DIAGNÓSTICO PRECOZ DE ENFERMEDAD CRÓNICA DE RIÑÓN.

Comprensión del negocio

Actualmente, hay muchas personas en el mundo que padecen enfermedades renales. Debido a varios factores de riesgo como la alimentación, el medio ambiente y el nivel de vida, muchas personas contraen enfermedades repentinamente sin comprender su condición. El diagnóstico de enfermedades renales crónicas es generalmente invasivo, costoso, lento y a menudo riesgoso. Es por eso que muchos pacientes llegan a etapas tardías sin tratamiento, especialmente en aquellos países donde los recursos son limitados. Por lo tanto, la estrategia de detección temprana de la enfermedad sigue siendo importante, particularmente en los países en desarrollo, donde las enfermedades generalmente se diagnostican en etapas tardías.

Enfermedad Renal Crónica (ERC):

La enfermedad renal crónica es una afección médica caracterizada por una disminución progresiva de la función renal a lo largo del tiempo. Puede estar en etapas tempranas, como la insuficiencia renal leve, o en etapas avanzadas, como la insuficiencia renal crónica. La ERC puede ser causada por diversas enfermedades subyacentes, como la diabetes, la hipertensión y enfermedades autoinmunitarias.

Estadísticas Globales:

Las estadísticas sobre la ERC varían según la región y el país. Sin embargo, en todo el mundo, la enfermedad renal crónica es un problema de salud pública creciente. Según la Organización Mundial de la Salud (OMS), se estima que alrededor del 10% de la población mundial tiene ERC, y millones de personas mueren cada año debido a complicaciones relacionadas con esta enfermedad.

Impacto Económico y Social:

La ERC tiene un impacto económico significativo en los sistemas de atención médica y en la calidad de vida de las personas. Los costos asociados con el tratamiento de la ERC, que puede incluir diálisis o trasplante renal, son sustanciales. Además, la enfermedad puede llevar a una disminución en la productividad laboral y una mayor carga para las familias.

Otros Problemas Relacionados:

La ERC también está relacionada con una serie de otros problemas de salud, como enfermedades cardiovasculares, anemia, desequilibrios electrolíticos y trastornos óseos y minerales. La gestión de la ERC implica no solo tratar la disfunción renal, sino también abordar estas complicaciones.

Se trata de un problema de clasificación supervisado, esto se debe a que nosotros tenemos datos con un dato objetivo (class).

Comprensión de los datos

Es un dataset que se compone de 25 atributos

Número	Variable	Descripción	Categorías

1	Age	Edad en años	Numerical (Edad en años)
2	Blood Pressure	Presión arterial en mm/Hg	Numerical (Presión arterial en mm/Hg)
3	Specific Gravity	Gravedad específica	Nominal (1.005, 1.010, 1.015, 1.020, 1.025)
4	Albumin	Albúmina	Nominal (0, 1, 2, 3, 4, 5)
5	Sugar	Azúcar en orina	Nominal (0, 1, 2, 3, 4, 5)
6	Red Blood Cells	Glóbulos rojos en sangre	Nominal (Normal, Anormal)
7	Pus Cell	Células de pus	Nominal (Normal, Anormal)
8	Pus Cell Clumps	Coágulos de células de pus	Nominal (Presente, No Presente)
9	Bacteria	Bacterias en orina	Nominal (Presente, No Presente)
10	Blood Glucose Random	Glucosa en sangre aleatoria	Numerical (bgr en mg/dl)
11	Blood Urea	Urea en sangre	Numerical (bu en mg/dl)
12	Serum Creatinine	Creatinina en suero	Numerical (sc en mg/dl)
13	Sodium	Sodio en suero	Numerical (sod en mEq/L)
14	Potassium	Potasio en suero	Numerical (pot en mEq/L)
15	Hemoglobin	Hemoglobina	Numerical (hemo en gms)
16	Packed Cell Volume	Volumen de células empaquetadas	Numerical (Valor en %)
17	White Blood Cell Count	Recuento de glóbulos blancos	Numerical (wc en células/cumm)
18	Red Blood Cell Count	Recuento de glóbulos rojos	Numerical (rc en millones/cmm)
19	Hypertension	Hipertensión	Nominal (Sí, No)
20	Diabetes Mellitus	Diabetes Mellitus	Nominal (Sí, No)
21	Coronary Artery Disease	Enfermedad de las arterias coronarias	Nominal (Sí, No)
22	Appetite	Apetito	Nominal (Bueno, Pobre)
23	Pedal Edema	Edema en los pies	Nominal (Sí, No)
24	Anemia	Anemia	Nominal (Sí, No)
25	Class	Clase	Nominal (CKD, No CKD)

