CODER HOUSE

SMOKING DETECTION

Proyecto final curso Data Science - Comisión 19145

Fecha de entrega 15/08/2022

Integrantes:

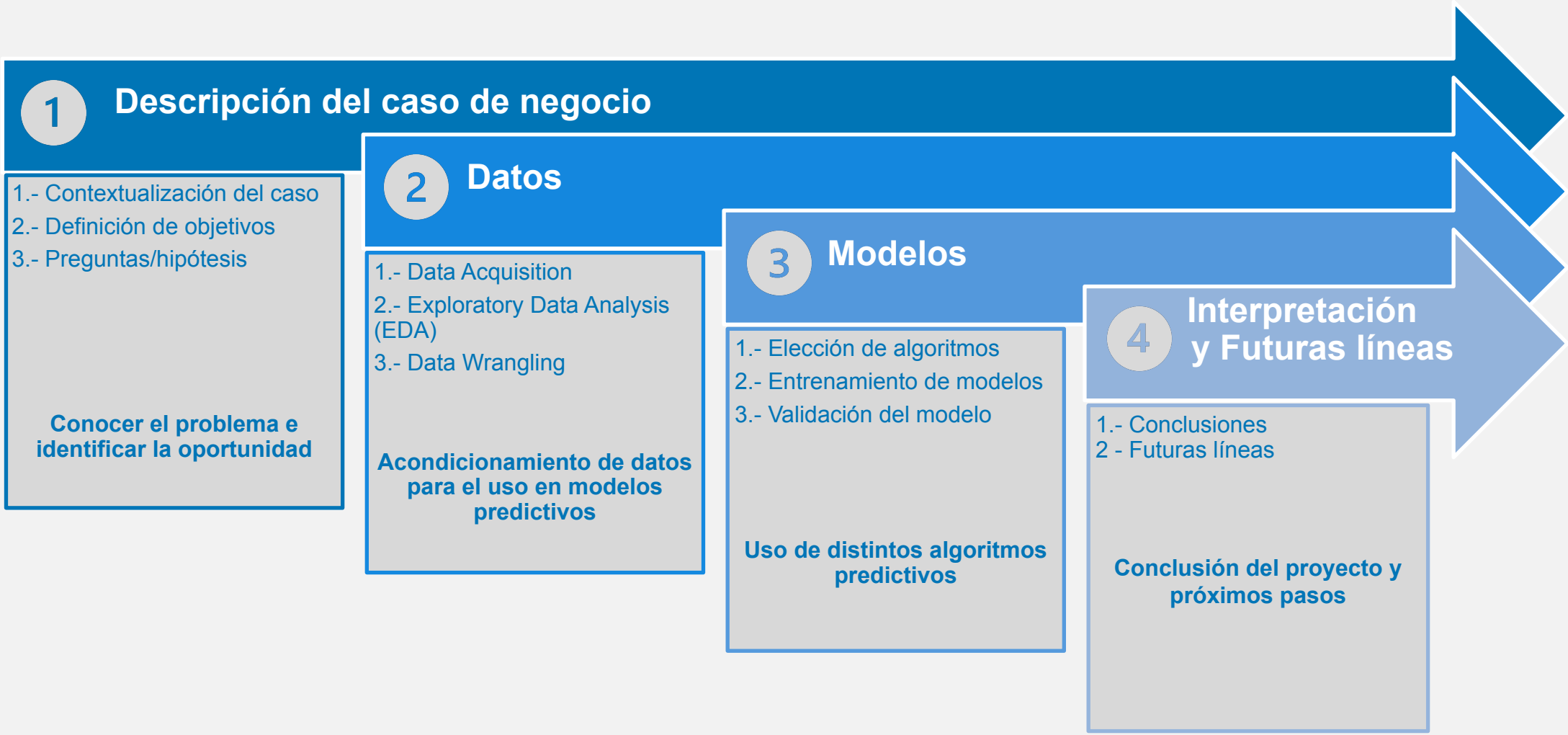
Pablo Perera

Leonardo Rosa

Christian Aldana

Eduardo Gonzalez

CONTENIDOS


1
Descripción del caso de negocio
1.- Contextualización del caso
2.- Definición de objetivos
3.- Preguntas/hipótesis
Conocer el problema e identificar la oportunidad
2
Datos
1.- Data Acquisition
2.- Exploratory Data Analysis (EDA)
3.- Data Wrangling
Acondicionamiento de datos para el uso en modelos predictivos
3
Modelos
1.- Elección de algoritmos
2.- Entrenamiento de modelos
3.- Validación del modelo
Uso de distintos algoritmos predictivos
4
Interpretación y Futuras líneas
1.- Conclusiones
2 - Futuras líneas
Conclusión del proyecto y próximos pasos

CODER HOUSE

1. Descripción del caso de negocio

Dedicación y problema abordado

CONTEXTUALIZACIÓN

Información general del tabaco y sus consecuencias negativas

Estadísticas de consumo:

- 1.300 millones de consumidores de tabaco en el mundo
- El costo económico total del tabaquismo a nivel mundial (derivado de los gastos en salud y las pérdidas de productividad asociados), es estimado a $ 1,4 billones de dólares
- Las compañías de productos de tabaco gastan al año más de 8.000 millones de dólares en mercadeo y publicidad
- Los niños y adolescentes que utilizan cigarrillos electrónicos tiene al menos el doble de probabilidades de fumar cigarrillos más tarde en su vida

Consecuencias en la salud:

- La mortalidad atribuible al consumo de tabaco en la Región representa el 15% de las defunciones por enfermedades cardiovasculares, el 24% por cáncer y 45% por enfermedades crónicas respiratorias.
- El tabaco mata a 8 millones de personas cada año
- La esperanza de vida de los fumadores es al menos 10 años menor que la de los no fumadores.

Fuente: OPS, Organización Panamericana de la Salud

CONTEXTUALIZACIÓN

Oportunidad para mejorar las consecuencias negativas del tabaco en la sociedad

Algunas estrategias para disminuir el consumo del tabaco:

- Monitor: vigilar el consumo de tabaco
- Protect: proteger a la población del humo de tabaco
- Offer: ofrecer ayuda para dejar de fumar
- Warn: advertir de los peligros del tabaco
- Enforce: hacer cumplir las prohibiciones sobre publicidad, promoción y patrocinio
- Raise: aumentar los impuestos al tabaco

Algunas ventajas con la disminución del consumo de tabaco:

- Mejora el estado de salud y la calidad de vida.
- Reduce el riesgo de muerte prematura y puede aumentar en hasta 10 años la expectativa de vida.
- Reduce el riesgo en cuanto a muchos efectos adversos de salud, los cuales incluyen malos resultados en la salud reproductiva, enfermedades cardiovasculares, enfermedad pulmonar obstructiva crónica (epoc) y cáncer.
- Beneficia la salud de las mujeres embarazadas, el feto y el bebé.

Fuente: OPS, Organización Panamericana de la Salud

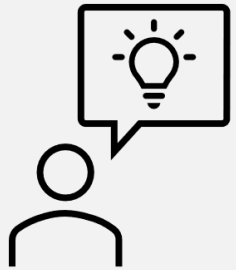
Fuente: Centers for Disease Control and Prevention

CONTEXTUALIZACIÓN

Propuesta de solución con analítica predictiva

Descripción del caso:

- Contamos con una base de datos de exámenes de hemograma de pacientes fumadores y con otra base de datos de exámenes no identificados si son o no fumadores. En base a lo anterior aplicaremos un proceso completo para aplicar Machine Learning y así predecir quienes son fumadores.



