Universitat Oberta de Catalunya

Análisis de la Encuesta de Salud Nacional y Examen de Nutrición de Estados Unidos (NHANES) usando Machine Learning

María José Crespo Estévez

Máster Universitario en Ciencias de Datos (Data Science)

Área: Minería de datos y Machine Learning

Tutora**: Laia Subirats Maté**

Director**: Jordi Casas Roma**

Junio de 2019

Índice

1. Introducción

1.1. Contexto y Justificación del Trabajo

- Área de la medicina.
  - Diagnósticos precoces.
  - Visibilidad a enfermedades raras
  - Predicción de epidemias, creación de alertas.
  - Apoyo a diagnósticos y tratamientos personalizados.
- País más saludable del mundo [1]
- Aportación a mejorar la calidad de vida y la salud.
- Inspiración para investigaciones futuras.
- Conjunto de datos *National Health and Nutrition Examination Survey* en Kaggle [2]
- Programa de estudios diseñados para evaluar la salud y estado de nutrición de adultos y niños de Estados Unidos [3]

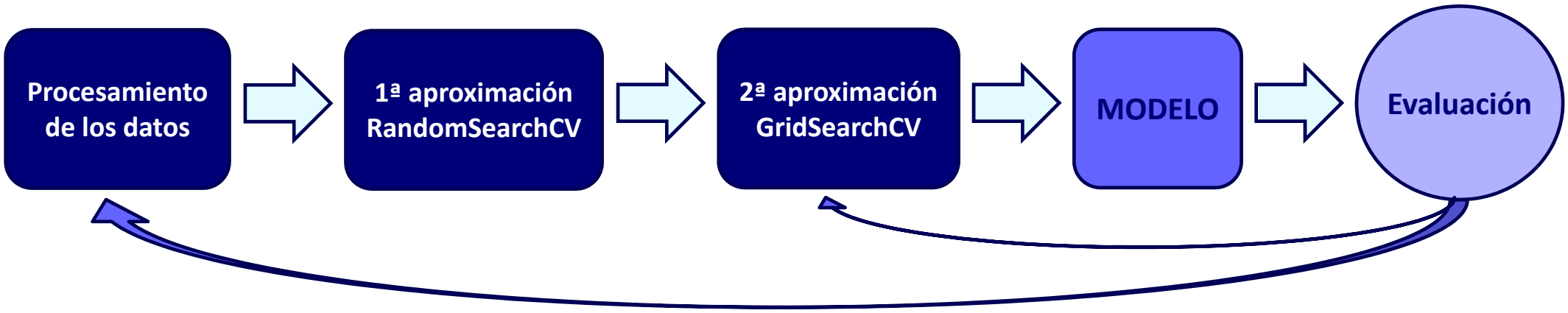
1.2. Objetivos del Trabajo

- Objetivos principales.
  - ✓ Modelo no supervisado y descubrir patrones.
  - ✓ Modelo predictivo sobre una enfermedad del dataset.
  - ✓ Comorbilidades entre enfermedades.

- Objetivos secundarios.
  - ✓ Reducir la dimensionalidad.
  - ✓ Interpretar los resultados.
  - ✓ Comprobar que no se haya producido sobreentrenamiento.
  - ✓ Evaluar y mejorar la precisión mediante un proceso iterativo.

1.3. Enfoque y Método Seguido

- Desarrollo de un producto nuevo.
- Proceso iterativo.



