Conjunto de Datos: Australia Weather Data.

Jonathan Andrés Jiménez Trujillo
Sebastian Prada Padilla

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Contenido

UNIVERSIDAD NACIONAL DE COLOMBIA

1. Introducción
2. Objetivos
3. Descripción de los datos.
4. Exploración de datos.
5. Preprocesamiento.
6. Asociación
7. Agrupación
8. Clasificación
9. Conclusiones

Introducción

UNIVERSIDAD NACIONAL DE COLOMBIA

El conjunto de datos describe mediciones climáticas durante 10 años en diferentes ciudades de Australia. A partir de estos registros se busca conseguir información que nos ayude a dar con los objetivos, todo esto aplicando diferentes técnicas de minería de datos, empezando desde análisis exploratorio a preprocesamiento para después hacer asociación, clasificación y agrupación.

Objetivos

UNIVERSIDAD NACIONAL DE COLOMBIA

- Desarrollar un modelo para predecir si al día siguiente llueve basado en las variables de un día en específico haciendo uso de modelos de clasificación binaria.
- Aplicar las técnicas de análisis exploratorio, preprocesamiento, agrupación, asociación y clasificación al conjunto de datos.

Análisis exploratorio

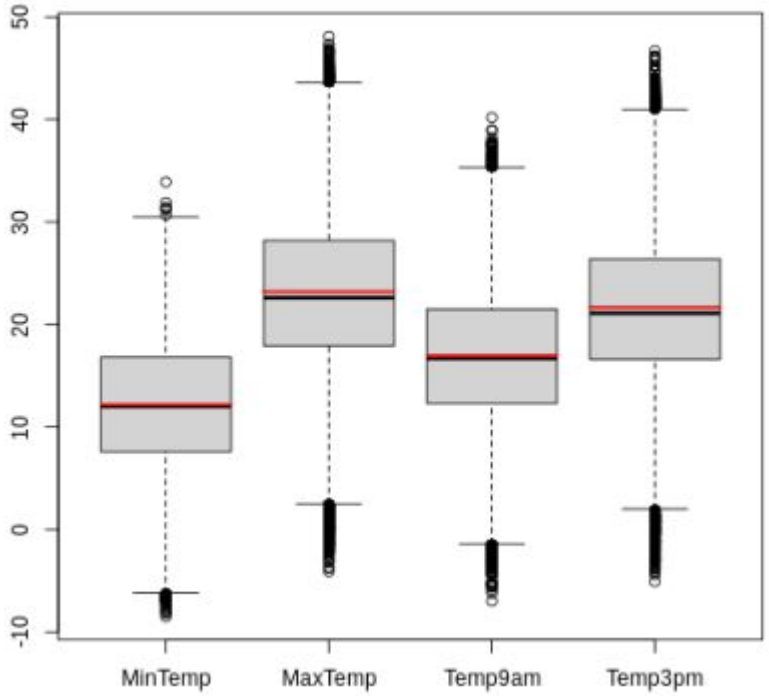
UNIVERSIDAD
NACIONAL
DE COLOMBIA

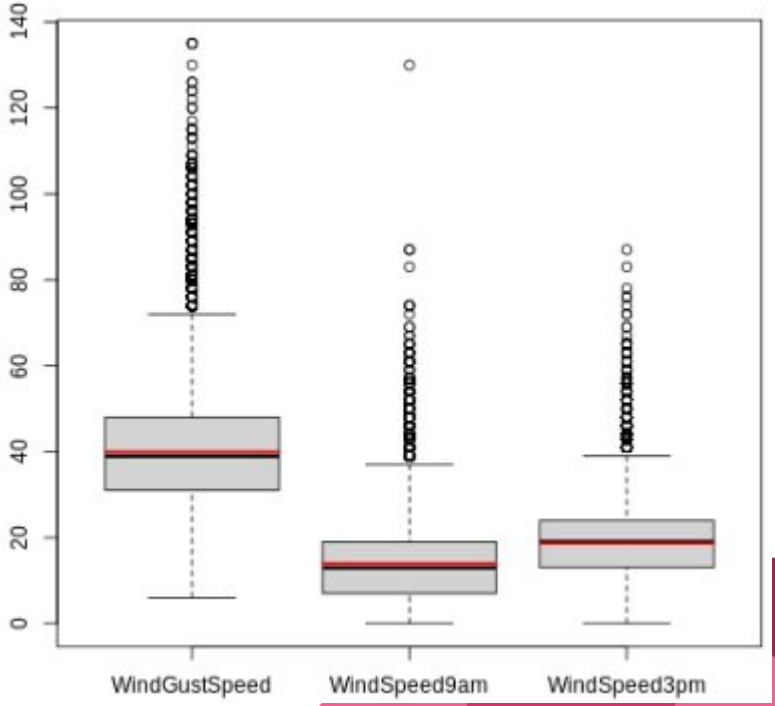
Descripción de los datos

UNIVERSIDAD NACIONAL DE COLOMBIA

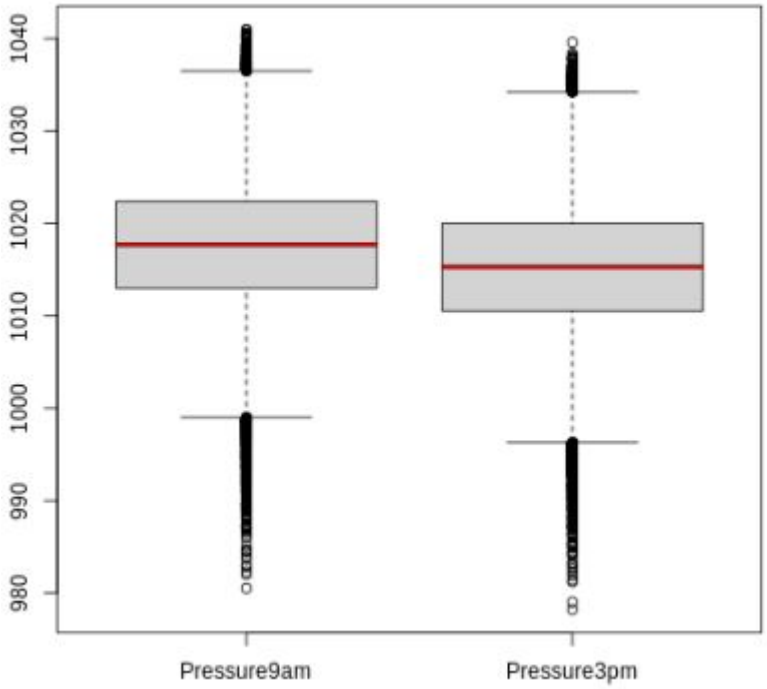
Variable	Descripción	Tipo	Rango	Unidad
Row ID	Identificador del registro	Nominal Discreto	99516 valores unicos	-
Location	Nombre de la ciudad de Australia	Nominal Discreto	45 valores unicos	-
MinTemp	Temperatura mínima durante el día	Proporción Continuo	[-8.5, 33.9]	Grados Celsius
MaxTemp	Temperatura máxima durante el día	Proporción Continuo	[-4.1, 48.1]	Grados Celsius