Objetivo principal:

- Optimizar el esfuerzo de campañas anti-tabaco en la sociedad mediante el análisis predictivo de fumadores por medio de exámenes típicos de sangre y características físicas de un paciente.

Utilidades futuras:

- Predecir a los fumadores con datos proporcionados por las instituciones de salud y así en conjunto a los gobiernos enfocar esfuerzos para combatir el consumo en la sociedad.
- Los beneficios finalmente se verán reflejados en la disminución de costos en salud, esfuerzos anti-tabaco y mejoramiento en la calidad de vida.

ETAPAS DEL DESARROLLO

Hipótesis y preguntas asociadas

Preguntas:

- ¿Con qué precisión podemos determinar si una persona es fumadora?
- ¿Es factible poder predecir a personas fumadoras a partir de exámenes hematológicos?
- ¿Qué características del examen hematológico no aportan datos significativos para predecir si es o no fumador una persona?
- ¿Cuáles son las variables principales para los análisis predictivos con este tipo de exámenes?

Hipótesis:

- Las características físicas como el peso, la altura, la visión, la edad o la audición de una persona influyen directamente en si es fumador o no.
- Con exámenes hematológicos podemos predecir si una persona es fumadora o no
- Una persona con caries es muy probable de que sea fumadora
- Los fumadores tienen una alta presión arterial

CODER HOUSE

2. DATOS

ACONDICIONAMIENTO PARA EL USO EN MODELOS DE CLASIFICACIÓN

Descripción del Data Set

ID : Indice

gender : Género (M/F)

age : Edad, categórica (rangos de 5 años)

height(cm) : Altura

weight(kg) : Peso

waist(cm) : Diámetro de circunferencia del abdomen

eyesight(left) : Visión izquierda

eyesight(right) : Visión derecha

hearing(left) : Audición izquierda

hearing(right) : Audición derecha

systolic : Presión arterial

relaxation : Pulso

fasting blood sugar : Azúcar en sangre (glicemia)

Cholesterol : Colesterol

triglyceride : Triglicéridos

HDL : Tipo de colesterol HDL

LDL : Tipo de colesterol LDL

hemoglobin : Hemoglobina

Urine protein : Proteína en orina

serum creatinine : Suero de creatinina

AST : Tipo transaminasa glutámico oxalacético AST

ALT : Tipo transaminasa glutámico oxalacético ALT

Gtp : Trifosfato de guanosina

oral : Tiene examinación oral

dental caries : Tiene caries

tartar : Tiene sarro

Variable target:

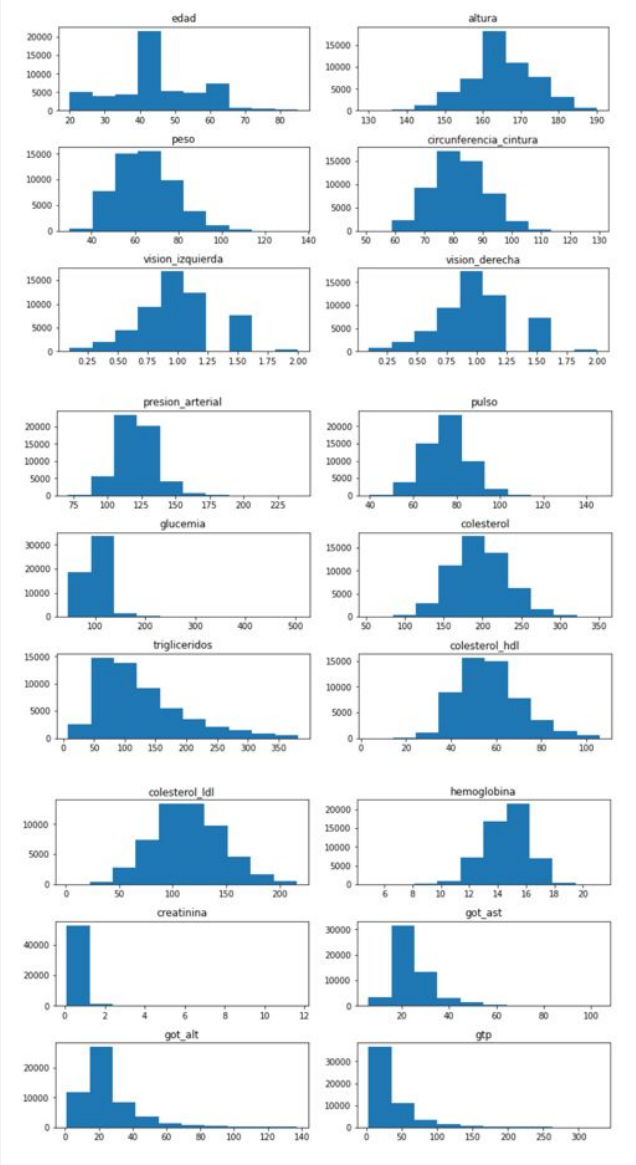
smoking : Es fumador.

ETAPAS DEL DESARROLLO

Exploratory Data Analysis (EDA)

Objetivos de la investigación:

Para el dataset levantado se tiene un punto muy importante a favor, el que considera que no existen datos nulos para cada una de las variables o atributos considerados dentro del análisis. Lo que implica que no tenemos el inconveniente de tomar decisiones respecto a eliminar muestras o variables faltantes y/o realizar estimaciones para su sustitución. Seguidamente realizamos un análisis estadístico de las variables del dataset de manera que tengamos una idea general de la información contenida en los datos. Al realizar lo anterior fue más fácil identificar la distribución que tienen las variables numéricas y cuáles de ellas son las que entregan una mayor información y cuáles no, de manera que se pueda realizar ajustes y eliminando algunas variables de nuestro análisis o datos outliers muy alejados del percentil 99.5 .



