



Accidente Cerebrovasculares

Julian Mónaco - Estudiante - Data Science

Jorge Ruiz - Profesor - Comisión 42390

Aldana Ruscitti - Tutora - Comisión 42390



Indice

| | |
|-------------------------------------------------|----|
| Motivo de la Investigación | 3 |
| Interrogantes Problema | 4 |
| Es la edad un factor clave? | 5 |
| Hipertensión y afecciones cardíacas | 6 |
| Influye la zona de residencia en los ACV? | 7 |
| El azúcar en sangre aumenta los riesgos? | 8 |
| Algoritmos de Clasificación..... | 9 |
| Random Forest | 10 |
| Regresión Logística | 11 |
| KNN | 12 |
| Conclusiones | 13 |



Motivo de la Investigación

NovaCare, líder en medicina prepaga, busca reducir costos de internación y tratamiento ocasionados por cuadros de accidente cerebrovasculares (ACV) entendiendo que la prevención es la clave no solo para abaratar costos sino que también esto significa una mejor calidad de vida para el paciente, para ello empresa se enfoca en entender que condiciones clínicas y factores sociales aumentan el riesgo de ACV y de esta forma impulsar programas de concientización y prevención para pacientes de alto riesgo promoviendo un estilo de vida saludable.



Un problema en crecimiento



Según un reporte de la Asociación Americana del Corazón de Estados Unidos, en el país norteamericano se gastaron anualmente USD 43.5 billones entre 2014 y 2015 en costos totales por accidentes cerebro vasculares. De ese monto, USD 28 billones fueron para internaciones hospitalarias, visitas a urgencias, medicamentos recetados y atención médica a domicilio. El gasto por paciente fue de USD 7.902.

Dicha cifra, según prevé el informe, se duplicará entre 2015 y 2035 debido al incremento en la aparición de casos de ACV en la población.

Impacto de los ACV en los costos





- Internaciones extendidas
- Requerimiento de medicación post internación
- Solicitud de profesionales para rehabilitación
- Tratamientos ambulatorios
- Mayor requerimiento de profesionales



Interrogantes Problema

NovaCare nos ha dispuesto 4 líneas de investigación que han detectado como posibles causas contributivas a cuadros que derivan en cuadros de ACV y consecuentemente incrementando los costos y recursos utilizados por el afiliado.

Luego de la reunión ejecutiva hemos dispuesto los siguientes interrogantes a investigar:

-  Es la edad un factor clave en el desarrollo de un cuadro positivo?
-  Las personas con problemas de hipertensión y afecciones cardíacas son mas propensas a desarrollar síntomas?
-  Influye la zona de residencia en los pacientes positivos?
-  El azúcar en sangre aumenta los riesgos?



Datos

La investigación se realizó sobre la información proporcionada por NovaCare que cuenta con registros de pacientes únicos que cubren un rango etario que va desde los neonatales hasta adultos mayores, incluyendo aspectos sociales como lugar de residencia, tipo de trabajo, situación frente al tabaquismo, etc. Por el lado clínico se nos proporciona información sobre problemas cardiacos, hipertensión, índice de masa corporal, etc.



