CODER HOUSE
Curso DataScience
Comisión 32835
Autora Inés Llambías

# VENTA DE SEGUROS DE VIAJERO

## Podremos predecir la próxima venta?

# Índice

# Resumen

Un tour operador y agencia de viajes ofrece a sus clientes un paquete de seguros de viaje.

Basándose en el historial de su base de datos, la empresa necesita saber qué clientes estarían interesados en comprarlo.

Los datos se tomaron del rendimiento y las ventas del paquete durante un período de tiempo. El seguro se ofreció a algunos clientes.

Objetivo del proyecto

# Determinar si un cliente de paquetes de viajes compraría o no un seguro de viaje.

# Variables del dataset

| | |
|---|---|
| Age | Edad del individuo |
| Employment Type | Tipo de empleo (sector gubernamental o sector privado/autónomo) |
| GraduateOrNot | Indicador de si el individuo es graduado o no |
| AnnualIncome | Ingreso anual del individuo |
| FamilyMembers | Número de miembros en la familia |
| ChronicDiseases | Indicador de si el individuo tiene enfermedades crónicas |
| FrequentFlyer | Indicador de si el individuo es un viajero frecuente |
| EverTravelledAbroad | Indicador de si el individuo ha viajado al extranjero alguna vez |
| TravelInsurance | Indicador de si el individuo tiene seguro de viaje |

# Insights generales

- Vemos que el ds no tiene nulos, se nota que al tomar el dataset directo de un ejercicio de Kaggle, el mismo ya fue trabajado anteriormente.
- El dataset no requiere relleno de nulos, interpolación ni borrado de columnas.
- Solo hay un duplicado que sera eliminado.