- Muchos datos tienen valores con “?” lo cual significa que es un valor faltante. Esto se pueden descartar con el fin de tener un dataset más limpio. Esto no va a afectar los datos puesto que son pocos valores en general
- Valores faltantes:

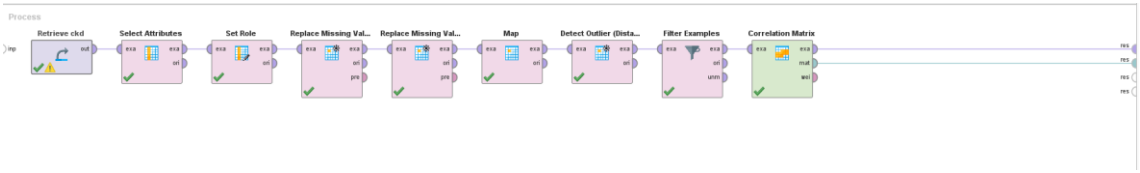
Variable	**Valores Faltantes**
Age	9
Blood Pressure	12
Blood Glucose Random	44
Blood Urea	19
Serum Creatinine	17
Sodium	88
Potassium	88
Hemoglobin	52
Packed Cell Volume	72
White Blood Cell Count	109
Red Blood Cell Count	134
Hypertension	6
Coronary Artery Disease	13
Appetite	13
Pedal Edema	13
Anemia	13
Class	13

Se puede ver que el dataset tiene unos cuantos valores faltantes.

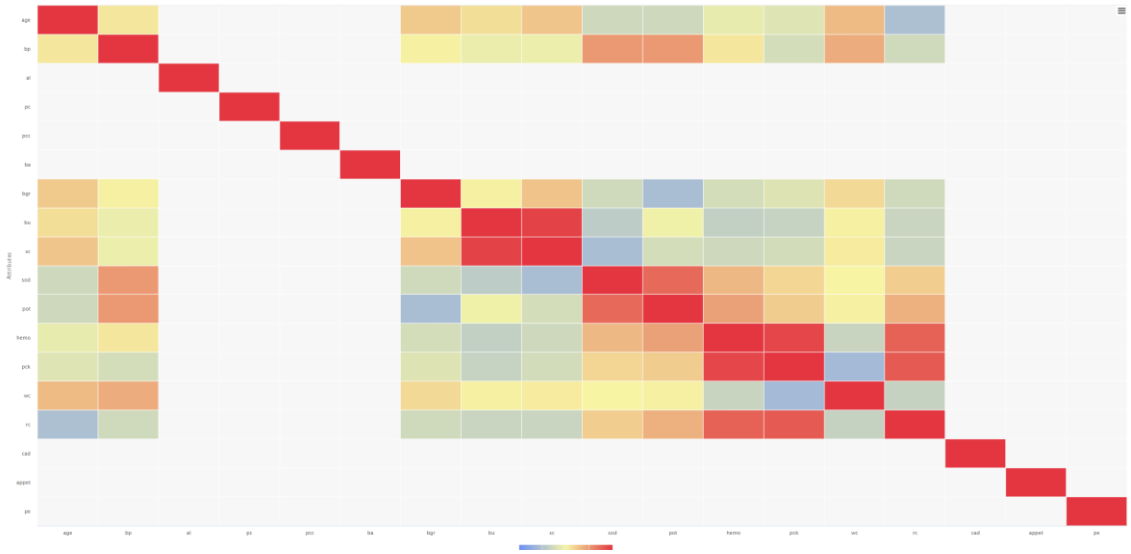
Class: tiene 13 valores faltantes pero debido a la letra del info (250 valores ckd) podemos entenderlos como ckd. Por lo tanto Class tiene 0 valores faltantes.

Dos datos para considerar son rbcc y wbcc los cuales tienen más de ¼ de los atributos. Debido a que estos son muy importantes para los problemas renales los missing van a ser remplazados por la media del atributo.

Como consideración para reducir la dimensionalidad de los atributos se va a quitar anemia, rbc pues rbcc abarca ambas. Mismo con hipertensión y presión de sangre. Diabetes y glucosa. Gravedad no se encuentra en la bibliografía.