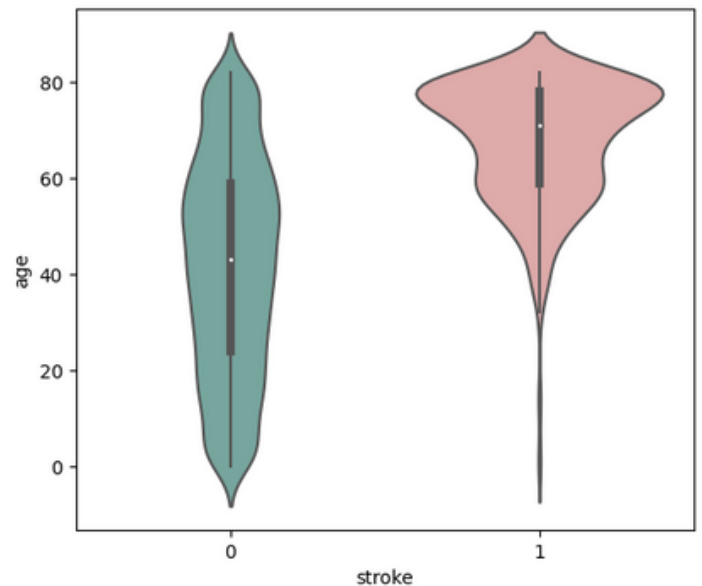
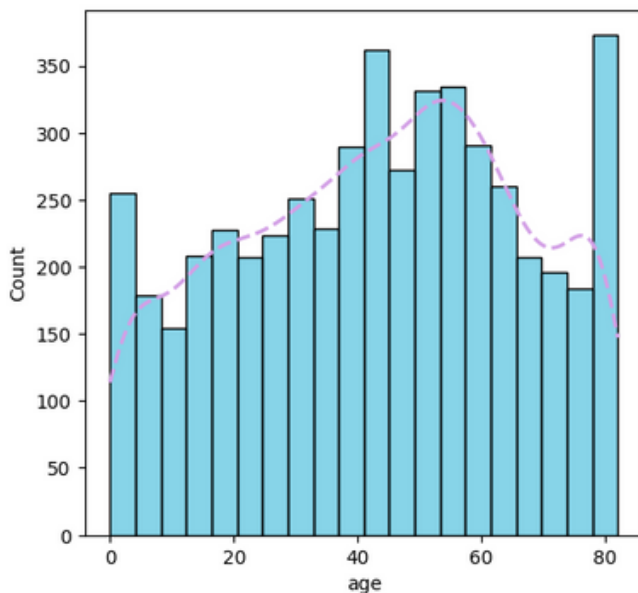
5110

REGISTROS ÚNICOS ANALIZADOS

Es la edad un factor clave en el desarrollo de un cuadro positivo?



Los registros fueron clasificados por grupos etarios que representen las etapas de la vida de una persona, desde su Juventud, pasando al joven adulto, etapas de adultez y persona mayor, en estas ultimas etapas es donde se ve aumentada la cantidad de casos positivos de ACV.



| Grupo de Edad | Casos Positivos | Porcentaje |
|---------------|-----------------|------------|
| (0, 18] | 2 | 0,8% |
| (18, 35] | 1 | 0,4% |
| (35, 55] | 35 | 14,1% |
| (55, 70] | 82 | 33,1% |
| (70, 82] | 128 | 51,6% |

*Comprobación estadística realizada con el coeficiente Biserial puntual que arroja un resultado de 0.24 y un P Valor menor a 0.05

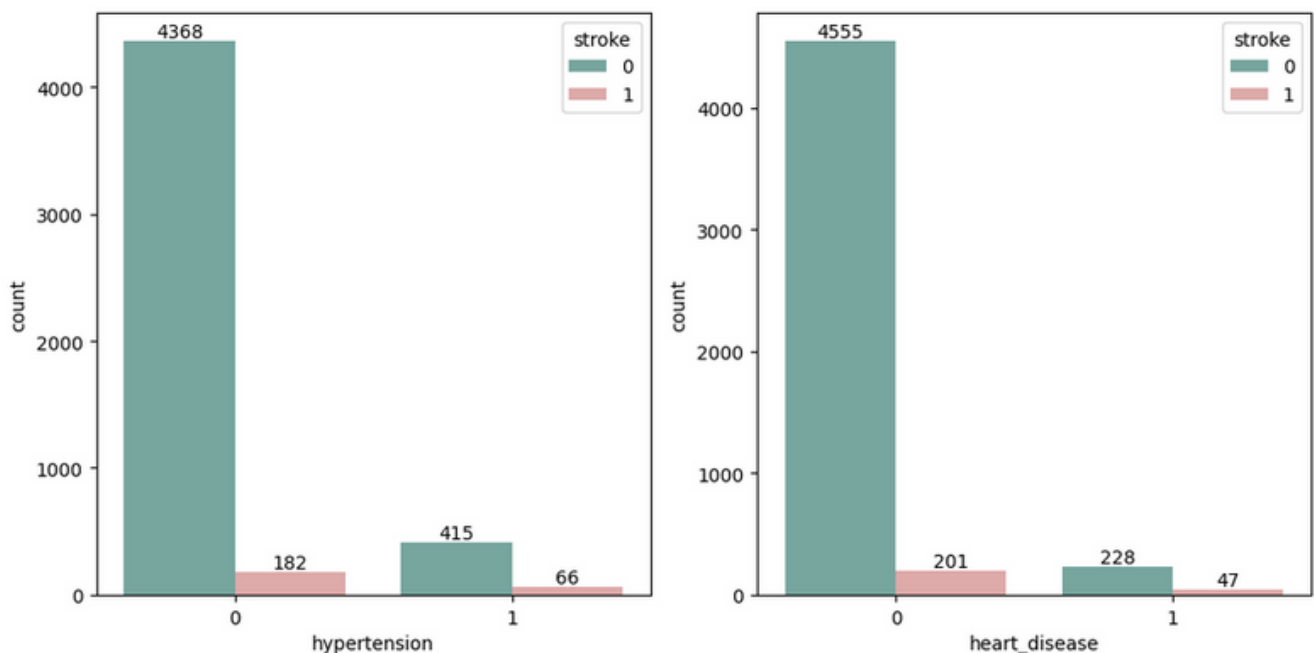
Los resultados obtenidos confirman que la edad es un factor determinante en estos cuadros ya que entran en juego una multiplicidad de factores que pueden predisponer al individuo a generar coágulos que viajen por el sistema sanguíneo y terminen alojados en el cerebro ocasionando la interrupción total o parcial del flujo, entre los principales factores podemos identificar:

