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

Práctico N°4 Ejercicios de Aprendizaje Supervisado

Integrantes del grupo X:

Basmadjian, Osvaldo Martín

Fernández, María Emilia

Romero, Fernando

Metodología

Selección de features

La selección de features se realizó en base a la relación visual y a la esperada por bibliografía para cada una de ellas con la variable target HTA, además de las exploraciones realizadas en prácticos anteriores. A continuación se indican las predictoras finalmente incluidas:

edad: edad en años. La probabilidad de padecer hipertensión aumenta con la edad debido a la interacción de numerosos factores como la pérdida de elasticidad de grandes arterias, el incremento de estímulos vasoconstrictores y aumento de grosor de la capa media arterial entre otros.

Factores relacionados al síndrome metabólico: este síndrome incluye varios factores interrelacionados que en su conjunto aumentan el riesgo de desarrollo de diabetes e hipertensión por lo que los consideraremos para generar el modelo predictivo de hipertensión arterial. Entre estos factores se encuentran la insulinorresistencia, diabetes, obesidad, alteración de los niveles sanguíneos de colesterol y triglicéridos. Dada la disponibilidad de datos, utilizaremos los siguientes features:

• **dbt**: presencia o no de diabetes.

• **tipo1**: padece diabetes insulinodependiente.

• **tipo2**: padece diabetes no insulinodependiente.

• **tratadbt**: tipo de tratamiento de diabetes.

• **dlp**: presencia de dislipemia.

• **cc**: circunferencia de cintura.

Actividad física: dado que la actividad física desencadena eventos neurohumorales que llevan a la reducción posterior del gasto cardíaco y de la resistencia vascular periférica, su práctica frecuente genera una disminución de la probabilidad de padecer hipertensión arterial. Por ello vamos a tener en cuenta los siguientes features:

• **actfis**: realiza actividad física.

• **nombre**: nombre de la actividad que realiza.

• **act**: actividad que realiza.

• **durac**: duración de la actividad física en minutos .

• **mets**: nivel de actividad física desde sedentarios a muy activos

stress: dado que el estrés genera un estado de hiperactivación simpática con el consiguiente aumento en la liberación de corticoides endógenos estos factores aumentan el riesgo de hipertensión y resistencia a la insulina por lo que vamos a tener en cuenta la variable estrés que computa el estrés percibido por el individuo.

fuma: tendremos en cuenta este feature debido a que el tabaquismo aumenta el riesgo de sufrir hipertensión arterial por múltiples mecanismos como el aumento del tono simpático y formación de radicales libres con la consiguiente disminución de los niveles de óxido nítrico (uno de los principales mediadores vasodiladores).

Features asociados a la alimentación que pueden aumentar el riesgo de dislipemias, obesidad y la consiguiente generación de hipertensión arterial:

• **fve1 y fve2**: valor energético.

• **fgr1**: ingesta de lípidos.

• **fgs1 y fgs2**: ingesta de grasas saturadas.

• **fgins1 y fgins2**: ingesta de grasas insaturadas.

• **fgp1 y fgp2**: ingesta de grasas poliinsaturadas.

• **fgm1 y fgm2**: ingesta de grasas monoinsaturadas.

• **fcol1 y fcol2**: ingesta de colesterol.

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

Construimos dos datasets independientes, uno para toda la población y otro para las personas de sexo femenino, separando, en cada caso, del dataset la variable target HTA.

Elección de modelos y meta-modelos de clasificación

Modelos de clasificación

Evaluamos el desempeño de cada uno de los siguientes modelos de clasificación binaria con sus parámetros por defecto:

- **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.
- **Multinomial Naive Bayes**: Basado en el teorema de Bayes, con la premisa de que las features predictoras son independientes entre sí

- **Perceptron**: permite resolver problemas de clasificación linealmente separables, empleando poco tiempo de entrenamiento y evaluación.
- **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.
- **Dummy Classifier:** lo utilizamos como baseline para comparar su rendimiento con los demás clasificadores. Este clasificador bajo la estrategia **"Dummy_stratified"** genera predicciones respetando la distribución de clases en el conjunto de entrenamiento. Mientras que en la estrategia **"Dummy_frequent"** siempre predice la etiqueta más frecuente en el conjunto de entrenamiento.

Meta-modelos de clasificación

- **Random Forest:** es un meta-estimador que ajusta varios clasificadores de árboles de decisión en varias submuestras del conjunto de datos y usa promedios para mejorar la exactitud (accuracy) de la predicción y controlar el sobreajuste.
- **Voting classifier:** este clasificador combina clasificadores diferentes conceptualmente (en nuestro caso árboles de decisión, gradiente estocástico de descenso y random forest) y utiliza las probabilidades promedio predichas (con la estrategia soft que nosotros utilizamos) para predecir las etiquetas. Un clasificador de este tipo puede ser útil para un conjunto de modelos con rendimientos buenos y semejantes permitiendo equilibrar sus debilidades individuales.

