Trabajo Practico Final Especialización en Explotación de Datos y Descubrimiento del Conocimiento

Junio 2016

Lopez Maria Lucrecia

Contents

Introducción

En el presente trabajo práctico se analizaran los resultados del nivel de ozono según las condiciones meteorológicas en la ciudad de Houston entre los años 1998 y 2004[1]. A partir del dataset se pretende comprender cuales son las variables más influyentes en el nivel de ozono de la zona y si, las condiciones fueron variando con el paso de los años. Además, se pretende determinar la probabilidad de que un día presente elevadas concentraciones de ozono, dadas sus condiciones meteorológicas.

En la primera parte del trabajo se realizará un análisis descriptivo del problema; para ello se hará un análisis exploratorio inicial de las variables, a fin de conocer los datos. Luego se aplicaran técnicas de reducción de variables (análisis de componentes principales) con el fin de observar si es posible armar grupos de condiciones meteorológicas. Finalmente, se hará un análisis discriminante por año. También, se usaran técnicas de clusterización a fin de ver si existen grupos con características propias asociadas a los niveles de ozono.

En la segunda parte del informe, se construirá un modelo predictivo mediante técnicas de clasificación para poder decidir si, dadas ciertas condiciones meteorológicas, el nivel de ozono será elevado. Para ellos se utilizaran KNN, regresión logística y arboles de decisión.

Software utilizado

El presente trabajo se realizo con más de un software, a fin de efectuar cada proceso en la aplicación que se considera más útil, en términos de tiempo de programación/ preparación y de procesamiento y salida de resultados. Entonces se utilizó

- Infostav y SPSS en el análisis exploratorio de datos; componentes principales y clusterización.
- R para el análisis de discriminante y generación de dataset con datos desbalanceados
- Weka y SPSS para arboles de decisión.
- Rapidminer para KNN

Dataset

El dataset cuenta con 73 atributos (72 numéricos y 1 fecha) y 2534 registros relacionados a las diferentes condiciones meteorológicas de cada día en Houston entre los años 1998 y 2004. La base consiste en un conjunto de datos de las temperaturas y velocidad del viento a cada hora del día; los valores de velocidad viento, humedad relativa, temperatura, velocidad del viento en dirección este-oeste, velocidad del viento en dirección norte-sur y altura geopotencial a diferentes alturas; el nivel de precipitaciones e índices específicos de tormentas.

[1] Houston tiene excesivos niveles de ozono y se encuentra entre las ciudades más contaminadas de los Estados Unidos. El ozono a nivel del suelo, o la niebla tóxica, es uno de los problemas predominantes de contaminación atmosférica en Houston

La clase solo puede tomar dos valores:

- 0: indica niveles normales de ozono
- 1: indica niveles elevados de ozono

En el anexo 1, se puede observar el listado de variables con su tipo y descripción.

Análisis exploratorio

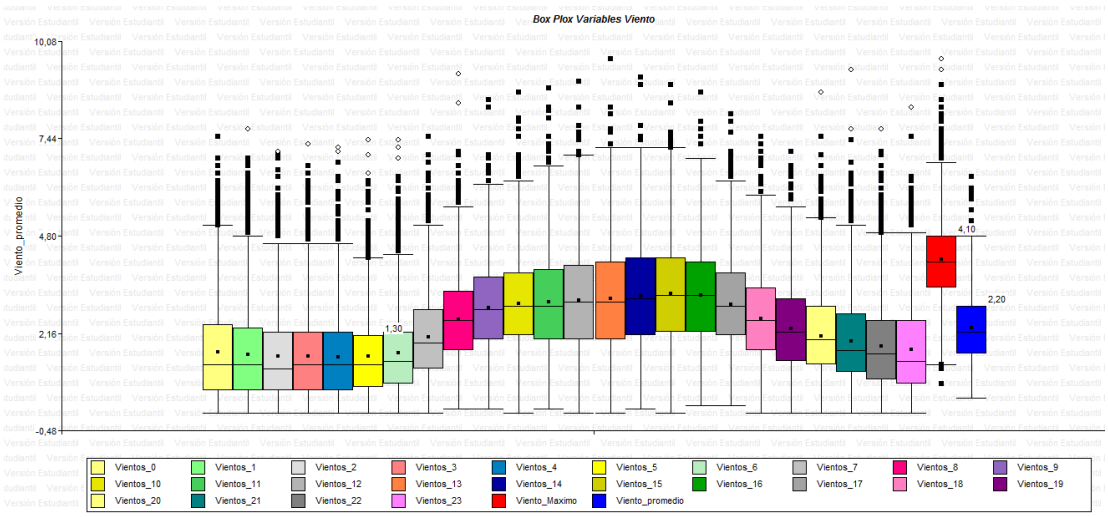
En primer lugar se realizo un análisis descriptivo de las variables y un box plot con InfoStav dividido por los atributos relacionados, a fin de facilitar el análisis.

Análisis de las variables relacionadas a vientos

Análisis de Descriptivos

Variable	N	Datos Faltantes	Mínimo	Máximo	Media	Desv. típ.	Mediana
Vientos_0	2235	299	0,0	7,5	1,64	1,27	1,3
Vientos_1	2242	292	0,0	7,7	1,59	1,27	1,3
Vientos_2	2240	294	0,0	7,1	1,55	1,24	1,2
Vientos_3	2242	292	0,0	7,3	1,53	1,21	1,3
Vientos_4	2241	293	0,0	7,2	1,52	1,2	1,3
Vientos_5	2242	292	0,0	7,4	1,54	1,17	1,3
Vientos_6	2243	291	0,0	7,4	1,64	1,16	1,4
Vientos_7	2245	289	0,0	7,5	2,05	1,16	1,9
Vientos_8	2244	290	0,1	9,2	2,54	1,19	2,5
Vientos_9	2247	287	0,1	8,5	2,85	1,22	2,8
Vientos_10	2246	288	0,0	8,7	2,97	1,3	2,9
Vientos_11	2242	292	0,1	8,8	3,02	1,39	2,9
Vientos_12	2247	287	0,0	9	3,04	1,42	3,0
Vientos_13	2246	288	0,0	9,6	3,11	1,44	3,0
Vientos_14	2246	288	0,1	9,1	3,18	1,42	3,1
Vientos_15	2248	286	0,0	8,9	3,23	1,37	3,2
Vientos_16	2250	284	0,2	8,7	3,19	1,28	3,2
Vientos_17	2251	283	0,2	8,1	2,93	1,23	2,9
Vientos_18	2248	286	0,0	7,5	2,56	1,24	2,5
Vientos_19	2242	292	0,0	7,1	2,29	1,21	2,2
Vientos_20	2240	294	0,0	8,7	2,09	1,21	2,0
Vientos_21	2241	293	0,0	9,3	1,94	1,21	1,7
Vientos_22	2234	300	0,0	7,7	1,8	1,23	1,6
Vientos_23	2237	297	0,0	8,3	1,71	1,26	1,4
Viento Máximo	2261	273	0,8	9,6	4,17	1,17	4,1
Viento_promedio	2261	273	0,4	6,4	2,31	0,92	2,2
Promedio	**2245**	**289**	**0,1**	**8,2**	**2,38**	**1,25**	**2,2**

Box plot



De ambos análisis surge que la velocidad del viento disminuye durante las horas de la noche y comienza a incrementarse a partir de 7 de la mañana, alcanzando sus velocidades máximas alrededor del mediodía; y disminuyendo a partir de las 18 horas. En todas las horas se observa una alta dispersión de valores y muchos outliers sobre todo a las horas del mediodía. Se aprecia también que la media de los valores máximos es casi el doble que la media del promedio de viento en todas las horas.

En lo que respecta a la varianza, presenta valores similares de dispersión durante todo el día (aproximadamente 1,2 km/h) con excepción de las horas del mediodía donde la varianza es mayor. Además, se aprecia que la varianza es aproximadamente la mitad del valor de la media, lo que muestra la alta dispersión de valores.

También se observa que en todos los casos hay aproximadamente entre un 10-15% de valores perdidos.

Análisis de las variables relacionadas a Temperatura

Análisis de Descriptivos

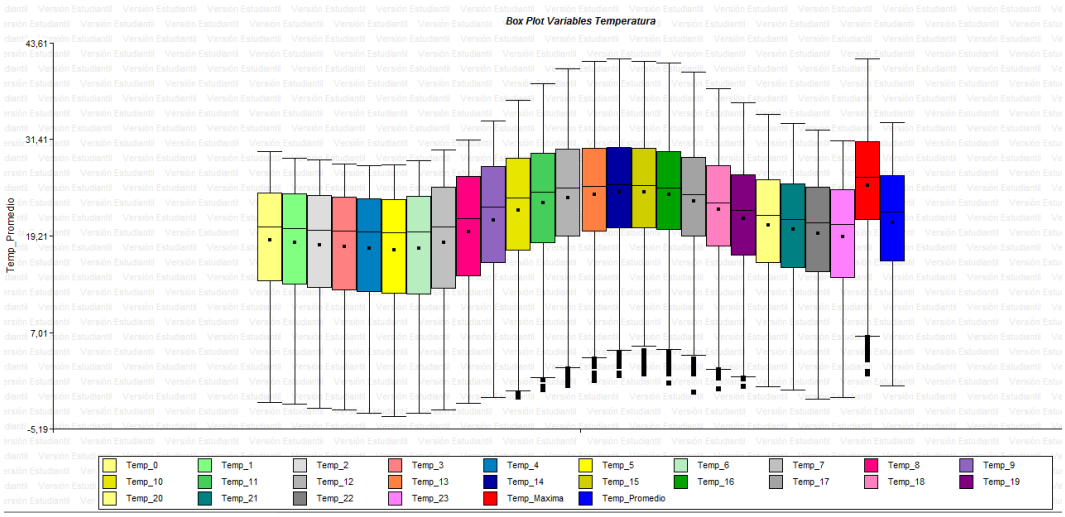
Variable	N	Datos Faltantes	Mínimo	Máximo	Media	Desv. típ.	Mediana
Temp_0	2344	190	-1,8	29,9	18,65	7,02	20,4
Temp_1	2349	185	-2,1	29	18,35	7,09	20,2
Temp_2	2347	187	-2,6	28,8	18,06	7,15	19,9
Temp_3	2350	184	-2,8	28,3	17,82	7,21	19,9
Temp_4	2350	184	-3,2	28,1	17,61	7,25	19,7
Temp_5	2351	183	-3,6	28,2	17,48	7,31	19,6
Temp_6	2351	183	-3,2	28,7	17,59	7,52	19,7
Temp_7	2351	183	-2,8	30,1	18,42	7,87	20,4
Temp_8	2349	185	-1,9	31,4	19,78	7,88	21,4
Temp_9	2349	185	-1,2	33,8	21,22	7,76	22,9
Temp_10	2346	188	-1,2	36,4	22,46	7,69	24,0
Temp_11	2342	192	-0,3	38,5	23,39	7,63	24,8
Temp_12	2345	189	0,3	40,4	24,03	7,56	25,3
Temp_13	2343	191	0,9	41,3	24,43	7,47	25,5
Temp_14	2342	192	1,5	41,6	24,71	7,38	25,7
Temp_15	2347	187	1,7	41,3	24,72	7,3	25,6
Temp_16	2350	184	0,6	41,1	24,4	7,24	25,3
Temp_17	2352	182	-0,6	39,9	23,63	7,18	24,4
Temp_18	2350	184	-0,2	37,8	22,51	7,09	23,4
Temp_19	2346	188	0,1	36,1	21,43	6,92	22,5
Temp_20	2345	189	0,2	34,6	20,62	6,87	21,8
Temp_21	2349	185	-0,3	33,4	20,03	6,86	21,3
Temp_22	2342	192	-1,4	32,6	19,5	6,9	20,9
Temp_23	2345	189	-1,2	31,3	19,06	6,96	20,7
Temp_Máxima	2359	175	1,7	41,6	25,58	7,15	26,6
Temp_Promedio	2359	175	0,3	33,6	20,84	7,01	22,2
Promedio	**2348**	**186**	**-0,9**	**34,5**	**21,01**	**7,28**	**22,5**

Box plot

De ambos análisis surge que la temperatura disminuye durante las horas de la noche y comienza a incrementarse a partir de 7 de la mañana, alcanzando sus valores máximas alrededor del mediodía; y disminuyendo a partir de las 19/20 horas. En todas las horas se observa una alta dispersión de valores y muchos outliers sobre todo a las horas del mediodía.

En lo que respecta a la varianza, presenta valores similares de dispersión durante todo el día con excepción de las horas entre las 7-12 donde la varianza es mayor. Además, se aprecia que la varianza es aproximadamente un tercio del valor de la media, lo que muestra la alta dispersión de valores.

También se observa que en todos los casos hay aproximadamente entre un 7% de valores perdidos.

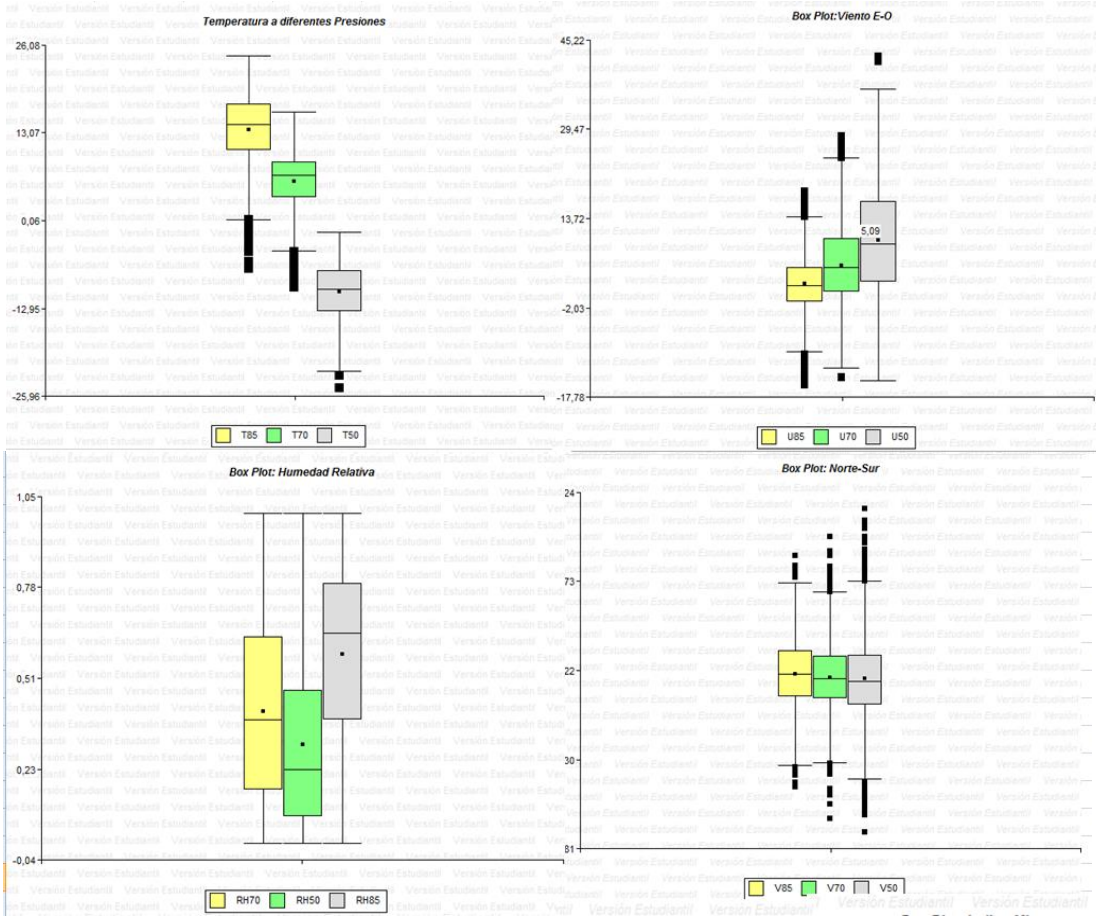


Análisis del resto de las variables

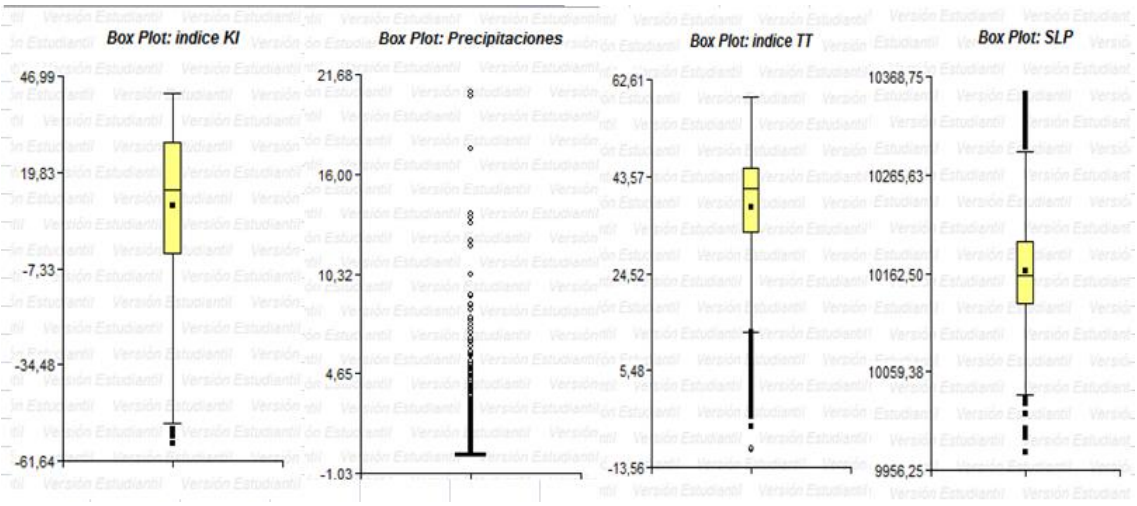
Análisis de Descriptivos

Variable	N	Datos Faltantes	Mínimo	Máximo	Media	Desv. típ.	Mediana
HT50	2422	112	5480	5965	5818,82	79,18	5835
HT70	2434	100	2919	3249	3145,42	49,18	3153,5
HT85	2439	95	1351	1642	1531,49	36,69	1535
KI	2398	136	-56,7	42,05	10,51	20,72	14,93
Precipitaciones	2532	2	0	20,65	0,37	1,32	0
RH50	2409	125	0,01	1	0,3	0,25	0,23
RH70	2419	115	0,01	1	0,41	0,27	0,38
RH85	2429	105	0,01	1	0,58	0,26	0,64
SLP	2439	95	9975	10350	10164,2	52,42	10160
SLP_	2376	158	-135	140	-0,12	35,83	0
T50	2419	115	-24,8	-1,7	-10,51	3,88	-10,1
T70	2427	107	-9,9	16,2	5,93	3,87	6,8
T85	2435	99	-7,1	24,5	13,58	4,87	14,3
TT	2409	125	-10,1	59,15	37,39	11,23	41,1
U50	2324	210	-14,92	42,36	9,87	9,53	9,25
U70	2377	157	-14,37	28,21	5,46	6,68	5,09
U85	2354	180	-15,77	18,56	2,14	4,73	1,88
V50	2324	210	-25,99	30,42	0,83	7,36	0,36
V70	2377	157	-23,68	25,54	0,99	6,19	0,86
V85	2354	180	-18,1	22,16	1,66	6,13	1,55

Box plot

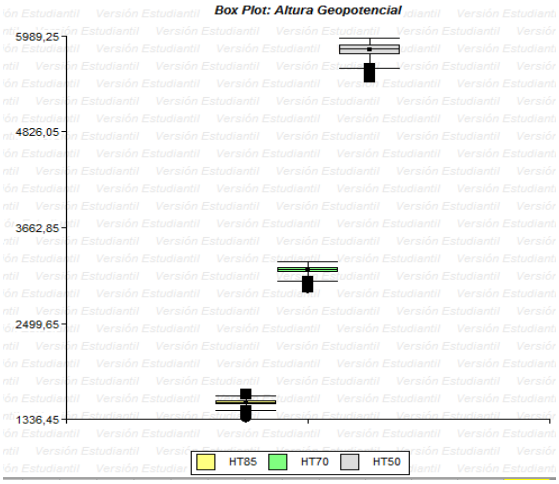


En las variables altura geopotencial y velocidad del viento este-oeste a medida que aumenta la altura disminuye la mediana. Mientras que para las variables Temperatura, Humedad Relativa y viento norte-sur a medida que aumenta la altura se incrementan los valores de la mediana.



