COLEGIO DE CIENCIAS E INGENIERÍA
INGENIERIA INDUSTRIAL
IIN-3007-ANALITICA DE DATOS

NRC: 3007

Proyecto Final

SEMESTRE: Segundo Semestre 2023/2024

NOMBRE Y CÓDIGO DEL ESTUDIANTE:

Antony Navarrete (00213945)
Fernando Molina (00209868)

PROFESOR(A): María Baldeon Calisto

FECHA DE ENTREGA: 22/04/2024

Proyecto Final

1. Introducción

Los ataques cerebrales y accidentes cerebrovasculares, devastadores en su impacto y alcance, constituyen una de las mayores preocupaciones de salud pública en todo el mundo. Responsables de aproximadamente el 11% de las muertes a nivel global anualmente, estas condiciones no solo representan una carga significativa para los sistemas de salud, sino que también generan un tremendo sufrimiento humano y familiar.

En este contexto crítico, surge la necesidad imperante de abordar estos eventos cerebrovasculares con estrategias preventivas y predictivas sólidas. Es por ello que este proyecto se propone como una respuesta proactiva a este desafío, enfocándose en el desarrollo y la implementación de un modelo predictivo de vanguardia. Este modelo tiene como objetivo principal determinar con precisión y anticipación si un paciente presenta una alta o baja probabilidad de sufrir un ataque cerebral.

Al emplear tecnologías innovadoras, análisis de datos avanzados y técnicas de aprendizaje automático, buscamos no solo identificar los factores de riesgo y las señales tempranas de un posible evento cerebrovascular, sino también brindar a los profesionales de la salud una herramienta efectiva para la toma de decisiones informadas y la intervención preventiva.

Al hacerlo, no solo aspiramos a reducir la incidencia y la mortalidad asociadas con los ataques cerebrales, sino también a mejorar la calidad de vida de millones de personas en todo el mundo, ofreciéndoles una oportunidad vital de detección temprana y prevención de esta grave enfermedad.

2. Explicación EDA

a. Valores Nulos

Podríamos encontrar valores nulos, en blanco o atípicos en ciertas variables. Estos valores inconsistentes pueden afectar la calidad de nuestro análisis y la precisión de nuestros modelos predictivos. Para abordar el problema que tienen ciertas filas de la base de datos donde encontramos datos con valores nulos, en blanco o atípicos, hemos usado ciertos criterios.

Por ejemplo para la variable de índice de masa corporal BMI hemos reemplazado los valores vacíos con la media de los datos proporcionados para no perder el dato. Por otra parte para otras variables como 'Heart Disease' encontramos que habían valores tanto numéricos de 0 y 1 como valores de tipo string "Yes" y "No", en este caso procedimos a reemplazar "Yes" : 1 y "No": 0. Además para ciertas columnas donde no se tenía tantos datos inconsistentes se procedió directamente a eliminar la fila de datos.

b. Valores Atipicos

Al revisar la base de datos, podemos identificar los diferentes tipos de datos contenidos en cada variable o columna. Algunas variables, como Hypertension (hypertension), Heart Disease (enfermedad cardíaca), Married (estado civil), Work (estado laboral), Residence (lugar de residencia), Smoking (historial de tabaquismo), Gender (género), Income (ingreso), Children (número de hijos) y Stroke (ataque cerebral), son principalmente categóricas o binarias, mientras que otras como Avg Glucose Level (nivel promedio de glucosa en sangre), Bmi (índice de masa corporal) y Age (edad) son variables numéricas. Al comparar estos datos con lo que esperaríamos encontrar, podemos notar que algunas variables pueden requerir limpieza de datos.

La decisión de eliminar datos atípicos basada en una prueba de porcentaje de aparición es una estrategia común para mejorar la calidad de los datos y garantizar la precisión de los análisis posteriores. Al observar que los datos atípicos representaban menos del 5% de la muestra, se tomó la decisión de eliminarlos para mitigar su posible impacto en los resultados del análisis.

Una vez eliminados los datos atípicos, se realiza una nueva exploración de los datos y visualizaciones para verificar cómo afectó esto a la distribución y la relación entre las variables.