- Cambios en el corazón propios de la edad (fibrilación auricular, arritmia, etc)
- Fragilidad vascular: Con la edad, las paredes de los vasos sanguíneos pueden volverse más frágiles y propensas a la ruptura
- Aterosclerosis: La acumulación de placa en las arterias
- Estilo de vida poco saludables

Las personas con problemas de hipertensión y afecciones cardíacas son mas propensas a desarrollar síntomas?



Ya hemos visto en el punto anterior que la edad es un factor clave en el desarrollo de cuadros de ACV ya que el cuerpo sufre modificaciones sobre todo aquellas que afectan al corazón y al flujo sanguíneo, por lo tanto ahora el interrogante es confirmar si dichas percepciones son correctas y se confirman con los datos que hemos recibido.



| Hipertensión | Afecciones Cardíacas | ACV | Total registros | % registros |
|--------------|----------------------|------------|-----------------|-------------|
| Negativo | Negativo | Negativo | 4191 | 96,6% |
| | | Confirmado | 148 | 3,4% |
| | Confirmado | Negativo | 177 | 83,9% |
| | | Confirmado | 34 | 16,1% |
| Confirmado | Negativo | Negativo | 364 | 87,3% |
| | | Confirmado | 53 | 12,7% |
| | Confirmado | Negativo | 51 | 79,7% |
| | | Confirmado | 13 | 20,3% |

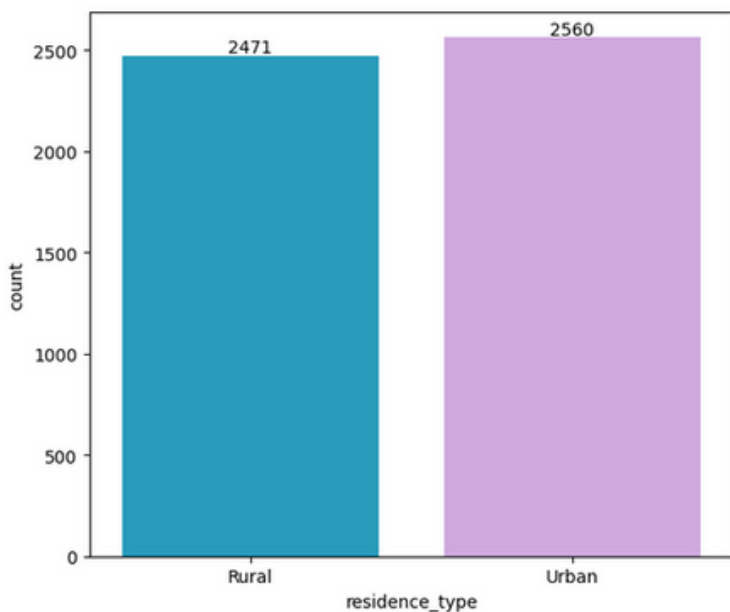
*Comprobación estadística realizada con el coeficiente CHI2 que arroja un resultado de 163.8 y un P Valor menor a 0.05

Aquellos registros que presentan alguna de las dos patologías mencionadas tienen un mayor porcentaje de sufrir un ACV que aquellas que no, en las personas sin ninguna de las dos patologías el sufrir un ACV es muy poco probable según la investigación realizada.