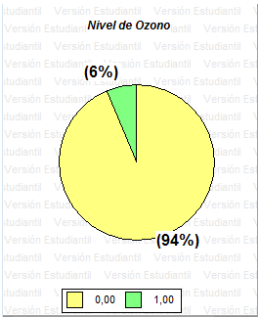
A continuación un pequeño análisis de cada variable:

- Altura Geopotencial: se observa que el desvío es pequeño en relación a la media. Se aprecia que según la altura, las variables son muy diferentes y que la cantidad de datos faltantes es alrededor del 5%



- Índice KI: este atributo presenta una amplia dispersión de valores con aproximadamente un 5% de valores perdidos.
- Precipitaciones: la mediana es 0 y presenta un gran número de outliers. Lo cual es lógico ya que la mayoría de los días no llueve.
- Humedad Relativa: aquí se muestra una alta dispersión de valores, con una mediana creciente a medida que aumenta la altura.
- Presión al nivel del mar: en este caso el desvió es pequeño en relación a la media. En el box plot se observa un gran número de outliers.
- Temperatura: las medias varían significativamente según la altura. Además, la dispersión es elevada.
- Índice TT: presenta una alta dispersión, y un gran número de outliers.
- Velocidad del viento E-O: muestran una alta dispersión y un cantidad de valores perdidos cercana al 10%
- Velocidad del viento N-S: la dispersión es superior a la media y tiene un gran número de outliers.

Análisis de la clase



En el grafico se observa que en la mayoría de los registros (94%) la clase corresponde a un nivel normal de ozono. Esto nos muestra que las clases están desbalanceadas, por lo tanto, debemos tratar el dataset a fin de corregir este efecto.

Tratamiento de Datos Faltantes

El dataset cuenta con 184982 celdas de las cuales 14937 son datos faltantes, es decir, 8% de los valores presentan valores perdidos. Sin embargo, si lo analizamos a nivel de registros, se observa que 687 días presentan algún dato faltante; es decir el 25 % del total.

Los datos faltantes podrían imputarse por un método “cold deck”, es decir, se toman valores de otras bases de datos o se calculan a partir de relaciones también obtenidos de otras

fuentes. En este caso al ser variables climáticas tienen que existir en la web otros dataset que permitan imputar los datos faltantes de la misma, ya que contamos con la información del día y la hora de cada valor perdido.

Para los fines de este trabajo practico, no se realizara dicha imputación. Pero si se tratara de un caso real, sería muy útil poder realizarla; a pesar de que es un trabajo minucioso de búsqueda de información.

Análisis de Componentes Principales

Aunque el análisis de componentes principales, no tiene supuestos previos y siempre es posible aplicarlo, previo a efectuar el mismo se analizara si tiene sentido realizarlo; para ello se estudiará la correlación entre las variables de a pares.

Debido a que tenemos 72 variables continuas, y no tendría sentido mostrar una matriz de 72x72, ni un grafico de correlación entre variables de a pares; se construyo una tabla donde se indica la cantidad de pares (sin considerar los pares repetidos ni cada variables contra sí misma) que tienen un índice de correlación mayor que un determinado número (cuando se refiere a mayor, se habla en valor absoluto; correlación tanto positiva (>0) o negativa (<0)).

% Correlación	Cantidad de Pares	% Pares
>0.5	733	29%
>0.6	638	25%
>0.7	522	20%
>0.8	382	15%
>0.9	213	8%

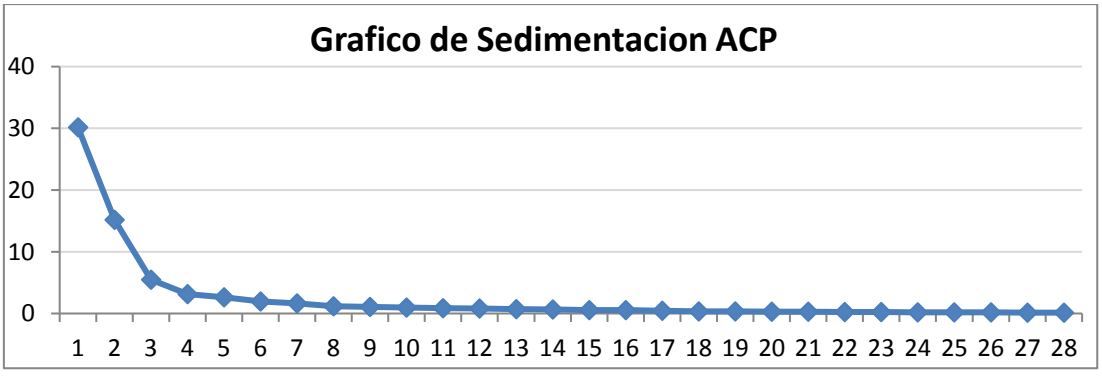
En la tabla de observa que más del 25% de los pares tiene una correlación del 60% y un 15% de los mismos presentan una índice de regresión mayor al 80%. Estos nos indicarían que muchos pares de variables están fuertemente relacionados entre sí; entonces, sería lógico pensar en efectuar un análisis de componentes principales.

Análisis de Componentes Principales

Luego de realizar la corrida del ACP con las variables estandarizadas en InfoStav se obtuvieron los siguientes resultados

Autovalor	Valor	Proporción	Prop Acum
1	30,19	0,42	0,42
2	15,2	0,21	0,63
3	5,47	0,08	0,71
4	3,17	0,04	0,75
5	2,62	0,04	0,79
6	1,93	0,03	0,81
7	1,66	0,02	0,84
8	1,18	0,02	0,85
9	1,06	0,01	0,87
10	0,95	0,01	0,88
11	0,86	0,01	0,89
12	0,79	0,01	0,9
13	0,69	0,01	0,91
14	0,64	0,01	0,92
15	0,56	0,01	0,93
16	0,53	0,01	0,94
17	0,43	0,01	0,94
18	0,36	0,01	0,95
19	0,32	0,00	0,95
20	0,29	0,00	0,96
21	0,27	0,00	0,96
22	0,24	0,00	0,96
23	0,23	0,00	0,97
24	0,21	0,00	0,97
25	0,18	0,00	0,97
26	0,17	0,00	0,97
27	0,16	0,00	0,98
28	0,14	0,00	0,98

Como se observa en la tabla con 10 componentes se explica casi el 90% de la variabilidad del sistema, este número parece elevado pero si tenemos en cuenta que el dataset cuenta con 73 atributos, parece razonable trabajar con 10 variables. A continuación se muestran los mismos resultados en un grafico de sedimentación.

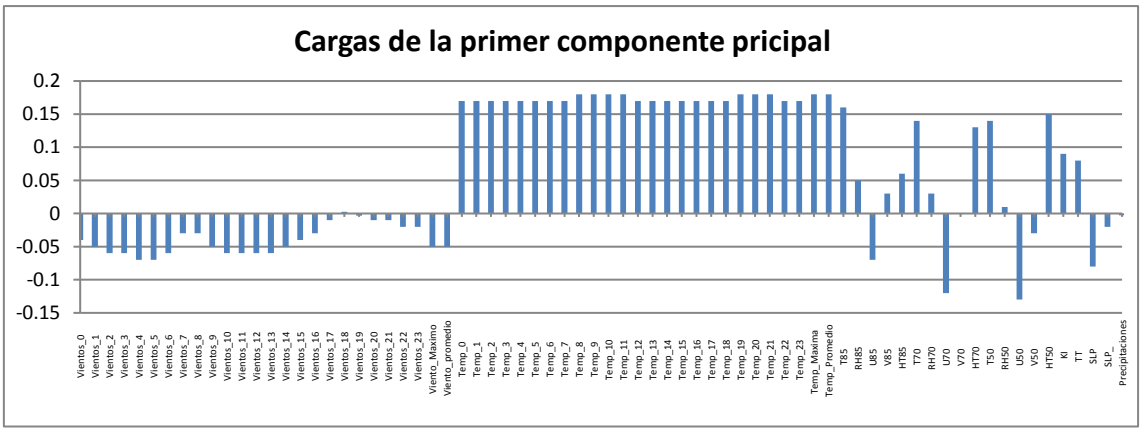


También se aprecia que con tres componentes se explica el 71% de la variabilidad del sistema (en el grafico se aprecia el cambio en la pendiente a partir de la tercera componente). Mientras que con 5 componentes explicamos casi el 80% de la variabilidad. A fines de facilitar el análisis, trabajaremos con 3 componentes, ya que nos permite graficarlas y observar resultados.

Analizaremos los autovectores de las 3 componentes (en el Anexo II se observa la tabla con los resultados).

Cargas primera componente principal

Esta componente tiene coeficientes positivos y negativos; entonces lo interpretamos como una componente de forma. Se observa que todas las variables relacionadas a velocidad del viento en cada hora presentan valores negativos mientras que las relacionadas a temperaturas horarias son todas positivas. En lo que respecta al resto de las variables se observa que la mayoría presentan coeficientes positivos con excepción de las relacionadas a la velocidad del viento en dirección este-oeste y las relacionadas al nivel del mar y precipitaciones.

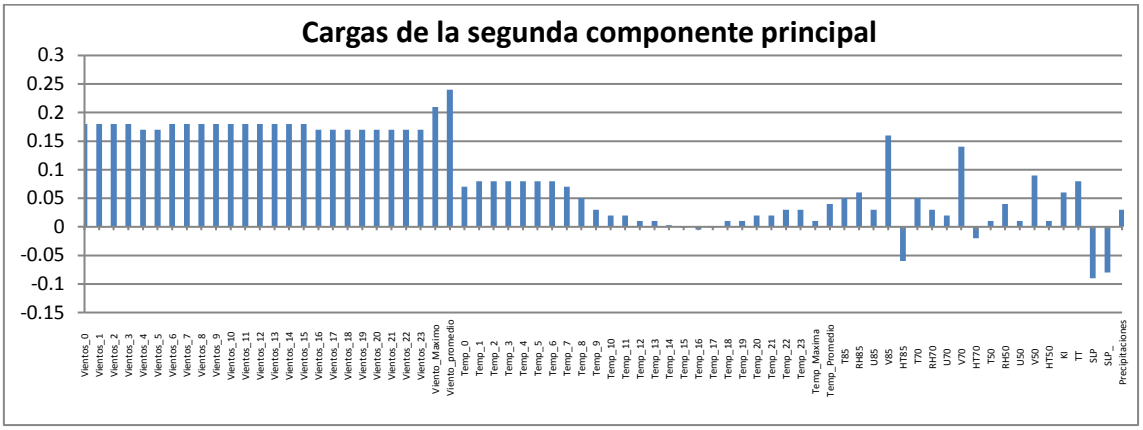


Con el análisis de coeficientes se podría interpretar esta componente como una descripción de las condiciones climáticas generales del día: tomando como positivo que el día estaría presentando buenas condiciones y negativo que no; ya que a mayor temperatura el valor se incrementa y a mayor velocidad del viento el mismo disminuye. Con respecto a las otras variables se puede decir que a mayor nivel de precipitaciones el valor de la componente decrece, al igual que con el viento este-oeste.

Entonces la primera componente podría explicarse como el nivel de ozono, ya que las temperaturas elevadas promueven la acumulación de gases, mientras que los vientos y precipitaciones dispersan los gases, por lo tanto, el nivel de ozono, disminuye.

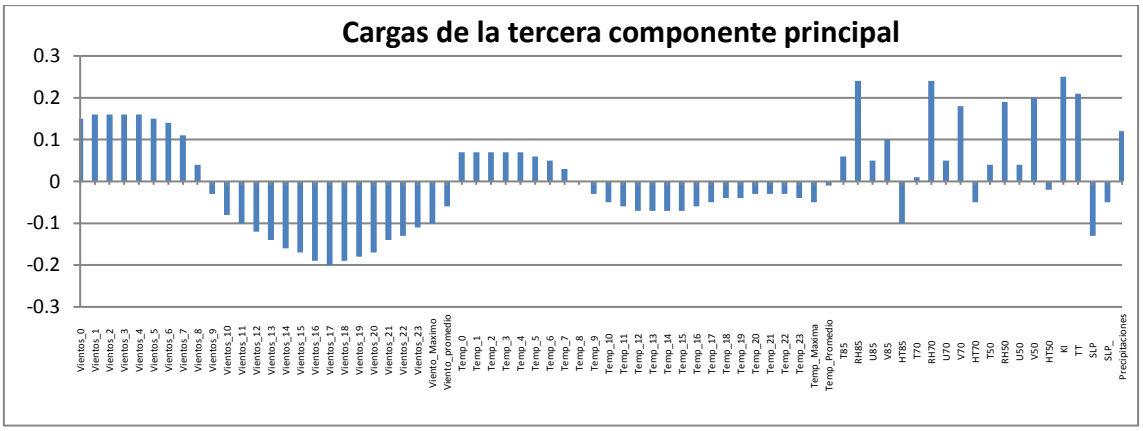
Cargas segunda componente principal

Esta también es una componente de forma pero en este caso considerando tanto a las velocidades del viento como a las temperaturas como positivos. En este caso solo considera con coeficientes apenas menores que 0 a las alturas geopotenciales y a las presiones sobre el nivel del mar.

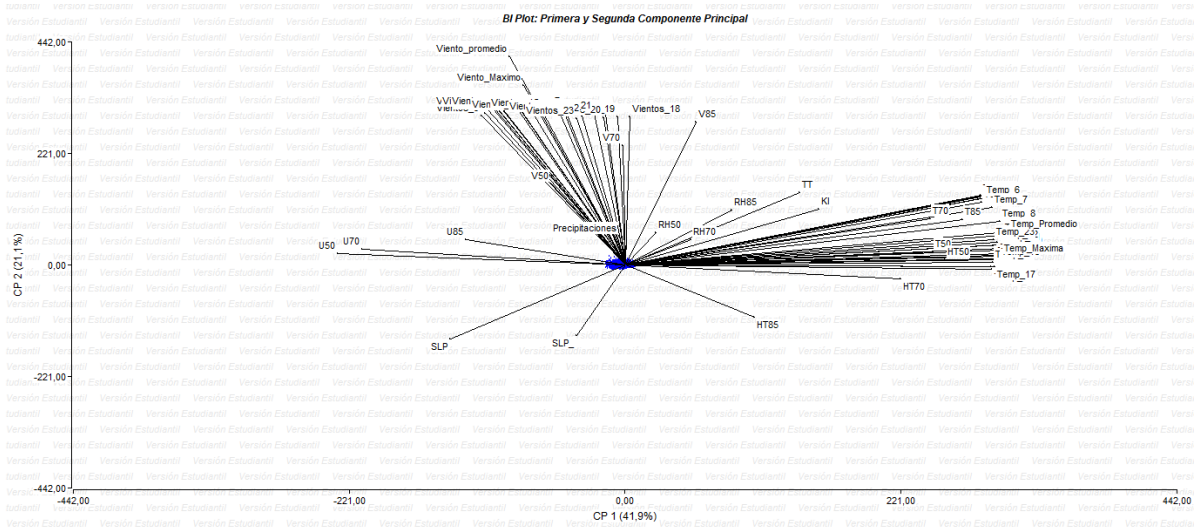


Cargas tercera componente principal

Esta componente que también es de forma por presentar coeficientes tanto positivos como negativos, considera al grupo de velocidades del viento hasta las 9 horas como positivas y luego como negativas. Lo mismo ocurre con la temperatura, quedando como positivas desde las 0 hasta las 9 horas y, negativas para el resto de las horas de día.



Biplot: Primera y Segunda Componente



En el biplot se observa que todas las variables relacionadas a la temperatura forman entre sí ángulos muy pequeños, y que todas las variables relacionadas a la velocidad del viento en las

diferentes horas de día también; estos ángulos pequeños indican que las variables de temperatura están fuertemente correlacionadas entre sí (situación que era esperable); y lo mismo sucede con las variables vinculadas a la velocidad horaria del viento.

Por otro lado se observa que el conjunto de variables de velocidad del viento están prácticamente ortogonales al conjunto de variables de temperatura; lo cual indica que ambos grupos no están correlacionados.

Adicionalmente se pueden mencionar los siguientes puntos:

1. Los pares de variables V85 con HT85, V70 con HT70 y V50 con HT50, presentan ortogonalidad entre sí; esto nos indica que independientemente del nivel de hpa la velocidad del viento norte-sur, no está relacionada con la altura geopotencial.
2. Los pares de variables U85 con HT85, U70 con HT70 y U50 con HT50, presentan entre sí ángulos de casi 180º; esto nos indica que independientemente del nivel de hpa la velocidad del viento este-oeste esta correlacionada negativamente con la altura geopotencial
3. Las variables U50, U70 y U85; presentan ángulos muy pequeños entre sí, indicando una correlación positiva. Lo mismo ocurre con V50, V70 y V85.
4. Los coeficientes KI y TT muestran ángulos pequeños entre sí sugiriendo una correlación entre lo mismo; lo cual resultaría razonable ya que ambos son índices de tormenta. Lo que sorprende es que ambos son casi ortogonales con el atributo precipitaciones, ya que intuitivamente relacionamos las tormentas con las lluvias.
5. Las variables de presión al nivel del mar (SLP y SLP_1) están correlacionadas entre sí.

Conclusiones ACP

Del análisis de componentes principales podemos concluir que las variables de viento y temperatura están muy relacionadas entre sí; y que es posible hacer una reducción del variables (De los 73 atributos iníciales podemos trabajar con entre 3 y 5 componentes los cuales nos dan brindan el 80% de la información del sistema). También pudimos observar relaciones entre las precipitaciones y los índices de tormentas y entre las alturas geopotenciales y los vientos, que son útiles para la interpretación de los resultados.

