"Ciencia de datos aplicada al estudio de la Obesidad y otras enfermedades crónicas en Córdoba"

Práctico N°3 Ejercicios de Introducción al aprendizaje supervisado

Integrantes del grupo X:

Basmadjian, Osvaldo Martín

Fernández, María Emilia

Romero, Fernando

Metodología

Creación de la variable TARGET

En primer lugar, creamos una variable capaz de codificar como 1 los valores de IMC mayores a 29.9 (personas con cualquier grado de obesidad) y como 0 los valores por debajo de dicho umbral (persona no obesa). De modo que los valores predichos como positivos corresponderían a la etiqueta obeso y los negativos a no obeso. Previo a esta clasificación eliminamos valores NaN de la variable IMC (114).

Selección de features

La selección de features se realizó en base a la relación existente entre la variable IMC (como variable continua, no codificada en 0 y 1) y cada una de las posibles predictoras (features). Dicha relación se exploró visualmente a través de gráficos de puntos para toda la población y para cada uno de los sexos por separado. Las predictoras incluidas fueron aquellas que guardaban relación lineal (visual) con IMC o bien, cuyos valores se concentraban marcadamente por encima o debajo del umbral de IMC >29.9. No hicimos una limpieza de las predictoras que pudieran estar altamente correlacionadas. Las predictoras finalmente incluidas fueron:

<u>Para toda la población:</u>

ic, dbt, tipo1, tipo2, tratadbt, mets, imagen_percibida, cc, ten2max, ten2min, vis, frus, frud, leg, past, azuc, agua, cpre

<u>Para personas de sexo femenino:</u>

dbt, tipo1, tipo2, tratadbt, mets, imagen_percibida, cc, ten2max, ten2min, vis, frus, frud, leg, agua, cpre

<u>Para personas de sexo masculino:</u>

ec, anther, dbt, tipo1, tipo2, tratadbt, tumben, mets, imagen_percibida, cc, ten2max, ten2min, vis , fvis, fibfrua, frus, frud, leg, (past), azuc, agua, cpre.

Evaluamos la cantidad de valores NaN por clase de la variable TARGET para cada uno de los atributos seleccionados. La mayoría de ellos presentaban a lo sumo un valor NaN, excepto por las variables mets y Gcarhue que presentaron 362 y 74 valores NaN, respectivamente. Como estos valores representaban menos del 8 % y 2% del dataset, consideramos conveniente eliminar dichas filas del dataset para poder utilizarlas como predictoras (en vez de no utilizar esas variables como predictoras por presentar algunos pocos valores NaN).

Construcción de los datasets y división de los mismos en conjuntos de entrenamiento y evaluación

Construimos tres datasets independientes, uno para toda la población, uno para las personas de sexo femenino y un tercero para las de sexo masculino.

Dado que encontramos asimetría en la cantidad de personas pertenecientes a cada una de la clases de la variable TARGET (3564 registros para la clase 0 y 614 registros para la clase 1), utilizamos el argumento “stratify” en el método de división del dataset para que se mantenga dicha proporción entre clases en los conjuntos de entrenamiento y evaluación. Luego, estuvimos en condiciones de dividir el dataset de toda la población, el de personas de sexo femenino y el de personas de sexo masculino en los conjuntos de entrenamiento y evaluación (0.80 de los datos se utilizaron para el entrenamiento y 0.20 para la evaluación, en todos los casos).

Elección de modelos de clasificación

Existen varios algoritmos de clasificación binaria, los que consideramos fueron los siguientes:

- **Descenso de Gradiente Estocástico**: Es simple y una muy buena aproximación para modelos lineales. Muy utilizado cuando el número de muestras es muy grande, y soporta un conjunto de parámetros para ajustar el modelo como corresponda.
- **Naive Bayes**: Basado en el teorema de Bayes, con la premisa de que las features predictoras son independientes entre sí
- **Regresión Logística**: El modelo se basa en probabilidades utilizando un modelo de regresión para predecir el valor.
- **Árboles de Decisión**: Genera un conjunto de reglas secuenciales para clasificar los datos.
- **K-Vecinos más cercanos**: No intenta generar un modelo, sino que intenta predecir el valor de una entrada en base a los valores de los vecinos más cercanos.