## TravelInsurance Rate

**33,72%**

Cuantificamos la tasa de compra el seguro

## Edad Promedio

**29,65**

Indica que la población del dataset es muy joven

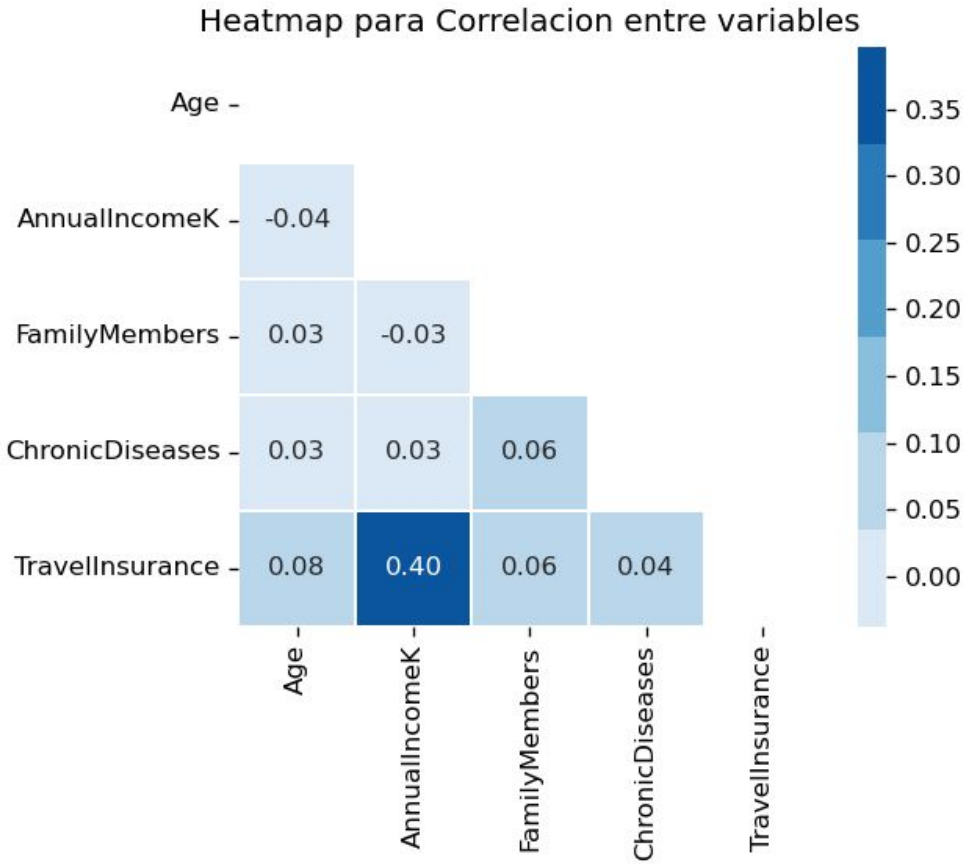
## Familias numerosas

**4,83**

La media de FamilyMembers es elevada

# Análisis de las correlaciones

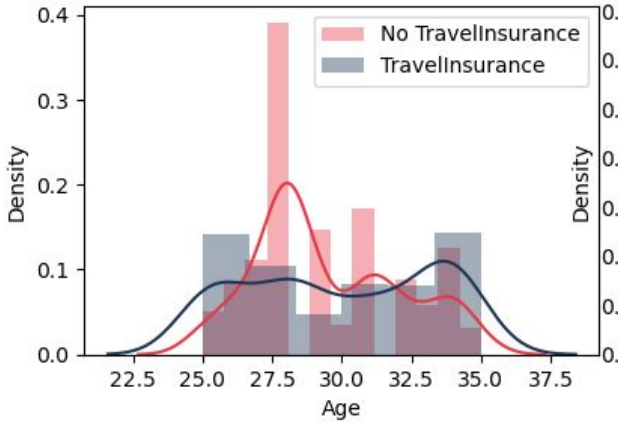
- Se ve que la relacion mas relevante y positiva es entre AnnualIncomeK Y TravelInsurance (0.40).
- En segundo lugar se detecta una relacion muy suve y negativa entre AnnualIncomeK y Age.

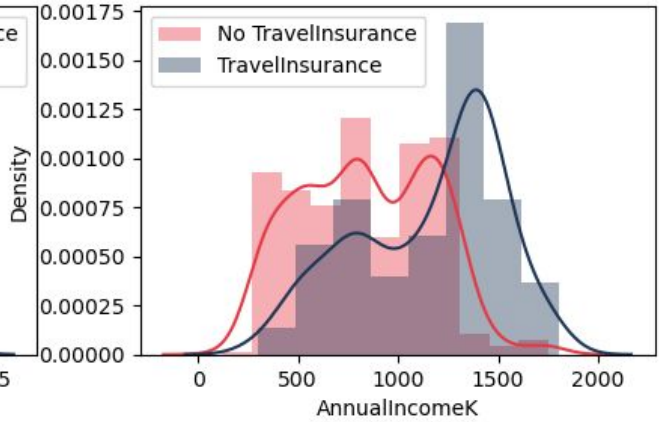


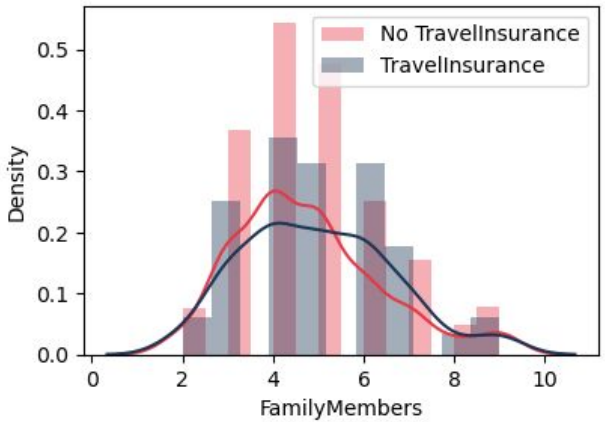
Considerando las variables numericas y el valor target, los graficos  revelan lo siguiente:

* La poblacion de 27 años tiene poca propension a comprar un seguro de viaje, mientras que los mayores a 30, son sutilmente mas propensos.

* La contratacion de seguro crece sustancialmente ente la poblacion que gana mas de 1,2M, mienrtas que disminuye para los que cobran 1.8M.



