CAPSTONE PROJECT

ALCANCES DEL PROYECTO

El grupo de cardiología del Hospital Metropolitano reporta la necesidad de poder predecir de mejor manera si un paciente tiene una enfermedad cardiovascular o no. En los últimos años se ha reportado una dificultad en el diagnóstico de enfermedades cardiovasculares por lo sutiles que pueden ser estos síntomas y no lograr predecir que la combinación de muchos factores pueden ser indicativos de una enfermedad. Por esto, han contactado a Data Analytics con el fin de entender mejor las características que definen estas enfermedades y las características de la población que llega al hospital presentado diversos síntomas. También quisieran tener un modelo predictivo que les ayude a analizar el conjunto de pruebas médicas realizadas en el hospital para definir la posibilidad de que el paciente tenga o no una enfermedad cardiovascular. El hospital ha entregado un set de datos con distintos atributos por paciente donde se define su sexo, edad y los resultados de ciertos exámenes médicos realizados por los médicos.

ANÁLISIS EXPLORATORIO DE DATOS

El proceso de Data Science debe empezar por un análisis exploratorio de los datos que se tiene para trabajar.

Primer análisis de los datos

Los datos entregados por el hospital cuentan con los siguientes atributos por paciente:

- Age: Edad
- Sex: Sexo
- Cp: Dolor de pecho
- Trestbps: Presión arterial en reposo (en mm Hg al ingreso al hospital)
- Chol: colesterol sérico (en mg/dl)
- Fbs: Azúcar en sangre en ayunas > 120 mg/dl (1 = true; 0 = false)
- Restecg: Resultados electrocardiográficos en reposo
- Thalach: Frecuencia cardíaca máxima alcanzada

- Exang: Angina Inducida por Ejercicio (1 = true; 0 = false)
- Oldpeak: Depresión del ST inducida por el ejercicio relativo al descanso
- Slope: La pendiente del segmento ST de ejercicio pico
- Ca: Número de vasos principales (0-3) coloreados por fluoroscopia
- Thal: (1 = normal; 2 = defecto fijo; 3 = defecto reversible)
- Target: (0 = enfermedad del corazón; 1 = asintomático)

Al observar la descripción de los datos destaca que existen 303 entradas para realizar el análisis. Cada una pertenece a un paciente distinto y se han eliminado los nombres para la confidencialidad con los mismos. Al tener 13 atributos de variables independientes confirmamos que tenemos la cantidad de entradas necesaria para poder realizar un análisis confiable de los datos.

En esta primera observación es importante también notar que se cuenta con varias variables categóricas que tienen un tipo de integer en la importación inicial de datos, debemos tomar esto en cuenta para los siguientes pasos del análisis.

Limpieza de los datos

A continuación se describen los pasos tomados para limpiar los datos y preparalos para la visualización y futura creación de modelos predictivos.

1. Fue primordial primero eliminar cualquier fila de datos que tuviera variables NA. Esto es posible dado que la cantidad de observaciones es suficiente para el estudio. En el caso de contar con una cantidad más limitada de datos entonces se debería utilizar un método que haga un cálculo de los datos NA y los reemplace.
2. Luego se eliminan todas las líneas duplicadas en el set de datos.
3. Los datos Thal = 0 no agregan ningún significado al set, por lo tanto se eliminan todas las líneas que contengan este valor.
4. Con estos cambios se considera limpio el set de datos para continuar con los siguientes pasos. Luego de estos pequeños cambios se tiene un set de datos de 300 entradas.

Creación de variables Dummy

Luego de limpiar los datos en cuanto a la eliminación de entradas duplicadas o entradas con variables que no aportan información al sistema se deben manejar los atributos que deben ser categóricos y por lo tanto, que son mejor manejados por medio de las variables Dummy. Antes de crear las variables dummy se le asigna a cada variable de cada atributo un nombre correspondiente. Por ejemplo, para sexo = 1 se asigna el nombre "male" y para sexo = 0 se asigna el nombre "female". Así se hace con el resto de los atributos de forma que haya un entendimiento más claro del significado. Una vez completado este paso se crean variables dummy que crean columnas por cada valor de los atributos. Continuando con el ejemplo de sexo tendríamos dos columnas en vez de una, en donde un valor = 1 significa que el atributo es verdadero. Los atributos aún no se convierten a variables categóricas dado que el análisis de matriz de correlación se debe hacer con integers.