Análisis Discriminante

El objetivo es ver si las condiciones climáticas variaron a lo largo de los años, entonces, se quiere ver si año es una variable con poder discriminante.

En primer lugar debemos verificar si se cumplen los supuestos del modelo:

1) Homogeneidad de Matrices de covarianza: como el test da significativo, se rechaza la hipótesis nula, entonces no se cumple el supuesto de que las varianza de los grupos sean iguales.

Prueba de Homogeneidad de Matrices de Covarianzas

Grupos	N	Estadístico	gl	p-valor
7	1847	23593,87	15768	<0,0001

2) Normalidad: aplicamos el test de Shapiro-Wilks como todas las variables dan significativas, se rechaza la hipótesis de de normalidad de las variables

Shapiro-Wilks (modificado)

Variable	n	Media	D.E.	W*	p(Unilateral D)
Vientos_0	2235	1,64	1,27	0,91	<0,0001
Vientos_1	2242	1,59	1,27	0,90	<0,0001
Vientos_2	2240	1,55	1,24	0,90	<0,0001
Vientos_3	2242	1,53	1,21	0,90	<0,0001
Vientos_4	2241	1,52	1,20	0,90	<0,0001
Vientos_5	2242	1,54	1,17	0,90	<0,0001
Vientos_6	2243	1,64	1,16	0,92	<0,0001
Vientos_7	2245	2,05	1,16	0,96	<0,0001
Vientos_8	2244	2,54	1,19	0,98	<0,0001
Vientos_9	2247	2,85	1,22	0,99	<0,0001
Vientos_10	2246	2,97	1,30	0,99	<0,0001
Vientos_11	2242	3,02	1,39	0,98	<0,0001
Vientos_12	2247	3,04	1,42	0,98	<0,0001
Vientos_13	2246	3,11	1,44	0,98	<0,0001
Vientos_14	2246	3,18	1,42	0,99	<0,0001
Vientos_15	2248	3,23	1,37	0,99	<0,0001
Vientos_16	2250	3,19	1,28	0,99	<0,0001
Vientos_17	2251	2,93	1,23	0,99	<0,0001
Vientos_18	2248	2,56	1,24	0,98	<0,0001
Vientos_19	2242	2,29	1,21	0,98	<0,0001
Vientos_20	2240	2,09	1,21	0,97	<0,0001
Vientos_21	2241	1,94	1,21	0,95	<0,0001
Vientos_22	2234	1,80	1,23	0,93	<0,0001
Vientos_23	2237	1,71	1,26	0,92	<0,0001
Viento_Maximo	2261	4,17	1,17	0,98	<0,0001
Viento_promedio	2261	2,31	0,92	0,96	<0,0001
Temp_0	2344	18,65	7,02	0,93	<0,0001
Temp_1	2349	18,35	7,09	0,93	<0,0001
Temp_2	2347	18,06	7,15	0,92	<0,0001
Temp_3	2350	17,82	7,21	0,92	<0,0001
Temp_4	2350	17,61	7,25	0,92	<0,0001
Temp_5	2351	17,48	7,31	0,92	<0,0001
Temp_6	2351	17,59	7,52	0,92	<0,0001
Temp_7	2351	18,42	7,87	0,93	<0,0001
Temp_8	2349	19,78	7,88	0,93	<0,0001
Temp_9	2349	21,22	7,76	0,94	<0,0001

Variable	n	Media	D.E.	W*	p(Unilateral D)
Temp_10	2346	22,46	7,69	0,94	<0,0001
Temp_11	2342	23,39	7,63	0,95	<0,0001
Temp_12	2345	24,03	7,56	0,95	<0,0001
Temp_13	2343	24,43	7,47	0,96	<0,0001
Temp_14	2342	24,71	7,38	0,96	<0,0001
Temp_15	2347	24,72	7,30	0,97	<0,0001
Temp_16	2350	24,40	7,24	0,97	<0,0001
Temp_17	2352	23,63	7,18	0,97	<0,0001
Temp_18	2350	22,51	7,09	0,97	<0,0001
Temp_19	2346	21,43	6,92	0,96	<0,0001
Temp_20	2345	20,62	6,87	0,95	<0,0001
Temp_21	2349	20,03	6,86	0,94	<0,0001
Temp_22	2342	19,50	6,90	0,94	<0,0001
Temp_23	2345	19,06	6,96	0,93	<0,0001
Temp_Maxima	2359	25,58	7,15	0,96	<0,0001
Temp_Promedio	2359	20,84	7,01	0,94	<0,0001
T85	2435	13,58	4,87	0,96	<0,0001
RH85	2429	0,58	0,26	0,94	<0,0001
U85	2354	2,14	4,73	0,99	<0,0001
V85	2354	1,66	6,13	1,00	0,1446
HT85	2439	1531,49	36,69	0,97	<0,0001
T70	2427	5,93	3,87	0,95	<0,0001
RH70	2419	0,41	0,27	0,95	<0,0001
U70	2377	5,46	6,68	0,99	<0,0001
V70	2377	0,99	6,19	1,00	<0,0001
HT70	2434	3145,42	49,18	0,96	<0,0001
T50	2419	-10,51	3,88	0,97	<0,0001
RH50	2409	0,30	0,25	0,90	<0,0001
U50	2324	9,87	9,53	0,99	<0,0001
V50	2324	0,83	7,36	0,99	<0,0001
HT50	2422	5818,82	79,18	0,95	<0,0001
KI	2398	10,51	20,72	0,93	<0,0001
TT	2409	37,39	11,23	0,88	<0,0001
SLP	2439	10164,20	52,42	0,98	<0,0001
SLP_	2376	-0,12	35,83	0,97	<0,0001
Precipitaciones	2532	0,37	1,32	0,31	<0,0001

A pesar de que no se cumplen los supuestos del modelo, aplicaremos análisis discriminante. A continuación se observan los resultados obtenidos

Classification Results[a,b]

			Clase	Predicted Group Membership 0	Predicted Group Membership 1	Total
Cases Selected	Original	Count	0	1126	322	1448
			1	12	98	110
		%	0	77,8	22,2	100,0
			1	10,9	89,1	100,0
Cases Not Selected	Original	Count	0	454	130	584
			1	6	31	37
		%	0	77,7	22,3	100,0
			1	16,2	83,8	100,0

a. 78,6% of selected original grouped cases correctly classified.

b. 78,1% of unselected original grouped cases correctly classified.

Se puede apreciar que se puede discriminar los grupos mediante una función discriminante, ahora vamos a ver si el año es una variable que sea significativa para poder discriminar los grupos.

Como se observa la variable año presenta un valor no significativo en el test de igualdad de medias, por lo tanto, no es una variable que tenga poder discriminante en el modelo.

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
año	1,000	,004	1	1556	,951
WSR0	,966	54,919	1	1556	,000
WSR1	,965	56,984	1	1556	,000
WSR2	,964	57,354	1	1556	,000
WSR3	,967	52,580	1	1556	,000
WSR4	,970	48,747	1	1556	,000
WSR5	,971	46,609	1	1556	,000
WSR7	,982	28,492	1	1556	,000
WSR8	,985	24,292	1	1556	,000
WSR12	,952	79,233	1	1556	,000
WSR17	1,000	,095	1	1556	,759
WSR18	1,000	,022	1	1556	,881

Clustering

Se aplicaran técnicas de clustering para ver cuales variables/ condiciones climáticas están relacionadas a altos niveles de ozono.

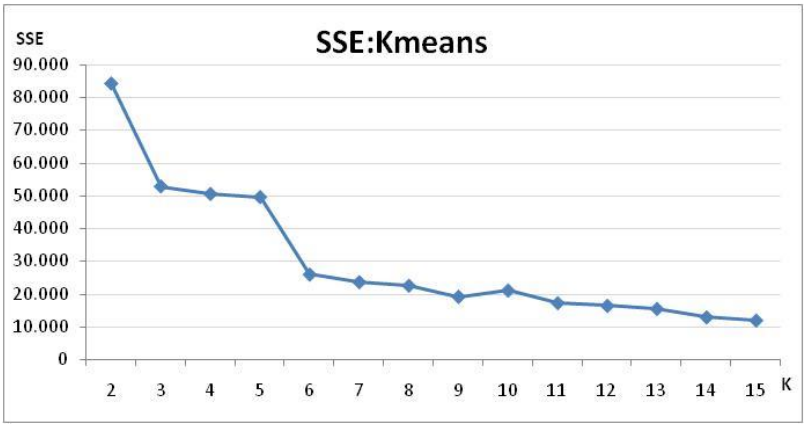
Kmeans

Se aplico esta técnica corriendo diferentes clúster en R desde k=2 a K=15 y se aplicaran las medidas estadísticas de distancia de Silhouette y SSE para determinar cuál es el mejor K a seleccionar.

Debido a que las variables presentan distintas unidades de medidas, se estandarizaron las variables.

SSE: Análisis de Resultados

Siendo el SSE la suma de los errores al cuadrado, entre más chico sea el SSE tendré mejores clústeres, sin embargo, tener una gran numero de conglomerados no es algo bueno para el análisis. Por lo tanto debo elegir un K tal que disminuya el SSE.



En el grafico de K vs SSE se observa que pasando de K=2 a 3; el SSE disminuye en casi un 50%. Luego se mantiene constantes con K=3,4 o 5. Para disminuir nuevamente de manera abrupta para K=6. Donde si comparamos la disminución del SSE desde K=2 a K=6 la misma es de un 70%.

En la tabla se observa la disminución del SSE para cada K vs K=2.

K	Disminucion del SSE
3	37%
4	40%
5	41%
6	69%
7	72%
8	73%
9	77%
10	75%
11	79%
12	81%
13	82%
14	85%
15	86%

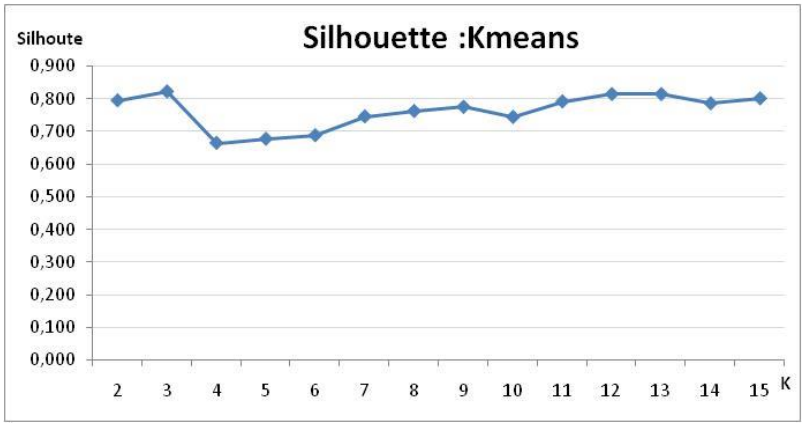
En la tabla se observa que para lograr un disminución significativa desde K=6, debemos pasar a k=9, lo cual implicaría introducir un gran número de clúster. También, se aprecia que desde K=9 a k =15, el SSE no disminuye demasiado, sino que lo hace muy lentamente.

Por lo tanto, seleccionaremos un número de entre 6-8.

Ahora analizaremos los resultados para Silhouette para ver cuál es el mejor clúster, en función de dos medidas para su validación.

Silhouette: Análisis de Resultados

En el grafico se observan los resultados del cálculo de la distancia de Silhouette, recordemos que entre mayor sea el valor del Silhoutte en valor absoluto, mejor será el clúster. Aquí se observa que, a diferencia de los resultados de SSE, los mejores clusters se encuentran con k=3. Luego disminuye con K=4,5 y 6; para luego mejorar nuevamente con K mayores a 7.



Algo interesante es que los valores de Silhouette, se encuentran siempre por encima del 60%, indicando que, para la mayoría de los K, los conglomerados son adecuados.

En función de los resultados de ambas medidas estadísticas, se seleccionara K=7, ya que con el mismo la distancia de Silhouette es del 75% y, el SSE disminuye en más de un 70% con respecto a K=2.

Clúster: Análisis de Conglomerados

K	Nivel de Ozono	Total
1	Alto	145
1 Total		145
2	Bajo	103
	Alto	2
2 Total		105
3	Bajo	170
	Alto	6
3 Total		176
4	Bajo	83
	Alto	5
4 Total		88
5	Bajo	1794
5 Total		1794
6	Bajo	168
	Alto	2
6 Total		170
7	Bajo	56
7 Total		56

Analizaremos los resultados de K=7. En la tabla se observa cuantos casos se encuentran agrupados en cada clúster, y de los mismos, cuales corresponden a un alto nivel de ozono y cuáles no.

Se observa que en todos los clúster, predomina una de las dos opciones del nivel de ozono. Entonces tenemos los siguientes tres casos.

1. Clúster 1: Exclusivo de nivel alto de ozono

2. Clusters 5 y 7: Exclusivos de nivel bajo de ozono
3. Clúster 2,3,4 y 5: Presentan muy pocos casos con alto nivel de ozono pero en su mayoría son bajos.

Ahora analizaremos cuales son las características predominantes de cada variable en cada uno de los grupos.

Temperatura en las distintas horas del día

La siguiente tabla muestra las medias de las temperaturas horarias para cada uno de los grupos. Se indicó el rojo el valor máximo y en verde el valor mínimo de cada atributo en relación a los otros clúster.

Atributo	K						
	1	2	3	4	5	6	7
Temp_0	22	12	23	21	19	16	19
Temp_1	22	12	23	21	19	16	19
Temp_10	29	15	0	25	23	20	23
Temp_11	30	16	0	26	23	21	24
Temp_12	31	17	0	27	24	21	24
Temp_13	31	18	0	27	24	22	25
Temp_14	32	18	0	27	25	22	25
Temp_15	32	18	0	27	25	22	25
Temp_16	31	18	0	27	24	22	25
Temp_17	30	17	0	26	24	21	24
Temp_18	29	16	0	25	22	20	23
Temp_19	27	15	0	24	21	19	22
Temp_2	21	12	24	20	18	16	18
Temp_20	26	14	0	23	21	18	21
Temp_21	25	14	0	22	20	18	21
Temp_22	24	13	0	22	20	17	20
Temp_23	24	13	0	21	19	17	20
Temp_3	21	11	24	20	18	15	18
Temp_4	21	11	23	20	18	15	18
Temp_5	21	11	24	20	18	15	18
Temp_6	21	11	24	20	18	15	18
Temp_7	23	11	25	21	19	16	19
Temp_8	25	12	0	22	20	17	20
Temp_9	27	14	0	24	21	18	22
Temp_Maxima	32	19	25	28	26	23	26
Temp_Promedio	26	14	24	23	21	18	21

Se aprecia para el grupo de K=1 (único grupo con altos niveles de ozono) que presenta las máximas temperaturas durante las horas del día (desde 8 hs hasta 23 horas). Mientras que el conglomerado de K=3, presenta la situación inversa: las menores temperaturas durante el día y las mayores temperaturas durante la noche. El grupo 2 presenta las menores temperaturas durante la noche.

Viento en las distintas horas del día

La siguiente tabla muestra las medias de la velocidad del viento a distintas horas para cada uno de los grupos. Se indicó el rojo el valor máximo y en verde el valor mínimo de cada atributo en relación a los otros clúster.

Atributo	K						
	1	2	3	4	5	6	7
Viento_Maximo	3,5	3,6	2,3	4,2	4,2	4,2	4,3
Viento_promedio	1,7	2,6	1,4	2,4	2,4	2,3	2,4
Vientos_0	0,8	2,2	0,5	1,8	1,7	1,7	1,9
Vientos_1	0,7	2,0	0,7	1,7	1,6	1,7	1,9
Vientos_10	1,9	2,5	0,0	3,1	3,1	3,0	3,0
Vientos_11	1,9	2,9	0,0	3,1	3,1	3,0	3,2
Vientos_12	1,9	4,9	0,0	3,1	3,1	3,0	3,1
Vientos_13	2,1	3,9	0,0	3,2	3,2	3,1	3,0
Vientos_14	2,3	4,2	0,0	3,3	3,2	3,2	3,3
Vientos_15	2,6	4,1	0,0	3,4	3,3	3,2	3,3
Vientos_16	2,8	4,1	0,0	3,2	3,2	3,2	3,2
Vientos_17	2,9	3,0	0,0	3,0	2,9	2,9	3,0
Vientos_18	2,5	2,0	0,0	2,7	2,6	2,3	2,6
Vientos_19	2,1	1,9	0,0	2,5	2,3	2,0	2,3
Vientos_2	0,7	1,8	1,0	1,6	1,6	1,6	1,8
Vientos_20	1,8	1,8	0,0	2,2	2,1	1,8	2,2
Vientos_21	1,5	2,1	0,0	2,1	2,0	1,7	2,1
Vientos_22	1,3	2,1	0,0	2,0	1,9	1,6	1,9
Vientos_23	1,1	2,3	0,0	1,8	1,8	1,5	1,8
Vientos_3	0,8	1,5	1,4	1,6	1,6	1,6	1,8
Vientos_4	0,8	1,4	1,8	1,5	1,6	1,6	1,7
Vientos_5	0,9	1,4	2,3	1,6	1,6	1,7	1,7
Vientos_6	1,0	0,9	1,8	1,6	1,7	1,8	1,8
Vientos_7	1,5	2,6	2,0	2,2	2,1	2,2	2,1
Vientos_8	2,0	3,2	0,0	2,7	2,6	2,7	2,8
Vientos_9	2,0	3,3	0,0	3,0	2,9	3,0	2,8

Se aprecia que el grupo 3, es quien presenta las menores velocidades de viento durante el día. Mientras que el conglomerado de K=1 presenta las menores velocidades durante la noche. El grupo 2 presenta velocidades elevadas durante gran parte de las horas del día.

Resto de los atributos

En la tabla se muestran las medias para el resto de los atributos. A continuación se detallan las conclusiones de la misma

- Las HT son máximas para el conglomerado 1
- El índice KI de tormentas es máximo para el conglomerado 3 y mínimo para el 2.
- El nivel de precipitaciones es máximo en el grupo 3 y mínimo en el 1.
- Los vientos en dirección este-oeste son mínimos en el grupos 3.
- El índice TT es máximo en el grupo 1 y mínimo en el grupo 4.