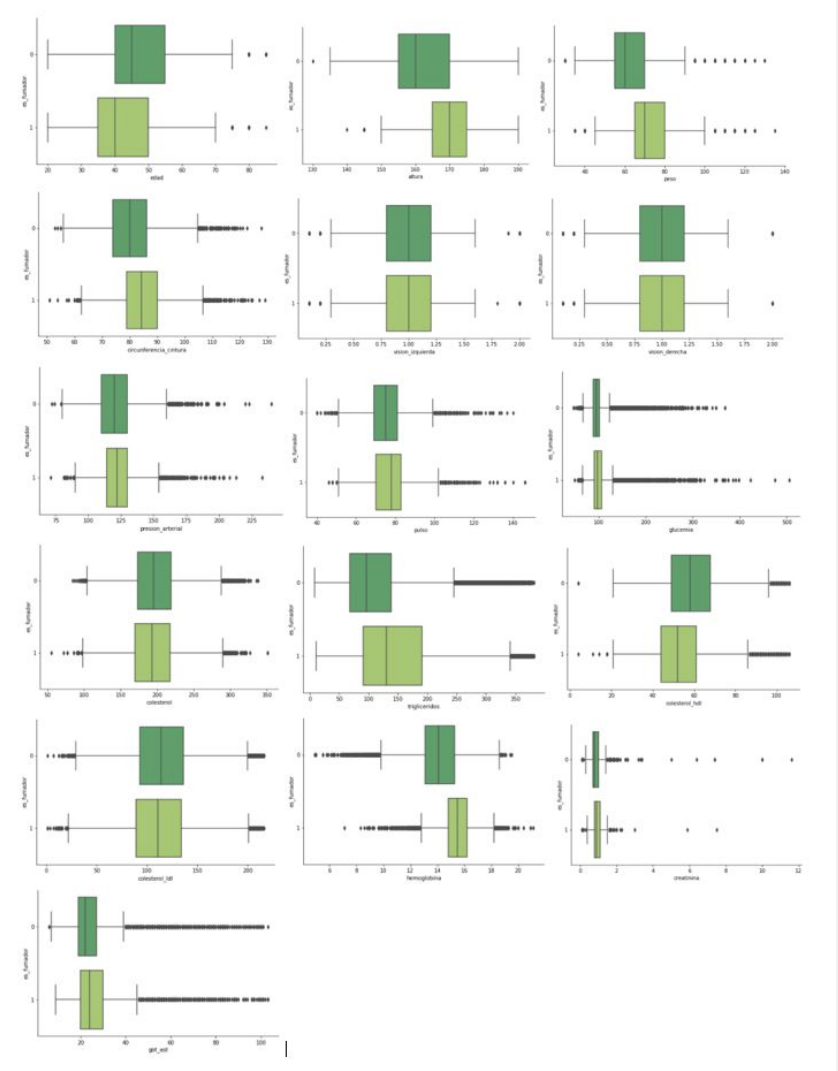
ETAPAS DEL DESARROLLO

Exploratory Data Analysis (EDA)

La variable objetivo:

Como nuestro objetivo apunta a la predicción de los casos para fumadores, nuestro target es la variable “es_fumador” en base a la información de la otras variables que son básicamente las que entregan los exámenes realizados en la institución de salud.

Con lo anterior en mente seguimos el análisis bivariado tomando en cuenta la variable objetivo.



ETAPAS DEL DESARROLLO
Data Wrangling

- Decidimos traducir los nombres de las variables y a algunas asignarle nombres más amigables o descriptivos, también aplicamos formato snake_case.
- Cambiamos valores de Y or N por 0 y 1, para facilitar los procesos y análisis posteriores
- Verificamos con un df.info() que no hubieran valores nulos en nuestro Dataset y efectivamente, no encontramos valores nulos.
- Transformamos el dataset, eliminando aquellas columnas que no son relevantes para el estudio.

```
df.rename(columns = {'ID':'id',
                     'gender':'genero',
                     'age':'edad',
                     'height(cm)':'altura',
```

```
[ ]  # Reemplazo data de dos variables, para que sea 1 o 0.
     df['sarro'] = df['sarro'].map({'Y': 1, 'N': 0})
     df['revision_oral'] = df['revision_oral'].map({'Y': 1, 'N': 0})
```

```
[ ]  df = df.drop(['audicion_izquierda', 'audicion_derecha','proteinuria'], axis = 1)

     num_list = [
                 'edad',
                 'altura',
                 'peso',
                 'circunferencia_cintura',
                 'vision_izquierda',
                 'vision_derecha',
                 'presion_arterial',
                 'pulso',
                 'glucemia',
                 'colesterol',
                 'trigliceridos',
                 'colesterol_hdl',
                 'colesterol_ldl',
                 'hemoglobina',
                 'creatinina',
                 'got_ast',
```

```
df = df.drop(['audicion_izquierda', 'audicion_derecha','proteinuria'], axis = 1

num_list = [
            'edad',
            'altura',
            'peso',
            'circunferencia_cintura',
            'vision_izquierda',
            'vision_derecha',
            'presion_arterial',
            'pulso',
```

CODER HOUSE

3. MODELOS

USO DE DISTINTOS MODELOS PREDICTIVOS

Algoritmos (modelos)

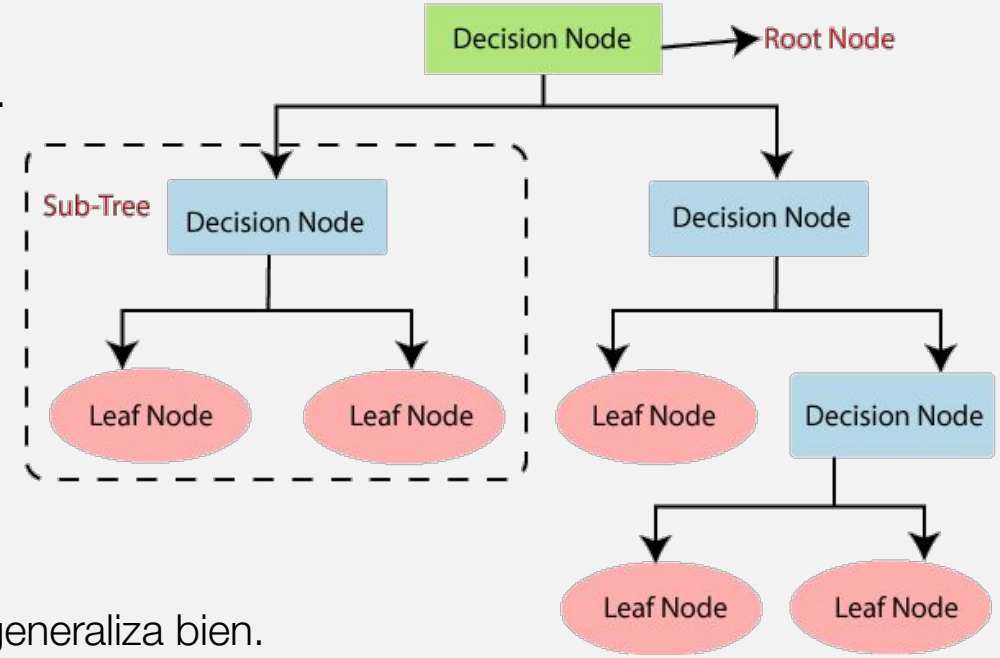
MODELO 1: Decision Tree

Los modelos Decision Tree aprenden generando reglas del tipo if-else. Separan los datos en grupos cada vez más pequeños.

Cada división es un nodo y cuando un nodo no conduce a nuevas divisiones, se genera una hoja.

Los beneficios de este modelo son:

- Es una caja blanca, sus resultados son fáciles de entender e interpretar.
- Relativamente robusto.
- Funciona bien con grandes conjuntos de datos
- Combinaciones de este modelo pueden dar buenos resultados sin perder escalabilidad, por ejemplo el Random Forest.