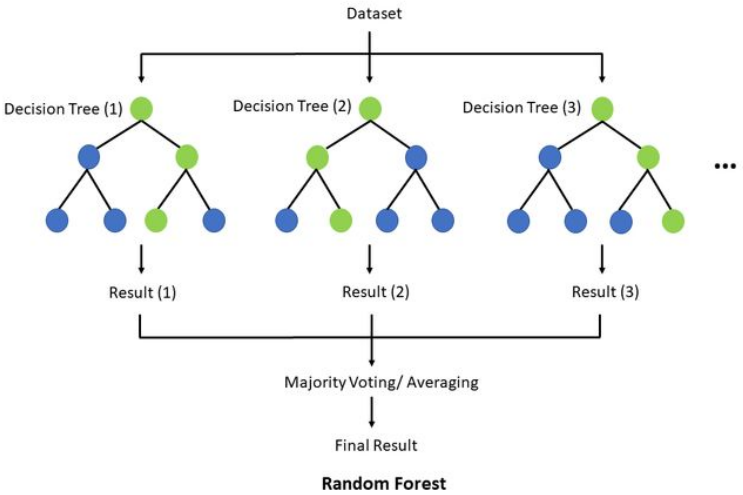



# Algoritmos de clasificación

Para el analisis que vamos a realizar ya contamos con la variable objetivo, que es la variable booleana "TravelInsurance" y esa es la respuesta a lo que queremos saber.

De este modo, entendemos que se trata de un problema de clasificacion binaria.

01 | Arbol de decisión
02 | Regresión logistica
03 | Random Forest



# Arbol de desición

## Presicion Rate positivos

**88.23%**

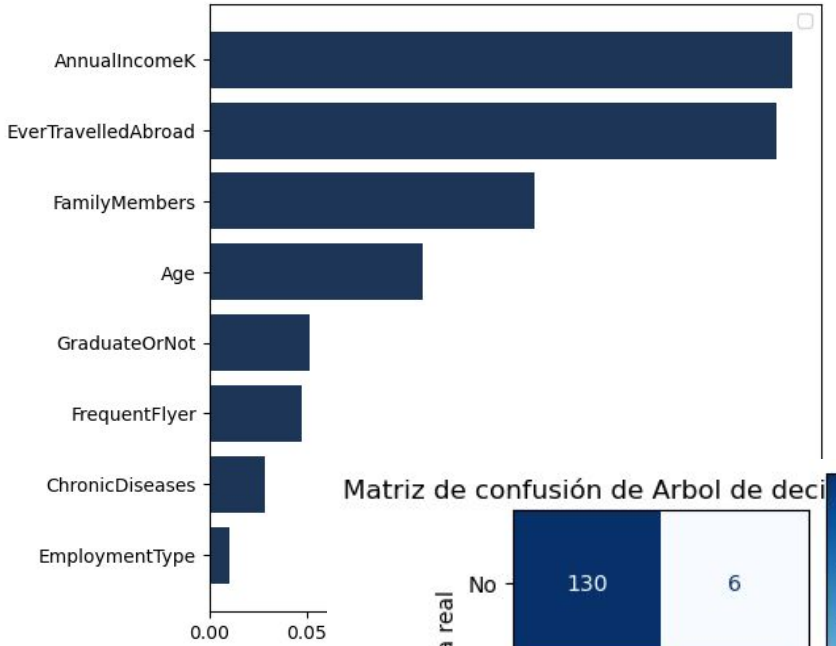
El modelo clasifica correctamente alrededor del 88.78% de las instancias positivas