Tras el primer filtrado se realiza una matriz de correlación y se obtienen los siguientes resultados:



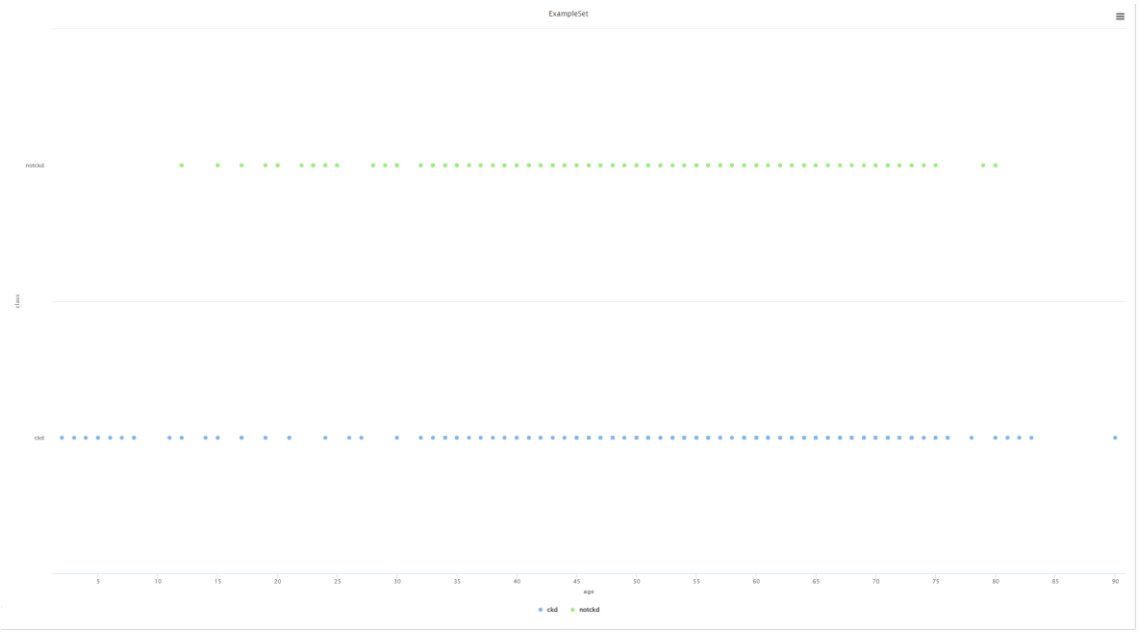
Attribut...	age	bp	al	pc	pcc	ba	bgr	bu	sc	sod	pot	hemo	pck	wc	rc	cad	appet	pe
age	1	0.083	?	?	?	?	0.243	0.131	0.264	-0.301	-0.300	-0.106	-0.174	0.319	-0.534	?	?	?
bp	0.083	1	?	?	?	?	0.030	-0.082	-0.078	0.496	0.496	0.083	-0.250	0.401	-0.287	?	?	?
al	?	?	1	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
pc	?	?	?	1	?	?	?	?	?	?	?	?	?	?	?	?	?	?
pcc	?	?	?	?	1	?	?	?	?	?	?	?	?	?	?	?	?	?
ba	?	?	?	?	?	1	?	?	?	?	?	?	?	?	?	?	?	?
bgr	0.243	0.030	?	?	?	?	1	0.034	0.273	-0.286	-0.565	-0.248	-0.183	0.151	-0.285	?	?	?
bu	0.131	-0.082	?	?	?	?	0.034	1	0.940	-0.411	-0.047	-0.379	-0.354	0.032	-0.328	?	?	?
sc	0.264	-0.078	?	?	?	?	0.273	0.940	1	-0.561	-0.247	-0.296	-0.263	0.059	-0.331	?	?	?
sod	-0.301	0.496	?	?	?	?	-0.286	-0.411	-0.561	1	0.751	0.332	0.174	0.009	0.217	?	?	?
pot	-0.300	0.496	?	?	?	?	-0.565	-0.047	-0.247	0.751	1	0.453	0.222	0.027	0.378	?	?	?
hemo	-0.106	0.083	?	?	?	?	-0.248	-0.379	-0.296	0.332	0.453	1	0.921	-0.335	0.790	?	?	?
pck	-0.174	-0.250	?	?	?	?	-0.183	-0.354	-0.263	0.174	0.222	0.921	1	-0.598	0.817	?	?	?
wc	0.319	0.401	?	?	?	?	0.151	0.032	0.059	0.009	0.027	-0.335	-0.598	1	-0.359	?	?	?
rc	-0.534	-0.287	?	?	?	?	-0.285	-0.328	-0.331	0.217	0.378	0.790	0.817	-0.359	1	?	?	?
cad	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	1	?	?
appet	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	1	?
pe	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	1