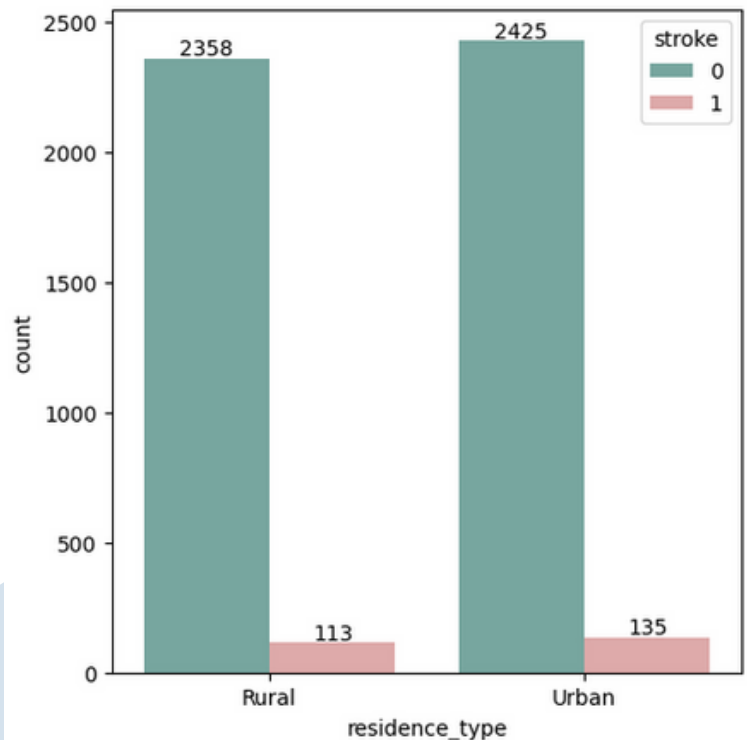
Influye la zona de residencia en los pacientes positivos?



El entorno es parte fundamental en la vida de una persona, nuestro entorno nos afecta directamente y por sobre todo condiciona nuestros estilos de vida, costumbres y vivencias, por eso se quiere determinar si el lugar de residencia de los pacientes puede ser un factor determinante en la aparición de cuadros positivos de ACV. En los registros recibidos contamos con una clasificación binaria en donde se identifica si los individuos habitan en zonas urbanas o rurales.



*A diferencia de otras variables, la muestra cuenta con un balance entre las clases



| Residencia | ACV | Cantidad | Porcentaje |
|------------|------------|----------|------------|
| Rural | Negativo | 2358 | 95.4% |
| | Confirmado | 113 | 4.6% |
| Urban | Negativo | 2425 | 94.7% |
| | Confirmado | 135 | 5.3% |

