interfere'* con 'Don't know'
- **2.5 Codificación categórica:**
  - Se transformaron las variables categóricas a formato numérico

3. Modelos Comparados

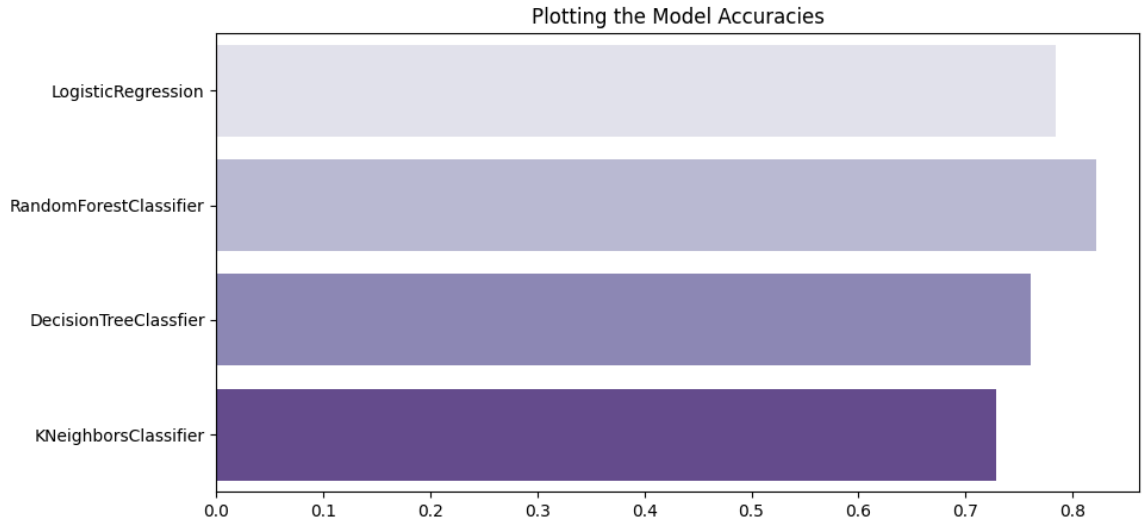
Se entrenaron y evaluaron los siguientes modelos de clasificación con el objetivo de predecir la variable *'treatment':*

- **Logistic Regression:** Clasificador lineal binario.
- **Random Forest:** Ensamble de árboles de decisión.
- **Decision Tree:** Árbol único de decisiones.
- **K-Nearest Neighbors:** Clasificación basada en cercanía de características.