Atributo	K						
	1	2	3	4	5	6	7
HT50	5862	5774	5833	0	5820	5774	5825
HT70	3169	3117	3150	0	3147	3119	3148
HT85	1541	1518	1532	1531	1533	1516	1531
KI	12,49	-4,66	17,23	0	10,9	7,73	10,15
Nivel de Ozono	1	0,02	0,03	0,06	0	0,01	0
Precipitaciones	0,06	0,36	0,65	0,18	0,39	0,28	0,26
RH50	0,21	0,28	0,37	0	0,31	0,29	0,28
RH70	0,35	0,29	0,52	0	0,41	0,42	0,41
RH85	0,54	0,48	0,65	0,72	0,58	0,57	0,51
SLP	10145	10186	10156	10124	10166	10159	10160
SLP_	-1,38	-0,14	-0,66	-4	-1,11	12,13	-5
T50	-8,48	-12,14	-9,24	0	-10,58	-11,91	-10,14
T70	7,98	4,16	6,3	0	5,99	4,23	6,22
T85	16,8	9,94	14,52	18,04	13,58	11,73	14,43
TT	39,27	29,8	39,04	0	37,72	35,5	36,55
U50	4,0	16,6	7,9	0,0	10,1	10,4	9,3
U70	1,0	8,7	3,8	0,0	5,5	10,1	6,0
U85	-0,4	3,8	1,1	-0,7	2,2	5,1	2,7
V50	-3,2	2,1	1,6	0,0	1,0	2,1	0,4
V70	-2,6	0,8	1,1	0,0	1,3	0,6	0,4
V85	-1,3	0,4	0,9	3,8	2,2	-2,0	1,6

Análisis de grupos

El grupo que más nos interesa es el 1, por ser el que presenta las características asociadas a elevados niveles de ozono. El mismo nos indica que altas temperaturas durante el día junto con un bajo índice de precipitaciones, elevadas alturas geopotenciales, y bajas velocidades de viento durante las horas de las noche, son condiciones que generan un elevado nivel de ozono en la ciudad.

Los resultados del conglomerado 1 son acordes a lo que dicen los expertos: que en los días despejados (lo cual indica un bajo índice de precipitaciones) los niveles de ozono se incrementan. Dado que el calor, la luz solar incrementan la concentración de gases durante el día, la formación de ozono también se incrementa. Adicionalmente, los especialistas afirman que la luz y el calor son los mejores motores que ayudan a la formación de ozono, los días cálidos y soleados deberían tener más ozono que los días frescos y nublados (lo cual en nuestro clúster se representa con elevadas temperaturas durante el día). El viento, también puede jugar un papel importante. En días ventosos, el viento puede dispersar el ozono causando que los niveles bajen (en nuestros clúster 1 de elevados niveles de ozono, la velocidades del viento son bajas). La contaminación con ozono puede ser mala, especialmente durante las olas de calor del verano ya que el aire no se mezcla muy bien y la contaminación no se dispersa.

El resto de los grupos, con bajos niveles de ozono, están más asociados a niveles bajos o medios de temperatura; y mayores velocidades de vientos, los cuales dispersan el cumulo de gases e impiden la formación de ozono.

Año	Total
1998	28
1999	27
2000	38
2001	20
2002	4
2003	18
2004	10

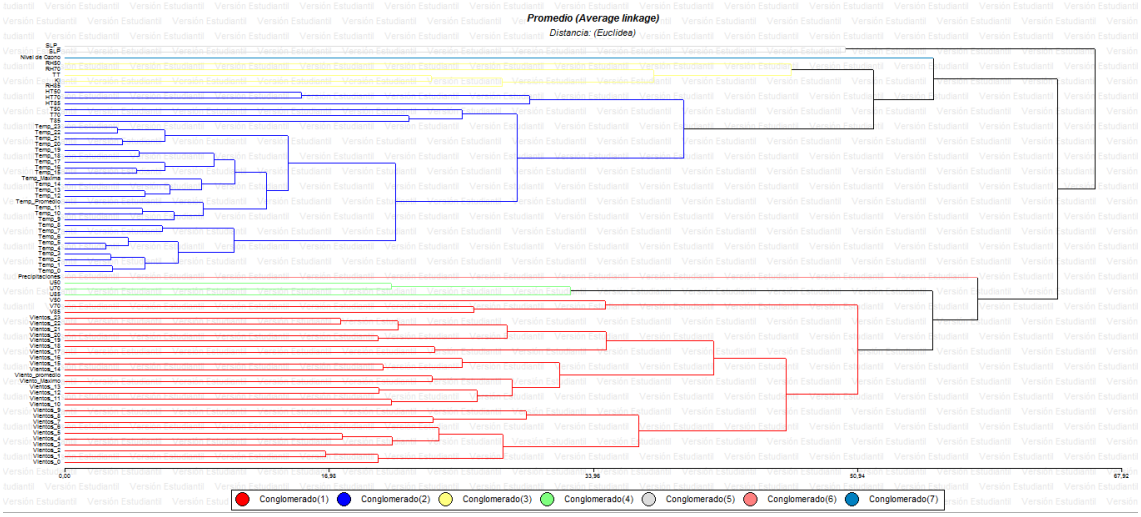
Finalmente analizaremos si la cantidad de días con alto niveles de ozono se incremento o disminuyo con el paso de los años, para ellos veremos la cantidad de casos de cada año en el conglomerado 1. En la tabla se observa que la cantidad de días con altos niveles de ozono se fue incrementando hasta el 2000, con una leve disminución a partir del 2001.

Clúster Jerárquico

Se realizaron distintos clúster jerárquico de las variables variando los métodos y las distancias y para cada uno de ellos se guardo el coeficiente de correlación cofenético para validar cual es el mejor clúster

Para todos los métodos se estandarizaron las variables.

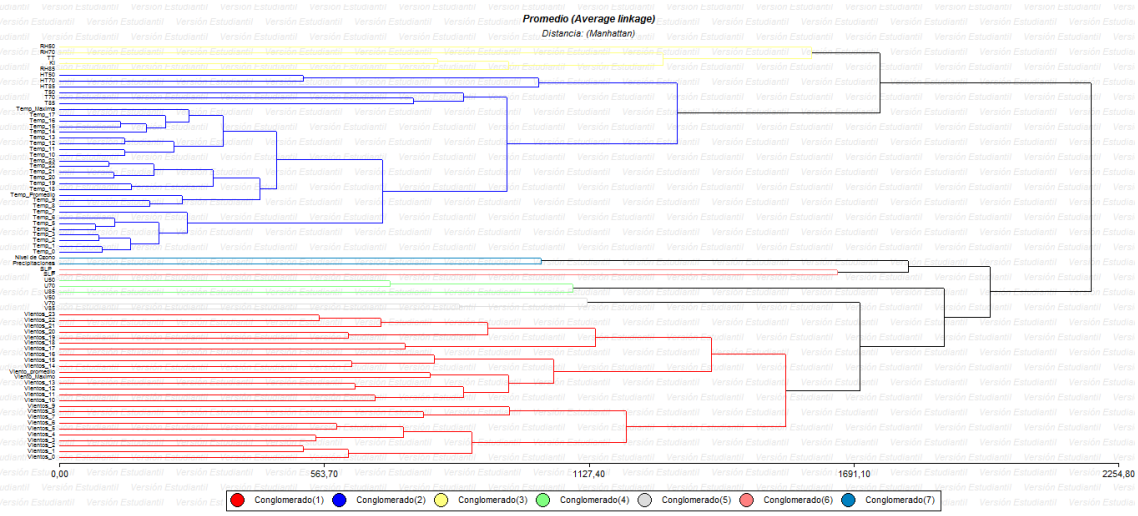
Average Linkage (Distancia Euclidea)



Correlación cofenética= 0,956

Aquí se observa que la variables de las velocidades del viento horarias esta relacionadas a las variables de viendo N-S y E-O, además de las precipitaciones. Mientras que los atributos de temperatura son más cercanos a las humedades relativas y a los niveles de ozono.

Average Linkage (Distancia Manhathan)

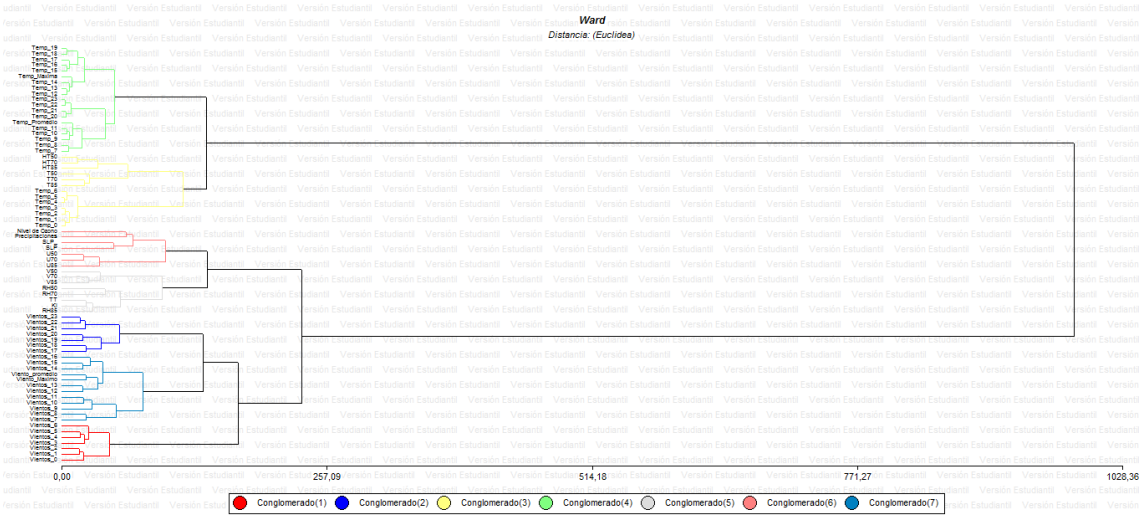


Correlación cofenética= 0,949

En la figura se aprecia que los niveles de ozono están fuertemente relacionados con las precipitaciones y, al igual que en el caso previo las variables de temperatura y velocidad de

viento horarias están relacionadas. Por otra parte, se evidencia una relación entre los atributos de humedades relativas y los índices de tormentas.

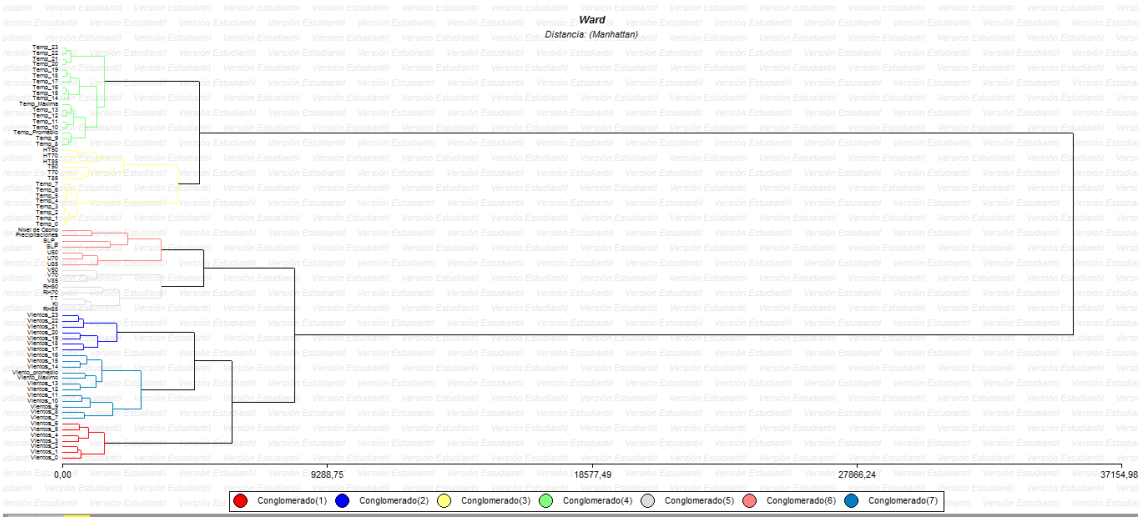
Ward (Distancia Euclidea)



Correlación cofenética= 0,701

En este caso, a diferencia de los casos anteriores, el grupo de temperatura lo abrió en dos, el primero se relacionan entre si las temperaturas durante el día; mientras que en el segundo, relaciones las temperaturas durante las horas nocturnas con la altura geopotencial.

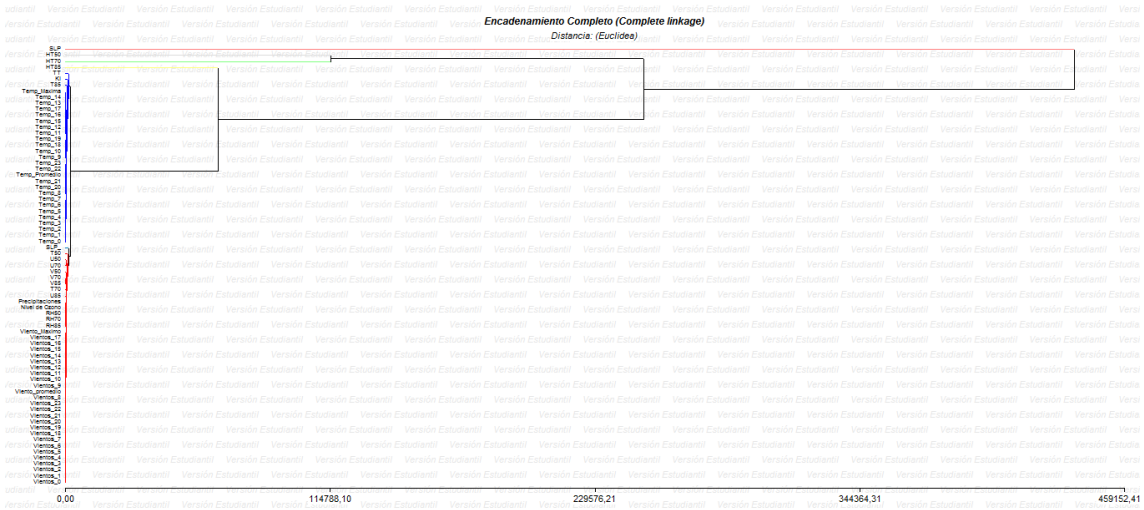
Ward (Distancia Manhathan)



Correlación cofenética= 0,720

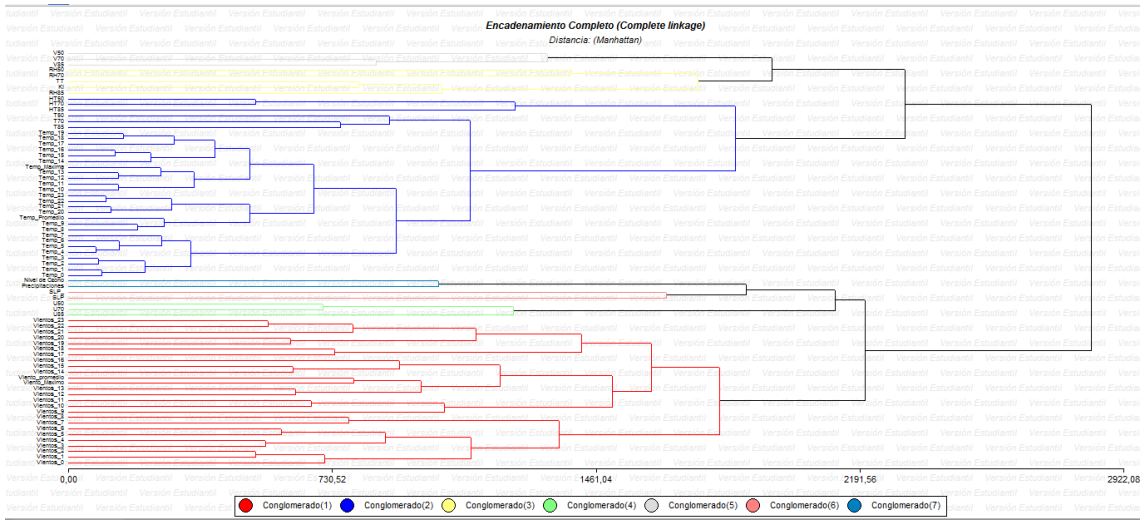
Los resultados son similares a los obtenidos con el método de Ward con la distancia Euclidea.

Complete Linkage (Distancia Euclidea)



Correlación cofenética= 0,977

Complete Linkage (Distancia Manhathan)



Correlación cofenética= 0,929

Nuevamente se observa que el nivel de ozono está fuertemente relacionado con las precipitaciones; los índices de tormenta (KI y TT) están vinculados con las humedades relativas; y las temperaturas y velocidades horarias están vinculadas entre sí.

Análisis

Complete Linkage y Average linkage mostraron coeficientes de correlación cofenética superiores al 90% independientemente del tipo de distancias adoptada. Mientras que para Ward, los resultados fueron del orden del 70%.

En todos los casos las relaciones entre variables fueron similares, mostrando una relación entre el nivel de precipitaciones y el nivel del ozono.

Modelos Predictivos: Clasificación

Se van a aplicar tres técnicas de clasificación, en primer lugar una muy simple como KNN (vecinos más cercanos), y luego arboles de decisión y regresión logística. Se comparan los resultados de las mismas para ver la precisión de cada una y determinar cuál es el mejor método de clasificación.

KNN: Vecinos más cercanos

Como se menciono previamente, el dataset presenta clase desbalanceada, entonces vamos a aplicar el algoritmo a diferentes conjuntos de datos generados a partir del dataset original luego de aplicarle distintas técnicas para el tratamiento de clase desbalanceada. Adicionalmente para cada caso compararemos los resultados obtenidos trabajando con todas las variables contra los resultados que surge de trabajar solo con las componentes principales.

Dataset Original

En la siguiente tabla se observan los resultados obtenidos luego de aplicar la técnica de KNN para 20 diferentes K entre 1 y 100 utilizando cross validation como método de validación.

K	Dataset Original					
	Variables Originales			Componentes Principales		
	Exactitud	% Clasificación 1	% Clasificación 0	Exactitud	% Clasificación 1	% Clasificación 0
1	90.7%	18.2%	95.8%	89.2%	15.2%	94.3%
6	93.7%	15.2%	99.2%	92.9%	9.1%	98.7%
11	93.7%	3.0%	100.0%	93.5%	6.1%	99.6%
16	93.7%	3.0%	100.0%	93.5%	3.0%	99.8%
21	93.5%	0.0%	100.0%	93.7%	3.0%	100.0%
26	93.5%	0.0%	100.0%	93.5%	0.0%	100.0%
31	93.5%	0.0%	100.0%	93.5%	0.0%	100.0%
36	93.5%	0.0%	100.0%	93.5%	0.0%	100.0%
41	93.5%	0.0%	100.0%	93.5%	0.0%	100.0%
46	93.5%	0.0%	100.0%	93.5%	0.0%	100.0%
51	93.5%	0.0%	100.0%	93.5%	0.0%	100.0%
55	93.5%	0.0%	100.0%	93.5%	0.0%	100.0%
60	93.5%	0.0%	100.0%	93.5%	0.0%	100.0%
65	93.5%	0.0%	100.0%	93.5%	0.0%	100.0%
70	93.5%	0.0%	100.0%	93.5%	0.0%	100.0%
75	93.5%	0.0%	100.0%	93.5%	0.0%	100.0%
80	93.5%	0.0%	100.0%	93.5%	0.0%	100.0%
85	93.5%	0.0%	100.0%	93.5%	0.0%	100.0%
90	93.5%	0.0%	100.0%	93.5%	0.0%	100.0%
95	93.5%	0.0%	100.0%	93.5%	0.0%	100.0%
100	93.5%	0.0%	100.0%	93.5%	0.0%	100.0%

En tabla se observa que independientemente del K que utilizamos la exactitud es cercana al 90%; y que los resultados no varían al aumentar el k, es decir, los resultados permanecen invariables con k=21 en el caso en que utilizamos las variables originales para armar el modelo

y con k=26 para el caso que se usaron las componentes principales. También se aprecia que con k>21 en el primer caso y k>26 en el segundo, la performance de los modelos es la misma.