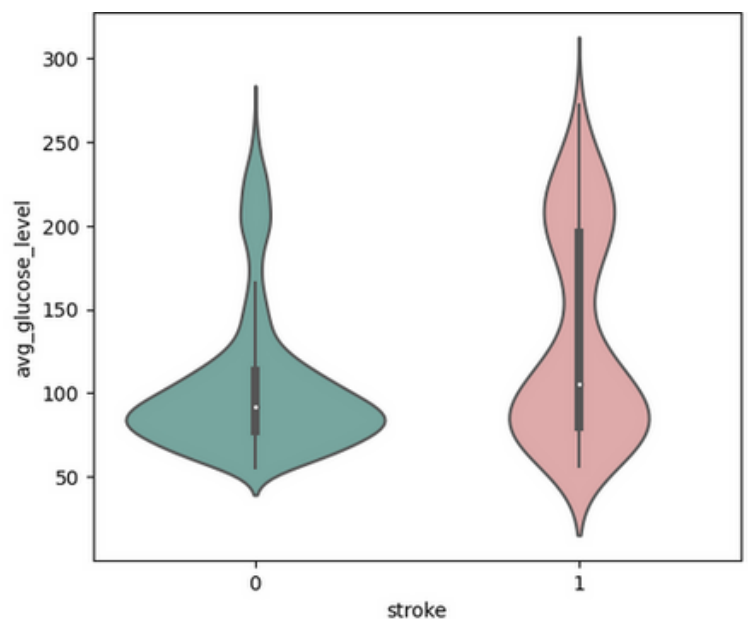
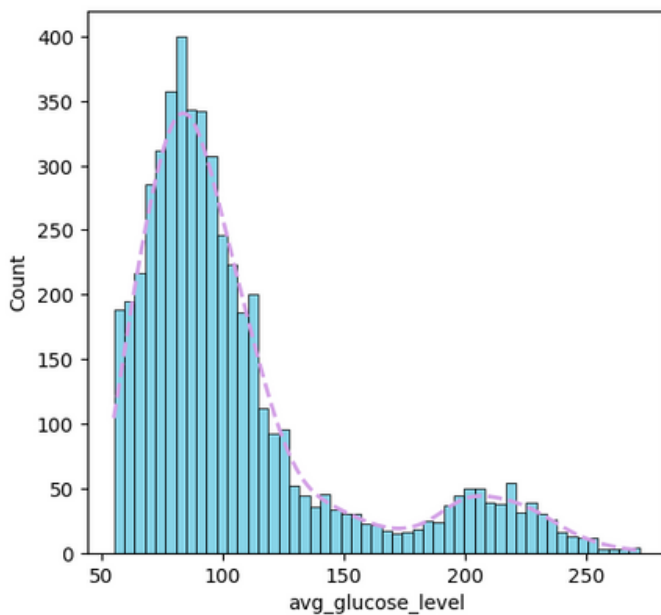
*Comprobación estadística realizada con el coeficiente CHI2 que arroja un resultado de 1.17 y un P Valor superior a 0.05

El entorno donde reside el paciente en este caso no pareciera tener incidencia sobre aumentar los riesgos de sufrir un ACV basados en los datos de muestra que hemos recibido. Podría haberse intuido a priori que una persona en una ciudad desarrolla hábitos menos saludables que aumenten el riesgo del cuadro, pero no es conclusivo y dependerá de otros factores a tener en cuenta.

El azúcar en sangre aumenta los riesgos?



A lo largo del informe hemos discutido como los hábitos poco saludables pueden tener un impacto directo en la salud del individuo, el consumo de azúcar en sangre tiene un impacto directo en los niveles de presión que ya confirmamos tiene una relación directa con el ACV, buscamos confirmar esta teoría de como altos niveles de azúcar en sangre pueden hacer mas propenso al individuo a sufrir un caso positivo.

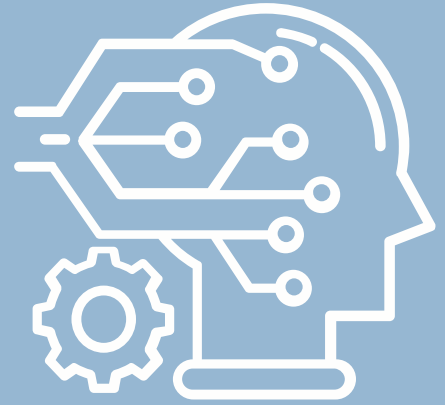


| Azucar en Sangre | Casos Positivos | Porcentaje |
|------------------|-----------------|------------|
| (50, 100] | 112 | 45.2% |
| (100, 150] | 47 | 19.0% |
| (150, 200] | 34 | 13.7% |
| (200, 250] | 50 | 20.2% |
| (250, 300] | 5 | 2.0% |

*Comprobación estadística realizada con el coeficiente Biserial puntual que arroja un resultado de 0.13 y un P Valor menor a 0.05

Efectivamente como devuelven los resultados de la investigación niveles altos de azúcar en sangre pueden fomentar el desarrollo de cuadros que deriven en ACV ya que puede dañar los vasos sanguíneos a lo largo del tiempo, debilitándolos y volviéndolos más propensos a la acumulación de placa (aterosclerosis). Esto estrecha las arterias y dificulta la circulación sanguínea, lo que aumenta el riesgo de coágulos y bloqueos

Algoritmos de Clasificación -Machine Learning-



En esta sección del documento abordaremos el uso de algoritmos de clasificación para intentar generar un modelo que nos permita identificar en qué casos NovaCare debe comenzar a implementar el protocolo de prevención de ACV según el paciente. Para esto someteremos los datos a diferentes pruebas e identificaremos cuál modelo obtiene un mejor rendimiento.