Bu sc –pck – pck rc – rc – sod pot

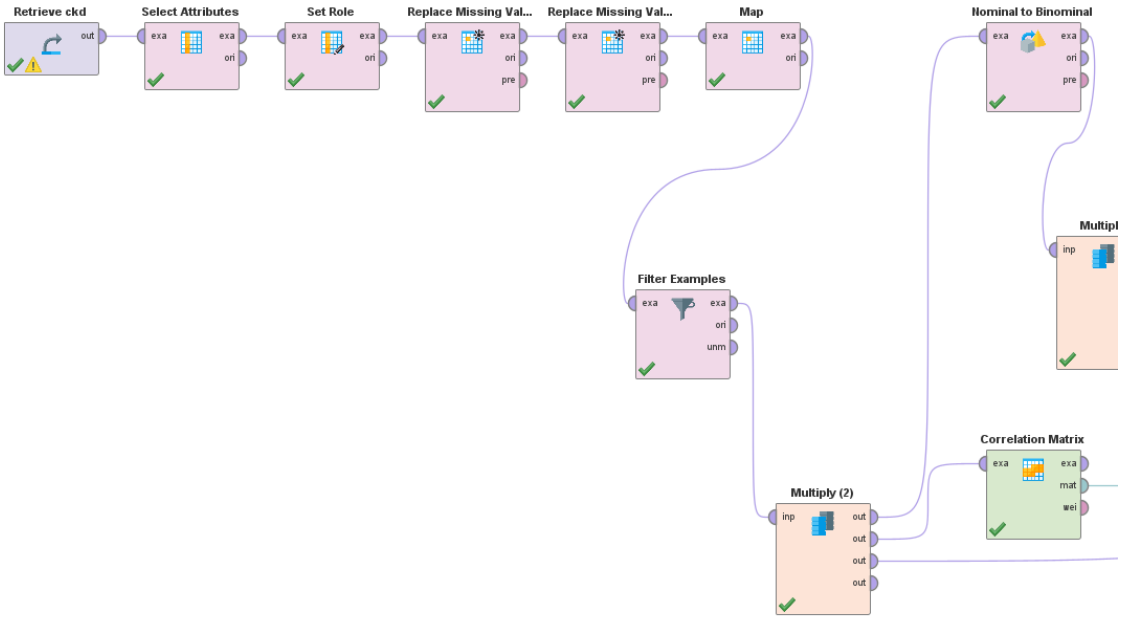
sod	pot	0.751
hemo	rc	0.790
pck	rc	0.817
hemo	pck	0.921
bu	sc	0.940