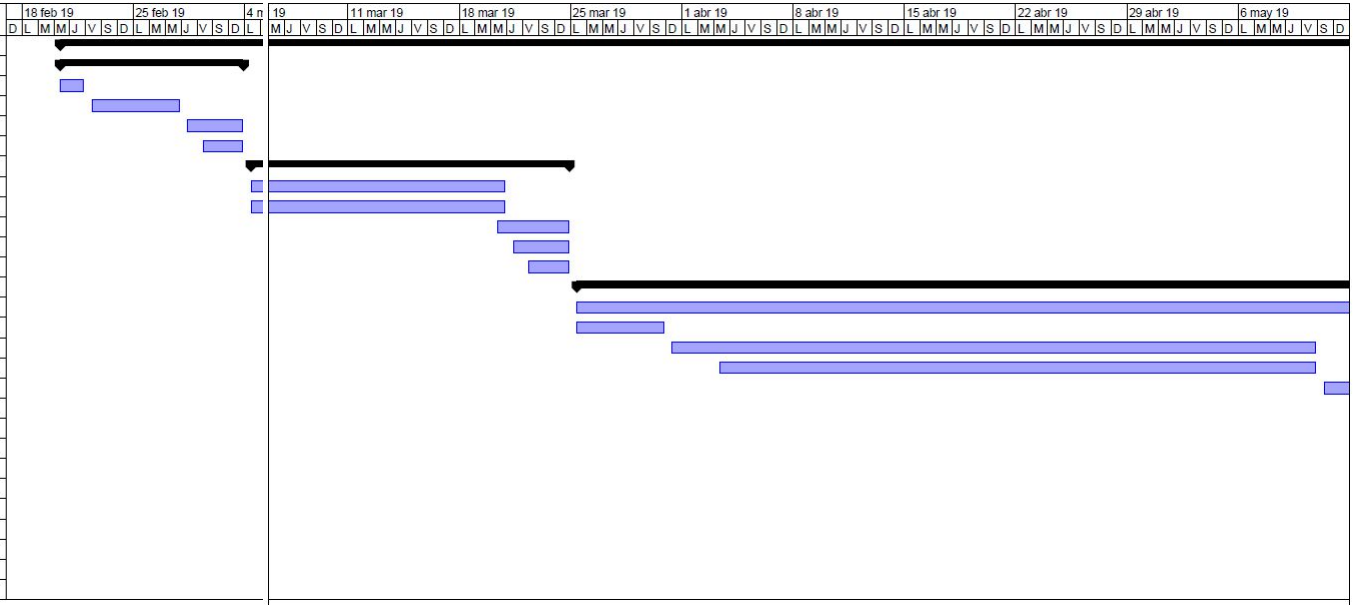
1.4. Planificación del Trabajo

- 36 horas semanales.
- 6 días a la semana.

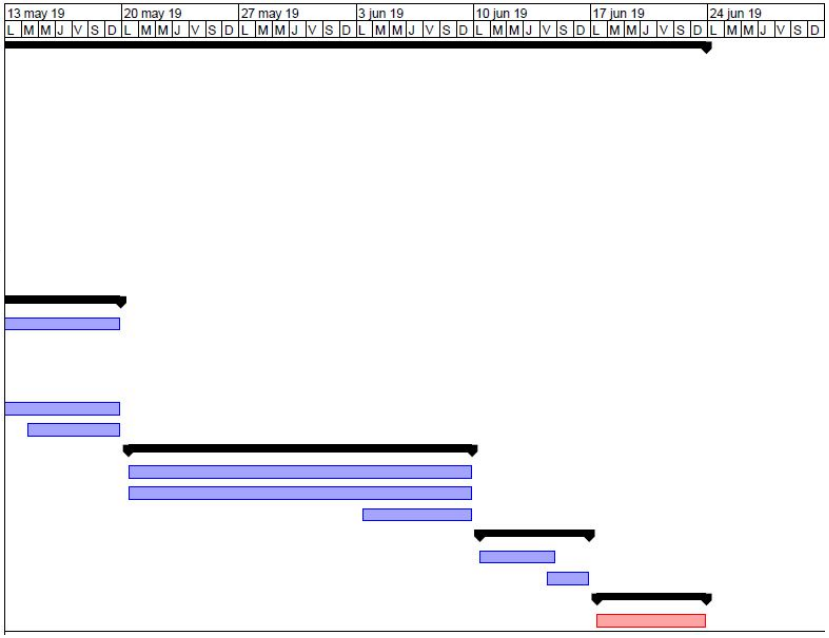
	DESCRIPCIÓN	FECHA	HORAS
PEC 1	Definición y planificación del trabajo final	03/03/19	72
PEC 2	Estado del arte o análisis de mercado del proyecto	24/03/19	108
PEC 3	Diseño e implementación del trabajo	19/05/19	288
PEC 4	Redacción de la memoria	09/06/19	108
PEC 5	Presentación y defensa del proyecto	16/06/19	36
			612

- Definición de las tareas en cada hito en diagrama de Gantt con ProjectLibre [4]

	Nombre	Duracion	Inicio	Terminado
1	**Trabajo Final Máster**	**124 days**	**20/02/19 9:00**	**23/06/19 22:00**
2	**PEC1 - Definición y planificación del trabajo final**	**12 days**	**20/02/19 9:00**	**3/03/19 22:00**
3	Comprensión de los datos	2 days	20/02/19 9:00	21/02/19 22:00
4	Definición del contexto, del enfoque y de objetivos	6 days	22/02/19 9:00	27/02/19 22:00
5	Planificación del trabajo	4 days	28/02/19 9:00	3/03/19 22:00
6	Redacción de la entrega	3 days	1/03/19 9:00	3/03/19 22:00
7	**PEC2 - Estado del arte o análisis de mercado del proyecto**	**21 days**	**4/03/19 9:00**	**24/03/19 22:00**
8	Recolección de resultados de otros trabajos	17 days	4/03/19 9:00	20/03/19 22:00
9	Búsqueda de bibliografía	17 days	4/03/19 9:00	20/03/19 22:00
10	Selección inicial de herramientas y formación	5 days	20/03/19 9:00	24/03/19 22:00
11	Corrección detalles anteriores y revisión	4 days	21/03/19 9:00	24/03/19 22:00
12	Redacción de la entrega	3 days	22/03/19 9:00	24/03/19 22:00
13	**PEC3 - Diseño e implementación del trabajo**	**56 days**	**25/03/19 9:00**	**19/05/19 22:00**
14	Búsqueda de bibliografía	56 days	25/03/19 9:00	19/05/19 22:00
15	Preparación de los datos	6 days	25/03/19 9:00	30/03/19 22:00
16	Modelado de los datos	41 days	31/03/19 9:00	10/05/19 22:00
17	Evaluación de los modelos	38 days	3/04/19 9:00	10/05/19 22:00
18	Corrección detalles anteriores y revisión	9 days	11/05/19 9:00	19/05/19 22:00
19	Redacción de la entrega	6 days	14/05/19 9:00	19/05/19 22:00
20	**PEC4 - Redacción de la memoria**	**21 days**	**20/05/19 9:00**	**9/06/19 22:00**
21	Búsqueda de bibliografía	21 days	20/05/19 9:00	9/06/19 22:00
22	Desarrollo completo de la memoria y entrega final	21 days	20/05/19 9:00	9/06/19 22:00
23	Corrección detalles anteriores y revisión	7 days	3/06/19 9:00	9/06/19 22:00
24	**PEC5 - Presentación y defensa del proyecto**	**7 days**	**10/06/19 9:00**	**16/06/19 22:00**
25	Realización de la presentación	5 days	10/06/19 9:00	14/06/19 22:00
26	Revisión y entrega	3 days	14/06/19 9:00	16/06/19 22:00
27	**Defensa pública**	**7 days**	**17/06/19 9:00**	**23/06/19 22:00**
28	Responder a las preguntas del tribunal	7 days	17/06/19 9:00	23/06/19 22:00