Esta etapa de validación es crucial para garantizar la integridad de los resultados y la interpretación adecuada de los hallazgos obtenidos.

c. Visualización de Datos

Para visualizar las variables mencionadas, se pueden emplear distintos tipos de gráficos según su naturaleza. Las variables categóricas como Hypertension, Heart Disease, Married, Work, Residence,

Smoking, Gender and Children pueden representarse mediante gráficos de barras, mostrando la frecuencia de cada categoría para una comparación clara. Por otro lado, las variables numéricas como Avg Glucose Level, Bmi, Age e Income pueden visualizarse con histogramas para examinar la distribución de los valores. Además, la variable binaria Stroke puede representarse también con un gráfico de barras para mostrar la proporción de pacientes que han experimentado un ataque cerebral. Estas técnicas de visualización proporcionan una manera efectiva de explorar la distribución y las relaciones entre las variables, facilitando la identificación de patrones y tendencias relevantes en los datos.

d. Tipo de datos

Tras llevar a cabo un análisis exhaustivo de los datos, que incluyó la eliminación de valores atípicos y nulos, así como la conversión de variables en variables ficticias para su posterior procesamiento, se obtiene:

```
Información general sobre la base de datos:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5112 entries, 0 to 5111
Data columns (total 13 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Hypertension       5112 non-null   int64
 1   Heart Disease      5112 non-null   object
 2   Married            5112 non-null   object
 3   Work               5112 non-null   object
 4   Residence          5112 non-null   object
 5   Avg_glucose_level  5112 non-null   float64
 6   Bmi                4911 non-null   float64
 7   Smoking            5112 non-null   object
 8   Gender             5112 non-null   object
 9   Age                5112 non-null   float64
 10  Income             5085 non-null   float64
 11  Children           5112 non-null   int64
 12  Stroke             5112 non-null   int64
dtypes: float64(4), int64(3), object(6)
memory usage: 519.3+ KB
```

3. **Explicar la correlación entre las variables predictivas y la importancia de la variable de respuesta.**

En el caso de un modelo predictivo para determinar la probabilidad de sufrir un ataque cerebral, la correlación entre variables como la hipertensión, el historial de tabaquismo, el índice de masa corporal (IMC) y los niveles de glucosa en sangre con la ocurrencia de un ataque cerebral es de gran

importancia. Si se observa una correlación positiva significativa entre la hipertensión y los ataques cerebrales, esto podría indicar que la hipertensión es un factor de riesgo importante. Del mismo modo, una correlación negativa entre el consumo regular de frutas y verduras y los ataques cerebrales podría sugerir un efecto protector de una dieta saludable.

Por lo tanto, al comprender la correlación entre las variables predictivas y la variable respuesta, podemos identificar los factores clave que influyen en el evento de interés y desarrollar modelos predictivos más precisos y efectivos. Esto a su vez nos permite tomar medidas preventivas más específicas y personalizadas para reducir el riesgo de ocurrencia del evento adverso.

La variable respuesta es fundamental en el análisis de datos, ya que representa lo que se busca predecir o entender en un estudio. Su importancia se refleja en la evaluación de modelos, toma de decisiones y comprensión de los factores que influyen en el fenómeno estudiado. Es el criterio principal para medir la eficacia de los modelos y guía la selección de variables predictoras. En fin, la variable respuesta es clave para interpretar y aplicar los resultados del análisis de datos.

