Análisis Predictivo - Final

Ezequiel Shinzato

Detección de fumadores

Oportunidad

- Comprender los efectos que produce el tabaquismo en el cuerpo humano.

- Contribución a la detección de acción de fumar en tiempo real (aplicaciones, sensores).

Objetivo

Generar un modelo predictivo que se base en datos de la salud del individuo para detectar si es fumador o no.

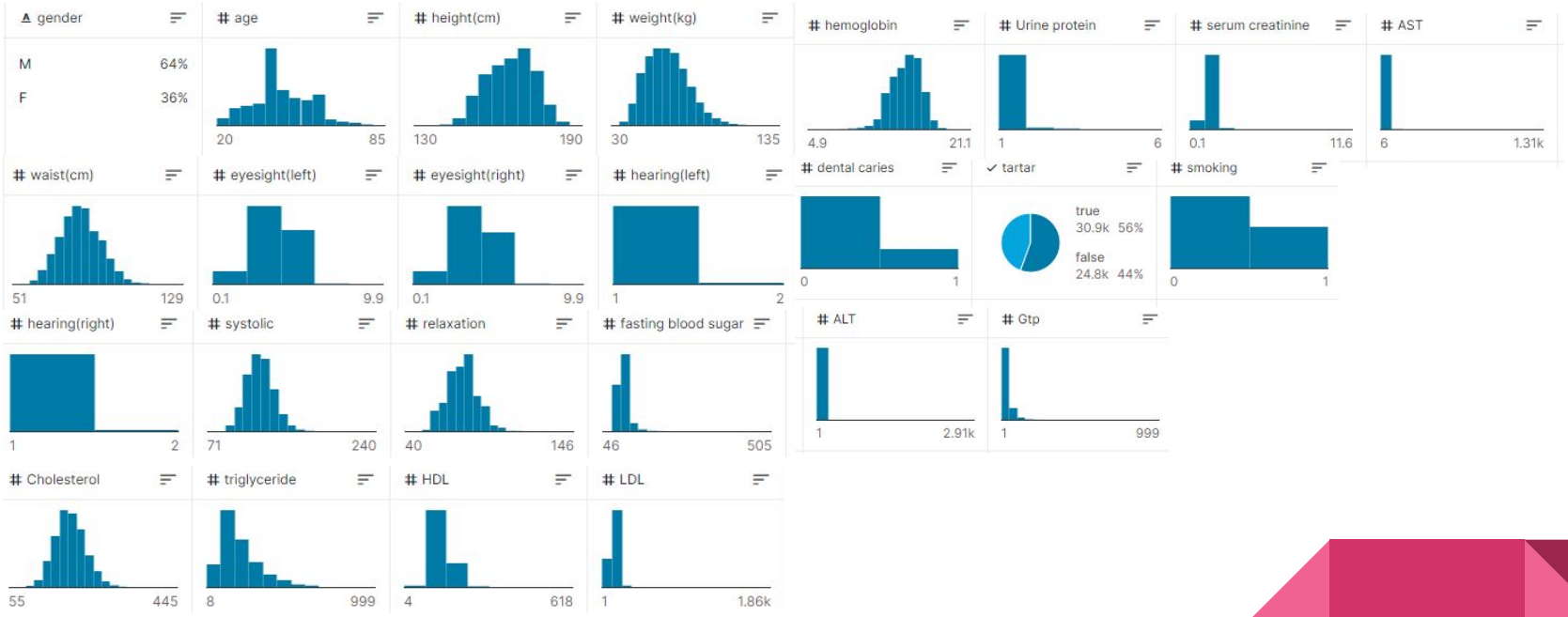
Dataset

Variable a predecir: Smoking (Boolean)