	Nombre	Duracion	Inicio	Terminado
13	**PEC3 - Diseño e implementación del trabajo**	**56 days**	**25/03/19 9:00**	**19/05/19 22:00**
14	Búsqueda de bibliografía	56 days	25/03/19 9:00	19/05/19 22:00
15	Preparación de los datos	6 days	25/03/19 9:00	30/03/19 22:00
16	Modelado de los datos	41 days	31/03/19 9:00	10/05/19 22:00
17	Evaluación de los modelos	38 days	3/04/19 9:00	10/05/19 22:00
18	Corrección detalles anteriores y revisión	9 days	11/05/19 9:00	19/05/19 22:00
19	Redacción de la entrega	6 days	14/05/19 9:00	19/05/19 22:00
20	**PEC4 - Redacción de la memoria**	**21 days**	**20/05/19 9:00**	**9/06/19 22:00**
21	Búsqueda de bibliografía	21 days	20/05/19 9:00	9/06/19 22:00
22	Desarrollo completo de la memoria y entrega final	21 days	20/05/19 9:00	9/06/19 22:00
23	Corrección detalles anteriores y revisión	7 days	3/06/19 9:00	9/06/19 22:00
24	**PEC5 - Presentación y defensa del proyecto**	**7 days**	**10/06/19 9:00**	**16/06/19 22:00**
25	Realización de la presentación	5 days	10/06/19 9:00	14/06/19 22:00
26	Revisión y entrega	3 days	14/06/19 9:00	16/06/19 22:00
27	**Defensa pública**	**7 days**	**17/06/19 9:00**	**23/06/19 22:00**
28	Responder a las preguntas del tribunal	7 days	17/06/19 9:00	23/06/19 22:00



2. Estado del Arte

2.1 Estudios con el conjunto de datos NHANES

- *Machine Learning Classification Analysis for a Hypertensive Population as a Function of Several Risk Factors [5]*
  - NHANES 2007-2016
  - Variables independientes: edad, sexo, raza, IMC, consumo de tabaco , enfermedades renales y diabetes.
  - Regresión logística en sklearn de Python.
  - SEN = 77%, SPE = 68%.
  - Obesidad, edad entre 71 y 80 años, raza negra no hispana y hombres.

- *Early Diabetes Prediction Using Voting Based Ensemble Learning [6]*
  - NHANES 2013-2014.
  - 10172 muestras y 54 variables de *questionnaire.*
  - Reducción de dimensiones a 24 variables.
  - Ensemble Model: regresión logística, kNN, Random Forest y Gradient Boosting en sklearn de Python.

- *Associations of cumulative exposure to heavy metal mixtures with obesity and its comorbidities among U.S. adults in NHANES 2003–2014 [7]*
  - Hipertensión y diabetes tipo 2.
  - ElasticNet flexible.
  - Modelo de riesgo predictivo como suma con pesos de las sustancias.
  - Regresión lineal y regresión logística en R.
  - La exposición a metales pesados está relacionado con obesidad, hipertensión y diabetes tipo 2.

2.2 Técnicas de Interés en otros Conjuntos de Datos

- *Quantification of tongue colour using machine learning in Kampo medicine [8]*
  - Color de la lengua indica problemas físicos o mentales.
  - K-Means k = 4 en MATLAB.
  - Cuantifica la media del color en CIELAB.
  - La media del color obtenida es insuficiente para la evaluación clínica.

- *Comparison of Machine Learning Methods for the Arterial Hypertension Diagnostics [9]*
  - Frecuencia cardíaca.
  - 30 voluntarios sanos y 41 con hipertensión.
  - Posición horizontal 300 s y 70º 300 s.
  - LDA, QDA, kNN, SVM, árboles de decisión y Naïve Bayes en sklearn de Python.
  - Selección de 53 características y se escogen las que tengan correlación inferior a 0.25.
  - LDA y QDA con 4 variables independientes.

- *Risk-Predicting Model for Incident of Essential Hypertension Based on Environmental and Genetic Factors with Support Vector Machine [10]*
  - Cuestionario de investigación epidemiológica de la población Han china de Beijing.
  - 9 factores ambientales y 12 genéticos, 559 muestras pacientes con hipertensión y 641 que no.
  - Tres SVM con función kernel radial con factores ambientales, con factores genéticos y con ambos factores.
  - Tres SVM con función kernel laplaciana con factores ambientales, con factores genéticos y con ambos factores.
  - Mejor resultado con kernel laplaciana. ACC=80.1%, SEN=63.3% y SPE=86.7%

- *Elevated High Density Lipoprotein Cholesterol is associated with hyponatremia in hypertensive patients [11]*
  - Demografía, clínico y de laboratiorio de SPRINT.
  - 9361 muestras con hipertensión y sin diabetes.
  - Gradient Boosting Machine en R.
  - Evaluación de resultados con NHANES 2005-2010.
  - HDL-C elevado es factor de riesgo para hiponametria.