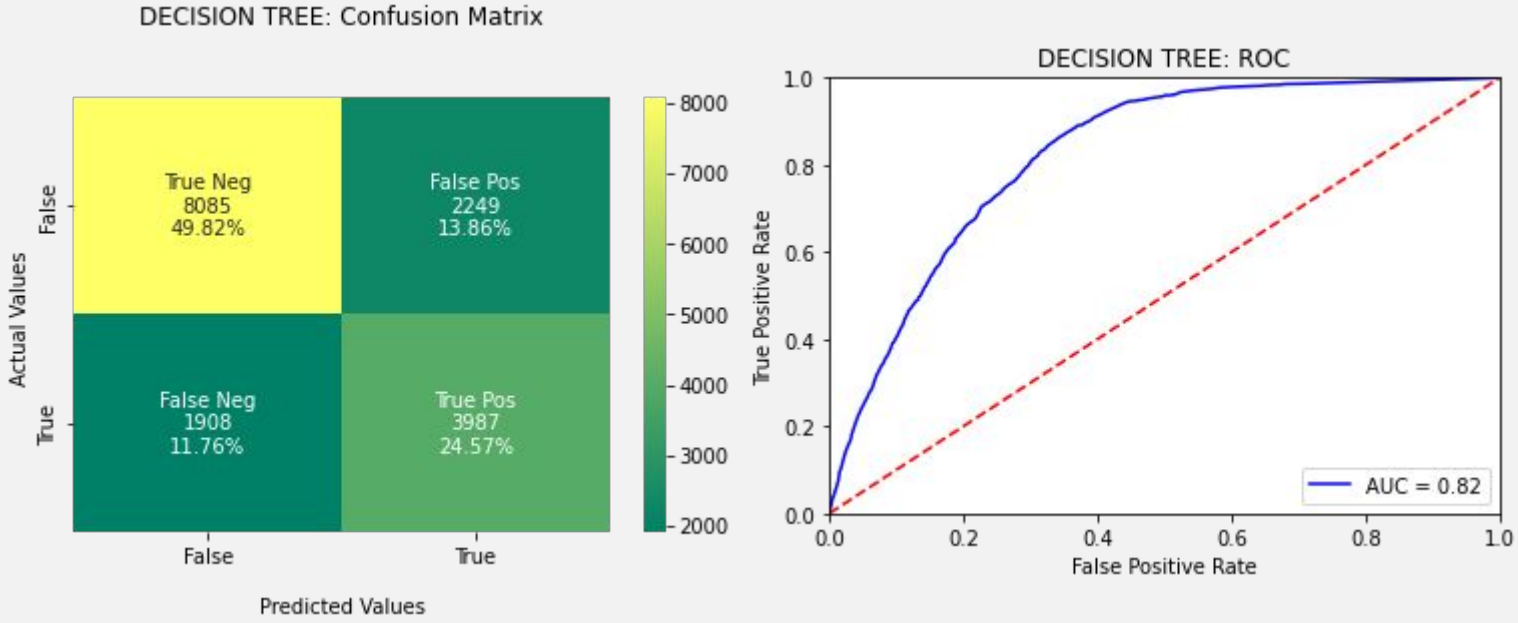


Las contras del modelo son:

- Tiende al overfitting, por lo que el modelo al predecir nuevos datos no generaliza bien.
- Está influenciado por datos outliers, creando árboles con ramas muy profundas que no predicen bien datos nuevos.
- Se puede sesgar si hay desbalance de clases

Algoritmos (modelos)

MODELO 1: Decision Tree Resultados

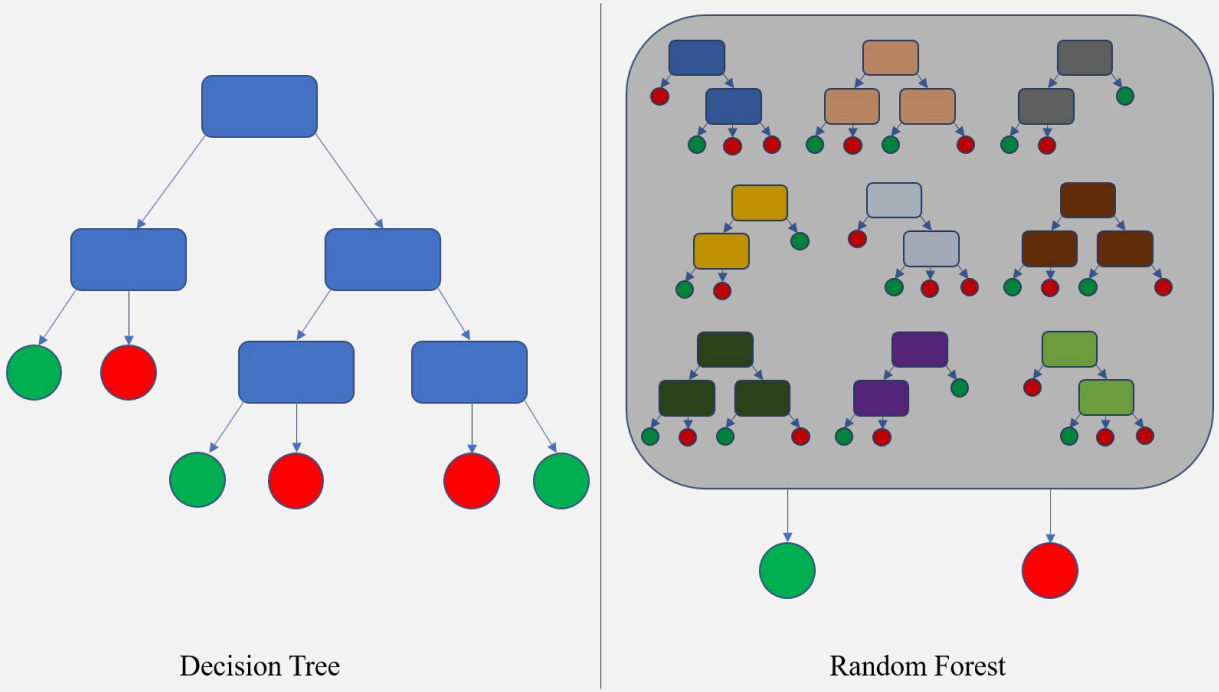



```
Accuracy Decision Tree: 0.743853595415614
Presición Decision Tree: 0.6393521488133419
Recall Decision Tree: 0.6763358778625954
F1 Score Decision Tree: 0.6573242106998599
```

Algoritmos (modelos)

MODELO 2 : Random Forest

Uno de los problemas que se genera con la creación de un árbol de decisión, es que si le damos la profundidad suficiente, el árbol tiende a “memorizar” las soluciones en vez de generalizar el aprendizaje. Es decir, a padecer de overfitting. La solución para evitar esto es la de crear muchos árboles y que trabajen en conjunto.