4. Bases de datos No-Balanceadas

Un conjunto de datos desbalanceado, conocido como unbalanced dataset, exhibe una distribución de clases objetivo sesgada hacia una clase en particular en comparación con otras. Por ejemplo, en problemas de clasificación binaria, donde una clase representa eventos raros como enfermedades, fraudes o fallos, mientras que la otra clase representa eventos comunes, la clase de eventos raros puede tener menos ejemplos que la clase de eventos comunes. Esto puede llevar a problemas como el desempeño sesgado del modelo, donde los algoritmos de aprendizaje automático tienden a favorecer la clase mayoritaria, y la evaluación engañosa del modelo, donde métricas como la precisión pueden ser engañosas debido al desbalance de clases. Para abordar estos problemas, se pueden emplear técnicas como el sobre y submuestreo, que ajustan la distribución de clases mediante la generación de ejemplos sintéticos o la eliminación de ejemplos de la clase mayoritaria, y técnicas híbridas que combinan ambos enfoques. Además, el aprendizaje sensible al costo y los métodos de ensamble también pueden mejorar la precisión de los modelos en conjuntos de datos desbalanceados. Es

fundamental evaluar varias técnicas y seleccionar la más adecuada según el contexto del problema y las características específicas del conjunto de datos.

5. División de la Base de datos

La división apropiada de los datos en conjuntos de entrenamiento, validación y prueba es esencial para el desarrollo de modelos de aprendizaje automático, asegurando una evaluación imparcial de su rendimiento y su capacidad para generalizar datos no observados. En esta estrategia, la elección precisa de los porcentajes asignados a cada conjunto es crítica, ya que puede afectar significativamente la capacidad del modelo para aprender patrones y adaptarse a nuevas instancias. Se seleccionaron los siguientes porcentajes: 60% para entrenamiento, 20% para validación y 20% para prueba. Esta elección se basa en proporcionar una cantidad suficiente de datos para el entrenamiento y evitar el sobreajuste, permitir la selección del mejor modelo y la optimización de hiper parámetros, y garantizar una evaluación imparcial del rendimiento final del modelo. Esta división equilibrada sigue las mejores prácticas establecidas en la literatura de aprendizaje automático y ha demostrado ser efectiva en diversas aplicaciones y conjuntos de datos.

6. Algoritmos utilizados

Random Forest es un algoritmo de aprendizaje supervisado ampliamente utilizado en problemas de clasificación y regresión. Este método se basa en la construcción de múltiples árboles de decisión durante el entrenamiento y la combinación de sus resultados para mejorar la precisión y generalización del modelo. Cada árbol de decisión en un Random Forest se entrena con una muestra aleatoria del conjunto de datos y utiliza un subconjunto aleatorio de características para tomar decisiones. Luego, la predicción final se realiza promediando las predicciones de todos los árboles individuales (en el caso de clasificación) o tomando el promedio ponderado (en el caso de regresión).

SMOTE (Synthetic Minority Over-sampling Technique) es una técnica de remuestreo que aborda el desbalance de clases al sintetizar ejemplos nuevos de la clase minoritaria. En lugar de replicar

directamente ejemplos existentes, SMOTE genera muestras sintéticas interpolando entre ejemplos de la clase minoritaria que son similares. Esto ayuda a evitar el sobreajuste al replicar datos existentes y mejora la capacidad del modelo para generalizar a datos no vistos. Al equilibrar la distribución de clases, SMOTE contribuye a mejorar el rendimiento del modelo en la detección de la clase minoritaria y reduce el sesgo hacia la clase mayoritaria durante el entrenamiento.

La Máquina de Vectores de Soporte (SVM) es una herramienta de aprendizaje automático diseñada tanto para clasificación como para predicción de regresión, con una destacada capacidad para maximizar la precisión predictiva y evitar el sobreajuste a los datos. Esta técnica tiene como objetivo encontrar un hiperplano óptimo que separe eficientemente los datos pertenecientes a diferentes clases. El objetivo fundamental es identificar una función que divida los datos en dos clases y, a su vez, pueda clasificar con precisión nuevas muestras del conjunto de prueba. La SVM ha demostrado ser una herramienta valiosa en la identificación de enfermedades basadas en datos médicos. En cuanto a los hiper parámetros utilizados, destaca el kernel, que se elige en su forma lineal. Esta elección se justifica por la utilidad del kernel lineal al trabajar con datos dispersos, una situación común en problemas médicos, como la evaluación de las probabilidades de enfermedades cardíacas.