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.
- Random Forest: Al ser un conjunto de árboles consideramos que el la función de regularización para este modelo es n_estimators, que es la cantidad de árboles presente en el modelo.

Selección de Hiperparametros

Luego de explorar varios modelos con hiperparámetros por defecto, elegimos aquellos que mostraron los mejores rendimiento para realizar la selección de sus hiperparametros y poder obtener rendimientos aún mejores. Para este fin, revisamos la documentación disponible de cada modelo de clasificación, y seleccionamos aquellos hiperparámetros que creímos más convenientes por su efecto en el rendimiento del modelo cuando los evaluamos individualmente. Luego realizamos tanto RandomizedSearch como GridSearch (con 5-fold cross validation), para obtener múltiples combinaciones de parámetros y seleccionar aquel modelo y conjunto de hiperparámetros con las 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

En este modelo, 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.
- **Alpha**: constante que multiplica el término de regularización.
- **Warm_start**: Inicializa las variables con el resultado de la predicción del modelo anterior.
- **Average**: Calcula el promedio de todas las actualizaciones de las iteraciones y almacena el resultado.

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

Tabla 1. Hiperparámetros del mejor modelo de descenso por gradiente estocástico para cada una de las poblaciones estudiadas

Hiperparametros SDG		
	Mujeres	Todos
Loss	log	squared_hinge
Penalty	L1	L1
alpha	1,00E-06	1,00E-06
warm_start	True	True
average	True	True

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.
- **Min_samples_split**: La cantidad mínima de samples para subdividir un nodo interno.
- **Max_features**: La cantidad máxima de features al considerar para hacer la división.

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

Tabla 2. Hiperparámetros del mejor modelo de árboles de decisión para cada una de las poblaciones estudiadas

Hiperparametros Desicion Tree		
	Mujeres	Todos
criterion	gini	gini
splitter	random	random
max_depth	6	5
min_samples_split	17	1
min_samples_leaf	100	2
max_features	None	None

Random Forest

Los hiperparametros variados fueron:

- **n_estimators:** La cantidad de árboles a generar.
- **criterion:** La función para medir la calidad de la división
- **warm_start:** Inicializa las variables con el resultado de la predicción del modelo anterior.
- **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_samples_split:** La cantidad mínima de samples para subdividir un nodo interno.
- **min_samples_leaf:** La cantidad mínima de samples para subdividir un nodo interno.

- **max_features:** La cantidad máxima de features al considerar para hacer la división.

Los mejores resultados se obtuvieron con los siguientes hiperparametros:

Tabla 3. Hiperparámetros del mejor meta-modelo de Random Forest para cada una de las poblaciones estudiadas

Hiperparametros Random Forest		
	Mujeres	Todos
n_estimators	120	130
criterion	entropy	entropy
warm_start	True	True
max_depth	20	30
min_samples_split	2	2
min_samples_leaf	3	3
max_features	auto	log2

Voting

Dada la naturaleza del meta-modelo voting, decidimos no hacer ajuste de hiperparámetros para el mismo

Resultados

A continuación presentamos los resultados de las predicciones de las 2 poblaciones.

Población Mujeres

En las tablas 4-7 se presentan las métricas correspondientes a los mejores modelos y meta-modelos obtenidos por defecto y con selección de hiperparámetros por GridSearch para las personas de sexo femenino.

Como se esperaba, en los modelos de clasificación de árboles de decisión (Tabla 4) y de descenso por gradiente estocástico (Tabla 5), la selección de hiperparámetros condujo a un incremento del rendimiento de los modelos. Interesantemente, el rendimiento del meta-modelo random forest no mejoró tras la selección de hiperparámetros.

Considerando los hiperparámetros por defecto, los mejores valores de accuracy se obtuvieron con los meta-modelos, tomando valores de 0.77 y 0.76 en el caso del random forest y el voting, respectivamente (Tablas 6 y 7 respectivamente). Mientras que con selección de hiperparámetros, el mejor modelo fue el random forest, seguido por el voting y el modelo de descenso por gradiente estocástico que mostraron valores de accuracy idénticos. A su vez, los modelos que presentaron una menor proporción de falsos negativos (i.e. probabilidad de ser hipertenso pero ser clasificado como no hipertenso) tanto utilizando

los hiperparámetros por defecto como luego de la selección de los mismos fueron el modelo de descenso por gradiente estocástico y el meta-modelo de voting.