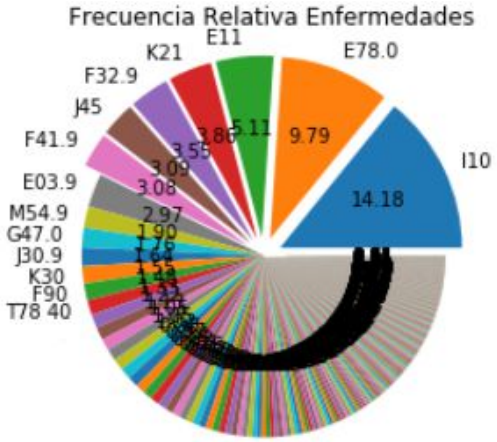
3. Proceso de Implementación

3.1. Exploración y Procesado de los datos

- 6 archivos separados por comas: *demographic, diet, examination, labs, medications y questionnaire*.
- *Demographic*
  - 10175 registros y 47 variables.
  - Información individual y familia, composición familiar, género, edad, etnia, educación, estado civil,...
- *Diet*
  - 9813 registros y 168 variables.
  - Dieta de los participantes y nutrientes.
- *Examination*
  - 9813 registros y 224 variables.
  - Datos clínicos, medidas corporales,...
- *Labs*
  - 9813 registros y 424 variables.
  - Pruebas analíticas.
- *Medications*
  - 20194 registros y 13 variables.
  - Medicaciones, enfermedad, ...
  - Eliminación de los registros con la variable RXDRSD1 nula.
- *Questionnaire*
  - 10175 registros y 953 variables
  - Idioma, consumo de alcohol y tabaco, hábitos alimenticios y ejercicio físico,...

- Resumen de cada variable y contabilización de los valores nulos.
- Imputación de los valores nulos.
  - *RIDAGEMN* y *RIDEXAGM* se les imputa *RIDAGEYR* multiplicada por 12.
  - Mediana para las variables numéricas y el más frecuente para las categóricas.
- *Medications* se agrupa por *SEQN* tras descategorizar.
- Enfermedades con mayor frecuencia.

ENFERMEDAD	CÓDIGO	FRECUENCIA RELATIVA
Essential (primary) hypertension	I10	14.18
Pure hypercholesterolemia	E78.0	9.79
Type 2 diabetes mellitus	E11	5.11
Gastro-esophageal reflux disease	K21	3.86
Major depressive disorder, single episode, unspecified	F32.9	3.55
Asthma	J45	3.09
Anxiety disorder, unspecified	F41.9	3.08