Como están altamente relacionadas se va a quitar hemo sc y sod

Visualizacion



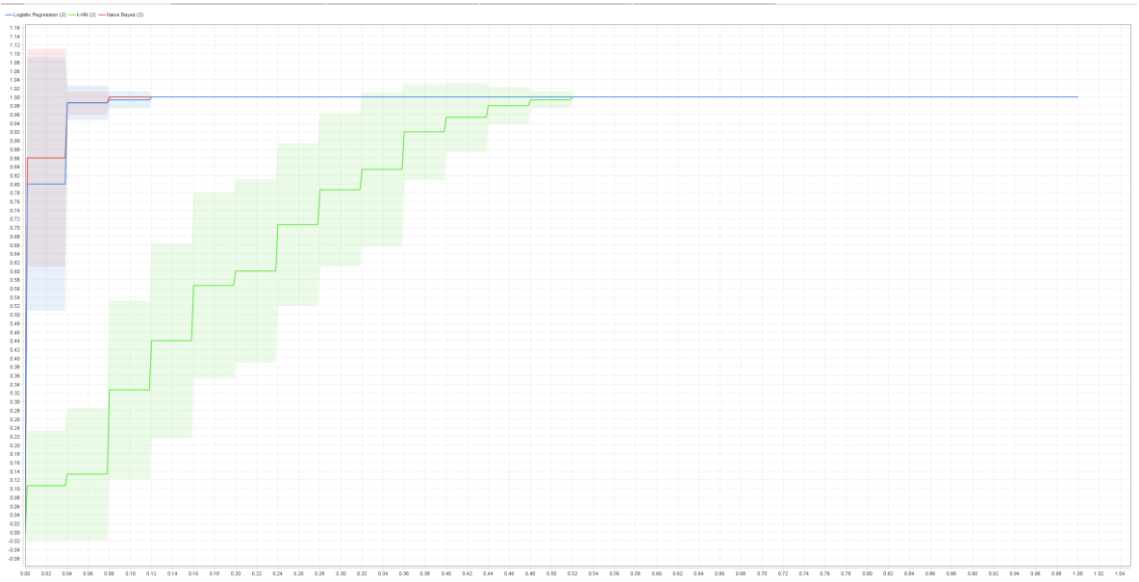
Tratamiento de los datos

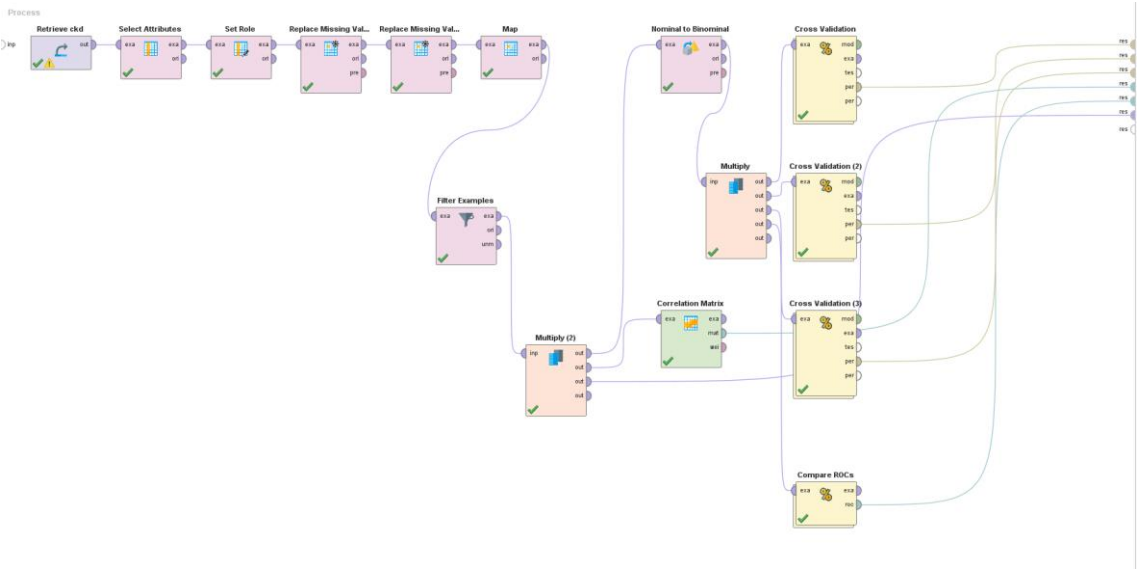


1. Se seleccionan los atributos que se desean quitar
2. Se setea el rol de label
3. Se remplazan los valores wc y rbc por el promedio
4. Se remplaza el no de class por notckd
5. Se filtran los “?”

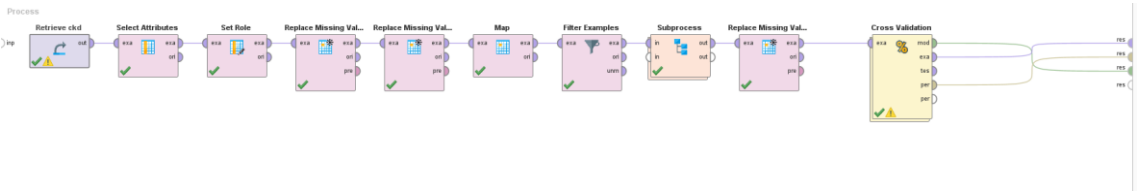
Modelado

Curva ROCs





LDA:



Evaluación

LR

accuracy: 96.25% +/- 3.39% (micro average: 96.25%)

	true ckd	true notckd	class precision
pred. ckd	241	6	97.57%
pred. notckd	9	144	94.12%
class recall	96.40%	96.00%	

KNN

accuracy: 74.50% +/- 7.25% (micro average: 74.50%)

	true ckd	true notckd	class precision
pred. ckd	170	22	88.54%
pred. notckd	80	128	61.54%
class recall	68.00%	85.33%	

NB

accuracy: 95.25% +/- 4.16% (micro average: 95.25%)

	true ckd	true notckd	class precision
pred. ckd	231	0	100.00%
pred. notckd	19	150	88.76%
class recall	92.40%	100.00%	

LDA

accuracy: 95.50% +/- 2.84% (micro average: 95.50%)

	true ckd	true notckd	class precision
pred. ckd	241	9	96.40%
pred. notckd	9	141	94.00%
class recall	96.40%	94.00%	