Tabla 4. Métricas por defecto y con selección de hiperparámetros por GridSearch del mejor modelo de árboles de decisión obtenido para las personas de sexo femenino de la población

Decision Tree Mujeres		
	Por defecto	Con selección de Hiperparámetros
Accuracy	0.68	0.74
Precision	0.8	0.82
Recall	0.74	0.8
F1 Score	0.77	0.81
True Negatives	244	263
False Negatives	85	66
True Positives	61	55
False Positives	78	84

Tabla 5. Métricas por defecto y con selección de hiperparámetros del mejor modelo de descenso por gradiente estocástico obtenido para las personas de sexo femenino de la población

SGD Mujeres		
	Por defecto	Con selección de Hiperparámetros
Accuracy	0.70	0.76
Precision	0.70	0.79
Recall	0.99	0.89
F1 Score	0.82	0.84
True Negatives	327	295
False Negatives	2	34
True Positives	134	78
False Positives	5	61

Tabla 6. Métricas por defecto y con selección de hiperparámetros del mejor meta-modelo de Random Forest obtenido para las personas de sexo femenino de la población

Random Forest Mujeres		
	Por defecto	Con selección de Hiperparámetros
Accuracy	0.77	0.77
Precision	0.82	0.81
Recall	0.87	0.87
F1 Score	0.84	0.84
True Negatives	288	289
False Negatives	41	40
True Positives	63	66
False Positives	76	73

Tabla 7. Métricas por defecto del meta-modelo de Voting obtenido para las personas de sexo femenino de la población

Voting	
	Mujeres
Accuracy	0.76
Precision	0.78
Recall	0.90
F1 Score	0.84
True Negatives	299
False Negatives	30
True Positives	81
False Positives	58

Toda la población

En las tablas 8-11 se presentan las métricas correspondientes a los mejores modelos y meta-modelos obtenidos por defecto y con selección de hiperparámetros por GridSearch para toda la población.

De modo semejante a lo descripto para la población de sexo femenino, los modelos de evaluados mostraron un mejor rendimiento tras la selección de hiperparámetros respecto al rendimiento obtenido con los hiperparámetros por defecto. Es interesante destacar que las diferencias en el rendimiento de los modelos con los hiperparámetros por defecto y luego de la selección de los mismos fueron de menor magnitud para toda la población respecto a las observadas para la población de sexo femenino (por ejemplo, el rango del incremento de accuracy en la población de sexo femenino observado con los diferentes modelos fue: 0.00-0.06; mientras que el rango del incremento de accuracy en toda la población observado con los diferentes modelos fue: 0.01-0.04).

Tras la selección de hiperparámetros, todos los modelos obtuvieron valores semejantes de accuracy, tomando valores de 0.72 en el random forest (Tabla 10), 0.71 en el voting y el descenso por gradiente (Tabla 11 y 9, respectivamente), y 0.70 en el árbol de decisión (Tabla 8). Cabe mencionar que la menor magnitud de los valores de accuracy observados para toda la población respecto a los obtenidos para las personas de sexo femenino podrían ser indicadores de que las features utilizadas son más eficientes para la predicción en la población de sexo femenino. De modo que podrían explorarse otras features para realizar la predicción en toda la población y evaluar si el rendimiento de los modelos mejora.

Por otro lado, los modelos que presentaron una menor proporción de falsos negativos (i.e. probabilidad de ser hipertenso pero ser clasificado como no hipertenso) tanto utilizando los hiperparámetros por defecto como luego de la selección de los mismos fueron el modelo de descenso por gradiente estocástico y el meta-modelo de voting, al igual que en las predicciones realizadas para las personas de sexo femenino.

Tabla 8. Métricas por defecto y con selección de hiperparámetros del mejor modelo de árboles de decisión obtenido para toda la población

Dicision Tree Todos		
	Por defecto	Con selección de Hiperparámetros
Accuracy	0.69	0.7
Precision	0.77	0.77
Recall	0.76	0.78
F1 Score	0.77	0.78
True Negatives	403	415
False Negatives	124	122

True Positives	117	120
False Positives	141	138

Tabla 9. Métricas por defecto y con selección de hiperparámetros del mejor modelo de descenso por gradiente estocástico obtenido para toda la población

SGD Todos		
	Por defecto	Con selección de Hiperparámetros
Accuracy	0.67	0.71
Precision	0.70	0.74
Recall	0.89	0.86
F1 Score	0.78	0.80
True Negatives	472	458
False Negatives	55	69
True Positives	199	156
False Positives	59	102

Tabla 10. Métricas por defecto y con selección de hiperparámetros del mejor modelo de random forest obtenido para toda la población

Random Forest Todos		
	Por defecto	Con selección de Hiperparámetros
Accuracy	0.71	0.72
Precision	0.76	0.77
Recall	0.83	0.84
F1 Score	0.79	0.80
True Negatives	439	447
False Negatives	88	80
True Positives	139	134
False Positives	119	124

Tabla 11. Métricas por defecto del meta-modelo de Voting obtenido para toda la población

Voting	
	Todos
Accuracy	0.71
Precision	0.74
Recall	0.87
F1 Score	0.80
True Negatives	459
False Negatives	68
True Positives	160
False Positives	98

Conclusiones

Si tuviéramos que seleccionar un único modelo, el de descenso por gradiente estocástico otenido tras la selección de hiperparámetros resultaría el más conveniente resolver el problema de clasificación tratado en este práctico para la población de sexo femenino a fin de. Mientras que para toda la población, el meta-modelo de voting sería el óptimo.