En base a esta lista decidimos utilizar los modelos de árboles de decisión y de descenso de gradiente Estocástico.

Elección de la función de regularización:

- Descenso de gradiente estocástico: dado que la regularización L1 tiende a desestimar aquellas features menos importantes y nuestro criterio de selección de features se realizó únicamente de manera visual consideramos que la regulación L1 era más adecuada para generar el modelo preliminar de descenso de gradiente estocástico.
- Árboles de decisión: existen dos métodos de regularización básicos para árboles de decisión: determinar la profundidad del árbol o la cantidad mínima de muestras por hoja. En este caso al no tener una noción aproximada de cuántas ramas se podrían generar a

partir del árbol de decisión con nuestro set datos, preferimos regular la cantidad mínima de muestras por hoja imponiendo un valor arbitrario de 100.

Selección de Hiperparametros 💬

Para la selección de hiperparametros, revisamos la documentación disponible de cada modelo de clasificación, seleccionando los que creímos más convenientes. Luego realizamos un GridSearch, que a su vez, de acuerdo a la documentación, realiza un 5-fold, variando los datos de entrenamiento y obteniendo mejores métricas. Utilizamos como estimador “accuracy”, que se define como la distancia entre el punto predicho y el punto real. Para problemas de clasificación binaria, “accuracy” se define como la cantidad de veces que el modelo produjo una etiqueta correcta.

Descenso por gradiente estocástico

Los hiperparametros variados fueron:

- **Loss**: determina la función de loss, pudiendo tomar los valores ‘**Hinge**’, que genera un support vector machine lineal, ‘**log**’ que genera un modelo de regresión logística, ‘**perceptrón**’ además de otras opciones.
- **Penalty**: conocido como el término de regularización.
- **Learning rate**: selecciona la tasa de aprendizaje.
- **Eta0**: selecciona la tasa de aprendizaje inicial.
- **Max_iter**: indica la cantidad máxima de iteraciones, después de este número, si el modelo no converge, arroja una advertencia.
- **Alpha**: constante que multiplica el término de regularización.

El mejor modelo tanto para toda la población como para mujeres y hombres se logró con los siguientes parámetros:

	Todos	**Mujeres**	**Hombres**
Loss	hinge	log	log
Penalty	elasticnet	l1	elasticnet
Learning rate	adaptive	adaptive	adaptive
Eta0	0.1	0.1	0.1
Max_iter	20000	20000	20000
Alpha	0.01	0.1	0.1

Árboles de decisión

Los hiperparametros variados fueron:

- **Criterion**: La función para medir la calidad de la división.
- **Splitter**: La estrategia utilizada para la separación en cada nodo.
- **Max depth**: La profundidad máxima del árbol, por defecto no tiene valor máximo, así que el árbol se extenderá lo que necesite para ajustar los datos.
- **Min_sample_leaf**: La cantidad mínima de muestras para que sea considerada una hoja.
- **Max_features**: La cantidad máxima de features al considerar para hacer la división.

El mejor modelo tanto para toda la población como para mujeres y hombres se logró con los siguientes parámetros:

	Todos	**Mujeres**	**Hombres**
Criterion	gini	gini	gini
Splitter	random	random	best
Max depth	11	11	4
Min_sample_leaf	31	13	17
max_features	None	None	auto

Resultados

A continuación presentamos los resultados de las predicciones de las 3 poblaciones.

Población Completa

Al comparar las métricas con hiperparámetros mejorados se observan resultados comparables para ambos modelos. El modelo de descenso de gradiente estocástico muestra un valor de “accuracy” ligeramente superior, junto con mayores valores en “recall” y “F1 Score” (tablas 1 y 2).