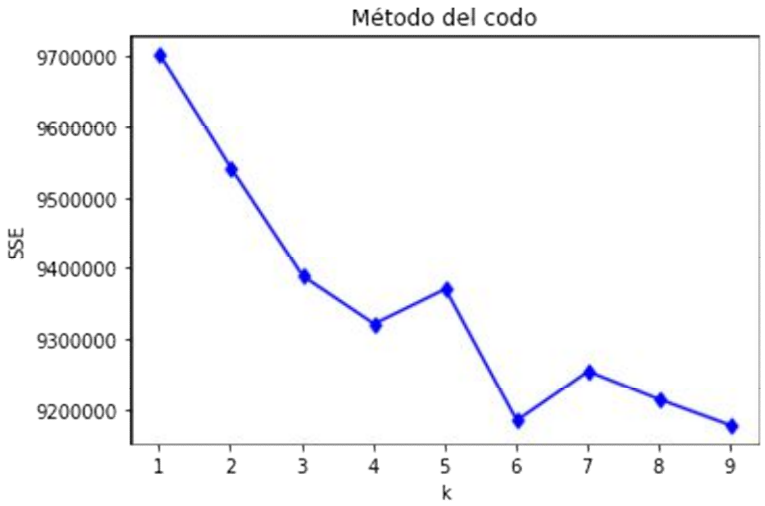


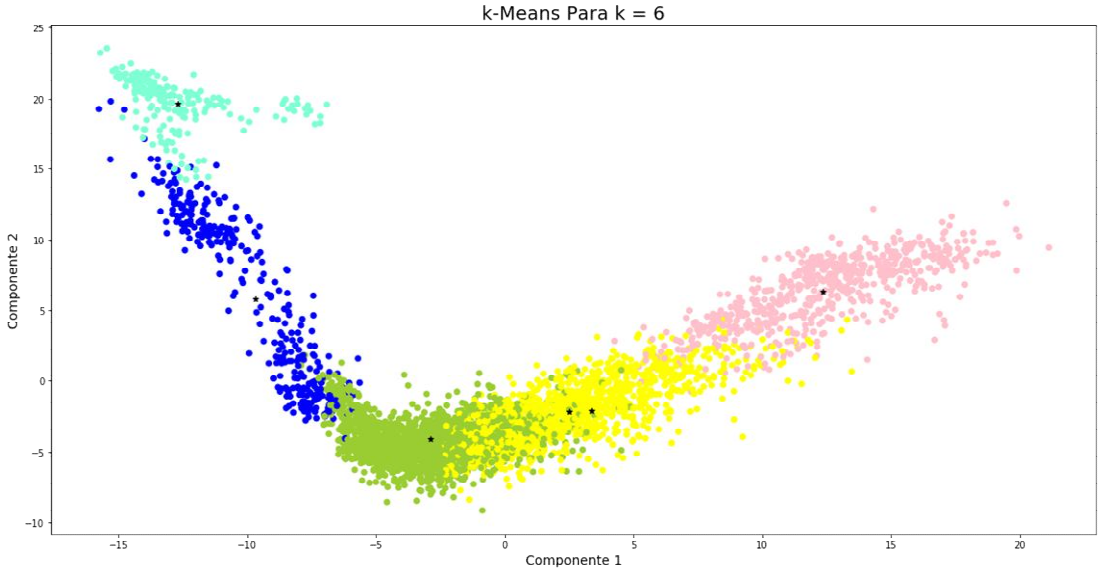
- Se eliminan variables con correlación mayor o igual a 0.8 o menor o igual a - 0.8, excepto si son enfermedades.
- Descategorización de variables categóricas con *get_dummies* de la librería *Pandas*.
- Normalización de los datos con *StandardScaler* de la librería *Sklearn*.

3.2. Clustering

3.2.1 Clustering con Todos los Datos

- Selección de 6 clústeres tras visualizar la gráfica del Método del Codo.
- Hallamos el modelo k-Means de Sklearn con 6 clústeres. El resto de valores del algoritmo por defecto [12]
- Mostramos gráfica en la proyección PCA con 2 componentes.

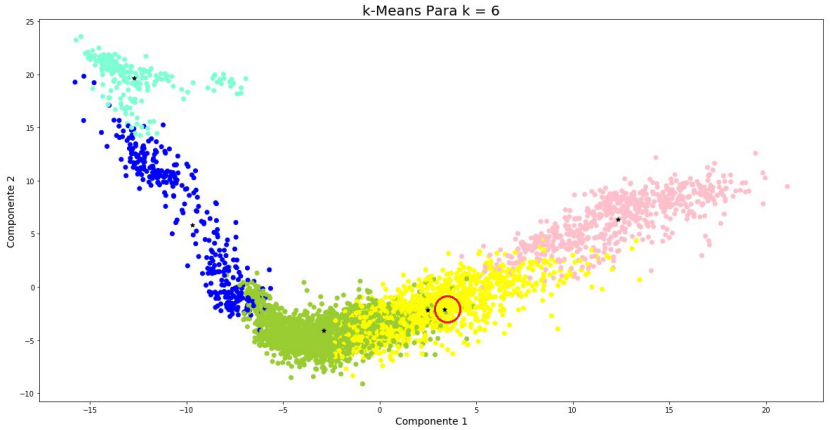




- Tamaño de los clústeres.

CLÚSTER	TAMAÑO
Clúster 1	1
Clúster 2	1683
Clúster 3	611
Clúster 4	1199
Clúster 5	330
Clúster 6	192

- El clúster 1 está formado por un único componente. Corresponde al centroide señalado .



- Predicción de las etiquetas de los clústeres con Random Forest de Sklearn con max_depth = 91 y n_estimators = 136 .Para el resto de parámetros los valores por defecto [13].

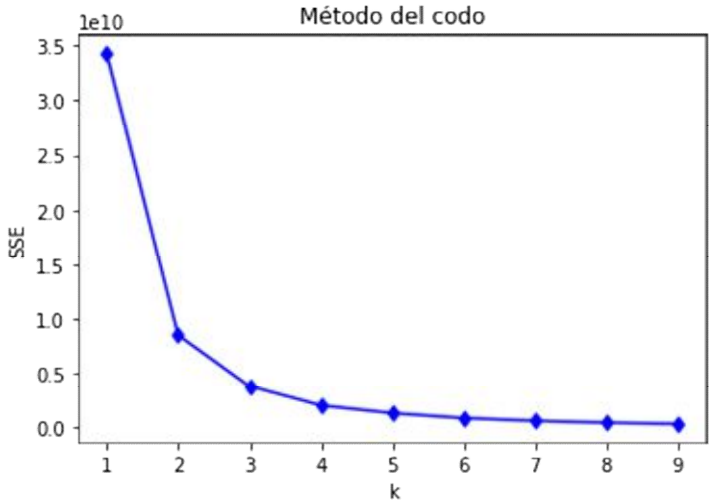
- Las 10 variables más importantes.

DESCRIPCIÓN	CÓDIGO VARIABLE	IMPORTANCIA
Edad en años	RIDAGEYR	0.016688
Buen estado del 2º premolar arriba derecha permanente	OHX04CTC_S	0.013943
Buen estado del 2º premolar abajo derecha permanente	OHX29CTC_S	0.01197
Buen estado del 2º premolar arriba izquierda permanente	OHX13CTC_S	0.011789
Buen estado del 2º premolar abajo izquierda permanente	OHX20CTC_S	0.011515
Estado 2º molar abajo izquierda	OHX18TC	0.010943
Canino arriba izquierda perdido por enfermedad dental y reemplazado	OHX11CTC_P	0.010511
Estado incisivo abajo izquierda	OHX23TC	0.010231
Buen estado del incisivo central arriba izquierda permanente	OHX09CTC_S	0.01016
2º molar arriba derecha sin erupciones	OHX02CTC_U	0.010122