Aunque la perfomance lograda es cercana al 90% en todos los casos, el modelo no es bueno para determinar los días en que el nivel de ozono será elevado, ya que los resultados de clasificación de la clase 1 – Nivel del Ozono Elevado – alcanzo como máximo una tasa del 18%. Entonces, no se cumple el objetivo de lo que queríamos discriminar.

Dataset con tratamiento de clase desbalanceada

Se utilizaron los paquetes de R "unbalanced" y "ROSE" para generar 4 dataset nuevos cada una con la aplicación de una de las técnicas para tratamiento de datos desbalanceados: undersampling, oversampling, undersampling & oversampling y SMOTE.

En el anexo III se encuentran las tablas con los resultados para los 4 dataset luego de aplicar la técnica de KNN para 20 diferentes K entre 1 y 100 utilizando cross validation como método de validación. Mientras, que en la siguiente tabla se observa el resumen de las performance obtenidas luego de las pruebas.

	Undersampling		Oversampling		Undersampling & Oversampling		SMOTE	
	Variables Originales		Variables Originales		Variables Originales		Variables Originales	
K	Exactitud	Clasificación	Exactitud	Clasificación	Exactitud	% Clasificación 1	Exactitud	Clasificación
1	66.7%	73.1%	94.6%	97.6%	96.2%	98.6%	63.4%	60.4%
6	74.5%	80.8%	87.6%	66.7%	88.9%	72.2%	66.5%	64.6%
11	72.5%	84.6%	83.2%	36.9%	83.8%	47.2%	71.4%	71.9%
16	70.6%	80.8%	83.2%	34.5%	82.7%	36.1%	70.5%	68.8%
21	70.6%	80.8%	82.7%	23.8%	83.0%	30.6%	67.0%	61.5%
26	66.7%	76.9%	83.4%	22.6%	84.1%	31.9%	67.9%	61.5%
31	66.7%	80.8%	82.0%	15.5%	83.5%	26.4%	64.7%	56.3%
36	66.7%	80.8%	82.9%	16.7%	82.7%	22.2%	63.8%	54.2%
41	68.6%	84.6%	81.5%	13.1%	83.8%	22.2%	62.9%	54.2%
46	66.7%	84.6%	82.7%	16.7%	84.1%	25.0%	62.1%	52.1%
51	64.7%	88.5%	82.0%	10.7%	84.3%	22.2%	62.1%	54.2%

Se observa que al aplicar técnicas de balanceo de datos, la performance de los modelos para clasificar los días con alto nivel de ozono mejora considerablemente con respecto a la aplicación de los modelos sobre el dataset original. De la tabla podemos extraer las siguientes conclusiones:

- Undersampling presenta una performance superior al 70% en la clasificación 1 para todos los k; con una leve tendencia creciente al aumentar el valor de k a partir de k= 26.
- Oversampling presenta una performance de clasificación elevada para k=1 pero cae abruptamente al subir el valor de k.
- Undersampling & Oversampling presenta un comportamiento similar a Oversampling
- SMOTE presenta % menores que las otras tres técnicas, pero con resultados que no varían significativamente al aumentar el valor de K.

- La mejor performance se obtiene con Undersampling & Oversampling pero con K=1, lo cual no es representativo. Entonces, consideramos como nuestro mejor resultado Undersampling con k=11

A continuación analizaremos los mismos resultados pero aplicando las técnicas de balanceo de datos, sobre las componentes principales.

K	Undersampling		Oversampling		Undersampling & Oversampling		SMOTE	
	Componentes Principales		Componentes Principales		Componentes Principales		Componentes Principales	
	Exactitud	% Clasificación 1	Exactitud	% Clasificación 1	Exactitud	% Clasificación 1	Exactitud	% Clasificación 1
1	74.50%	80.80%	94.6%	97.6%	95.9%	100.0%	84.8%	85.4%
6	78.40%	88.50%	88.8%	79.8%	89.5%	81.9%	82.1%	87.5%
11	82.40%	100.00%	86.9%	63.1%	86.2%	55.6%	79.9%	87.5%
16	80.40%	96.20%	85.5%	50.0%	85.4%	50.0%	81.7%	88.5%
21	78.40%	100.00%	85.7%	41.7%	84.3%	47.2%	79.0%	86.5%
26	76.50%	100.00%	87.6%	50.0%	85.1%	44.4%	81.3%	88.5%
31	74.50%	100.00%	86.0%	39.3%	84.1%	44.4%	79.0%	87.5%
36	74.50%	100.00%	84.8%	32.1%	86.5%	52.8%	79.5%	88.5%
41	74.50%	100.00%	85.7%	33.3%	85.1%	43.1%	77.7%	88.5%
46	74.50%	100.00%	85.3%	32.1%	84.6%	41.7%	78.1%	88.5%
51	72.50%	96.20%	85.7%	32.1%	83.0%	31.9%	79.0%	88.5%
55	72.50%	96.20%	85.0%	26.2%	83.0%	31.9%	77.7%	88.5%

De la tabla podemos extraer las siguientes conclusiones:

- Todas las técnicas de balanceo de datos muestran una performance superior al aplicarla sobre las componentes principales que sobre las variables originales.
- Undersampling muestra los mejores resultados con un % de clasificación cercana al 100% desde k=21 a k=46.
- Oversampling presenta una performance de clasificación elevada para k=1 pero cae abruptamente al subir el valor de k.
- Undersampling & Oversampling presenta un comportamiento similar a Oversampling
- SMOTE muestra una performance superior al 85% independientemente del k.

En resumen para crear nuestro modelo de clasificación de días con elevado nivel de ozono en el ambiente, la técnica KNN no nos brinda buenos resultados con el dataset original. Sin embargo, al trabajar con dataset que fueron sometidos a técnicas de balanceo de datos la performance mejora considerablemente.

Arboles de Decisión: Random Forest

Dataset Original

Se aplico random forest al dataset original variando la cantidad de arboles entre 50 y 1000 y la máxima altura del árbol entre 5 y 10. Al igual que en KNN, se utilizó cross validation para test los modelos.

En la siguiente tabla se pueden apreciar los resultados para el dataset original con las variables originales; en la misma se puede apreciar que la performance de clasificación no varía al incrementarse la máxima altura permitida ni la cantidad de arboles. También, se observa que

si bien el modelo es muy bueno clasificando los días con bajo nivel de ozono, no logra clasificar correctamente los días con elevado nivel de ozono en la atmosfera.

Cantidad Arboles	Altura											
	Clasificación_0						Clasificación_1					
	5	6	7	8	9	10	5	6	7	8	9	10
50	93%	93%	93%	93%	93%	93%	0%	0%	0%	0%	0%	0%
98	93%	93%	93%	93%	93%	93%	0%	0%	0%	0%	0%	0%
145	93%	93%	93%	93%	93%	93%	0%	0%	0%	0%	0%	0%
193	93%	93%	93%	93%	93%	93%	0%	0%	0%	0%	0%	0%
240	93%	93%	93%	93%	93%	93%	0%	0%	0%	0%	0%	0%
288	93%	93%	93%	93%	93%	93%	0%	0%	0%	0%	0%	0%
335	93%	93%	93%	93%	93%	93%	0%	0%	0%	0%	0%	0%
383	93%	93%	93%	93%	93%	93%	0%	0%	0%	0%	0%	0%
430	93%	93%	93%	93%	93%	93%	0%	0%	0%	0%	0%	0%
478	93%	93%	93%	93%	93%	93%	0%	0%	0%	0%	0%	0%
525	93%	93%	93%	93%	93%	93%	0%	0%	0%	0%	0%	0%
573	93%	93%	93%	93%	93%	93%	0%	0%	0%	0%	0%	0%
620	93%	93%	93%	93%	93%	93%	0%	0%	0%	0%	0%	0%
668	93%	93%	93%	93%	93%	93%	0%	0%	0%	0%	0%	0%
715	93%	93%	93%	93%	93%	93%	0%	0%	0%	0%	0%	0%
763	93%	93%	93%	93%	93%	93%	0%	0%	0%	0%	0%	0%
810	93%	93%	93%	93%	93%	93%	0%	0%	0%	0%	0%	0%
858	93%	93%	93%	93%	93%	93%	0%	0%	0%	0%	0%	0%
905	93%	93%	93%	93%	93%	93%	0%	0%	0%	0%	0%	0%
953	93%	93%	93%	93%	93%	93%	0%	0%	0%	0%	0%	0%
1000	93%	93%	93%	93%	93%	93%	0%	0%	0%	0%	0%	0%

A continuación se muestra el mismo análisis pero con la utilización de componentes principales; en donde se aprecia que los % de clasificación no varían al incrementarse la altura y la cantidad arboles. Además, se observa que si bien las performance son mejores que en el caso previo, la clasificación de elevado nivel de ozono es apenas un 6%.

Cantidad Arboles	Altura											
	Clasificación_0						Clasificación_1					
	5	6	7	8	9	10	5	6	7	8	9	10
50	94%	94%	94%	94%	94%	94%	6%	6%	6%	6%	6%	6%
98	94%	94%	94%	94%	94%	94%	6%	6%	6%	6%	6%	6%
145	94%	94%	94%	94%	94%	94%	6%	6%	6%	6%	6%	6%
193	94%	94%	94%	94%	94%	94%	6%	6%	6%	6%	6%	6%
240	94%	94%	94%	94%	94%	94%	6%	6%	6%	6%	6%	6%
288	94%	94%	94%	94%	94%	94%	6%	6%	6%	6%	6%	6%
335	94%	94%	94%	94%	94%	94%	6%	6%	6%	6%	6%	6%
383	94%	94%	94%	94%	94%	94%	6%	6%	6%	6%	6%	6%
430	94%	94%	94%	94%	94%	94%	6%	6%	6%	6%	6%	6%
478	94%	94%	94%	94%	94%	94%	6%	6%	6%	6%	6%	6%
525	94%	94%	94%	94%	94%	94%	6%	6%	6%	6%	6%	6%
573	94%	94%	94%	94%	94%	94%	6%	6%	6%	6%	6%	6%
620	94%	94%	94%	94%	94%	94%	6%	6%	6%	6%	6%	6%
668	94%	94%	94%	94%	94%	94%	6%	6%	6%	6%	6%	6%
715	94%	94%	94%	94%	94%	94%	6%	6%	6%	6%	6%	6%
763	94%	94%	94%	94%	94%	94%	6%	6%	6%	6%	6%	6%
810	94%	94%	94%	94%	94%	94%	6%	6%	6%	6%	6%	6%
858	94%	94%	94%	94%	94%	94%	6%	6%	6%	6%	6%	6%
905	94%	94%	94%	94%	94%	94%	6%	6%	6%	6%	6%	6%
953	94%	94%	94%	94%	94%	94%	6%	6%	6%	6%	6%	6%
1000	94%	94%	94%	94%	94%	94%	6%	6%	6%	6%	6%	6%

Dataset con tratamiento de datos desbalanceados

Se aplico random forest a los dataset con tratamiento de clase desbalanceada (métodos: oversampling, undersampling, oversampling & undersampling y SMOTE) variando la cantidad

de arboles entre 50 y 1000 y la máxima altura del árbol entre 5 y 10. Al igual que en KNN, se utilizó cross validation para test los modelos.

En el anexo IV se encuentran los resultados para cada modelo generado variando los parámetros aplicado sobre las variables originales y sobre las componentes principales para cada método de tratamiento de clase desabalanceada.

A modo de resumen para poder comparar los métodos la siguiente tabla muestra el mejor resultado obtenido para cada caso; de la comparación de resultados se puede apreciar que el dataset al que se aplico el método de undersampling muestra resultados muy superiores que los modelos generados a partir del dataset original y a partir de los otros métodos para tratamiento de clase desabalanceada. Adicionalmente, se desprende de la tabla que los resultados al utilizar las componentes principales son mejores que al utilizar las variables originales.

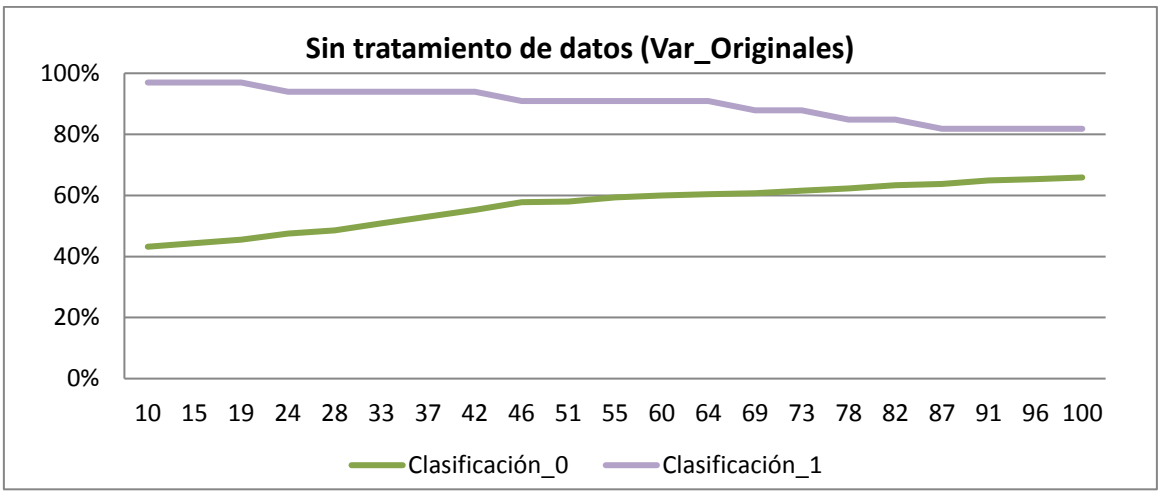
Dataset	Variables	Altura	Cantidad Arboles	Clasificación_0	Clasificaci	Comentario
Original	Originales	5	50	93%	0%	Resultado obtenido con mas de un conjunto de parámetros
Original	ACP	5	50	94%	6%	Resultado obtenido con mas de un conjunto de parámetros
Undesampling	Originales	5	50	73%	88%	Resultado obtenido con mas de un conjunto de parámetros
Undesampling	ACP	7	50	76%	92%	
Oversampling	Originales	6	50	82%	8%	Resultado obtenido con mas de un conjunto de parámetros
Oversampling	ACP	5	50	82%	6%	Resultado obtenido con mas de un conjunto de parámetros
Undersampling&Oversampling	Originales	5	193	83%	11%	Resultado obtenido con mas de un conjunto de parámetros
Undersampling&Oversampling	ACP	8	98	82%	10%	
SMOTE	Originales	5	50	57%	0%	Resultado obtenido con mas de un conjunto de parámetros
SMOTE	ACP	8	98	79%	70%	

Regresión Logística

Dataset original

Se aplicó regresión logística al dataset con las variables originales y con componentes principales variando el número de iteraciones del algoritmo iterativo de máxima verosimilitud. Previo a la aplicación de esta técnica se procedió a la imputación de datos faltantes.

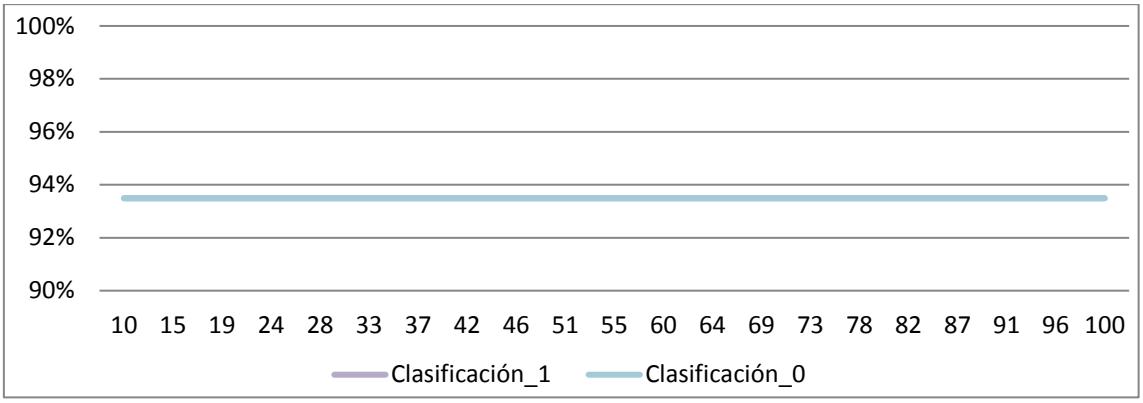
En el siguiente gráfico se observan la performance de clasificación de modelo con el dataset original y las variables originales al variar la cantidad de iteraciones.



En el dibujo se observa que la curva de clasificación de la clase_0 crece al crecer el número de iteraciones hasta alcanzar su máximo de 65% - 85%; en donde la curva se estabilizar

permaneciendo su performance constante. Mientras, que él % de clasificación de 1 disminuye al incrementarse la cantidad de iteraciones, permaneciendo constante para valores mayores a 85 en un 80%.

A continuación hicimos el mismo experimento utilizando componentes principales. Como observa en este caso, la clasificación de 0 es estable en torno al 93%; mientras que la clasificación de la clase 1 es nula.



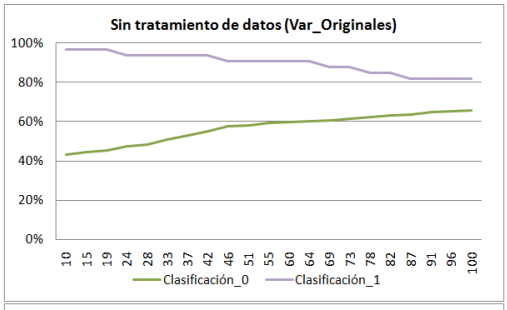
Dataset con tratamiento de datos desbalanceados

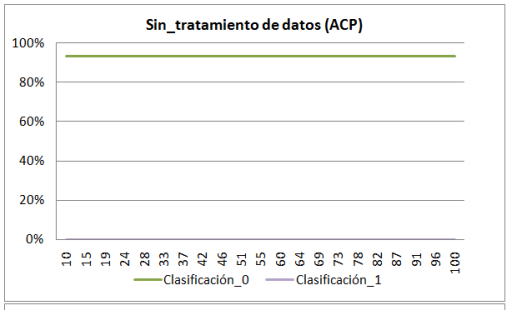
Se aplico regresión logística a los dataset con tratamiento de clase desbalanceada (métodos: oversampling, undersampling, oversampling & undersampling y SMOTE) variando el número de iteraciones del algoritmo iterativo de máxima verosimilitud. Al igual que en KNN y Random Forest, se utilizó cross validation para test los modelos.