Gradient Boost, un modelo de conjunto para datos numéricos o categóricos, se distingue por su enfoque de construcción secuencial de predicciones mediante árboles de decisión. En este proceso, cada nuevo modelo en la secuencia aprende de los errores del modelo previo, corrigiendolos de manera iterativa. Algunos hiper parámetros clave incluyen la tasa de aprendizaje, que regula la contribución de cada árbol, la profundidad máxima de los árboles, el número total de árboles en el conjunto y el número mínimo de muestras requeridas para la partición. La tasa de aprendizaje controla la velocidad de ajuste del modelo a los datos, mientras que la profundidad máxima de los árboles determina su complejidad. El número total de árboles y el número mínimo de muestras para particiones son cruciales para configurar el conjunto de manera eficaz. Estos hiper parámetros desempeñan un papel fundamental en la optimización del rendimiento del modelo, permitiendo

predicciones precisas en diversos conjuntos de datos. Gradient Boost se destaca como un modelo poderoso gracias a su capacidad para mejorar y corregir iterativamente errores anteriores, proporcionando predicciones sólidas y precisas en escenarios numéricos o categóricos.

7. Matriz de confusión

Se definieron dos modelos de aprendizaje automático: un clasificador Random Forest y un clasificador XG Boost. Posteriormente, se establecieron los rangos de hiper parámetros que serán explorados para cada modelo. Se emplearon tanto la búsqueda aleatoria como la búsqueda en la cuadrícula para encontrar la mejor combinación de hiper parámetros que maximice el área bajo la curva ROC (AUC-ROC), una métrica comúnmente utilizada para evaluar modelos de clasificación. Tras llevar a cabo la optimización de hiper parámetros, se identificaron los mejores modelos encontrados junto con sus respectivos puntajes AUC-ROC. Estos puntajes reflejan la capacidad de los modelos para clasificar correctamente las instancias positivas y negativas en los datos de entrenamiento.

Los resultados revelaron que el modelo XG Boost obtuvo una puntuación AUC-ROC más alta (0.857) en comparación con el modelo Random Forest (0.853). Esto sugiere que el modelo XG Boost exhibe una capacidad ligeramente superior para discriminar entre las clases positivas y negativas en los datos de entrenamiento en comparación con el modelo Random Forest.

8. Costos de Implementación

Desarrollar un modelo predictivo de ataques cerebrales implica una serie de costos que abarcan desde la adquisición y preparación de datos médicos hasta la infraestructura tecnológica necesaria para el procesamiento y análisis de datos. Los costos estimados incluyen entre $50,000 y $220,000, cubriendo áreas como la recolección de datos ($5,000 - $20,000), infraestructura tecnológica ($10,000 - $50,000), desarrollo del modelo ($20,000 - $100,000), evaluación y validación del modelo ($10,000 - $30,000), y finalmente, integración y despliegue ($5,000 - $20,000). Estas cifras pueden variar según la complejidad del proyecto y los recursos disponibles, pero proporcionan una estimación

general de los costos involucrados en la implementación de un modelo predictivo de ataques cerebrales.

9. Link al Repositorio de Github

https://github.com/Fercho20019/Analitica/settings

10. Conclusiones

En resumen, el proyecto se centra en la creación de un modelo predictivo para determinar la probabilidad de que un paciente sufra un ataque cerebral. Este enfoque proactivo busca mejorar la detección temprana y la intervención preventiva, dada la importancia crucial de abordar eficazmente estas afecciones, que representan una de las principales causas de muerte a nivel mundial. La base de datos utilizada, proveniente de un hospital público y anonimizada para la investigación, incluye una variedad de variables socio-demográficas y médicas de los pacientes, así como una variable respuesta que indica la ocurrencia de un ataque cerebral. El análisis de esta información puede ofrecer valiosos conocimientos sobre los factores de riesgo y ayudar a implementar medidas preventivas más efectivas, en última instancia, mejorando la salud y el bienestar de la población.

