Data Science
Coderhouse

Proyecto Final: Clasificador de enfermedades cardíacas a partir de indicadores

Profesore: Sebastian Ferraro
Tutores: Juan Manuel Romero

Alumnos:
Camila Baron
Erica Muller
Dimas Torres
Ignacio Silva
Daniela Flores

Fecha de Entrega: 16 / 09 / 2022

Índice

1. Introducción

1.1. Marco Teórico

Según la CDC (Center of Disease Control), las enfermedades cardíacas son una de las principales causas de muerte para las personas de la mayoría de las razas en los EE. UU. (afroamericanos, indios americanos y nativos de Alaska, y blancos). Aproximadamente la mitad de todos los estadounidenses (47 %) tienen al menos 1 de 3 factores de riesgo clave de enfermedad cardíaca: presión arterial alta, colesterol alto y tabaquismo. Otros indicadores clave incluyen el estado diabético, la obesidad (IMC alto), no realizar suficiente actividad física o beber demasiado alcohol. Esta problemática se presenta de forma similar en los demás países del mundo y es característica del modo de vida sedentario que llevamos en las sociedades modernas. Detectar y prevenir los factores que más inciden en las enfermedades del corazón es muy importante en el ámbito sanitario debido a que los tratamientos disponibles ejercen su efecto sobre los síntomas dado que las causas son conductuales. Los desarrollos computacionales, a su vez, permiten la aplicación de métodos de aprendizaje automático para detectar "patrones.ª partir de los datos que pueden predecir la condición de un paciente y corregir sus hábitos antes de que se presenten mayores dificultades.

Es por esta situación actual que decidimos enfocarnos en un data set el cual se concentre en indicadores relacionados a enfermedades cardíacas, de esta manera a partir de las herramientas vistas en clase lograremos estudiar y analizar la correlatividad e importancia de cada una de estas variables en la causa de las enfermedades cardiacas. Nuestro objetivo puntual en este trabajo se concentra en poder clasificar mediante estrategias de machine learning distintas enfermedades cardiacas relacionadas a nuestro data set. Idealmente este clasificador podría ser aplicado a dispositivos móviles para que usuarios puedan ingresar sus propios datos y la aplicación pueda indicarles un diagnostico inicial.

1.2. Adquisición de Datos

Originalmente, el conjunto de datos proviene del Sistema de Vigilancia de Factores de Riesgo del Comportamiento (BRFSS), que realiza encuestas telefónicas anuales para recopilar datos sobre el estado de salud de los residentes de EE. UU desde 1984.

El conjunto de datos más reciente (al 15 de febrero de 2022) incluye datos del 2020. Consta de 401.958 filas y 279 columnas. La gran mayoría de las columnas son preguntas que se hacen a los encuestados sobre su estado de salud, como "¿Tiene serias dificultades para caminar o subir escaleras?.º "¿Has fumado al menos 100 cigarrillos en toda tu vida?". Sobre este conjunto original nosotros tomamos una selección disponible en la plataforma Kaggle.

A diferencia del conjunto de datos de la BRFSS solo contiene 20 columnas con los factores de mayor relevancia según el curador de esta selección. Viendo los resultados obtenidos por el autor pensamos que podría refinarse su modelo para mejorar las métricas de calidad obtenidas por el mismo. Es interesante destacar que se desarrolló una app que permite interactuar con el modelo de forma amigable y que en caso de poder mejorarlo sería posible implementar una interfaz similar con el nuevo modelo.

2. Desarrollo

El siguiente trabajo se llevo a cabo a través del uso de Python en conjunto a todas sus librerías aprendidas durante el desarrollo de todo el curso. A través de estas herramientas se desarrollo el algoritmo de clasificación.

En primer lugar lo que debimos a hacer fue un pre procesamiento de datos, en donde nos encontramos en primera instancia con un data set con aproximadamente 40000 datos. Una vez que logramos visualizar el dataframe se implementaron diversas herramientas de limpieza de datos de manera que sea mas facil de trabajar, especificamente quitando datos vacios o fijandonos si habian datos faltantes, ademas de ver con que tipo de datos nos estabamos encontrando.

Una vez familiarizados con nuestro data set, empezamos viendo con que tipo de variables estaríamos trabajando, es decir variables categóricas y las variables continuas.