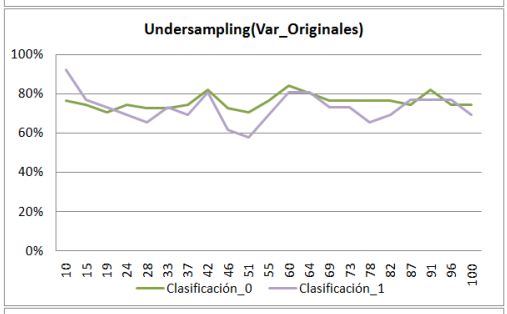
A continuación se observan las curvas con los resultados para cada modelo, del análisis de las mismas extraemos las siguientes conclusiones

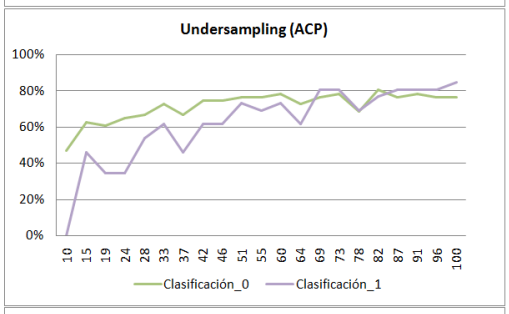
- En todos los casos los resultados de los modelos realizados con las variables originales son mejores que los obtenidos al utilizar las componentes principales
- **Undersampling:** el modelo creado a partir de este dataset presenta resultados con picos y valles al incrementar el número de iteraciones oscilando entre un 60 y 90% de performance en la clasificación de 1 cuando se lo trabaja con las variables originales. Mientras que muestra una tendencia creciente en la performance al aumentar el número de iteraciones cuando el modelo se realiza a partir de las componentes principales.
- **Oversampling:** el modelo creado a partir de las variables originales presenta una tendencia decreciente de 95 a 80% al incrementarse el número de iteraciones del algoritmo de máxima verosimilitud. Sin embargo, cuando se trabaja con las componentes principales los resultados son cercanos a cero independientemente del número de iteraciones.
- **Oversampling&Undersampling:** el comportamiento es similar al que obtuvimos con oversampling
- **SMOTE:** el modelo creado a partir de las variables originales no logra clasificar a la clase 1. Mientras que el modelo creado a partir de las componentes principales

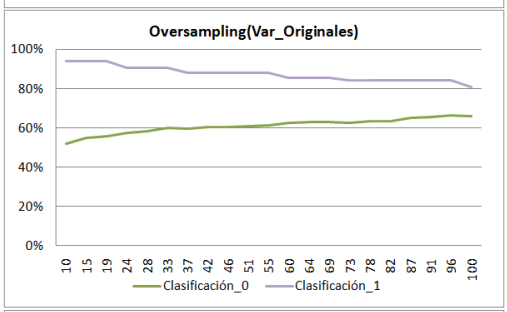
presenta una tendencia oscilante entre 0 y 20% en la clasificación del elevado nivel de ozono al variar el número de iteraciones.

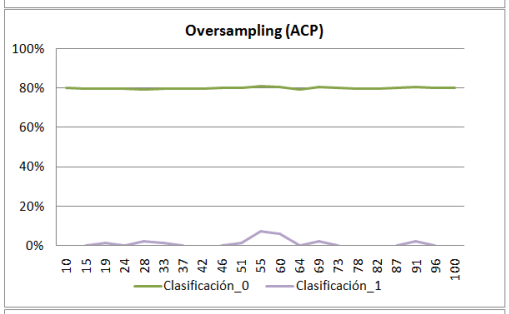


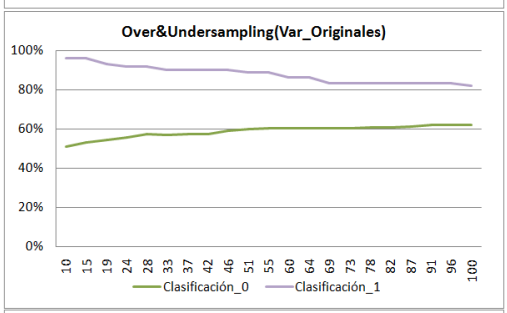


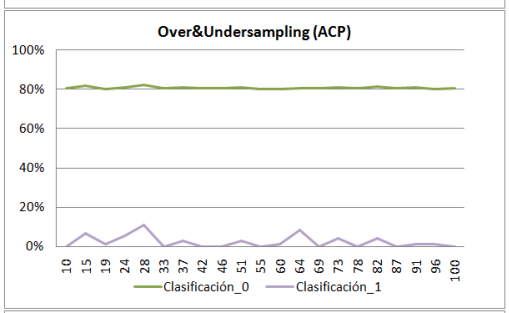


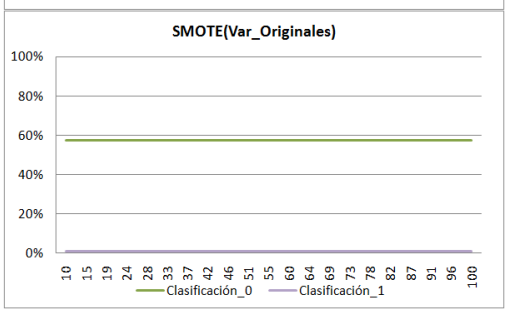


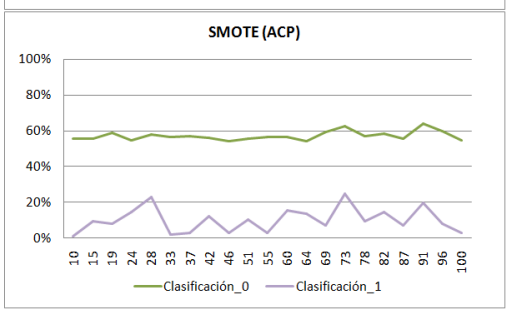












Comparación de modelos de clasificación

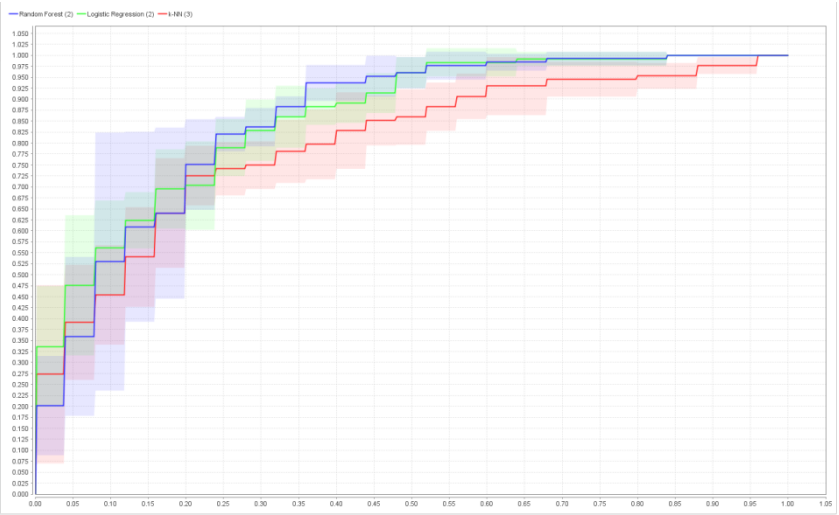
Se aplicaron los 3 métodos para crear modelo predictivos a fin de poder predecir si dadas la condiciones meteorológicas de un determinado día se espera un nivel de ozono elevado o normal. Como no se han obtenido buenos resultados con el dataset sin tratamiento de datos

desbalanceados, no analizaremos esos casos. Pero si compararemos las curvas ROC de los diferentes métodos aplicados con y sin componentes principales para las distintas técnicas de tratamiento de datos desbalanceados; cabe aclarar que cada una de ellas fue realizada sobre su respectivo dataset de prueba.

Undersampling

El siguiente gráfico muestra los resultados obtenidos al calcular la curva ROC del modelo que mejor clasifica la clase minoritaria del dataset tratado con undersampling con las variables originales.

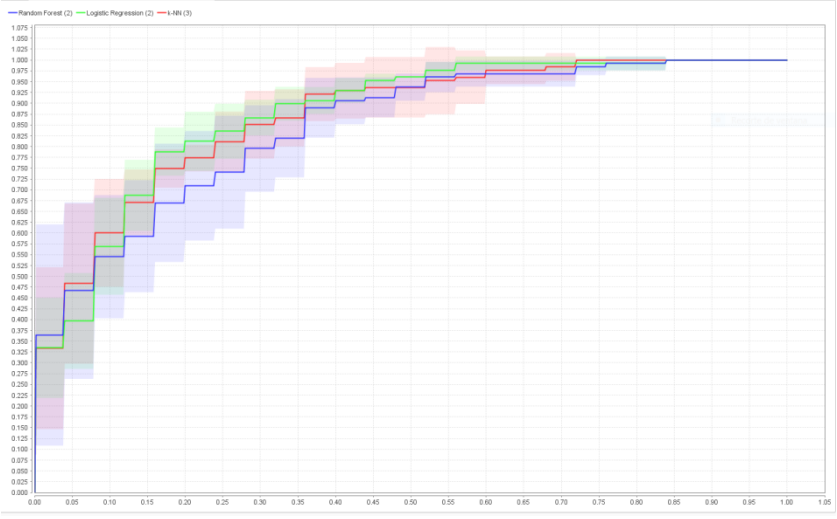
Como se observa en el gráfico, random forest y regresión logística presentan performance



similares (AUC RF = 0.77 vs AUC RL = 0.82); mientras que KNN está muy debajo en los resultados.

También se aprecia que random forest es la técnica que presenta mayor variabilidad de resultados (área azul) al aplicar la técnica de cross validation.

Si analizamos el mismo gráfico pero trabajando solo con componentes principales, se

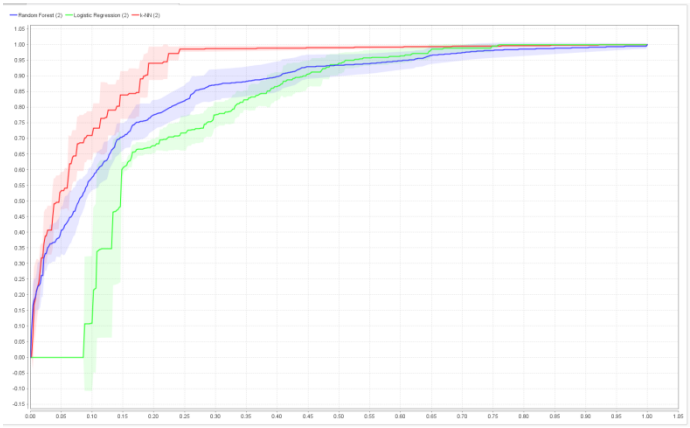


desprende que las tres técnicas tienen performance similares (AUC KNN=0.9, AUC RF=0.86 y AUC RL=0.88). En este caso, regresión logística y KNN, según el tramo de la curva que estemos analizando, una va teniendo mejor performance que la otra. Al igual que en las ROC previas, random forest es la técnica que muestra mayor variabilidad.

En este caso se observa que KNN y Regresión Logística muestran los mejores resultados al trabajar con componentes principales.

Oversampling

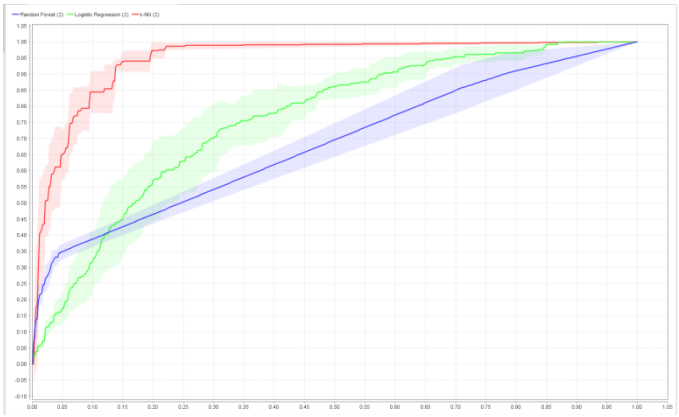
El siguiente gráfico muestra los resultados obtenidos al calcular la curva ROC del modelo que mejor clasifica la clase minoritaria del dataset tratado con oversampling con las variables originales.



En este caso, KNN presenta los mejores resultados, desprendiéndose de las otras técnicas (AUC KNN=0.92, AUC RF=0.79, AUC RL=0.68).

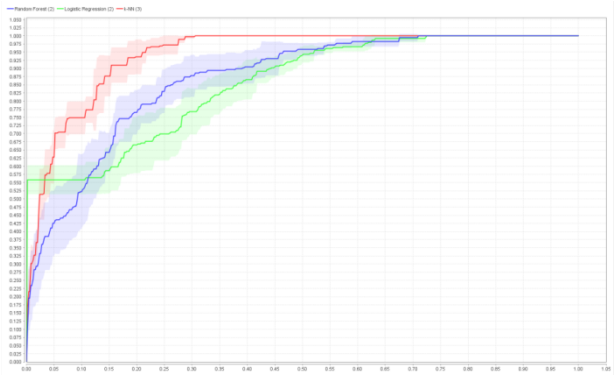
A diferencia del caso anterior, las curvas presentan mucha menor variabilidad.

A continuación se observa el mismo gráficos pero utilizando componentes principales. En el mismo se aprecia que Randon Forest no presenta una buena performance, lo cual confirma los resultados obtenidos previamente (AUC =0.6). En lo que respecta a regresión logística, también muestra resultados pobres con un AUC =0.67. Mientras, que KNN llega a un AUC de 0.93, tal como se muestra en el gráfico.



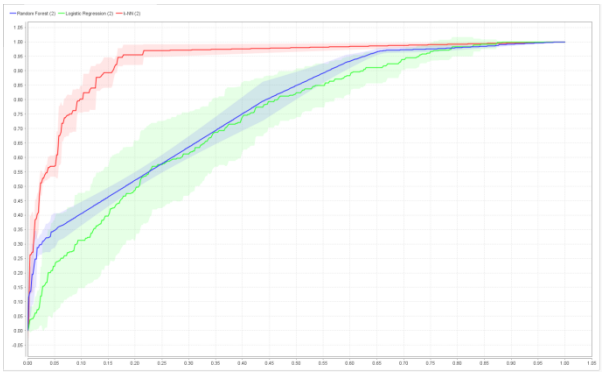
Por lo tanto, concluimos que para oversampling, la única técnica que muestra buenos resultados, utilizando o no componentes las componentes principales es KNN.

Undersampling & Oversampling



El siguiente gráfico muestra los resultados obtenidos al calcular la curva ROC del modelo que mejor clasifica la clase minoritaria del dataset tratado con undersampling&oversampling con las variables originales.

Al igual que en el caso anterior KNN se desprende del resto con un AUC de 0.93. Mientras que random forest, un poco por debajo con 0.86 y regresión logística alcanza apenas un AUC de 0.74, mostrando un modelo que no clasificación los casos con clase =1.



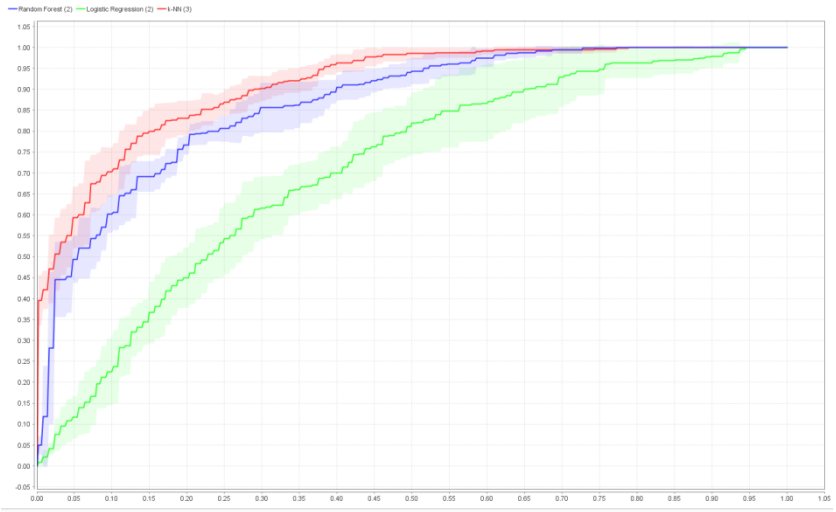
Analizando el mismo grafico pero con componentes principales, se observa en KNN una elevada performance de 0.95. Mientras que random forest y regresión logística se comportan de manera similar con 0.77 y 0.75 respectivamente. También se aprecia la elevada variabilidad de resultados en regresión logística al realizar cross validation.

Al igual que con oversampling los mejores resultados se obtienen con KNN.

SMOTE

No se analizara SMOTE con las variables originales, ya que ha mostrado resultados pobres en los diferentes modelos.

El siguiente gráfico muestra los resultados obtenidos al calcular la curva ROC del modelo que mejor clasifica la clase minoritaria del dataset tratado con SMOTE con componentes principales.



Como era de esperarse, regresión logística no presenta buenos resultados; mientras que KNN y Random Forest presentan performance similares, con una superioridad del primer modelo sobre el segundo.

Conclusiones comparación de técnicas.

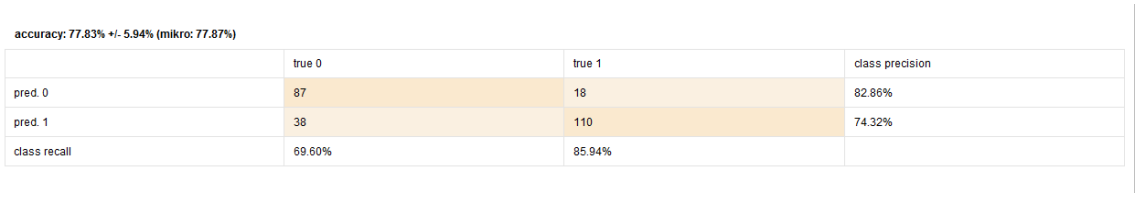
De los resultados se desprenden las siguientes conclusiones:

- KNN presenta performances superiores a Random Forest y Regresión Logística, para todos los métodos de tratamiento de datos menos para undersampling.
- Regresión logística presenta buenos resultados cuando se utiliza underampling.
- Random Forest y regresión logística muestran AUC similares en oversampling y en oversampling&undersampling.
- Undersampling con componentes principales muestra, en promedio, los mejores resultados con las tres técnicas.

Ensamble

De la ultima conclusión se desprende que sería convenientes probar un ensamble de las tres técnicas con el tratamiento de datos desbalanceados de undersampling y utilizando las componentes principales.