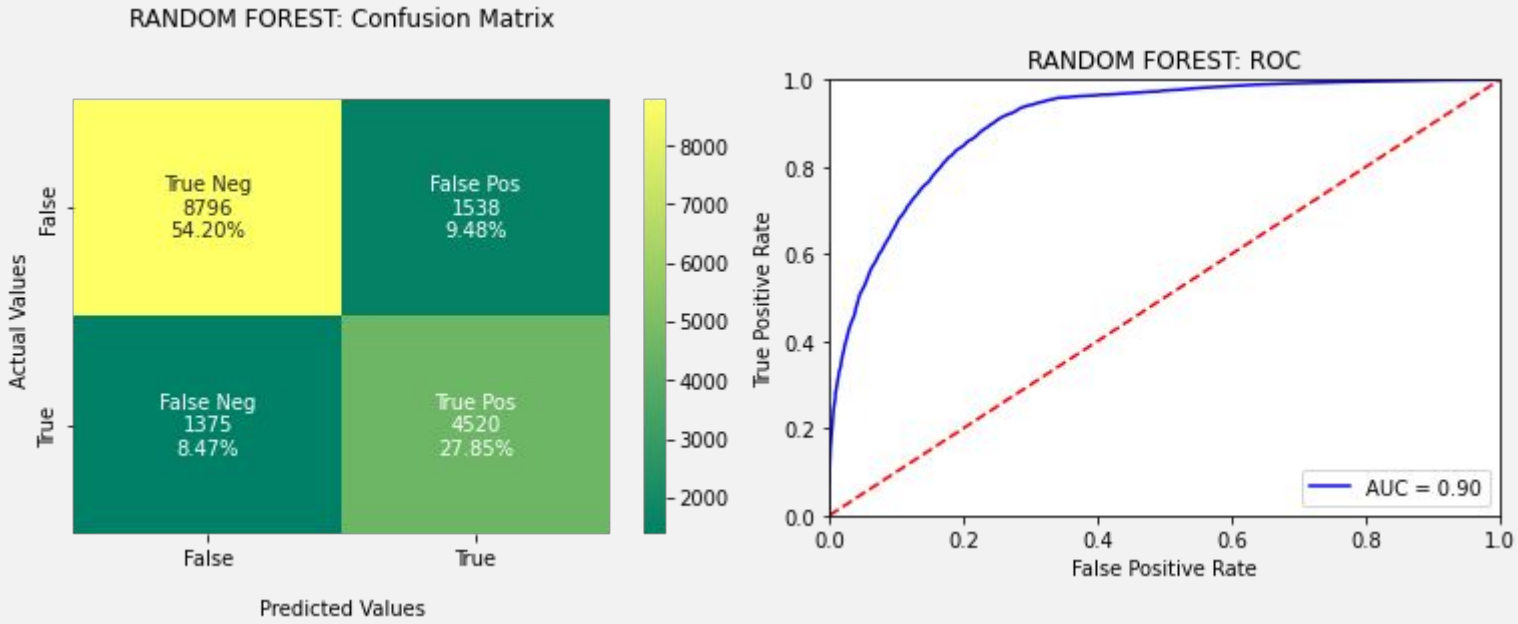


Decision Tree

Random Forest

Algoritmos (modelos)

MODELO 2 : Random Forest Resultados



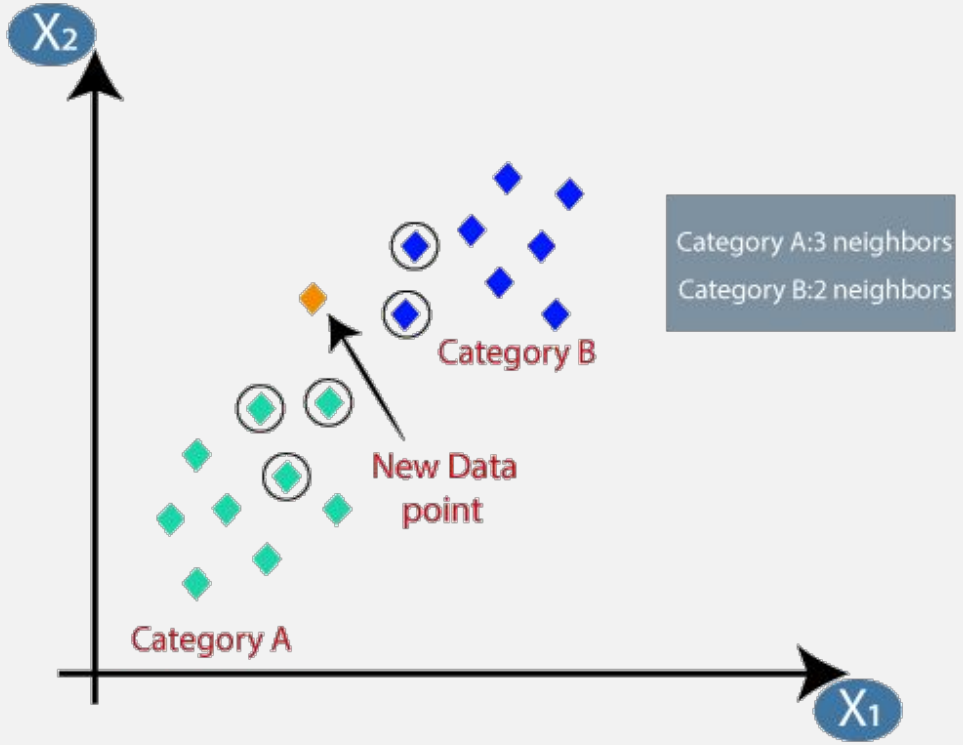

```
Accuracy Random Forest: 0.8205065007086081
Presición Random Forest: 0.7461208319577418
Recall Random Forest: 0.7667514843087362
F1 Score Random Forest: 0.7562954906717978
```

Algoritmos (modelos)

MODELO 3 : KNN

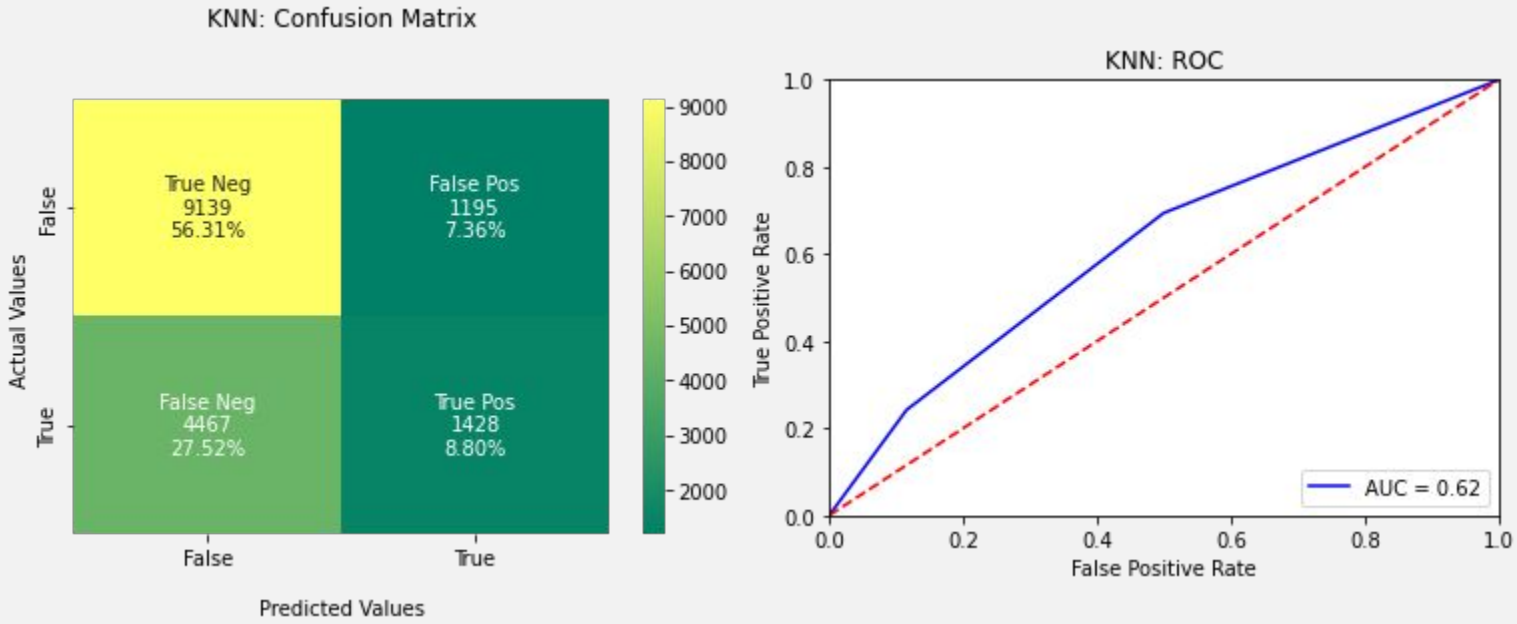
Puede usarse para clasificar nuevas muestras (valores discretos) o para predecir (regresión, valores continuos).

Sirve esencialmente para clasificar valores, buscando los puntos de datos “más similares” (por cercanía).