Variables categoricas:

- HeartDisease
- Smoking
- AlcoholDrinking
- Stroke
- DiffWalking
- Sex
- Race
- Diabetic
- PhysicalActivity
- GenHealth
- Asthma
- KidneyDisease
- SkinCancer

Variables continuas:

- BMI,
- PhysicalHealtH
- MentalHealth
- AgeCategory
- SleepTime

Luego se procedió a hacer el análisis exploratorio de datos, el cual nos permitió establecer la variable target la cual seria Heart Disease. Ademas se debio reducir el dataset a 4000 datos para poder trabajar con el.

Siguiendo con el desarrollo se continuo realizando un análisis univariado, bivariado y multivariado, de modo que podamos extraer mas información que nos ayude a futuro a realizar mejores decisiones cuando nos encontramos con el clasificador.

2.1. Análisis Univariado

En primer lugar quisimos mostrar un boxplot el cual muestre la variable sleeptime. En donde obtuvimos que el 50$ de las personas encuestadas duerme entre 6 y 8 horas, lo que era de esperarse teniendo en cuenta el ritmo circadiano normal. A pesar de esto la otra mitad de la muestra se encuentra por fuera de estos límites llegando a dormir 3 horas como mínimo y 11 horas como máximo. Siendo una enfermedad poco frecuente los valores fueras de la norma pueden ser valiosos para encontrar correlaciones con nuestro target.

Luego evaluamos la variable del Índice de masa corporal (BMI) tambien a través de un boxplot, en este encontramos que cada cuartil represantaba los rangos de BMI posibles. Esto tiene sentido ya que la determinacion de estos rangos fue precisamente basada en los cuartiles de una población general.

A continuación realizamos un análisis de los grupos según su rango etéreo, en este veíamos que la distribución de edades en las personas encuestadas es bastante pareja aunque se ve una preferencia por informantes cercanos a los 50, tal vez porque el formato de la encuesta es telefónica. Se podría atraer a las nuevas generaciones por un formulario virtual y más aún si esta acoplado a una red social (encuesta de Instagram).

Considerando las variables cualitativas es importante mencionar el análisis de grupos étnicos ya que en este de ve el claro predominio de personas blancas dentro de la muestra indica la distribución étnica del país donde se tomaron los datos (USA). Esta información sería relevante al querer aplicar el modelo obtenido con estos, sobretodo si se encuentra que esta variable tiene una alta correlación con nuestra variable target.

2.2. Análisis Bivariado

En esta sección nos concentramos principalmente en encontrar la correlación entre todas nuestras variables y ver cuales son las mejor a analizar. Para esto utilizamos el análisis de correlación de phiK

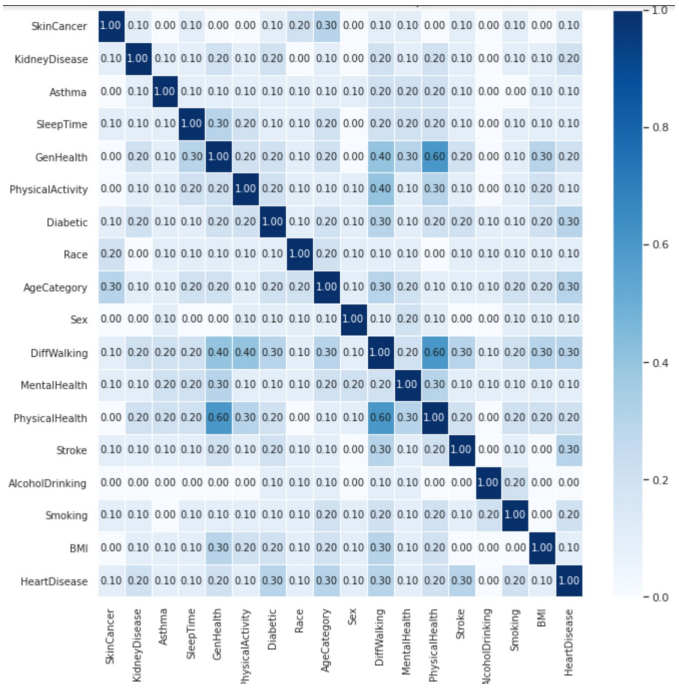

Figura 1: Análisis correlación

De este paneo general se puede ver que hay una correlación considerable entre "PhycalHealthz "GeneralHealth"pero no entre estas y "MentalHealth", lo que a primera instancia podría indicar que los encuestados valoran mas su salud física que mental al momento de evaluar su salud global. Por otro lado otra correlación significativa es entre la salud física y la tendencia a caminar, que era esperable. Algo que puede ser un poco desalentador es que ninguna de las variables tiene una correlación grande con la variable objetivo "HeartDisease"

Luego concentrándonos en la relación ente enfermedades cardiacas y edad encontramos que efectivamente las categorías que incluían las edades mas avanzadas tiene mayor proporción de personas con enfermedades cardiacas.