La siguiente tabla muestra los resultados obtenidos, donde se observa que el ensamble de los tres modelos por votación simple muestra buenos resultados en la clasificación de 1, pero no ocurre lo mismo en la clasificación de cero (AUC =0.77)



accuracy: 77.83% +/- 5.94% (mikro: 77.87%)

	true 0	true 1	class precision
pred. 0	87	18	82.86%
pred. 1	38	110	74.32%
class recall	69.60%	85.94%	

De aquí se concluye que este ensamble no mejora la performance de las técnicas originales aplicadas individualmente.

Conclusión

A lo largo del trabajo se observó que existe una alta correlación entre el nivel de ozono y las condiciones climáticas. En el ACP, se mostro que existen relaciones entre las variables, las cuales fueron ratificadas por los métodos de clustering (Kmeans y Jerárquico).

Se concluye que temperaturas elevadas durante las horas de luz solar, con velocidades del viento bajas y bajo nivel de precipitaciones son las condiciones que generan registros de elevados niveles de ozono; ya que la luz y el calor son los mejores motores que ayudan a la formación de ozono; pero cuando hay vientos, estos evitan la concentración de gases y los dispersan.

Conocer las condiciones bajo las cuales es más común encontrar elevados niveles de ozono, serviría para fijar políticas con respecto al uso de los automóviles en determinadas estaciones del año, las reglas ambientales para las empresas, etc. Las mismas son necesarias debido a que este gas es sumamente reactivo, siendo perjudicial para el medio ambiente y para la salud.

En la segunda parte del trabajo se intentó realizar un modelo predictivo para clasificar los días con elevado nivel de ozono, para ellos se aplicaron 3 técnicas para realizar el modelo KNN, Regresión Logística y Arboles de Decisión (Random Forest). El principal inconveniente que se presento durante el proceso fue trabajar con datos con clases desbalanceadas; entonces se debió trabajar con técnicas que permitían balancear la clase minoritaria. Finalmente del análisis de todos los modelos realizados, se concluyo que el modelo generado mediante KNN con undersampling y utilizando las componentes principales presenta la mejor performance de clasificación de días con elevado nivel de ozono, considerando los diferentes aspectos por los que fue analizado.

Próximos pasos

Como próximos pasos, el objetivo sería poder trabajar con los métodos de imputación de datos faltantes a fin de lograr obtener los valores que se asemejen lo mayor posible a la realidad. Adicionalmente, seria provechoso poder trabajar más exhaustivamente el ensamble de las técnicas predictivas a fin de mejorar la performance del modelo.

Otro punto que nos permitiría mejorar los modelos predictivos es intentar fijar costos para medir el impacto de clasificar erróneamente a un día que tendrá elevado nivel de ozono versus no clasificar correctamente una jornada con dichas condiciones ambientales.

Bibliografía

- http://www.airinfonow.org/: EPA: United States Environmental Protection Agency
- https://archive.ics.uci.edu: UCI: Machine Learning Repository
- https://www.r-project.org/
- http://www.inside-r.org
- Introduction to Datamining- Pang-Ning Tan, Michigan State University, Michael Steinbach, University of Minnesota.

Anexo I: Tabla Descripción de Variables

Variable	Descripción	Tipo
Date	Fecha	Date
WSR0	Velocidad del viento (Km/h) a las 0 Hs	continuos
WSR1	Velocidad del viento (Km/h) a las 1 Hs	continuos
WSR2	Velocidad del viento (Km/h) a las 2 Hs	continuos
WSR3	Velocidad del viento (Km/h) a las 3 Hs	continuos
WSR4	Velocidad del viento (Km/h) a las 4 Hs	continuos
WSR5	Velocidad del viento (Km/h) a las 5 Hs	continuos
WSR6	Velocidad del viento (Km/h) a las 6 Hs	continuos
WSR7	Velocidad del viento (Km/h) a las 7 Hs	continuos
WSR8	Velocidad del viento (Km/h) a las 8 Hs	continuos
WSR9	Velocidad del viento (Km/h) a las 9 Hs	continuos
WSR10	Velocidad del viento (Km/h) a las 10 Hs	continuos
WSR11	Velocidad del viento (Km/h) a las 11 Hs	continuos
WSR12	Velocidad del viento (Km/h) a las 12 Hs	continuos
WSR13	Velocidad del viento (Km/h) a las 13 Hs	continuos
WSR14	Velocidad del viento (Km/h) a las 14 Hs	continuos
WSR15	Velocidad del viento (Km/h) a las 15 Hs	continuos
WSR16	Velocidad del viento (Km/h) a las 16 Hs	continuos
WSR17	Velocidad del viento (Km/h) a las 17 Hs	continuos
WSR18	Velocidad del viento (Km/h) a las 18 Hs	continuos
WSR19	Velocidad del viento (Km/h) a las 19 Hs	continuos
WSR20	Velocidad del viento (Km/h) a las 20 Hs	continuos
WSR21	Velocidad del viento (Km/h) a las 21 Hs	continuos
WSR22	Velocidad del viento (Km/h) a las 22 Hs	continuos
WSR23	Velocidad del viento (Km/h) a las 23 Hs	continuos
WSR_PK	Velocidad del viento (Km/h) máxima	continuos
WSR_AV	Velocidad del viento (Km/h) promedio	continuos

Variable	Descripción	Tipo
T0	Temperatura (ºC) a las 0 Hs	continuos
T1	Temperatura (ºC) a las 1 Hs	continuos
T2	Temperatura (ºC) a las 2 Hs	continuos
T3	Temperatura (ºC) a las 3 Hs	continuos
T4	Temperatura (ºC) a las 4 Hs	continuos
T5	Temperatura (ºC) a las 5 Hs	continuos
T6	Temperatura (ºC) a las 6 Hs	continuos
T7	Temperatura (ºC) a las 7 Hs	continuos
T8	Temperatura (ºC) a las 8 Hs	continuos
T9	Temperatura (ºC) a las 9 Hs	continuos
T10	Temperatura (ºC) a las 10 Hs	continuos
T11	Temperatura (ºC) a las 11 Hs	continuos
T12	Temperatura (ºC) a las 12 Hs	continuos
T13	Temperatura (ºC) a las 13 Hs	continuos
T14	Temperatura (ºC) a las 14 Hs	continuos
T15	Temperatura (ºC) a las 15 Hs	continuos
T16	Temperatura (ºC) a las 16 Hs	continuos
T17	Temperatura (ºC) a las 17 Hs	continuos
T18	Temperatura (ºC) a las 18 Hs	continuos
T19	Temperatura (ºC) a las 19 Hs	continuos
T20	Temperatura (ºC) a las 20 Hs	continuos
T21	Temperatura (ºC) a las 21 Hs	continuos
T22	Temperatura (ºC) a las 22 Hs	continuos
T23	Temperatura (ºC) a las 23 Hs	continuos
T_PK	Temperatura (ºC) máxima	continuos
T_AV	Temperatura (ºC) promedio	continuos
T85	Temperatura (ºC) a un nivel de 850 hpa	continuos
RH85	Humedad Relativa a un nivel de 850 hpa	continuos
U85	Velocidad del viento (km/h) en dirección E-O a un nivel de 850 hpa	continuos
V85	Velocidad del viento (km/h) en dirección N-S a un nivel de 850 hpa	continuos
HT85	Altura geopotencial a un nivel de 850 hpa	continuos
T70	Temperatura (ºC) a un nivel de 700 hpa	continuos
RH70	Humedad Relativa a un nivel de 700 hpa	continuos
U70	Velocidad del viento (km/h) en dirección E-O a un nivel de 700 hpa	continuos
V70	Velocidad del viento (km/h) en dirección N-S a un nivel de 700 hpa	continuos
HT70	Altura geopotencial a un nivel de 700 hpa	continuos
T50	Temperatura (ºC) a un nivel de 500 hpa	continuos

Variable	Descripción	Tipo
T50	Temperatura (ºC) a un nivel de 500 hpa	continuos
RH50	Humedad Relativa a un nivel de 500 hpa	continuos
U50	Velocidad del viento (km/h) en dirección E-O a un nivel de 500 hpa	continuos
V50	Velocidad del viento (km/h) en dirección N-S a un nivel de 500 hpa	continuos
HT50	Altura geopotencial a un nivel de 500 hpa	continuos
KI	Una medida del potencial tormenta basado en gradiente vertical de temperatura, contenido de humedad de la atmósfera inferior y la extensión vertical de la capa húmeda	continuos
TT	Una medida de la fuerza de la tormenta	continuos
SLP	Presión al nivel del mar	continuos
SLP_1	SLP variación con respecto al día anterior	continuos
Precp	Nivel de Precipitaciones (mm)	continuos

Anexo II: ACP Auto vectores

Variables	e1	e2	e3	e4	e5
Vientos_0	-0,04	0,18	0,15	0,12	0,0018
Vientos_1	-0,05	0,18	0,16	0,14	0,01
Vientos_2	-0,06	0,18	0,16	0,15	-4,00E-03
Vientos_3	-0,06	0,18	0,16	0,16	-3,40E-03
Vientos_4	-0,07	0,17	0,16	0,18	0,01
Vientos_5	-0,07	0,17	0,15	0,17	0,01
Vientos_6	-0,06	0,18	0,14	0,19	0,01
Vientos_7	-0,03	0,18	0,11	0,22	-0,02
Vientos_8	-0,03	0,18	0,04	0,23	-0,07
Vientos_9	-0,05	0,18	-0,03	0,19	-0,12
Vientos_10	-0,06	0,18	-0,08	0,12	-0,14
Vientos_11	-0,06	0,18	-0,1	0,08	-0,13
Vientos_12	-0,06	0,18	-0,12	0,05	-0,11
Vientos_13	-0,06	0,18	-0,14	0,01	-0,1
Vientos_14	-0,05	0,18	-0,16	-0,01	-0,09
Vientos_15	-0,04	0,18	-0,17	-0,04	-0,07
Vientos_16	-0,03	0,17	-0,19	-0,07	-0,04
Vientos_17	-0,01	0,17	-0,2	-0,12	0,04
Vientos_18	2,40E-03	0,17	-0,19	-0,14	0,13
Vientos_19	-3,30E-03	0,17	-0,18	-0,15	0,18
Vientos_20	-0,01	0,17	-0,17	-0,17	0,2
Vientos_21	-0,01	0,17	-0,14	-0,16	0,21
Vientos_22	-0,02	0,17	-0,13	-0,16	0,21
Vientos_23	-0,02	0,17	-0,11	-0,14	0,19
Viento_Máximo	-0,05	0,21	-0,1	-0,03	-0,06
Viento_promedio	-0,05	0,24	-0,06	0,04	0,01
Temp_0	0,17	0,07	0,07	-1,70E-03	-0,01
Temp_1	0,17	0,08	0,07	-0,01	-0,01
Temp_2	0,17	0,08	0,07	-0,01	-4,10E-03
Temp_3	0,17	0,08	0,07	-0,01	-1,60E-03
Temp_4	0,17	0,08	0,07	-0,02	2,40E-03
Temp_5	0,17	0,08	0,06	-0,02	0,01
Temp_6	0,17	0,08	0,05	-0,02	0,01
Temp_7	0,17	0,07	0,03	-0,02	-2,10E-03
Temp_8	0,18	0,05	-1,10E-03	-0,02	-0,02
Temp_9	0,18	0,03	-0,03	-0,01	-0,04

Variables	e1	e2	e3	e4	e5
Temp_10	0,18	0,02	-0,05	-0,0024	-0,06
Temp_11	0,18	0,02	-0,06	0,003	-0,07
Temp_12	0,17	0,01	-0,07	0,01	-0,08
Temp_13	0,17	0,01	-0,07	0,02	-0,09
Temp_14	0,17	0,003	-0,07	0,02	-0,1
Temp_15	0,17	-1,70E-03	-0,07	0,02	-0,1
Temp_16	0,17	-4,60E-03	-0,06	0,03	-0,11
Temp_17	0,17	-1,60E-03	-0,05	0,03	-0,1
Temp_18	0,17	0,01	-0,04	0,02	-0,08
Temp_19	0,18	0,01	-0,04	0,01	-0,06
Temp_20	0,18	0,02	-0,03	0,01	-0,04
Temp_21	0,18	0,02	-0,03	2,50E-03	-0,02
Temp_22	0,17	0,03	-0,03	2,60E-04	-0,01
Temp_23	0,17	0,03	-0,04	-7,60E-04	3,30E-03
Temp_ Máxima	0,18	0,01	-0,05	0,01	-0,08
Temp_ Promedio	0,18	0,04	-0,01	1,20E-03	-0,04
T85	0,16	0,05	0,06	0,03	4,00E-03
RH85	0,05	0,06	0,24	-0,17	0,08
U85	-0,07	0,03	0,05	-0,16	-0,25
V85	0,03	0,16	0,1	-0,13	0,07
HT85	0,06	-0,06	-0,1	0,25	0,38
T70	0,14	0,05	0,01	0,09	0,092
RH70	0,03	0,03	0,24	-0,14	0,17
U70	-0,12	0,02	0,05	-0,14	-0,21
V70	-0,00092	0,14	0,18	-0,14	0,07
HT70	0,13	-0,02	-0,05	0,21	0,29
T50	0,14	0,01	0,04	0,1	0,06
RH50	0,01	0,04	0,19	-0,14	0,15
U50	-0,13	0,01	0,04	-0,12	-0,17
V50	-0,03	0,09	0,2	-0,15	0,02
HT50	0,15	0,01	-0,02	0,18	0,2
KI	0,09	0,06	0,25	-0,16	0,09
TT	0,08	0,08	0,21	-0,19	0,01
SLP	-0,08	-0,09	-0,13	0,21	0,31
SLP_	-0,02	-0,08	-0,05	0,18	0,07
Precipitaciones	-0.02	0,03	0,12	-0,09	0,17

Anexo III: Resultados KNN – Corrección Datos Desbalanceados

Undersampling

K	Undersampling					
	Variables Originales			Componentes Principales		
	Exactitud	% Clasificación 1	% Clasificación 0	Exactitud	% Clasificación 1	% Clasificación 0
1	66.7%	73.1%	60.0%	74.5%	80.8%	68.0%
6	74.5%	80.8%	68.0%	78.4%	88.5%	68.0%
11	72.5%	84.6%	60.0%	82.4%	100.0%	64.0%
16	70.6%	80.8%	60.0%	80.4%	96.2%	64.0%
21	70.6%	80.8%	60.0%	78.4%	100.0%	56.0%
26	66.7%	76.9%	56.0%	76.5%	100.0%	52.0%
31	66.7%	80.8%	52.0%	74.5%	100.0%	48.0%
36	66.7%	80.8%	52.0%	74.5%	100.0%	48.0%
41	68.6%	84.6%	52.0%	74.5%	100.0%	48.0%
46	66.7%	84.6%	48.0%	74.5%	100.0%	48.0%
51	64.7%	88.5%	40.0%	72.5%	96.2%	48.0%
55	62.7%	88.5%	36.0%	72.5%	96.2%	48.0%
60	64.7%	88.5%	40.0%	74.5%	100.0%	48.0%
65	64.7%	92.3%	36.0%	70.6%	100.0%	40.0%
70	66.7%	88.5%	44.0%	72.5%	100.0%	44.0%
75	64.7%	88.5%	40.0%	72.5%	100.0%	44.0%
80	64.7%	88.5%	40.0%	70.6%	100.0%	40.0%
85	64.7%	88.5%	40.0%	70.6%	100.0%	40.0%
90	58.8%	88.5%	28.0%	70.6%	100.0%	40.0%
95	58.8%	88.5%	28.0%	70.6%	100.0%	40.0%
100	58.8%	88.5%	28.0%	70.6%	100.0%	40.0%

Oversampling

K	Oversampling					
	Variables Originales			Componentes Principales		
	Exactitud	% Clasificación 1	% Clasificación 0	Exactitud	% Clasificación 1	% Clasificación 0
1	94.6%	97.6%	93.9%	94.6%	97.6%	93.9%
6	87.6%	66.7%	92.7%	88.8%	79.8%	91.0%
11	83.2%	36.9%	94.5%	86.9%	63.1%	92.7%
16	83.2%	34.5%	95.1%	85.5%	50.0%	94.2%
21	82.7%	23.8%	97.1%	85.7%	41.7%	96.5%
26	83.4%	22.6%	98.3%	87.6%	50.0%	96.8%
31	82.0%	15.5%	98.3%	86.0%	39.3%	97.4%
36	82.9%	16.7%	99.1%	84.8%	32.1%	97.7%
41	81.5%	13.1%	98.3%	85.7%	33.3%	98.5%
46	82.7%	16.7%	98.8%	85.3%	32.1%	98.3%
51	82.0%	10.7%	99.4%	85.7%	32.1%	98.8%
55	81.5%	10.7%	98.8%	85.0%	26.2%	99.4%
60	81.5%	10.7%	98.8%	85.7%	29.8%	99.4%
65	81.5%	9.5%	99.1%	85.3%	27.4%	99.4%
70	81.5%	9.5%	99.1%	85.5%	28.6%	99.4%
75	82.0%	9.5%	99.7%	84.6%	25.0%	99.1%
80	81.8%	9.5%	99.4%	84.3%	25.0%	98.8%
85	81.8%	8.3%	99.7%	84.3%	23.8%	99.1%
90	81.8%	8.3%	99.7%	84.6%	25.0%	99.1%
95	81.3%	6.0%	99.7%	84.1%	22.6%	99.1%
100	81.1%	3.6%	100.0%	84.8%	25.0%	99.4%

Oversampling & Undersampling

K	Undersampling & Oversampling					
	Variables Originales			Componentes Principales		
	Exactitud	% Clasificación 1	% Clasificación 0	Exactitud	% Clasificación 1	% Clasificación 0
1	96.2%	98.6%	95.6%	95.9%	100.0%	95.0%
6	88.9%	72.2%	93.0%	89.5%	81.9%	91.3%
11	83.8%	47.2%	92.6%	86.2%	55.6%	93.6%
16	82.7%	36.1%	94.0%	85.4%	50.0%	94.0%
21	83.0%	30.6%	95.6%	84.3%	47.2%	93.3%
26	84.1%	31.9%	96.6%	85.1%	44.4%	95.0%
31	83.5%	26.4%	97.3%	84.1%	44.4%	93.6%
36	82.7%	22.2%	97.3%	86.5%	52.8%	94.6%
41	83.8%	22.2%	98.7%	85.1%	43.1%	95.3%
46	84.1%	25.0%	98.3%	84.6%	41.7%	95.0%
51	84.3%	22.2%	99.3%	83.0%	31.9%	95.3%
55	84.3%	22.2%	99.3%	83.0%	31.9%	95.3%
60	84.9%	23.6%	99.7%	83.5%	31.9%	96.0%
65	84.6%	23.6%	99.3%	83.0%	30.6%	95.6%
70	84.9%	23.6%	99.7%	83.0%	30.6%	95.6%
75	83.2%	16.7%	99.3%	82.7%	29.2%	95.6%
80	84.1%	19.4%	99.7%	83.0%	30.6%	95.6%
85	83.5%	16.7%	99.7%	82.7%	27.8%	96.0%
90	84.1%	19.4%	99.7%	82.4%	26.4%	96.0%
95	83.0%	13.9%	99.7%	83.5%	30.6%	96.3%
100	83.5%	16.7%	99.7%	82.4%	26.4%	96.0%