Algoritmos (modelos)

MODELO 3 : KNN Resultados

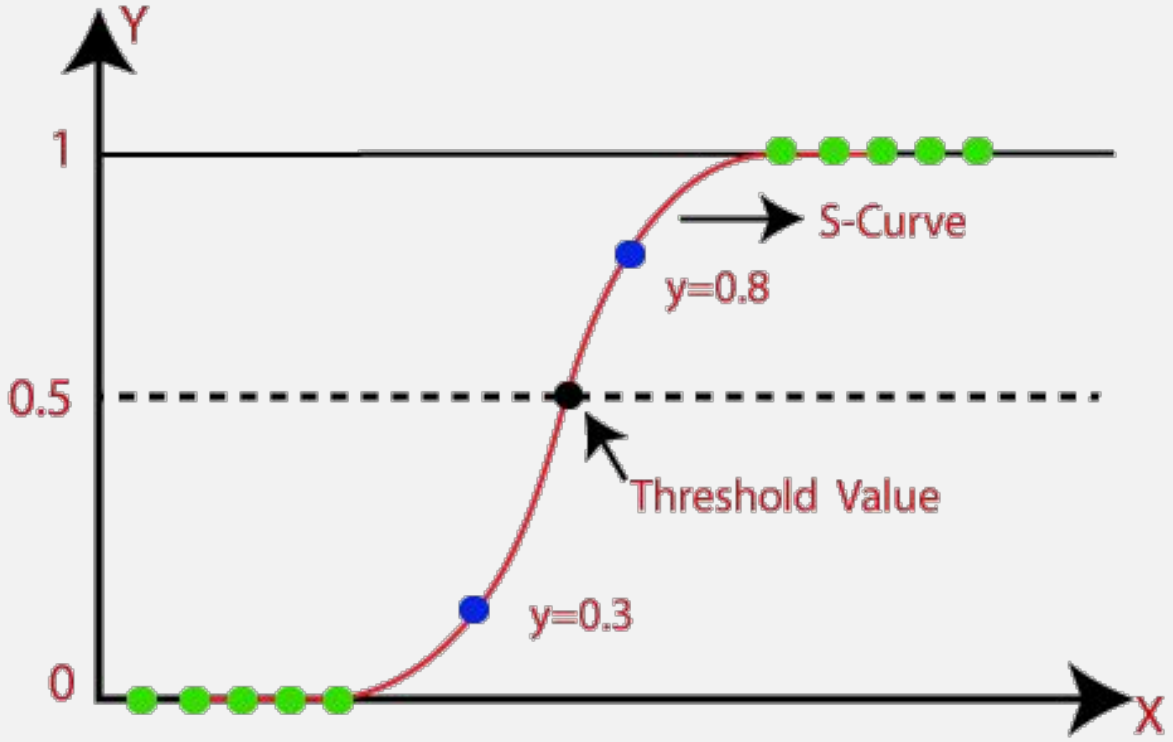



```
Accuracy KNN: 0.6511183683529485
Presición KNN: 0.5444147922226458
Recall KNN: 0.24223918575063613
F1 Score KNN: 0.33528997417234085
```

Algoritmos (modelos)

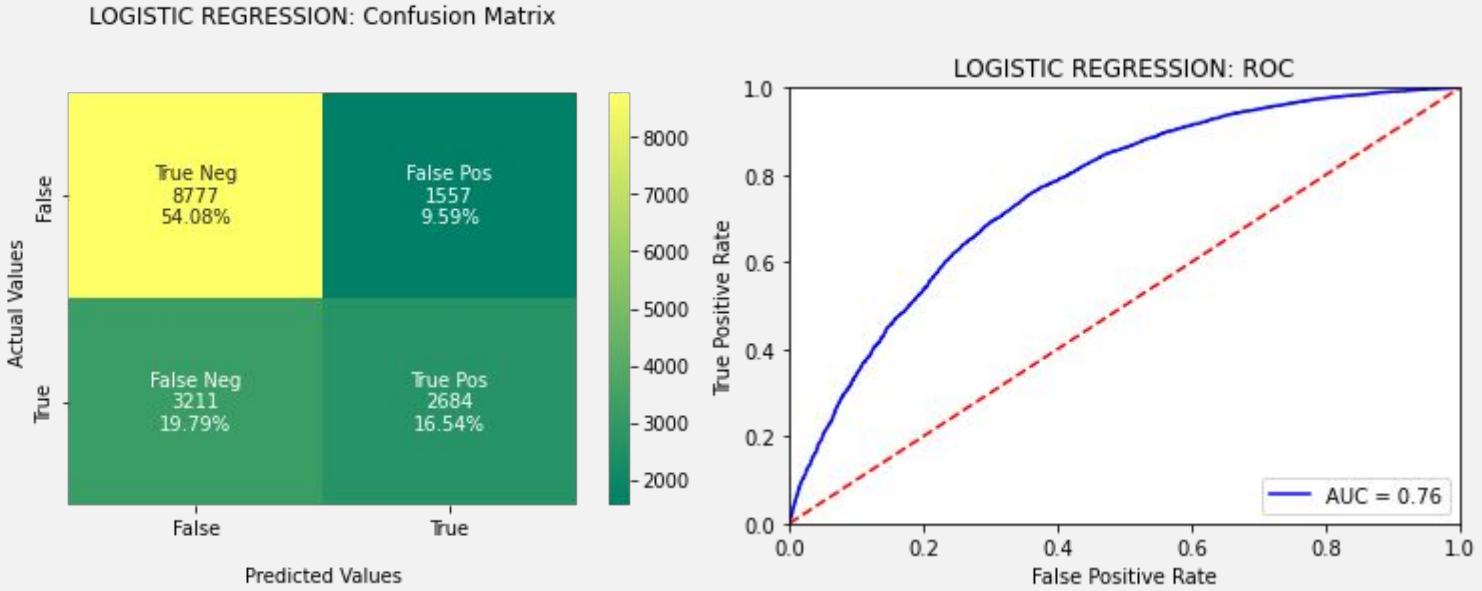
MODELO 4 : Logistic Regression

Se trata de una técnica de aprendizaje automático que proviene del campo de la estadística. A pesar de su nombre, no es un algoritmo, sino que es un método para problemas de clasificación, en los que se obtienen un valor binario entre 0 y 1.



Algoritmos (modelos)

MODELO 4 : Logistic Regression Resultados

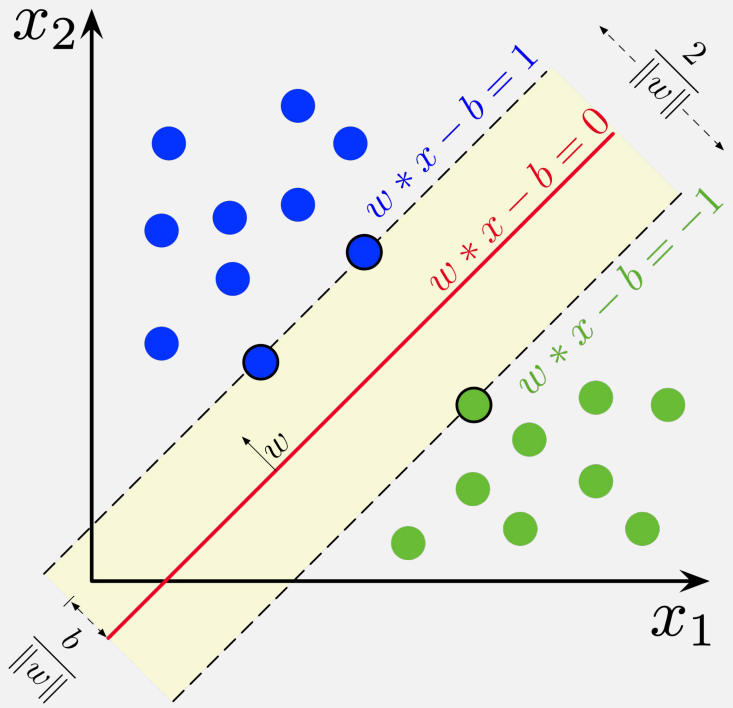



```
Accuracy Regresión Logística: 0.706204941770904
Presición Regresión Logística: 0.6328696062249469
Recall Regresión Logística: 0.4553011026293469
F1 Score Regresión Logística: 0.5295974743488556
```