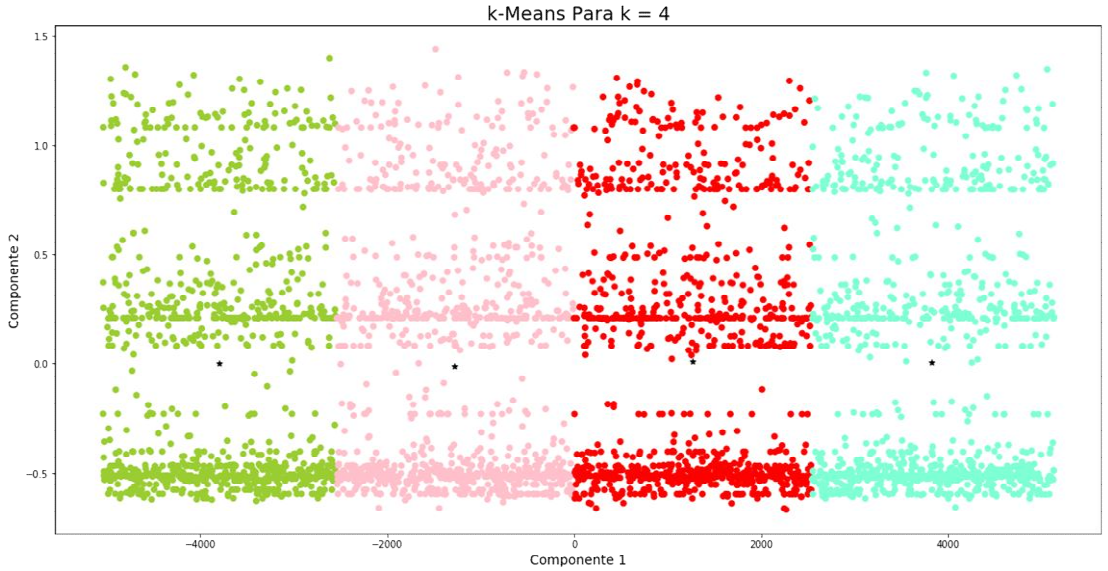
Los datos se organizan según la edad de los pacientes y el estado de salud dental de éstos.

3.2.2 Clustering de las Enfermedades

- Se usará solamente el fichero medications con las variables SEQN Y RXDRSC1 agrupadas por SEQN tras descategorizar.
- Selección de 4 clústeres con el método del codo.



- Hayamos el modelo k-Means de Sklearn con 4 clústeres. El resto de valores del algoritmo por defecto [12]
- Mostramos gráfica en la proyección PCA con 2 componentes.



- En la matriz de correlación se obtiene:

ENFERMEDAD 1	ENFERMEDAD 2	CORRELACIÓN
RXDRSC1_A44.9	RXDRSC1_A69.2	0.81639486
RXDRSC1_C74	RXDRSC1_R20.0	1
RXDRSC1_C74	RXDRSC1_Z51.11	1
RXDRSC1_D25.P	RXDRSC1_Z94.0	1
RXDRSC1_D35.2	RXDRSC1_E27.40	1
RXDRSC1_I49.1	RXDRSC1_K91.5	1
RXDRSC1_K12.2	RXDRSC1_M25.52	1
RXDRSC1_R20.0	RXDRSC1_Z51.11	1

- Se identifican y contabiliza las enfermedades por clúster. Comprobamos que las enfermedades correladas están en el mismo clúster.

✓ "Bartonellosis" (RXDRSC1_A44.9) y "Lyme disease" (RXDRSC1_A69.2) están relacionadas con el mismo paciente en el clúster 3 y clúster 4 de las dos y tres veces respectivas que están en los datos.

✓ "Malignant neoplasm of adrenal gland" (RXDRSC1_C74), "Anesthesia of skin" (RXDRSC1_R20.0) y "Encounter for antineoplastic chemotherapy" (RXDRSC1_Z51.11) están relacionadas con el mismo paciente en el clúster 2. "Anesthesia of skin" (RXDRSC1_R20.0) y "Encounter for antineoplastic chemotherapy" (RXDRSC1_Z51.11) están relacionadas con otro paciente también en el clúster 2.

✓ "Prevent uterine fibroids" (RXDRSC1_D25.P) y "Kidney transplant status" (RXDRSC1_Z94.0) están solamente en el clúster 1.

✓ "Benign neoplasm of pituitary gland" (RXDRSC1_D35.2) y "Unspecified adrenocortical insufficiency" (RXDRSC1_E27.40) en el clúster 2.

✓ "Atrial premature depolarization" (RXDRSC1_I49.1) y "Postcholecystectomy syndrome" (RXDRSC1_K91.5) en el clúster 3.

✓ "Cellulitis and abscess of mouth" (RXDRSC1_K12.2) y "Pain in elbow " (RXDRSC1_R20.0) en el clúster 4.