SMOTE

K	SMOTE					
	Variables Originales			Componentes Principales		
	Exactitud	% Clasificación 1	% Clasificación 0	Exactitud	% Clasificación 1	% Clasificación 0
1	63.4%	60.4%	65.6%	84.8%	85.4%	84.4%
6	66.5%	64.6%	68.0%	82.1%	87.5%	78.1%
11	71.4%	71.9%	71.1%	79.9%	87.5%	74.2%
16	70.5%	68.8%	71.9%	81.7%	88.5%	76.6%
21	67.0%	61.5%	71.1%	79.0%	86.5%	73.4%
26	67.9%	61.5%	72.7%	81.3%	88.5%	75.8%
31	64.7%	56.3%	71.1%	79.0%	87.5%	72.7%
36	63.8%	54.2%	71.1%	79.5%	88.5%	72.7%
41	62.9%	54.2%	69.5%	77.7%	88.5%	69.5%
46	62.1%	52.1%	69.5%	78.1%	88.5%	70.3%
51	62.1%	54.2%	68.0%	79.0%	88.5%	71.9%
55	62.9%	54.2%	69.5%	77.7%	88.5%	69.5%
60	62.1%	52.1%	69.5%	79.0%	90.6%	70.3%
65	61.2%	51.0%	68.8%	78.6%	90.6%	69.5%
70	61.2%	53.1%	67.2%	77.2%	90.6%	67.2%
75	62.1%	54.2%	68.0%	76.8%	90.6%	66.4%
80	62.5%	54.2%	68.8%	77.2%	90.6%	67.2%
85	62.5%	54.2%	68.8%	77.2%	89.6%	68.0%
90	62.1%	53.1%	68.8%	75.9%	90.6%	64.8%
95	62.9%	54.2%	69.5%	74.6%	89.6%	63.3%
100	62.5%	53.1%	69.5%	75.4%	90.6%	64.1%

Anexo IV: Random Forest con balanceo de clase

Undersampling con variables Originales

Cantidad Arboles	Altura											
	Clasificación_0						Clasificación_1					
	5	6	7	8	9	10	5	6	7	8	9	10
50	73%	71%	73%	73%	69%	73%	88%	88%	88%	88%	85%	85%
98	73%	69%	73%	71%	73%	73%	88%	85%	88%	88%	88%	88%
145	71%	71%	71%	73%	71%	71%	88%	88%	88%	88%	88%	88%
193	71%	73%	73%	71%	73%	71%	88%	88%	88%	88%	88%	88%
240	73%	71%	73%	71%	73%	73%	88%	88%	88%	88%	88%	88%
288	73%	71%	71%	73%	71%	75%	88%	88%	88%	88%	88%	88%
335	71%	73%	73%	71%	73%	75%	88%	88%	88%	88%	88%	88%
383	71%	73%	75%	73%	73%	71%	88%	88%	88%	88%	88%	88%
430	71%	73%	73%	73%	73%	73%	88%	88%	88%	88%	88%	88%
478	71%	71%	71%	71%	75%	73%	88%	88%	88%	88%	88%	88%
525	71%	73%	71%	75%	71%	73%	88%	88%	88%	88%	88%	88%
573	71%	71%	71%	71%	73%	73%	88%	88%	88%	88%	88%	88%
620	71%	71%	73%	71%	75%	73%	88%	88%	88%	88%	88%	88%
668	71%	73%	73%	71%	71%	73%	88%	88%	88%	88%	88%	88%
715	71%	71%	73%	71%	71%	75%	88%	88%	88%	88%	88%	88%
763	71%	71%	71%	71%	73%	71%	88%	88%	88%	88%	88%	88%
810	71%	73%	71%	71%	75%	75%	88%	88%	88%	88%	88%	88%
858	71%	73%	71%	73%	73%	73%	88%	88%	88%	88%	88%	88%
905	71%	71%	73%	73%	73%	73%	88%	88%	88%	88%	88%	88%
953	71%	73%	71%	73%	73%	71%	88%	88%	88%	88%	88%	88%
1000	71%	71%	73%	71%	71%	73%	88%	88%	88%	88%	88%	88%

Undersampling con ACP

Cantidad Arboles	Altura											
	Clasificación_0						Clasificación_1					
	5	6	7	8	9	10	5	6	7	8	9	10
50	78%	71%	76%	73%	73%	73%	88%	81%	92%	85%	88%	85%
98	71%	73%	71%	76%	75%	73%	81%	85%	85%	88%	85%	85%
145	73%	76%	73%	73%	69%	75%	85%	88%	85%	88%	81%	88%
193	75%	76%	75%	75%	73%	75%	85%	88%	85%	88%	85%	88%
240	75%	73%	75%	75%	69%	73%	85%	85%	88%	85%	81%	88%
288	76%	73%	75%	75%	73%	73%	88%	88%	88%	88%	85%	85%
335	73%	75%	73%	73%	75%	75%	88%	88%	85%	85%	88%	85%
383	75%	75%	73%	71%	73%	75%	88%	85%	85%	81%	85%	85%
430	75%	73%	75%	71%	75%	75%	88%	85%	88%	85%	88%	88%
478	73%	73%	76%	73%	73%	73%	88%	85%	88%	88%	85%	85%
525	75%	73%	75%	75%	73%	75%	88%	85%	88%	85%	85%	88%
573	75%	73%	73%	75%	75%	73%	88%	88%	85%	85%	85%	85%
620	73%	73%	71%	73%	75%	73%	88%	88%	85%	85%	88%	85%
668	75%	75%	75%	75%	73%	73%	88%	88%	88%	88%	85%	85%
715	73%	73%	75%	75%	75%	76%	88%	88%	88%	88%	88%	88%
763	75%	75%	73%	71%	73%	75%	88%	88%	85%	85%	85%	88%
810	75%	75%	75%	75%	73%	73%	88%	88%	85%	85%	85%	85%
858	75%	75%	73%	73%	73%	75%	88%	88%	85%	85%	85%	85%
905	73%	76%	73%	75%	71%	75%	88%	88%	85%	88%	85%	85%
953	76%	73%	73%	73%	73%	82%	88%	88%	85%	85%	85%	6%
1000	73%	75%	73%	75%	73%	82%	88%	88%	85%	85%	88%	6%

Oversampling con variables Originales

Cantidad Arboles	Altura											
	Clasificación_0						Clasificación_1					
	5	6	7	8	9	10	5	6	7	8	9	10
50	82%	82%	82%	82%	82%	82%	7%	8%	7%	7%	6%	8%
98	82%	82%	82%	82%	82%	82%	7%	7%	7%	8%	8%	7%
145	82%	82%	82%	82%	82%	82%	7%	7%	8%	7%	8%	7%
193	81%	82%	82%	82%	82%	82%	5%	7%	7%	7%	7%	7%
240	82%	82%	82%	82%	82%	82%	7%	7%	7%	7%	8%	7%
288	82%	82%	82%	82%	82%	82%	7%	7%	7%	7%	7%	7%
335	82%	82%	82%	82%	82%	82%	7%	7%	7%	7%	7%	8%
383	82%	82%	82%	82%	82%	82%	7%	7%	7%	6%	7%	7%
430	82%	82%	82%	82%	82%	82%	7%	7%	7%	7%	7%	7%
478	82%	82%	82%	82%	82%	82%	7%	7%	7%	7%	7%	7%
525	82%	82%	82%	82%	82%	82%	7%	7%	7%	7%	7%	7%
573	82%	82%	82%	82%	82%	82%	7%	7%	7%	7%	7%	7%
620	82%	82%	82%	82%	82%	82%	7%	7%	7%	7%	7%	7%
668	82%	82%	82%	82%	82%	82%	7%	7%	7%	7%	7%	7%
715	82%	82%	82%	82%	82%	82%	7%	7%	7%	7%	7%	7%
763	82%	82%	82%	82%	82%	82%	7%	7%	7%	7%	7%	7%
810	82%	82%	82%	82%	82%	82%	7%	7%	7%	7%	7%	7%
858	82%	82%	82%	82%	82%	82%	7%	7%	7%	7%	7%	7%
905	82%	82%	82%	82%	82%	82%	7%	7%	7%	7%	7%	7%
953	82%	82%	82%	82%	82%	82%	7%	7%	7%	7%	7%	7%
1000	82%	82%	82%	82%	82%	82%	7%	7%	7%	7%	7%	7%

Oversampling con ACP

Cantidad Arboles	Altura											
	Clasificación_0						Clasificación_1					
	5	6	7	8	9	10	5	6	7	8	9	10
50	82%	80%	82%	82%	80%	80%	6%	0%	6%	6%	0%	0%
98	81%	82%	80%	82%	82%	80%	5%	6%	0%	6%	6%	0%
145	82%	82%	82%	82%	81%	82%	6%	6%	6%	6%	5%	6%
193	82%	82%	82%	80%	82%	82%	6%	6%	6%	0%	6%	6%
240	82%	82%	82%	82%	82%	82%	6%	6%	6%	6%	6%	6%
288	80%	82%	82%	82%	82%	82%	0%	6%	6%	6%	6%	6%
335	82%	82%	82%	82%	82%	82%	6%	6%	6%	6%	6%	6%
383	82%	82%	82%	82%	82%	82%	6%	6%	6%	6%	6%	6%
430	82%	82%	82%	82%	82%	82%	6%	6%	6%	6%	6%	6%
478	82%	82%	82%	82%	82%	82%	6%	6%	6%	6%	6%	6%
525	81%	82%	82%	82%	82%	82%	5%	6%	6%	6%	6%	6%
573	82%	82%	82%	82%	80%	82%	6%	6%	6%	6%	0%	6%
620	82%	82%	82%	82%	82%	82%	6%	6%	6%	6%	6%	6%
668	82%	82%	82%	82%	82%	82%	6%	6%	6%	6%	6%	6%
715	82%	82%	82%	82%	82%	82%	6%	6%	6%	6%	6%	6%
763	82%	82%	82%	82%	82%	82%	6%	6%	6%	6%	6%	6%
810	82%	82%	82%	82%	81%	82%	6%	6%	6%	6%	5%	6%
858	82%	82%	82%	82%	82%	82%	6%	6%	6%	6%	6%	6%
905	82%	82%	82%	82%	82%	82%	6%	6%	6%	6%	6%	6%
953	82%	82%	82%	82%	82%	82%	6%	6%	6%	6%	6%	6%
1000	82%	82%	82%	82%	82%	82%	6%	6%	6%	6%	6%	6%

Undersampling & Oversampling con variables Originales

Cantidad Arboles	Altura											
	Clasificación_0						Clasificación_1					
	5	6	7	8	9	10	5	6	7	8	9	10
50	82%	83%	83%	82%	82%	82%	10%	11%	11%	10%	10%	10%
98	82%	82%	82%	82%	82%	82%	10%	10%	7%	10%	10%	10%
145	82%	83%	83%	83%	82%	83%	10%	11%	11%	11%	10%	11%
193	83%	83%	82%	82%	82%	82%	11%	11%	10%	10%	10%	10%
240	83%	83%	83%	82%	82%	83%	11%	11%	11%	10%	10%	11%
288	82%	82%	82%	82%	82%	82%	10%	10%	10%	10%	10%	10%
335	82%	82%	82%	82%	82%	82%	10%	10%	10%	10%	10%	10%
383	82%	82%	83%	83%	83%	83%	10%	10%	11%	11%	11%	11%
430	82%	82%	83%	83%	82%	82%	10%	10%	11%	11%	10%	10%
478	83%	83%	82%	82%	82%	83%	11%	11%	10%	10%	10%	11%
525	82%	82%	82%	82%	82%	83%	10%	10%	10%	10%	10%	11%
573	82%	82%	82%	82%	82%	83%	10%	10%	10%	10%	10%	11%
620	82%	83%	83%	82%	82%	82%	10%	11%	11%	10%	10%	10%
668	82%	83%	82%	82%	82%	82%	10%	11%	10%	10%	10%	10%
715	82%	82%	82%	83%	83%	82%	10%	10%	10%	11%	11%	10%
763	82%	82%	82%	82%	82%	82%	10%	10%	10%	10%	10%	10%
810	83%	82%	82%	82%	82%	83%	11%	10%	10%	10%	10%	11%
858	82%	82%	83%	83%	82%	82%	10%	10%	11%	11%	10%	10%
905	82%	82%	82%	82%	82%	82%	10%	10%	10%	10%	10%	10%
953	82%	82%	83%	82%	83%	83%	10%	10%	11%	10%	11%	11%
1000	82%	83%	82%	83%	83%	83%	10%	11%	10%	11%	11%	11%

Undersampling & Oversampling con ACP

Cantidad Arboles	Altura											
	Clasificación_0						Clasificación_1					
	5	6	7	8	9	10	5	6	7	8	9	10
50	82%	82%	82%	82%	82%	82%	7%	7%	7%	10%	7%	7%
98	82%	81%	82%	82%	82%	82%	7%	0%	7%	7%	7%	7%
145	82%	81%	82%	82%	82%	82%	7%	0%	7%	7%	7%	7%
193	82%	82%	82%	82%	82%	82%	7%	7%	7%	7%	7%	7%
240	82%	82%	82%	82%	82%	82%	7%	7%	7%	7%	7%	7%
288	82%	82%	82%	82%	82%	82%	7%	7%	7%	7%	7%	7%
335	82%	82%	82%	82%	82%	82%	7%	7%	7%	7%	7%	7%
383	82%	82%	82%	82%	82%	82%	7%	7%	7%	7%	7%	7%
430	82%	82%	82%	82%	82%	82%	7%	7%	7%	7%	7%	7%
478	82%	82%	82%	82%	82%	82%	7%	7%	7%	7%	7%	7%
525	82%	82%	82%	82%	82%	82%	7%	7%	7%	7%	7%	7%
573	82%	82%	82%	82%	82%	82%	7%	7%	7%	7%	7%	7%
620	82%	82%	82%	82%	82%	82%	7%	7%	7%	7%	7%	7%
668	82%	82%	82%	82%	82%	82%	7%	7%	7%	7%	7%	7%
715	82%	82%	82%	82%	82%	82%	7%	7%	7%	7%	7%	7%
763	82%	82%	82%	82%	82%	82%	7%	7%	7%	7%	7%	7%
810	82%	82%	82%	82%	82%	82%	7%	7%	7%	7%	7%	7%
858	82%	82%	82%	82%	82%	82%	7%	7%	7%	7%	7%	7%
905	82%	82%	82%	82%	82%	82%	7%	7%	7%	7%	7%	7%
953	82%	82%	82%	82%	82%	82%	7%	7%	7%	7%	7%	7%
1000	82%	82%	82%	82%	82%	82%	6%	6%	6%	6%	6%	6%

SMOTE con variables Originales

Cantidad Arboles	Altura											
	Clasificación_0						Clasificación_1					
	5	6	7	8	9	10	5	6	7	8	9	10
50	57%	57%	57%	57%	57%	57%	0%	0%	0%	0%	0%	0%
98	57%	57%	57%	57%	57%	57%	0%	0%	0%	0%	0%	0%
145	57%	57%	57%	57%	57%	57%	0%	0%	0%	0%	0%	0%
193	57%	57%	57%	57%	57%	57%	0%	0%	0%	0%	0%	0%
240	57%	57%	57%	57%	57%	57%	0%	0%	0%	0%	0%	0%
288	57%	57%	57%	57%	57%	57%	0%	0%	0%	0%	0%	0%
335	57%	57%	57%	57%	57%	57%	0%	0%	0%	0%	0%	0%
383	57%	57%	57%	57%	57%	57%	0%	0%	0%	0%	0%	0%
430	57%	57%	57%	57%	57%	57%	0%	0%	0%	0%	0%	0%
478	57%	57%	57%	57%	57%	57%	0%	0%	0%	0%	0%	0%
525	57%	57%	57%	57%	57%	57%	0%	0%	0%	0%	0%	0%
573	57%	57%	57%	57%	57%	57%	0%	0%	0%	0%	0%	0%
620	57%	57%	57%	57%	57%	57%	0%	0%	0%	0%	0%	0%
668	57%	57%	57%	57%	57%	57%	0%	0%	0%	0%	0%	0%
715	57%	57%	57%	57%	57%	57%	0%	0%	0%	0%	0%	0%
763	57%	57%	57%	57%	57%	57%	0%	0%	0%	0%	0%	0%
810	57%	57%	57%	57%	57%	57%	0%	0%	0%	0%	0%	0%
858	57%	57%	57%	57%	57%	57%	0%	0%	0%	0%	0%	0%
905	57%	57%	57%	57%	57%	57%	0%	0%	0%	0%	0%	0%
953	57%	57%	57%	57%	57%	57%	0%	0%	0%	0%	0%	0%
1000	57%	57%	57%	57%	57%	57%	0%	0%	0%	0%	0%	0%

SMOTE con ACP

Cantidad Arboles	Altura											
	Clasificación_0						Clasificación_1					
	5	6	7	8	9	10	5	6	7	8	9	10
50	58%	62%	78%	70%	79%	69%	8%	11%	63%	41%	65%	36%
98	59%	61%	78%	79%	79%	79%	5%	11%	64%	70%	64%	63%
145	60%	78%	79%	76%	79%	63%	8%	63%	63%	55%	65%	20%
193	61%	76%	78%	78%	77%	79%	11%	52%	64%	64%	56%	68%
240	60%	65%	76%	78%	80%	81%	7%	23%	54%	61%	65%	66%
288	61%	71%	79%	71%	79%	79%	10%	40%	63%	38%	61%	63%
335	59%	75%	75%	79%	79%	79%	7%	52%	52%	61%	64%	65%
383	60%	71%	66%	76%	80%	79%	8%	42%	29%	53%	66%	63%
430	60%	71%	76%	80%	75%	80%	8%	40%	51%	63%	51%	64%
478	60%	62%	76%	76%	80%	79%	8%	14%	54%	53%	61%	63%
525	60%	75%	75%	79%	79%	81%	9%	50%	52%	64%	63%	65%
573	59%	75%	79%	78%	79%	78%	6%	50%	63%	59%	64%	58%
620	60%	69%	76%	79%	79%	80%	9%	34%	54%	64%	61%	63%
668	60%	74%	75%	80%	78%	79%	7%	49%	52%	64%	59%	63%
715	59%	71%	72%	78%	76%	79%	6%	43%	42%	57%	54%	63%
763	59%	64%	77%	79%	77%	79%	6%	21%	55%	60%	57%	61%
810	60%	64%	75%	79%	80%	79%	8%	21%	50%	63%	63%	59%
858	61%	71%	70%	79%	80%	80%	9%	39%	36%	64%	65%	64%
905	59%	72%	80%	79%	79%	79%	6%	43%	64%	63%	64%	63%
953	60%	61%	77%	73%	80%	79%	7%	11%	58%	43%	64%	63%
1000	60%	71%	75%	75%	79%	80%	9%	39%	53%	53%	63%	64%