Algoritmos (modelos)

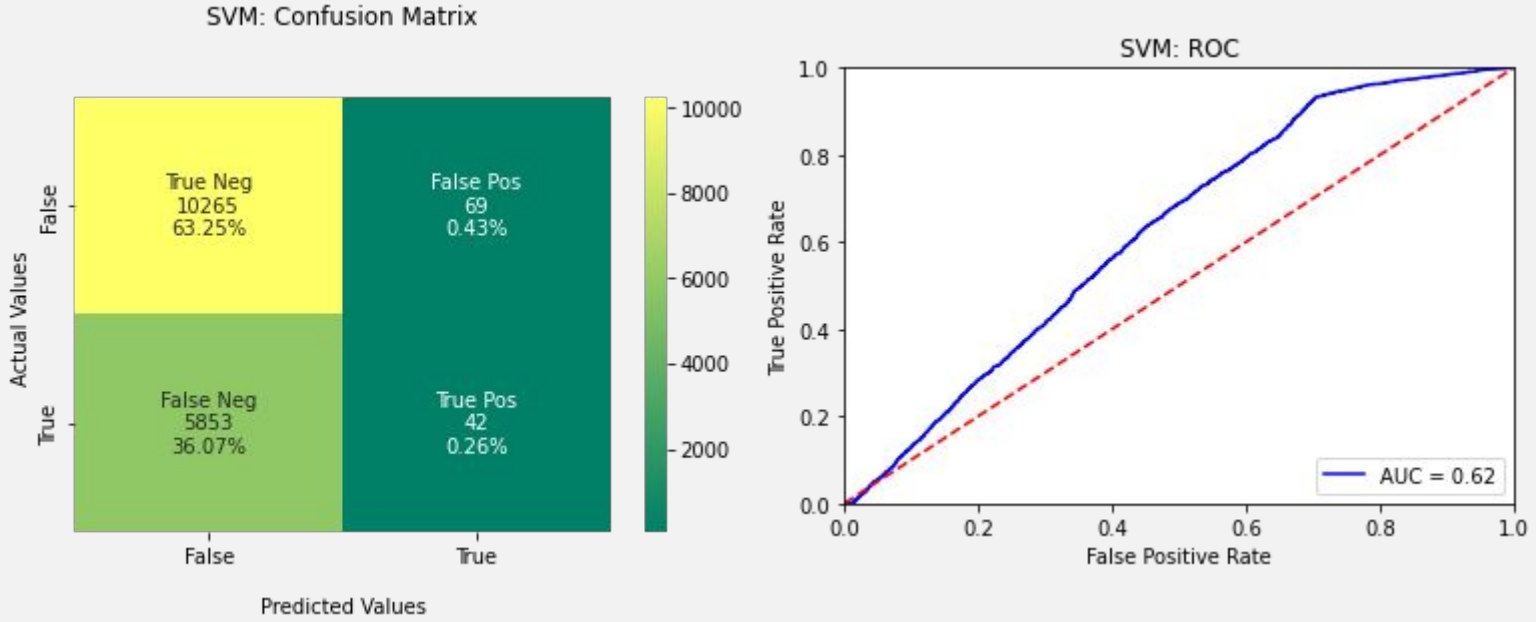
MODELO 4 : SVM

Las Máquinas de Vectores de Soporte (Support Vector Machines) permiten encontrar la forma óptima de clasificar entre varias clases. La clasificación óptima se realiza maximizando el margen de separación entre las clases. Los vectores que definen el borde de esta separación son los vectores de soporte. En el caso de que las clases no sean linealmente separables, podemos usar el truco del kernel para añadir una dimensión nueva donde sí lo sean.



Algoritmos (modelos)

MODELO 5 : SVM

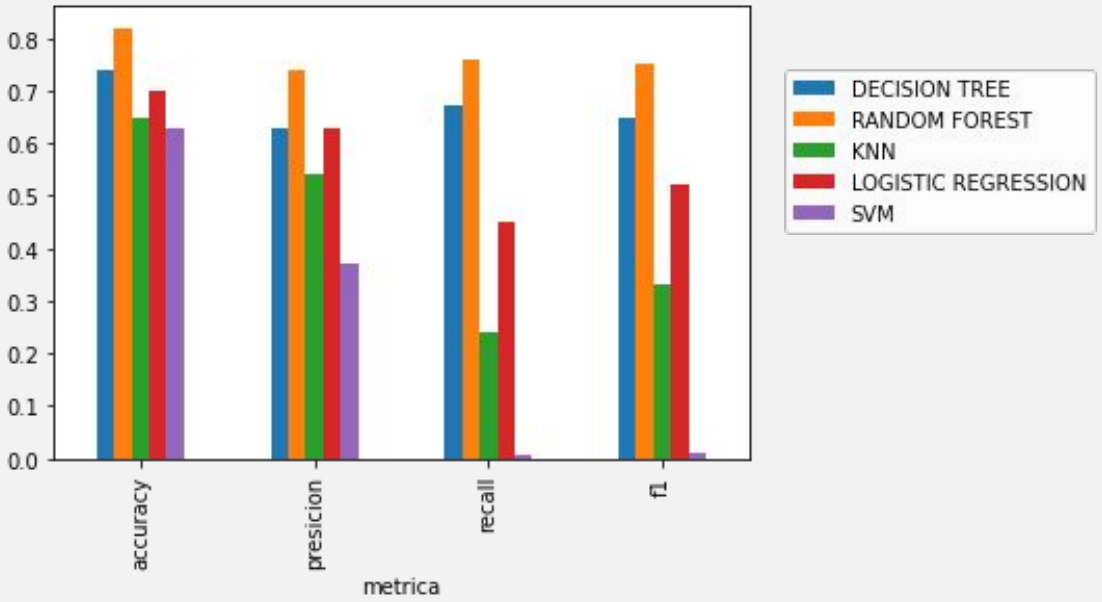



```
Accuracy SVM: 0.6350976646743484
Presición SVM: 0.3783783783783784
Recall SVM: 0.0071246819338422395
F1 Score SVM: 0.013986013986013986
```

Algoritmos (modelos)