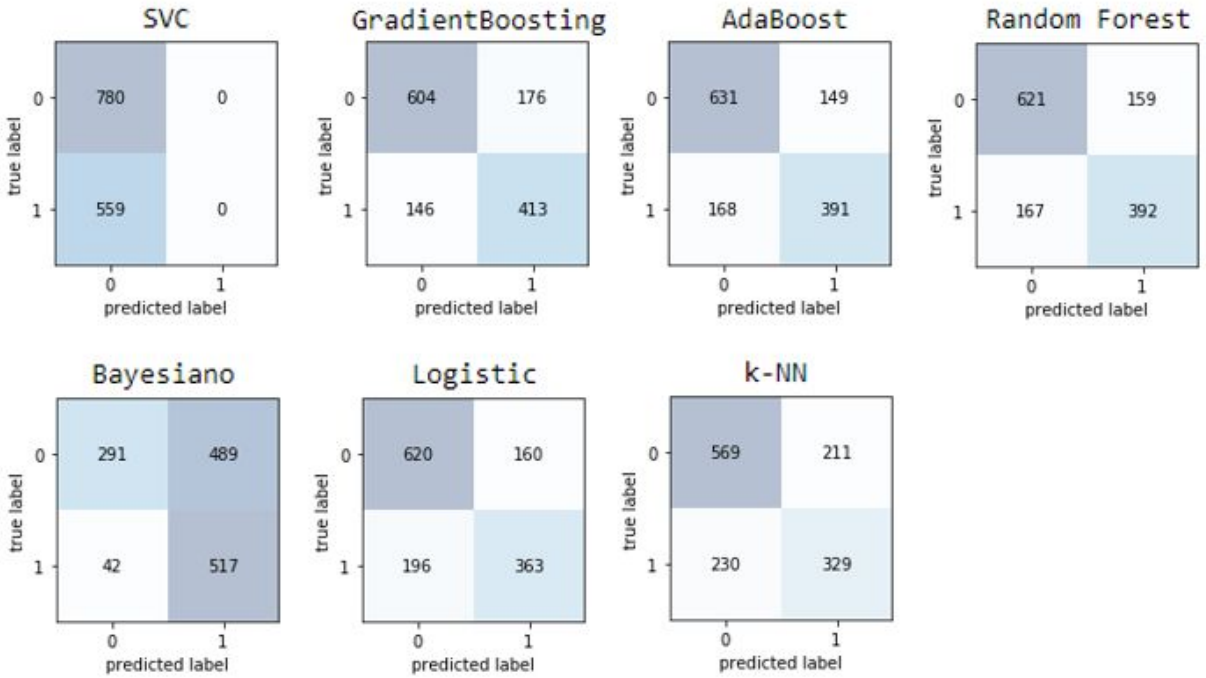
3.3. Modelo Predictivo

- Modelo predictivo para la hipertensión. La variable *RXDRSC1* corresponde a I10.
- Selección de variables según criterio de riesgos para la hipertensión de la OMS [14]
  - Enfermedades cardiovasculares, renales, diabetes e hipercolesterolemia.
  - Ingresos bajos.
  - Alimentación poco saludable e inactividad física.
  - Consumo de tabaco y alcohol.
  - Edad.
  - Género.
  - Sobrepeso.
  - Antecedentes familiares.
- Varios modelos predictivos. Se personalizan los siguientes parámetros, el resto de ellos serán tomados por defecto.

ALGORITMO	PARÁMETRO 1	PARÁMETRO 2	REFERENCIA
Support Vector Classification	C = 10	gamma = 20	[15]
Gradient Boosting Classifier	learning_rate = 0.1	n_estimators = 102	[16]
AdaBoost Classifier	learning_rate = 0.3	n_estimators = 107	[17]
Random Forest Classifier	max_depth = 50	n_estimators = 200	[13]
Naïve Bayes	var_smoothing = 9.81e-09	-------	[18]
Logistic Regression	penalty = l1	C = 0.16	[19]
KNeighbors Classifier	n_neighbors = 8	n_weights = distance	[20]

4. Evaluación

- Matriz de confusión de cada modelo.



- Métricas derivadas de la matriz de confusión.

MODELO	ACC	ERR	TPR	FPR	PRE	SPE	F1
Support Vector Classification	58.25 %	41.75 %	0	0	-	100 %	-
Gradient Boosting Classifier	75.95 %	24.05 %	73.88 %	22.56 %	70.12 %	77.44 %	71.95 %
AdaBoost Classifier	76.33 %	23.67 %	69.95 %	19.10 %	72.41 %	80.90 %	71.16 %
Random Forest Classifier	75.65 %	24.35 %	70.13 %	20.38 %	71.14 %	79.61 %	70.63 %
Naive Bayes	60.34 %	39.66 %	92.49 %	62.69 %	51.39 %	37.31 %	66.07 %
Logistic Regression	73.41 %	26.58 %	64.94 %	20.51 %	69.41 %	79.49 %	67.10 %
k-NN	67.06 %	32.94 %	58.86 %	27.05 %	60.93 %	72.95 %	59.87 %

5. Conclusiones

- ✓ Objetivos principales cumplidos.
  - Modelo de clustering.
  - Modelo clasificatorio para predecir la hipertensión de un paciente.
  - Modelo de clustering para hallar comorbilidades.
- ✓ Modelo de clustering.
  - Los datos se agrupan según la edad y variables relacionadas con el estado de salud dental.
  - Se diferencian clústeres según las variables más importantes, pero afecta la imputación de los valores nulos.
- ✓ Modelo de clustering de las enfermedades.
  - Posibles comorbilidades.
    - ➢ Malignant neoplasm of adrenal gland", "Anesthesia of skin" y "Encounter for antineoplastic chemotherapy".
    - ➢ "Bartonellosis" y "Lyme disease".
    - ➢ "Prevent uterine fibroids" y "Kidney transplant status".
    - ➢ "Benign neoplasm of pituitary gland" y "Unspecified adrenocortical insufficiency".
    - ➢ "Atrial premature depolarization" y "Postcholecystectomy syndrome".
    - ➢ "Cellulitis and abscess of mouth" y "Pain in elbow".