Dada la naturaleza de los datos y del objetivo de la investigación, los modelos elegidos para ser evaluados fueron los siguientes:

Cada modelo fue sometido a diversas pruebas para obtener el mejor resultado en cada iteración, los resultados mostrados en este documento corresponden a tres instancias, una utilizando variables originales, otra donde las variables continuas fueron convertidas a categóricas y finalmente utilizando un umbral personalizado para separar los casos positivos de los negativos.

Desbalance

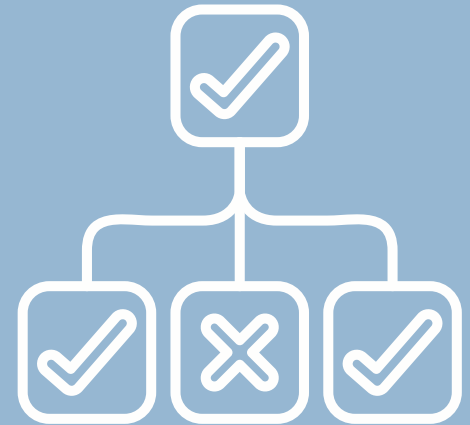
Hemos implementado técnicas de balanceo para evitar el sesgo, mejorar la capacidad predictiva, mejorar la detección de la clase minoritaria, evitar el sobre ajuste y no caer en un escenario de generalización deficiente. Las técnicas utilizadas fueron:

- Penalización de la clase mayoritaria
- Generación de datos sintéticos (SMOTE)
- Reducción de clase mayoritaria (Near Miss)



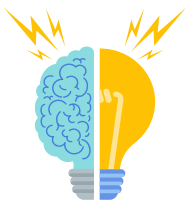
Algoritmos de Clasificación

-Random Forest-



| Random Forest | Casos | | | | | | | | |
|-------------------------------------------------------------------------------|-----------|-----|------------------|-----|-----------|-----|------------------|-----|-------------------------------------|
| Estrategia | Negativos | % | Falsos Negativos | % | Positivos | % | Falsos Positivos | % | Coficiente Predictivo (F1-Positivo) |
| Selección de variables , validación cruzada y optimización de hiperparámetros | | | | | | | | | |
| Balanced | 1031 | 72% | 17 | 21% | 64 | 79% | 398 | 28% | 24% |
| SMOTE | 1064 | 74% | 22 | 27% | 59 | 73% | 365 | 26% | 23% |
| Near Miss | 504 | 35% | 19 | 23% | 62 | 77% | 925 | 65% | 12% |
| Utilizando Categorización de variables | | | | | | | | | |
| Balanced | 966 | 68% | 13 | 16% | 68 | 84% | 463 | 32% | 22% |
| SMOTE | 971 | 68% | 13 | 16% | 68 | 84% | 458 | 32% | 22% |
| Near Miss | 787 | 55% | 10 | 12% | 71 | 88% | 642 | 45% | 18% |
| Utilizando Corte Optimo (Umbral personalizado) | | | | | | | | | |
| Balanced | 1164 | 81% | 22 | 27% | 59 | 73% | 265 | 19% | 29% |
| SMOTE | 1138 | 80% | 22 | 27% | 59 | 73% | 291 | 20% | 27% |
| Near Miss | 907 | 63% | 15 | 19% | 66 | 81% | 522 | 37% | 20% |

*Los % entre columnas estan calculados en base al total de negativos y positivos según corresponda, de tal manera, podemos interpretar los resultados como: "que % de los casos negativos o positivos representa dicho numero"



Insights

Utilizando el modelo de Random Forest observamos que se obtuvo el mejor resultado en la ultima iteración utilizando la penalidad sobre la clase mayoritaria, aunque esto esta sujeto a debate ya que si observamos la mejora proviene de un menor numero de falsos positivos, lo cual es un buen indicio pero sacrifica un aumento de 9 puntos en los falsos negativos, aunque este aumento esta sesgado si lo medimos en valores nominales ya que el numero es pequeño en comparación con el valor nominal en la reducción de los falsos positivos.