Matriz de Correlación

Se continúa con un análisis de la matriz de correlación para observar la correlación entre las variables independientes y también la correlación con la variable dependiente (Target). En el notebook de jupyter se puede observar el heatmap para visualizar estas correlaciones.

Al estudiar la correlación entre las variables independientes se puede observar que los 10 pares con mayores correlaciones son atributos de la misma variable original. Por lo tanto, se decide que a nivel de estos no se eliminará ningún atributo. Ahora al analizar la correlación de la variable dependiente con todas las variables independientes del set de datos se observa que no existe ningún atributo que tenga alta correlación con esta variable.

Últimos ajustes de los datos antes de la visualización

Quedan unos cuantos detalles con respecto a los datos. Primero se convierten todas las variables pertinentes en variables categóricas utilizando el astype('category'). También se ajusta la variable 'oldpeak' dado que maneja valores de "0.*' y se multiplican todos sus valores por 10 para poder tener integers. Esto es posible dado que luego al normalizar la data se podrá tener una relación de peso con respecto al atributo y multiplicarlo por 10 no afectará el resultado final.

Visualización de los datos

En el notebook de Jupyter se pueden observar varias gráficas que permiten visualizar el comportamiento de los datos. A continuación de detallan las observaciones principales:

GRÁFICOS DE BARRAS

- Los pacientes tienen edades entre los 20 y 77. La mayoría de estos pacientes se encuentra entre los 50s.
- Hay casi dos veces más pacientes hombres que mujeres.
- El data set tiene un poco más de datos para pacientes asintomáticos que pacientes con enfermedades del corazón.
- Si observamos una visualización de las edades y la cantidad de casos asintomáticos/con enfermedad cardiovascular por cada edad se observa lo siguiente:
    - De los 29 a los 54 años se presentan más personas asintomáticas que personas con enfermedad cardiovascular.
    - De los 55 a 63 años, a los 67 años y 77 años se presentan más personas con enfermedad del corazón que personas asintomáticas.
    - Las demás edades presentan más casos asintomáticos o la misma cantidad de asintomáticos/enfermos del corazón.
- Al observar los hombres y mujeres con respecto a los casos asintomáticos y personas con enfermedad cardiovascular se observa:
    - Hombres: Existen más casos con enfermedad cardiovascular que asintomáticos.
    - Mujeres: Se presentan pocos casos con enfermedad cardiovascular, la mayoría son asintomáticas.
- Cabe destacar que las personas que no presentan dolor de pecho pueden tener enfermedades cardiovasculares. Por lo tanto, se recomienda no ignorar estos casos y tratar de utilizar modelos predictivos.
- Las personas con ventricular hyperthrophy y STT wave abnormality son más propensas a tener enfermedades cardiovasculares.
- Las personas con defectos del corazón previamente conocidos tienen alta probabilidad de tener una enfermedad cardiovascular.

GRÁFICOS DE DISPERSIÓN

- Al graficar sexo vs edad coloreado por target podemos observar que los hombres con enfermedad cardiovascular presentan casos a cualquier edad. Por otro lado, las mujeres tienen menos casos de enfermedad y estas se concentran entre los 55 y 65 años.
- Al graficar la edad vs thalach (frecuencia cardiaca máxima alcanzada) se puede observar que la mayoría de csasos con enfermedad cardiovascular se presentan cuando los pacientes tienen los menores valores de thalach. Los pacientes asintomáticos presentaron los valores de thalach más altos (>140 latidos por segundo).