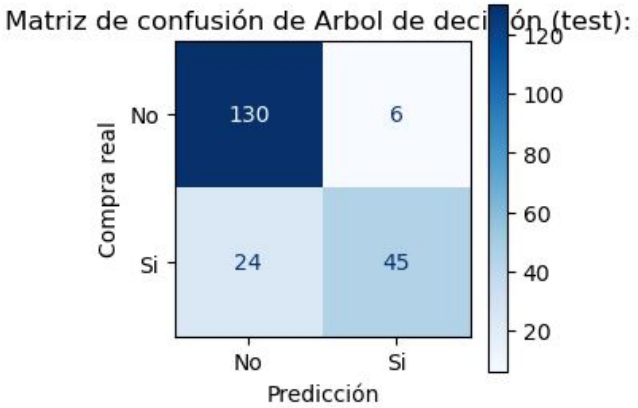
## Accuracy

**85.37%**

El modelo clasifica correctamente alrededor del 85.37% de todas las instancias

Feature Importance



# Regresión logistica

## Presicion Rate positivos

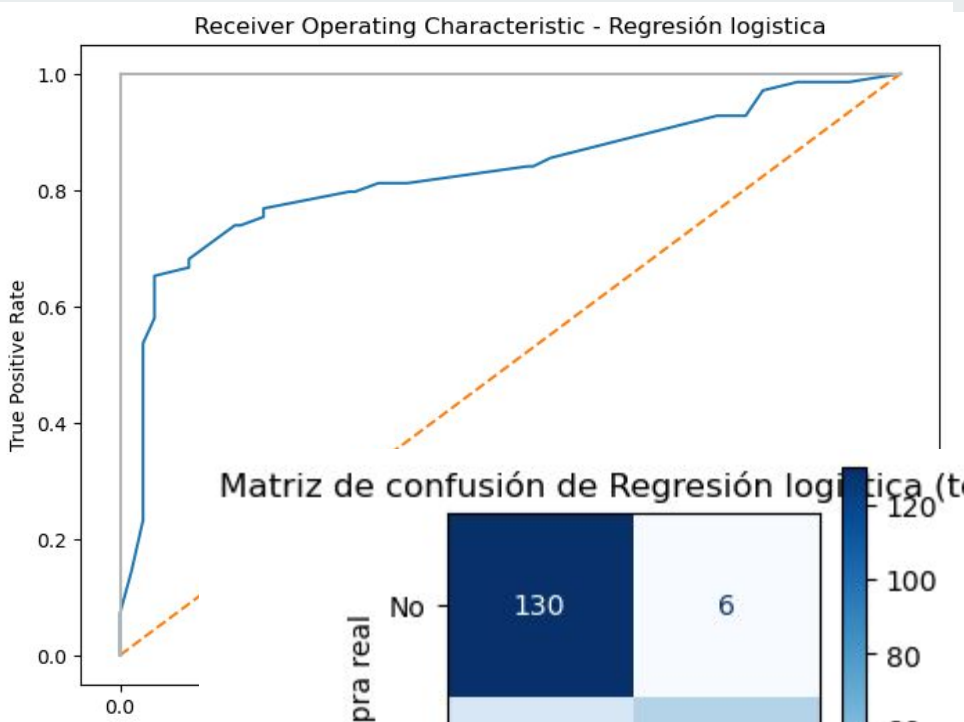
**88.00%**

El modelo clasifica correctamente alrededor del 88.78% de las instancias positivas