En consecuencia se deberá medir y ponderar que casos tiene un mayor impacto en los costos a futuro para la empresa.

Configuración del modelo

```

bootstrap: True
class_weight: None /Balanced
criterion: 'gini'
max_depth: 8
max_leaf_nodes: 11
min_samples_leaf: 8
min_samples_split: 2
n_estimators: 800
  
```



Algoritmos de Clasificación

-Regresión Logística-



| R. Logística | Casos | | | | | | | | |
|-------------------------------------------------------------------------------|-----------|-----|------------------|-----|-----------|-----|------------------|-----|---------------------------------------|
| Estrategia | Negativos | % | Falsos Negativos | % | Positivos | % | Falsos Positivos | % | Coefficiente Predictivo (F1-Positivo) |
| Selección de variables , validación cruzada y optimización de hiperparámetros | | | | | | | | | |
| Balanced | 1026 | 72% | 13 | 16% | 68 | 84% | 403 | 28% | 25% |
| SMOTE | 1026 | 72% | 15 | 19% | 66 | 81% | 403 | 28% | 24% |
| Near Miss | 987 | 69% | 22 | 27% | 59 | 73% | 442 | 31% | 20% |
| Utilizando Categorización de variables | | | | | | | | | |
| Balanced | 964 | 67% | 12 | 15% | 69 | 85% | 465 | 33% | 22% |
| SMOTE | 966 | 68% | 13 | 16% | 68 | 84% | 463 | 32% | 22% |
| Near Miss | 953 | 67% | 27 | 33% | 54 | 67% | 476 | 33% | 18% |
| Utilizando Corte Optimo (Umbral personalizado) | | | | | | | | | |
| Balanced | 1153 | 81% | 19 | 23% | 62 | 77% | 276 | 19% | 30% |
| SMOTE | 1174 | 82% | 23 | 28% | 58 | 72% | 255 | 18% | 29% |
| Near Miss | 980 | 69% | 28 | 35% | 53 | 65% | 449 | 31% | 18% |

*Los % entre columnas estan calculados en base al total de negativos y positivos según corresponda, de tal manera, podemos interpretar los resultados como: "que % de los casos negativos o positivos representa dicho numero"



Insights

En el caso del modelo de regresión logística vemos que no es sensible a la categorización de las variables ya que los resultados son los mismos, pero si presenta una mejoría al utilizar el corte óptimo en donde llega a tener una performance de 30% en el F1-Score de la clase Positiva superando al modelo Random Forest en su mejor resultado.

Cabe mencionar que no en todos los casos mejora el rendimiento ya que se observa que al utilizar Near Miss y corte optimo nos da como resultado la peor performance del modelo.

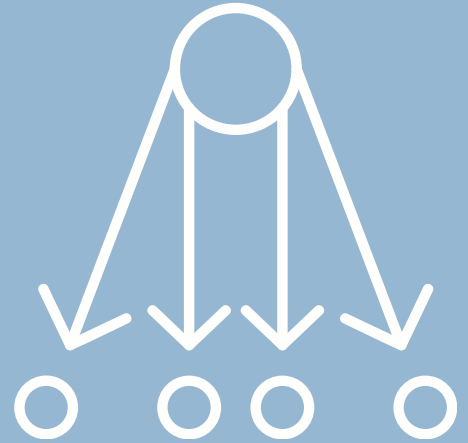
Configuración del modelo

```
C: 0.5
class_weight: None / Balanced
fit_intercept: True
intercept_scaling: 1
l1_ratio: None
max_iter: 1000
multi_class: 'auto'
penalty: 'l2'
solver: 'lbfgs'
```