MODELOS DE CLASIFICACIÓN

Una vez realizado el análisis exploratorio de datos se procede con la creación de modelos de clasificación que permitan predecir si una persona tendrá o no una enfermedad cardiovascular.

Antes de poder crear los modelos es importante normalizar o escalar los datos. En este caso se decide escalar los atributos continuos de forma que se puedan poner en una escala de 0 a 1 y poder ser comparados a nivel de peso con respecto a las variables categóricas. También se procede a implementar el algoritmo RFE para analizar si se debe eliminar alguna variable. A partir de este estudio y considerando que hay una opción limitada de atributos con los cuales trabajar se decide continuar con la totalidad de los datos para crear los modelos de clasificación.

Crear los sets de datos de entrenamiento y prueba

Para crear los modelos se dividen los datos 75% para entrenamiento y 25% para probar los modelos. Se establece también el Ground Truth con los valores conocidos de target por cada paciente.

Construyendo los modelos

Se construyen tres distintos modelos de clasificación bajo los parámetros default de scikit-learn. Los modelos elegidos son Random Forest, Support Vector Machine y Stochastic Gradient Descent.

Modelo	Puntaje del modelo
Random Forest	1.0
Support Vector Machine	0.917
Stochastic Gradient Descent	0.828

A partir de estos resultados se decide continuar con Support Vector Machine. La razón para no continuar con Random Forest es porque presenta un modelo perfecto que podría significar que el modelo ha aprendido la data de memoria y por lo tanto no sabría clasificar bien datos que no sean idénticos a los utilizados para crear el modelo. Support Vector Machine presenta el segundo puntaje y por lo tanto se decide continuar con el mismo.

Ajuste de parámetros de SVM

Se prueban varios ajustes de parámetros y se presentan los resultados a continuación:

Ajuste de parámetro	Puntaje del modelo
Kernel Lineal	0.904
C = 100	1.0
C = 50	0.994

Nuevamente se producen dos modelos casi perfectos que no son óptimos como modelos de clasificación. El tercer cambio hace que el modelo baje de puntaje y por lo tanto es descartado también. Se concluye que el mejor modelo es el generado con los parámetros default del paquete scikit-learn y es el modelo que se propondrá al Hospital Metropolitano.

Predicción y Evaluación de resultados

Los resultados obtenidos mediante la predicción y evaluación de resultados confirman el correcto funcionamiento del modelo y genera confianza para que el hospital pueda utilizarlo para predecir si sus pacientes tienen o no una enfermedad cardiovascular.

R Squared	RMSE
0.197	0.336

CONCLUSIONES Y RECOMENDACIONES

Se presentan las siguientes conclusiones y recomendaciones para el grupo de cardiología del Hospital Metropolitano.

- Se presenta un análisis de las características de los pacientes que se presentaron a hacer exámenes al centro de cardiología del Hospital Metropolitano. De este análisis y visualización de datos el hospital podrá tener un esquema general de sus pacientes. Entre las cosas que se determinan es la probabilidad de que el sexo infiera en una enfermedad cardiovascular, también su edad y el impacto que tienen los distintos resultados de los exámenes realizados.
- A partir del análisis dado se propone que sea utilizado para realizar campañas de protección contra la enfermedad cardiovascular a aquel sector de la población más propenso a tener una enfermedad de esta clase. Se recomienda compartir esta información con los servicios de salud del país para que cada persona pueda tomar las medidas necesarias para tener una mejor salud.
- Además de hacer un análisis general de las características de los pacientes se creo un modelo de clasificación que permitirá a los doctores tener un respaldo a sus diagnósticos. En cuyo caso el o la doctora tenga duda de si la persona podría tener una enfermedad cardiovascular puede recurrir a este modelo para cerciorarse y tener más confianza en su diagnóstico.
- Este análisis y la creación del modelo permite demostrar la gran utilidad del proceso de Data Science en el campo de la medicina. Se pueden generar conclusiones de impacto que permitan mejorar la salud de la población y mejorar la calidad de vida de la misma.