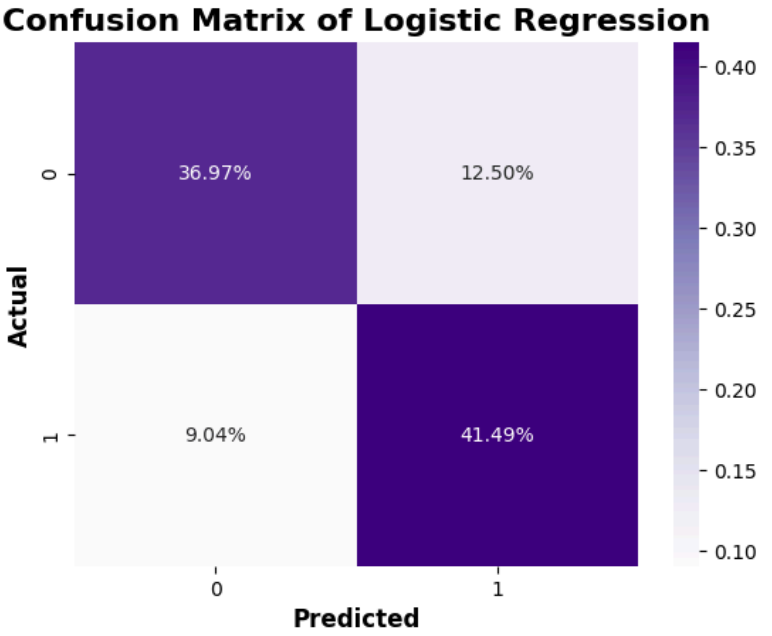
4. Resultados Obtenidos

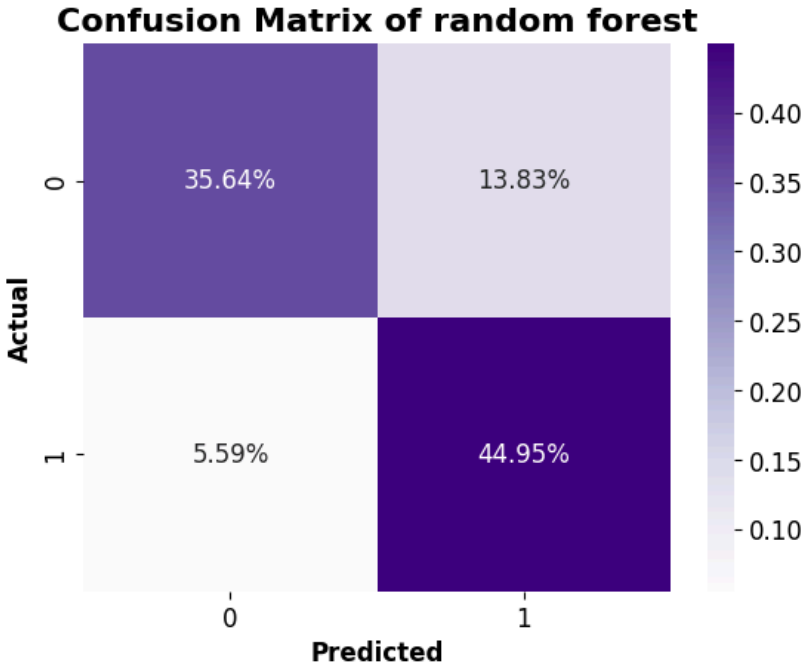
Métrica utilizada: *accuracy_score*

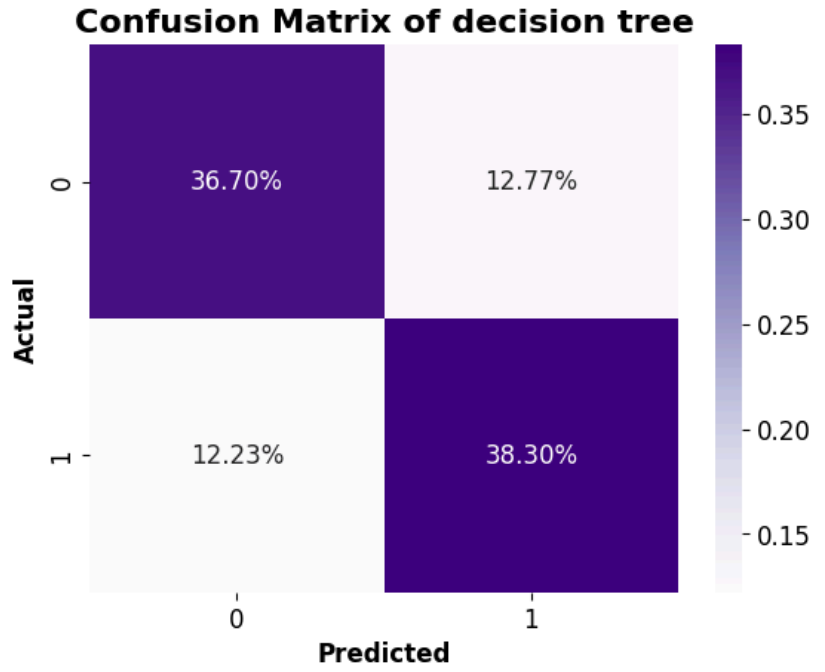
Modelo	Accuracy
Random Forest	82.18%
Logistic Regression	78.46%
Decision Tree	76.06%
K-Nearest Neighbors	72.87%

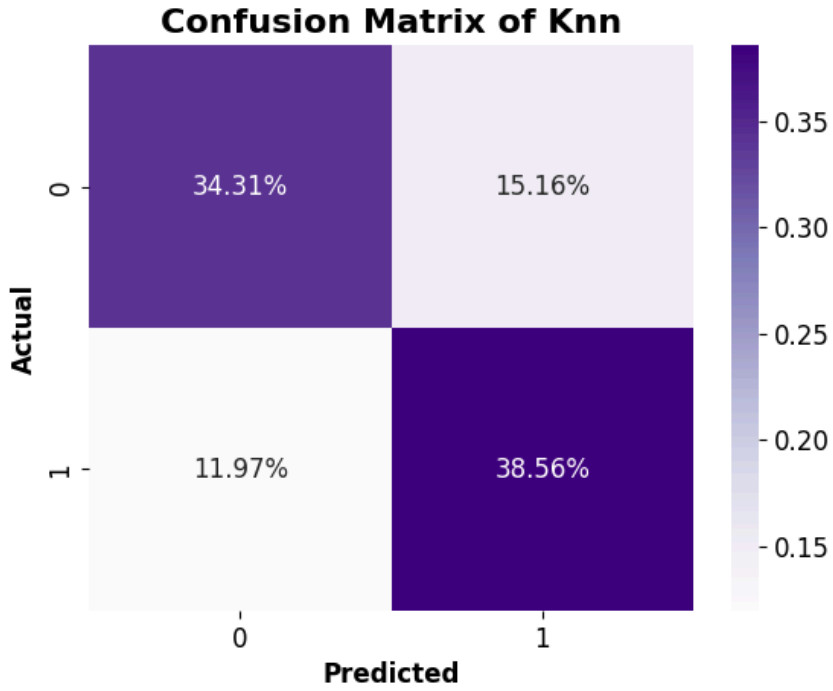


Matrices de confusión:





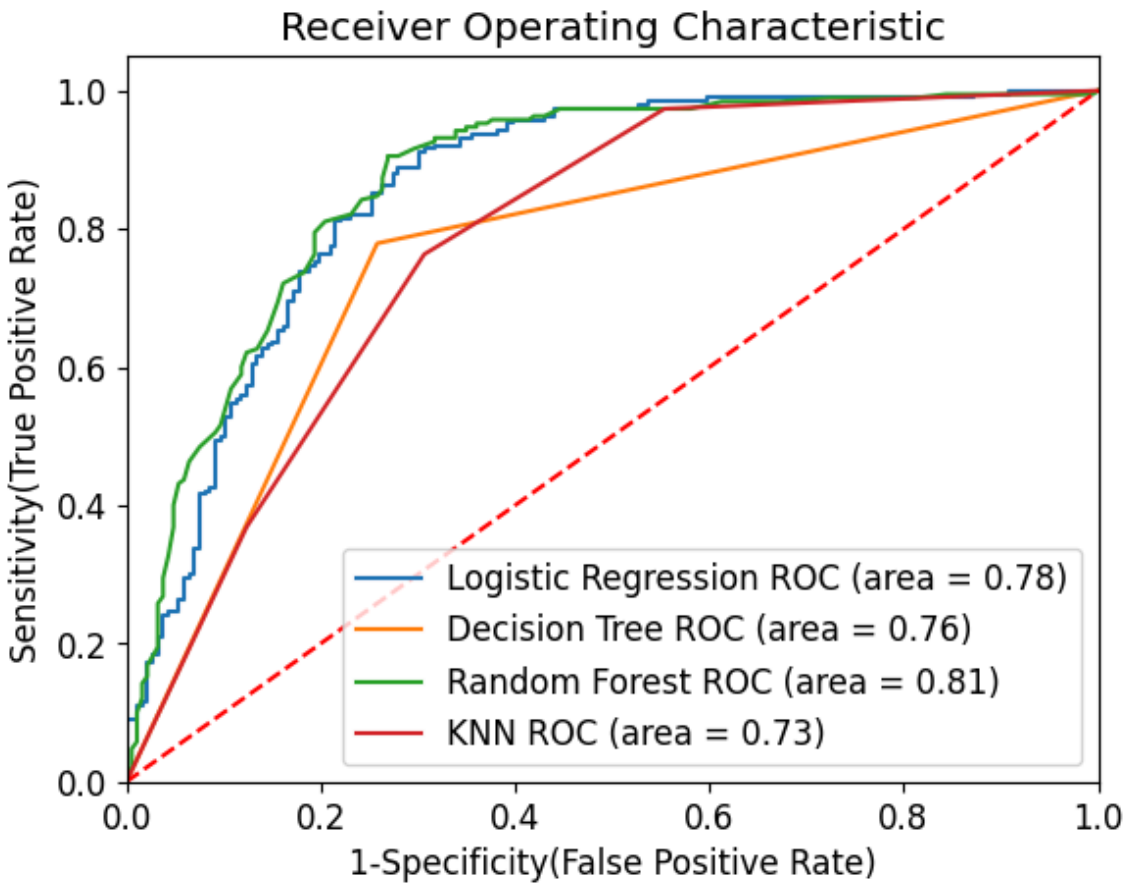




5. Análisis y Comparación

- **Random Forest** obtuvo la mayor precisión general, aunque con más falsos negativos que Logistic Regression.

- **Logistic Regression** presentó mejor capacidad para detectar correctamente los casos positivos (menor cantidad de falsos negativos).
- **Decision Tree y KNN** mostraron resultados más bajos, con mayor tendencia a clasificar incorrectamente ambos tipos de clases.
- **KNN** fue el modelo más débil, posiblemente afectado por la escala de datos o la elección de k.



6. Conclusión

El análisis muestra que el modelo Random Forest Classifier es el más efectivo para este conjunto de datos, logrando una excelente precisión general. No obstante, si el objetivo principal es minimizar los casos positivos no detectados, entonces el modelo Logistic Regression es más adecuado.

En general, se recomienda:

- Usar Random Forest como modelo de producción inicial.
- Considerar Logistic Regression si se requiere mayor sensibilidad.
- Explorar ajustes de hiperparámetros, validación cruzada y balanceo de clases para mejorar el rendimiento.