- ✓ Modelo predictivo.
  - Mayor exactitud con AdaBoost (76.33%)
  - Mayor tasa de verdaderos positivos con Naïve Bayes.
    - ➢ Buena métrica para el contexto.
    - ➢ Diagnóstico temprano.
  - Mayor tasa de falsos negativos, menor precisión y menor especificidad con Naïve Bayes.
    - ➢ El modelo tiende a predecir como positivo.
    - ➢ Diagnóstico temprano, pero supone costes médicos innecesarios.

6. Línea de Trabajo Futura

- Modelos predictivos para otras enfermedades: diabetes, hipercolesterolemia, ansiedad, depresión,...
- Modelo predictivo más complejo que generalizara a cualquier enfermedad.
- Modelo de clustering con la eliminación de los registros nulos o la imputación de valores segmentada.
- Evaluación de las comorbilidades obtenidas con otro conjunto de datos que pueda suponer una base para la investigación en el área de medicina.

7. Código

- El código implementado se puede ver en github [21]
- Cuatro archivos notebook de python:
  - Exploración y procesado de los datos.
  - Clustering.
  - Clustering de las enfermedades.
  - Modelo predictivo de la hipertensión.

9. Bibliografía y Referencias

[1] Bloomberg [Consulta: 29 de Febrero de 2019] <https://www.bloomberg.com/news/articles/2019-02-24/spain-tops-italy-as-world-s-healthiest-nation-while-u-s-slips?srnd=premium-europe>

[2] Kaggle [Consulta: 20 de Febrero de 2019]
<https://www.kaggle.com/cdc/national-health-and-nutrition-examination-survey>

[3] Center for Disease Control and Prevention [Consulta: 27 de Febrero de 2019]
<https://www.cdc.gov/Nchs/Nhanes/about_nhanes.htm>

[4] ProjectLibre [Consulta: 25 de Febrero]
< https://www.projectlibre.com/>

[5] López-Martínez F., Schwarcz.MD A., Núñez-Valdez E.R. & García-Díaz V. (2018). "Machine Learning Classification Analysis for a Hypertensive Population as a Function of Several Risk Factors". *Expert Systems With Applications* (vol. 110, pág. 206-215).

[6] Husain, A. & Khan M.H. (2018). "Early Diabetes Prediction Using Voting Based Ensemble Learning". *Communications in Computer and Information Science* (vol. 905, pág. 95-103). Singapur: Springer.

[7] Wang, X., Mukherjee, B. & Park, S.K. (2018). "Associations of cumulative exposure to heavy metal mixtures with obesity and its comorbidities among U.S. adults in NHANES 2003–2014". *Environment International* (vol. 121, pág 683–694).

[8] Kawanabe T., Kamarudin N.D., Ooi C.Y., Kobayashi F., Mi X., Sekine S., Wakasugi A., Odaguchi H. & Hanawa T. (2016). "Quantification of tongue colour using machine learning in Kampo medicine". *European Journal of Integrative Medicine* (núm. 8, pág. 932-941).

[9] Kublanov V.S., Dolganov A.Y, Belo D. & Gamboa H. (2017). "Comparison of Machine Learning Methods for the Arterial Hypertension Diagnostics". *Applied Bionics and Biomechanics* (vol. 2017, Article ID 5985479, 13 pages).

[10] Pei, Z., Liu, J., Liu, M., Zhou, W., Yan, P., Wen, S. & Chen, C. (2018). "Risk-Predicting Model for Incident of Essential Hypertension Based on Environmental and Genetic Factors with Support Vector Machine". *Interdisciplinary Sciences: Computational Life Sciences.* (vol. 10, pág. 126-130). Alemania.

[11] Israel, A. & Grossman, E. (2017). "Elevated High Density Lipoprotein Cholesterol is associated with hyponatremia in hypertensive patients". *The American Journal of Medicine* (vol. 130, pág. 1324.e7-1324.e13).

[12] Scikit-learn [Consulta: 7 de Junio de 2019] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

[13] Scikit-learn [Consulta: 7 de Junio de 2019] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[14] Organización Mundial de la Salud (2013). "Información general sobre la hipertensión en el mundo". Ginebra.

[15] Scikit-learn [Consulta: 7 de Junio de 2019] <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html >

[16] Scikit-learn [Consulta: 7 de Junio de 2019]<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html>

[17] Scikit-learn [Consulta: 7 de Junio de 2019] <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html>

[18] Scikit-learn [Consulta: 7 de Junio de 2019] <https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html>

[19] Scikit-learn [Consulta: 7 de Junio de 2019] <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html>

[20] Scikit-learn [Consulta: 7 de Junio de 2019] <https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html>

[21] Github [Consulta: 09 de Junio de 2019] <https://github.com/marcreest1/TFM>

Universitat Oberta
de Catalunya

FIN

Muchas gracias.

María José Crespo Estévez

UOC.universitat
@UOCuniversitat
UOCuniversitat