Comparación de modelos

```
Accuracy Decision Tree: 0.743853595415614
Presición Decision Tree: 0.6393521488133419
Recall Decision Tree: 0.6763358778625954
F1 Score Decision Tree: 0.6573242106998599
--------------------------------------------------
Accuracy Random Forest: 0.8205065007086081
Presición Random Forest: 0.7461208319577418
Recall Random Forest: 0.7667514843087362
F1 Score Random Forest: 0.7562954906717978
--------------------------------------------------
Accuracy KNN: 0.6511183683529485
Presición KNN: 0.5444147922226458
Recall KNN: 0.24223918575063613
F1 Score KNN: 0.33528997417234085
--------------------------------------------------
Accuracy Regresión Logística: 0.706204941770904
Presición Regresión Logística: 0.6328696062249469
Recall Regresión Logística: 0.4553011026293469
F1 Score Regresión Logística: 0.5295974743488556
--------------------------------------------------
Accuracy SVM: 0.6350976646743484
Presición SVM: 0.3783783783783784
Recall SVM: 0.0071246819338422395
F1 Score SVM: 0.013986013986013986
```



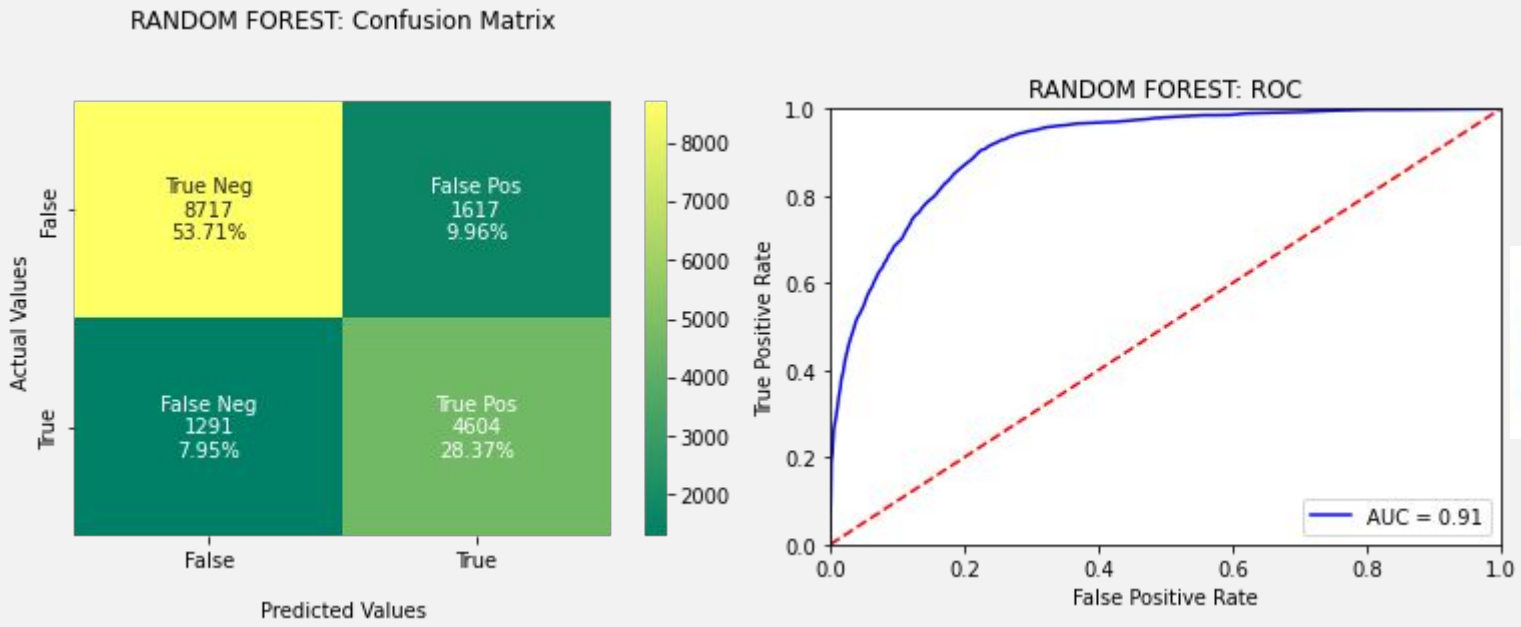
Podemos ver que el modelo que mejor se desempeña es el Random Forest, siendo superior en todas las metricas

Algoritmos (modelos)

Fine tuning Random Forest - RandomizedSearchCV

```
cv_rf = RandomizedSearchCV(h_rf, params1, n_iter=100, random_state = 42, cv=5,scoring='roc_auc',verbose=20)
```

Como vimos anteriormente, el modelo con mejor performance es el Random Forest, por ende lo seleccionamos para optimizar sus parámetros. Utilizando el RandomizedSearchCV logramos pasar de un AUC Score de 0.90 a 0.91, logrando una reducción de 0.52pp en los falsos negativos



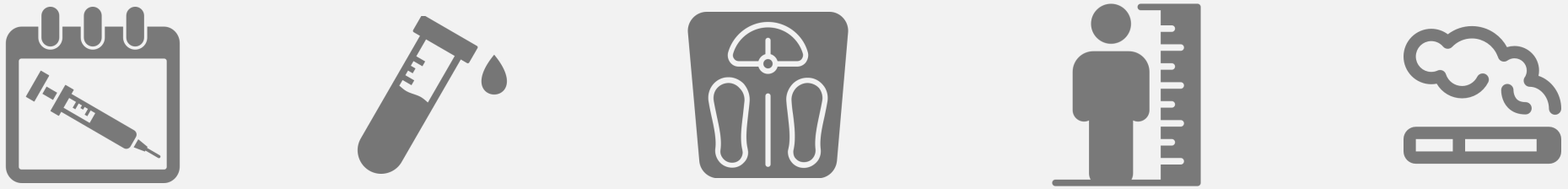

```
Accuracy Random Forest: 0.8304990757855822
Presición Random Forest: 0.759506421556283
Recall Random Forest: 0.7743260590500641
F1 Score Random Forest: 0.7668446478515129
```

4. CONCLUSIONES Y FUTURAS LÍNEAS

ETAPAS DEL DESARROLLO

Respuestas y conclusiones

- Pudimos determinar con un **74% de efectividad si una persona es fumadora o no, a partir del examen hematológico**
- No pudimos determinar correlaciones extremadamente fuertes, pero aplicamos un Random Forest y eso nos arrojó datos interesantes, pero no nos permite individualizar las variables, el resultado se obtiene del conjunto de todas las variables.
- Si bien al principio optamos por descartar algunas variables, luego concluimos que las características físicas de los fumadores, son datos importantes, ya que con variables como la edad, altura, peso, y otros datos, logramos acertar un **78%** de los pacientes que son fumadores.
  Por lo tanto concluimos que **SÍ, se puede detectar quien es fumador a partir de características físicas, no solo con los resultados de los exámenes de sangre.**



INTERPRETACIÓN Y DESPLIEGUE

CONCLUSIONES DEL ESTUDIO

Si bien el dataset utilizado en su estado original está en buenas condiciones sin necesidad de hacer una preparación muy extensa, no logramos identificar correlaciones muy marcadas en las variables.

Sin embargo en la utilización de modelos predictivos obtuvimos buenos resultados con Random Forest lo que hace viable la solución para futuros análisis predictivos de personas fumadoras.

FUTURAS LÍNEAS

Creemos rotundamente en el potencial de este modelo, si logramos adicionar más variables de valor, para maximizar el análisis y los resultados, como por ejemplo:

- Datos georeferenciados de pacientes, niveles de contaminación por ubicación y datos atmosféricos
- Históricos de exámenes y fichas de pacientes
- Otros tipos de exámenes asociados al paciente
- Históricos de defunción de los pacientes y sus causales

CODER HOUSE

¡MUCHAS GRACIAS!