Dado que por la naturaleza del problema las predicciones falsas negativas serían las más costosas (que se clasifique a una persona como no obesa cuando sí lo es), es valioso que sólo el 5% de los datos hayan sido mal clasificados como falsos negativos.

Descenso de Gradiente estocástico

	Train	Test	Test (hiperparámetros mejorados)
Accuracy	0.83	0.82	0.91
Precision	0.45	0.44	0.70
Recall	0.56	0.56	0.68
F1 Score	0.50	0.49	0.69
True Negatives	2259	562	612
False Negatives	207	51	38
True Positives	261	66	79
False Positives	324	84	34

Tabla 1. Resultados del modelo de descenso de gradiente estocástico en la población completa**.**

Árbol de decisión de toda la población			
	Train	Test	Test (hiperparámetros mejorados)
Accuracy	0.90	0.90	0.90
Precision	0.73	0.71	0.70
Recall	0.56	0.59	0.59
F1 Score	0.63	0.64	0.64
True Negatives	2486	618	616
False Negatives	206	48	48
True Positives	262	69	69
False Positives	97	28	30

Tabla 2. Resultados del modelo de árboles de decisión en la población completa**.**

Población Mujeres

De manera similar a lo observado en el análisis de la población completa, al analizar la muestra de mujeres se observan métricas ligeramente superiores con el modelo de descenso de gradiente estocástico. En este sentido, al evaluar ambos modelos con los hiperparámetros mejorados observamos un valor de “accuracy” levemente superior en el modelo de descenso por gradiente estocástico, junto con mayores valores en “recall” y F1 Score”.

Para esta población se observa un 7% de error de predicciones a falsos negativos.

Descenso de Gradiente estocástico (mujeres)			
	Train	Test	Test (hiperparámetros mejorados)
Accuracy	0.82	0.84	0.91
Precision	0.34	0.38	0.73
Recall	0.20	0.16	0.60
F1 Score	0.25	0.23	0.66
True Negatives	1438	367	370
False Negatives	214	56	27
True Positives	53	11	40
False Positives	103	18	15

Tabla 3. Resultados del modelo de descenso de gradiente estocástico en la población de mujeres**.**

Árbol de decisión (mujeres)			
	Train	Test	Test (hiperparámetros mejorados)
Accuracy	0.90	0.88	0.90
Precision	0.71	0.67	0.73
Recall	0.5	0.43	0.54
F1 Score	0.58	0.53	0.62

True Negatives	1486	371	372
False Negatives	134	38	31
True Positives	133	29	36
False Positives	55	14	13

Tabla 4. Resultados del modelo de árboles de decisión en la población de mujeres**.**

Población Hombres

A diferencia de lo analizado anteriormente, en la población masculina se observaron mejoras métricas con el modelo de árboles de decisión. En este sentido, el valor de “accuracy” fue ligeramente superior en dicho modelo y se observaron mayores valores de “recall” y “F1 score”.

Descenso de Gradiente estocástico (hombres)			
	Train	Test	Test (hiperparámetros mejorados)
Accuracy	0.83	0.81	0.88
Precision	0.48	0.39	0.63
Recall	0.32	0.29	0.63
F1 Score	0.39	0.33	0.63
True Negatives	951	233	237
False Negatives	132	35	18
True Positives	63	14	31
False Positives	69	22	18

Tabla 5. Resultados del modelo de descenso de gradiente estocástico en la población de hombres**.**

Árbol de decisión (hombres)			
	Train	Test	Test (hiperparámetros mejorados)

Accuracy	0.90	0.88	0.90
Precision	0.70	0.63	0.67
Recall	0.66	0.63	0.73
F1 Score	0.68	0.63	0.70
True Negatives	965	237	237
False Negatives	66	18	13
True Positives	129	31	36
False Positives	55	18	18

Tabla 6. Resultados del modelo de árboles de decisión en la población de hombres.