## Accuracy

**84.87%**

El modelo clasifica correctamente alrededor del 85.37% de todas las instancias



# Random Forest

## Presicion Rate positivos

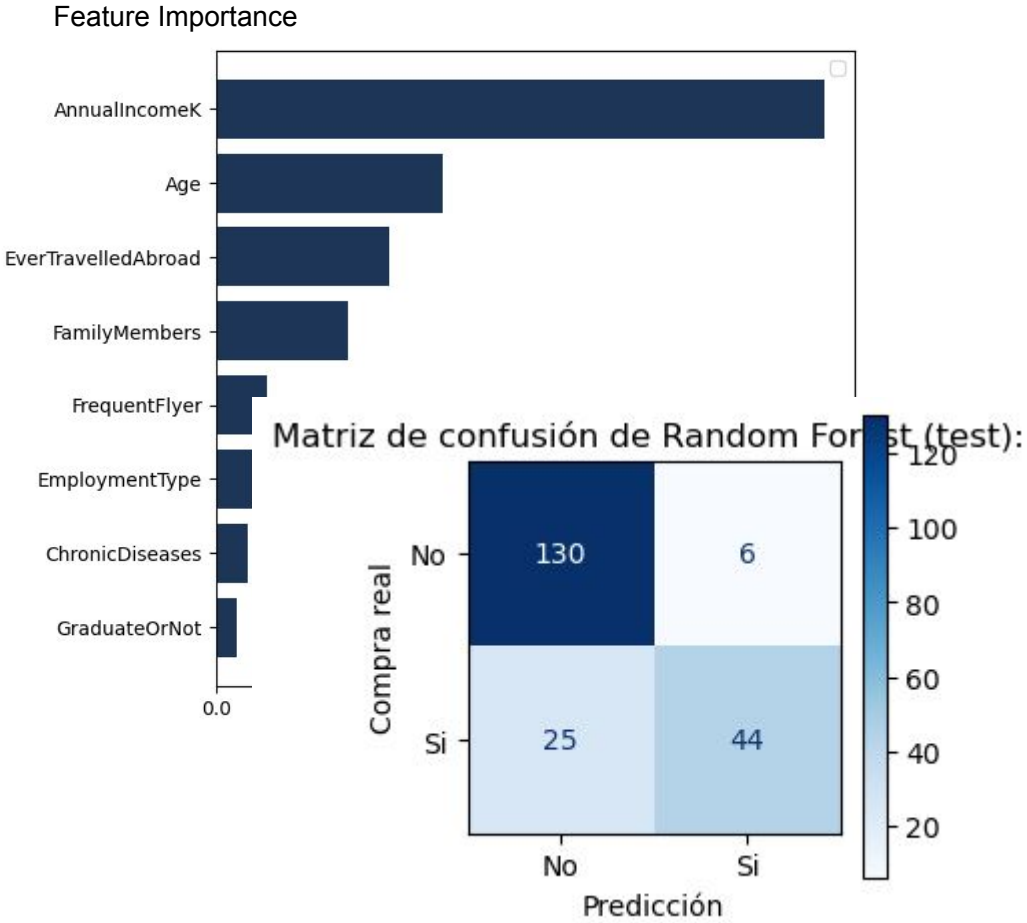
**88.00%**

El modelo clasifica correctamente alrededor del 88.78% de las instancias positivas

## Accuracy

**84.87%**

El modelo clasifica correctamente alrededor del 85.37% de todas las instancias



# Metricas

| Modelo | Precision (test) | Precision (train) | Accuracy (test) | Accuracy (train) | Recall (test) | Recall (train) | F1 (test) | F1 (train) |
|---|---|---|---|---|---|---|---|---|
| Arbol de decisión | 720.000 | 926.606 | 770.732 | 857.442 | 521.739 | 627.329 | 605.042 | 748.148 |
| Regresión logistica | 719.298 | 737.374 | 785.366 | 761.006 | 594.203 | 453.416 | 650.794 | 561.538 |
| Random Forest | 880.000 | 978.723 | 848.780 | 851.153 | 637.681 | 571.429 | 739.496 | 721.569 |

# Tiempos de ejecución

Con el fin de realizar una comparación entre los tiempos que demanda la utilización de cada uno de los algoritmos es que se realiza un cuadro comparativo de los tiempos

**0,0082**

Arbol de decisión

**10,6962**

Regresión logistica

**0,0796**

Random Forest

# Conclusiones

Si nos centramos en la métrica de precisión en el conjunto de prueba (precision_test), el mejor modelo sería Random Forest con una precisión del 88.00%. Esto significa que el 88.00% de las instancias clasificadas como positivas por el modelo realmente pertenecen a la clase positiva.

Sin embargo, si consideramos otras métricas, como el recall_test o el f1_test, el modelo de Random Forest puede no ser el mejor. Por ejemplo, el modelo de Regresión Logística tiene un recall_test del 59.42%, lo que indica que identifica un mayor porcentaje de instancias reales de la clase positiva en comparación con Random Forest.

El criterio para determinar el mejor modelo depende del contexto del problema y las necesidades específicas. Si se prioriza la precisión en la clasificación de la clase positiva, entonces Random Forest sería la elección. Si se busca un mejor equilibrio entre precisión y recall, el modelo de Regresión Logística podría ser preferible. Siempre es recomendable considerar múltiples métricas y realizar análisis adicionales para tomar la decisión más adecuada para el problema en cuestión.

# Responsable

**Ines Llambias**

illambi@gmail.com

https://www.linkedin.com/in/inesllambias/

# Gracias