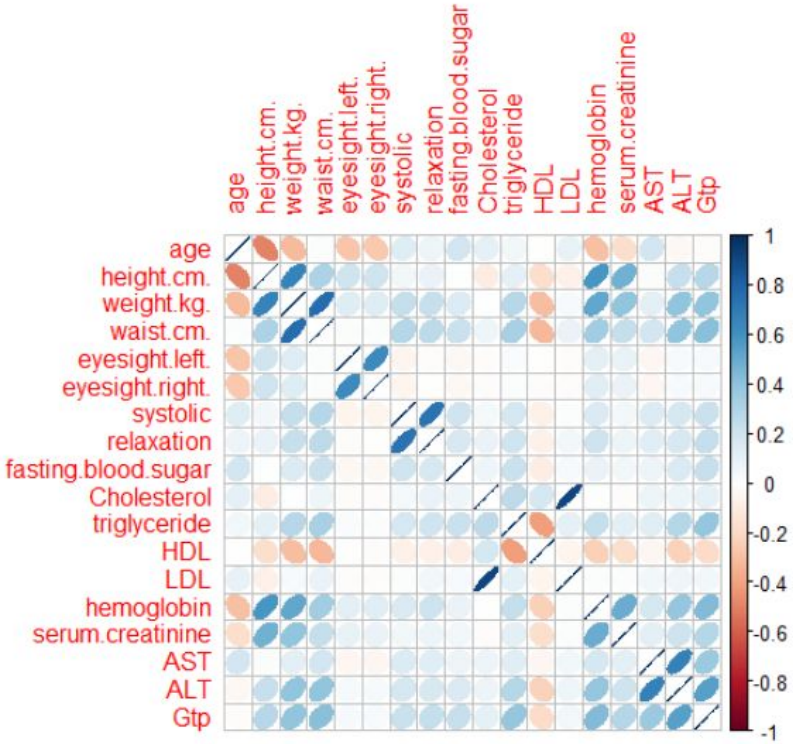
Variables predictoras (24): Género, Edad, Altura, Peso, Diámetro de cintura, Visión del ojo izq., Visión del ojo der., Audición del oído izq. (boolean), Audición del oído der. (boolean), presión sistólica, Presión diastólica, Azúcar en sangre, Colesterol, Triglocéridos, Colesterol bueno, Colesterol malo, Hemoglobina, Proteína en la orina, suero de creatinina, AST, ALT, GTP, Caries (boolean), sarro (boolean)

28279 registros de pacientes. 9322 fumadores.

Histogramas



Matriz de correlaciones



Modelos

- **Naive Bayes**
- **Red Neuronal**
- **Árbol de Decisión**
- **Random Forest**
- **Support Vector Machine**

Naive Bayes

Cross Validation (10 folds)

Laplace = 0

Usekernel = F

ACC: 0.72

SENS: 0.81

ESP: 0.67

		Esperado	
Predicho		**Fumador**	**No fumador**
	Fumador	26.8	21.9
	No fumador	6.2	45.2

Red Neuronal

Cross Validation (10 folds)

1 capa oculta, con 10 neuronas.

decay=0.5

ACC: 0.77

SENS: 0.68

ESP: 0.82

	Esperado		
Predicho		**Fumador**	**No fumador**
	Fumador	22.6	12.2
	No fumador	10.4	54.8

Support Vector Machine

Cross Validation (10 folds)

Kernel radial

Sigma=0.2

Cost=1.4

ACC: 0.8

SENS: 0.65

ESP: 0.87

	Esperado		
Predicho		**Fumador**	**No fumador**
	Fumador	21.3	8.4
	No fumador	11.7	58.7

Random Forest

Método OOB

MTRY=4

ACC: 0.86

SENS: 0.79

ESP: 0.9

		Esperado	
Predicho		**Fumador**	**No fumador**
	Fumador	7479	1843
	No fumador	1991	16966

Variables más importantes

- Género
- Hemoglobina
- GTP
- Altura
- Triglicéridos
- Ancho de cintura
- Colesterol
- Colesterol malo
- ALT

Árbol de Decisión

Split train/test 0.75/0.25

Árbol profundo, análisis de optimización de CP y poda.

CP=0.00014302059496567 5

ACC: 0.79

SENS: 0.66

ESP: 0.86

	Esperado		
Predicho		**Fumador**	**No fumador**
	Fumador	1532	669
	No fumador	798	4070

Comparación de modelos

	Modelos					
Resultados		**NV**	**RN**	**SVM**	**RF**	**AdD**
	ACC	0.72	0.77	0.8	0.86	0.79
	SENS	0.81	0.68	0.65	0.79	0.66
	ESP	0.67	0.82	0.87	0.9	0.86

Conclusiones

- Los modelos de árboles fueron los que obtuvieron mejores resultados.
- Naive Bayes tuvo mejores ratios de sensibilidad, pero peores de especificidad.
- Las variables género, hemoglobina, GTP y altura fueron las más importantes para el modelo final.

Muchas gracias!