Rainfall	Precipitación durante el día	Proporción Continuo	[0.0, 371.0]	milimetros
Evaporation	Evaporación durante el día	Proporción Continuo	[0.0, 86.2]	milimetros
Sunshine	Sol brillante durante el día	Proporción Continuo	[0.0, 14.5]	Horas
WindGusDir	Dirección de la rafaga de viento más fuerte durante el día	Nominal Discreto	16 valores unicos	puntos de compás
WindGuSpeed	Velocidad de la rafaga de viento más fuerte durante el día	Proporción Continuo	[6.0, 135.0]	Km/h
WindDir9am	Dirección del viento 10 minutos antes de las 9 am	Nominal Discreto	16 valores unicos	puntos de compás
WindDir3pm	Dirección del viento 10 minutos antes de las 3 pm	Nominal Discreto	16 valores unicos	puntos de compás
WindSpeed9am	Velociadad del viento 10 minutos antes de las 9 am	Proporción Continuo	[0.0, 130.0]	Km/h
WindSpeed3pm	Velocidad del viento 10 minutos antes de las 3 pm	Proporción Continuo	[0.0, 87.0]	Km/h
Humidity9am	Humedad del aire a las 9 am	Proporción Continuo	[0.0, 100.0]	Porcentaje
Humidity3pm	Humedad del aire a las 3 pm	Proporción Continuo	[0.0, 100.0]	Porcentaje
Pressure9am	Presión atmosférica a las 9 am	Proporción Continuo	[980.5, 1041.0]	Hectopascal
Pressure3pm	Presión atmosférica a las 3 pm	Proporción Continuo	[978.2, 1039.6]	Hectopascal
Cloud9am	Porción de nubes oscuras a las 9 am	Proporción Discreto	[0.0, 9.0]	Octavos
Cloud3pm	Porción de nubes oscuras a las 3 pm	Proporción Discreto	[0.0, 9.0]	Octavos
Temp9am	Temperatura a las 9 am	Proporción Continuo	[-7.0, 40.2]	Grados Celsius
Temp3pm	Temperatura a las 3 pm	Proporción Continuo	[-5.1, 46.7]	Grados Celsius
RainToday	El día de hoy llueve	Nominal Binario	2 valores	-
RainTomorrow	El día de mañana llueve, Si: 1 y No: 0	numerico Binario	2 valores	-