2.3. Análisis Multivariado

Nuevamente acá nos interesaba ver la correlación entre múltiples variables, para esto generamos un gráfico de correlaciones entre las variables numéricas.

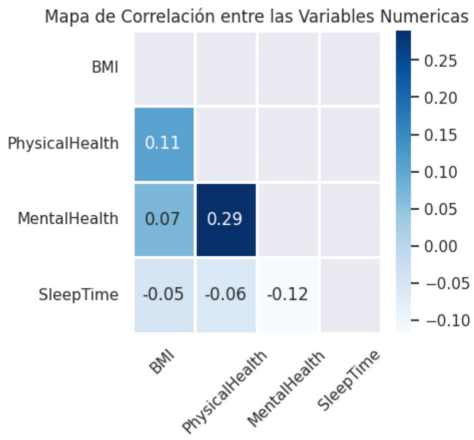

Figura 2: Análisis correlación multivariado

En este, vemos que la correlación entre las variables numéricas es bastante baja, donde la mayor de ellas es entre los dos parámetros de salud.

Como siguiente paso decidimos hacer a través de un gráfico de FacetGrid la relación entre las variables de Salud Mental y Salud Física sobre la variables target Heart Disease.

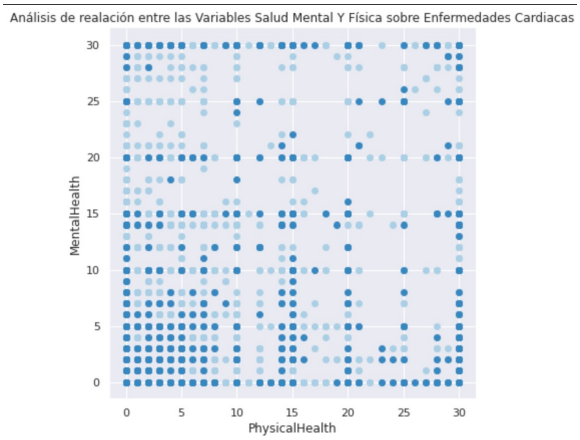

Figura 3: Relación Salud mental y física sobre enfermedades cardiacas

2.4. Desarrollo de los Modelos

Se probaron distintos modelos para tratar de predecir el riesgo de los pacientes de sufrir enfermedades cardiovasculares. Luego de seleccionar los mejores parámetros se midieron los indicadores de performance de cada uno de ellos y en base al área bajo la curva ROC se seleccionó a la regresión logística. Al ver que este modelo no tenía un desempeño satisfactorio se probó balancear las clases por medio del método de SMOTE. Luego de comparar el efecto de esta estrategia con los resultados sobre la clase objetivo desbalanceada recurrimos a un último recurso para poder evitar la gran cantidad de falsos negativos que estábamos teniendo; el cambio del umbral de rechazo. Para seleccionar el mejor umbral tomamos como métricas, la sensibilidad (capacidad de reconocer

verdaderos positivos) y la tasa de fallas (capacidad de reconocer verdaderos negativos). Siguiendo estos criterios llegamos a una situación de compromiso, con un umbral de 0.15, en la cual sólo 1 de cada 8 personas de alto riesgo serán clasificadas equivocadamente, un error tres veces menor que en el caso de pacientes de bajo riesgo. Esta situación se consideró óptima sopesando la gravedad de no atender a tiempo a una persona con riesgos de desarrollar una enfermedad cardiovascular contra el costo de una consulta médica preventiva.

3. Conclusión

En una primera instancia se analizaron las características de la población en estudio por métodos gráficos lo que permitió conocer algunos sesgos relevantes para nuestra investigación, como son la distribución racial de la muestra tomada y aún más importante la baja representación de las personas con enfermedades cardiovasculares. Esta asimetría dificulto el desarrollo de un modelo predictivo, llegando a una situación de compromiso donde se tuvo en cuenta el trasfondo sanitario de nuestro trabajo. En vistas a futuros trabajos de refinamiento se podría desarrollar una app on-line para facilitar tanto la recolección de datos como el despliegue ("deployment") del modelo similar al utilizado en uno de los trabajos originales sobre esta base de datos:

https://kamilpytlak-heart-condition-checker-app-2r42q4.streamlitapp.com/

Más allá de los resultados obtenidos consideramos que nuestra experiencia al llevar a cabo esté trabajo fue de fundamental importancia para adquirir las capacidades tanto técnicas como colaborativas para comenzar nuestro camino en el mundo de la ciencia de datos.