Algoritmos de Clasificación

-KNN-



| KNN | Casos | | | | | | | | |
|-------------------------------------------------------------------------------|-----------|-----|------------------|-----|-----------|-----|------------------|-----|--------------------------------------|
| Estrategia | Negativos | % | Falsos Negativos | % | Positivos | % | Falsos Positivos | % | Coeficiente Predictivo (F1-Positivo) |
| Selección de variables , validación cruzada y optimización de hiperparámetros | | | | | | | | | |
| No Penalty | 1353 | 95% | 67 | 83% | 14 | 17% | 76 | 5% | 16% |
| SMOTE | 1140 | 80% | 26 | 32% | 55 | 68% | 289 | 20% | 26% |
| Near Miss | 945 | 66% | 21 | 26% | 60 | 74% | 484 | 34% | 19% |
| Utilizando Categorización de variables | | | | | | | | | |
| No Penalty | 1394 | 98% | 73 | 90% | 8 | 10% | 35 | 2% | 14% |
| SMOTE | 1384 | 97% | 72 | 89% | 9 | 11% | 45 | 3% | 13% |
| Near Miss | 997 | 70% | 36 | 44% | 45 | 56% | 432 | 30% | 16% |
| Utilizando Corte Optimo (Umbral personalizado) | | | | | | | | | |
| No Penalty | 1323 | 93% | 70 | 86% | 11 | 14% | 106 | 7% | 11% |
| SMOTE | 735 | 51% | 42 | 52% | 39 | 48% | 694 | 49% | 10% |
| Near Miss | 491 | 34% | 25 | 31% | 56 | 69% | 938 | 66% | 10% |

*Los % entre columnas estan calculados en base al total de negativos y positivos según corresponda, de tal manera, podemos interpretar los resultados como: "que % de los casos negativos o positivos representa dicho numero"



Insights

KNN se ve altamente impactado por el cambio de enfoque al utilizar todas variables categóricas, incluso en el caso de SMOTE en donde había tenido un resultado aceptable incluso superando a los otros modelos en cuanto a rendimiento predictivo en la primera iteración, no se desempeña de la misma forma al cambiar el enfoque de las variables donde se ve altamente superado por los otros dos modelos que performan por encima incluso de su mejor valor.

No se recomendaría al menos en esta instancia y con estos datos avanzar con este modelo.

Configuración del modelo

No Penalty
leaf_size: 1
n_neighbors: 1
p: 1
weights: 'distance'

SMOTE
leaf_size: 1
n_neighbors: 10
p: 1
weights: 'distance'

Near Miss
leaf_size: 1
n_neighbors: 8
p: 2
weights: 'uniform'



Conclusiones



A lo largo de este informe pudimos comprobar como algunas de las hipótesis planteadas por NovaCare fueron confirmadas desde el punto de vista estadístico, con esta información es posible llevar a cabo un plan de prevención que tenga como objetivo atacar los puntos centrales que pueden generar cuadros de accidentes cerebrovasculares y en consecuencia un aumento en los gastos médicos por parte del paciente.

En este contexto recomendamos a NovaCare la implementación de un plan preventivo sobre los individuos que se ajusten al siguiente perfil:

- Personas mayores de 20 años (ya que los casos comienzan a partir de los 35 y es imperioso que los hábitos se desarrollen con tiempo)
- Residencia indistinta por lo que el plan puede implementarse en cada centro medico propio o afiliado
- Personas con cuadros de diabetes, problemas cardíacos y/o hipertensas
- Considerar además personas con antecedentes de las afecciones mencionadas en el punto anterior ya que esta probado que suelen ser hereditarias.

Se recomienda un plan integral que consista en mejorar la calidad de vida de las personas desde una temprana edad que consista en mejora en la alimentación, mejora en la actividad física y controles periódicos sobre posible aparición de afecciones cardíacas.

Conclusiones del modelo

Hablando de los hallazgos estadísticos y de modelo predictivo, el mismo necesita de una mayor cantidad y calidad de datos, se recomienda a NovaCare la recolección de datos mas granulares sobre todo en lo que concierne a los valores de presión y enfermedades y condiciones cardíacas de los pacientes ya que esto podría darnos mas herramientas para alimentar un modelo que mejore la predicción sin tener que lidiar con tantos casos de falsos positivos.

La prevención es el camino hacia la mejor calidad de vida, y juntos podemos hacer la diferencia.