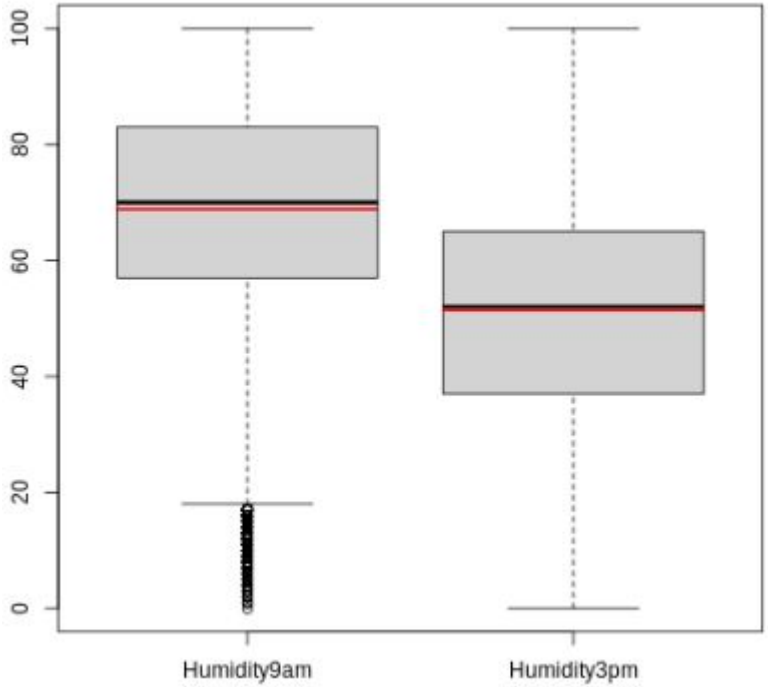
Boxplots de las variables numéricas



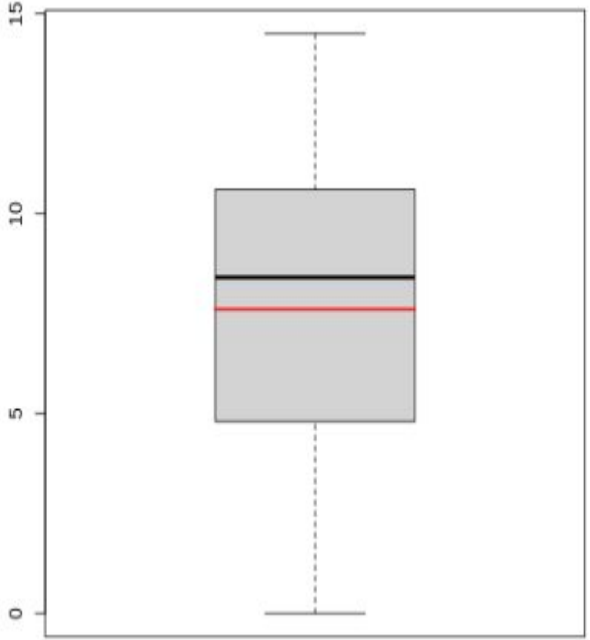


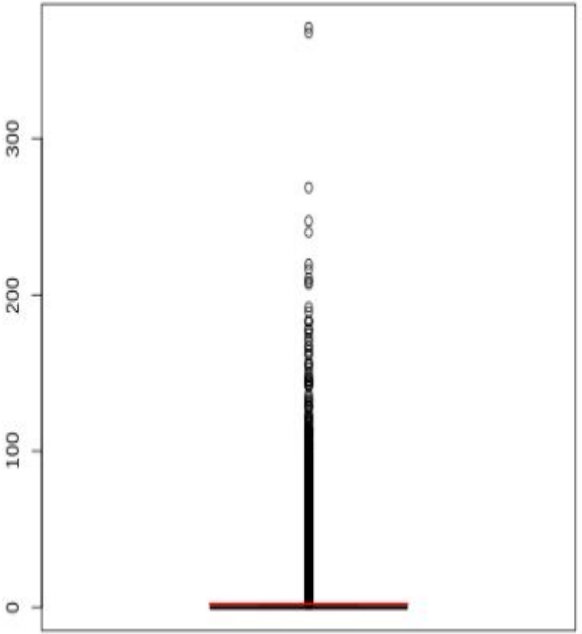
Boxplots de las variables numéricas