El algoritmo Random Forest ha demostrado ser altamente efectivo en la predicción de enfermedades cardíacas, mostrando puntajes F1 excepcionales en los conjuntos de entrenamiento, validación y prueba (0.99, 0.94 y 0.94, respectivamente). Estos resultados sugieren que el modelo ha aprendido de manera robusta los patrones presentes en los datos de entrenamiento y puede generalizar efectivamente a nuevos conjuntos de datos, logrando una precisión perfecta en la clasificación de casos positivos y negativos. El rendimiento perfecto del algoritmo KNN en el conjunto de validación, con un puntaje F1 de 0.90, es particularmente destacado, indicando su capacidad para clasificar con precisión todos los casos en un conjunto de datos independiente. La consistencia entre los altos puntajes en los conjuntos de entrenamiento y validación sugiere que el modelo no solo ha memorizado los datos de entrenamiento, sino que ha capturado de manera efectiva patrones subyacentes generalizables a nuevas instancias. Esta exitosa generalización es esencial para asegurar que el modelo sea

aplicable y preciso en situaciones del mundo real, reforzando la confianza en la capacidad del modelo KNN para realizar predicciones confiables en la clasificación de enfermedades cardíacas.

11. Referencias

Emanuel, M. J. (Argentina de 2022). CLASIFICACIÓN DE DATOS DESBALANCEADOS. Obtenido de https://sedici.unlp.edu.ar/bitstream/handle/10915/147410/Documento_completo.pdf?sequence=1&isAllowed=y

Miravet, B. A. (Junio de 2021). Mejora de las predicciones en muestras desbalanceadas. Obtenido de https://repositorio.uam.es/bitstream/handle/10486/697900/abella_miravet_blanca_tfg.pdf?sequence=1

Rodríguez, M. B. (septiembre de 2022). Desarrollo y validación de algoritmos de Deep Learning para la clasificación de fundiciones de hierro. Obtenido de https://repositorio.unican.es/xmlui/bitstream/handle/10902/26246/BarcenaRodriguezMarta-TFG-Matematicas.pdf?sequence=1&isAllowed=y

Sanz, F. (2020, noviembre 30). Cómo funciona el algoritmo XGBoost en Python. The Machine Learners. https://www.themachinelearners.com/xgboost-python/

Random forest. (s/f). Bing. Recuperado el 3 de abril de 2024, de https://www.bing.com/search?q=random+forest&cvid=aaf020ff497a49cdbbb439045dec17cb&gs_lcrp=EgZjaHJvbWUqBggAEAAYQDIGCAAQABhAMgYIARBFGDkyBggCEAAYQDIGCAMQABhAMgYIBBAAGEAyBggFEAAYQDIGCAYQABhAMgYIBxAAGEAyBggIEAAYQNIBCDMyOTFqMGo5qAIEsAIB&FORM=ANAB01&PC=LCTS

Overfitting vs. Underfitting: What is the difference? (2021, agosto 27). 365 Data Science. https://365datascience.com/tutorials/machine-learning-tutorials/overfitting-underfitting/

Marktab. (2023, 11 julio). Preparación de datos para ML Studio (clásico) - Azure Architecture Center. Microsoft Learn. https://learn.microsoft.com/es-es/azure/architecture/data-science-process/prepare-data

Acero, J., & Rojas, K. (2023). Optimización de Hiperparámetros en Algoritmos de Aprendizaje Automático. Universidad Industrial de Santander . Asif, D., Bibi, M., Shoaib, M., & Mukheimer, A. (2023). Enhancing Heart Disease Prediction through Ensemble Learning Techniques with Hyperparameter Optimization. Algorithms. https://creativecommons.org/licenses/by/4.

0/ Ansarullah, S. I., Mohsin Saif, S., Abdul Basit Andrabi, S., Kumhar, S. H., Kirmani, M. M., & Kumar, D. P. (2022). An Intelligent and Reliable Hyperparameter Optimization Machine Learning Model for Early Heart Disease Assessment Using Imperative Risk Attributes. Journal of healthcare engineering, 2022, 9882288. https://doi.org/10.1155/2022/9882288 (Retraction published J Healthc Eng. 2023 Oct 11;2023:9871962)

Powers, D. M. W. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies, 2(1), 37-63.

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861-874. Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(1), 29-36.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427-437.