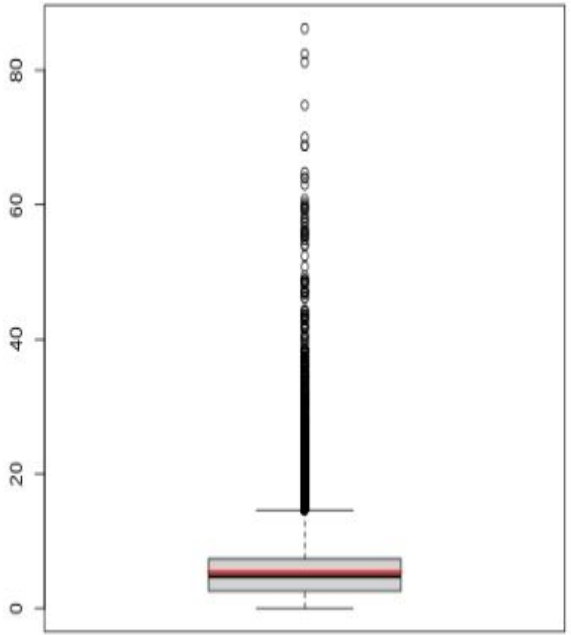




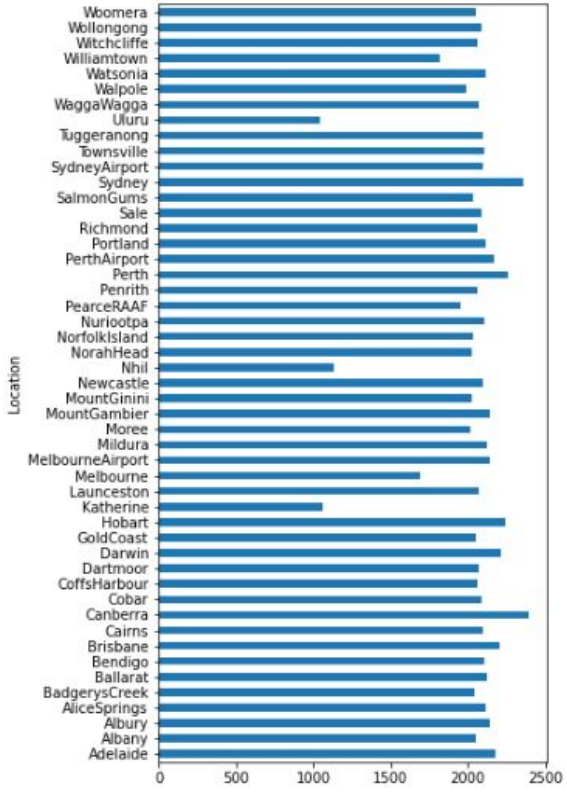
Sunshine, rainfall y evaporation

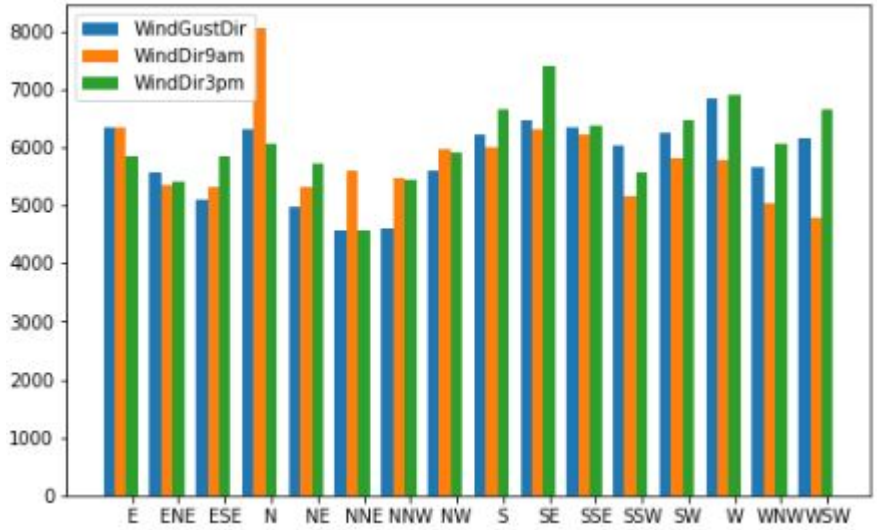




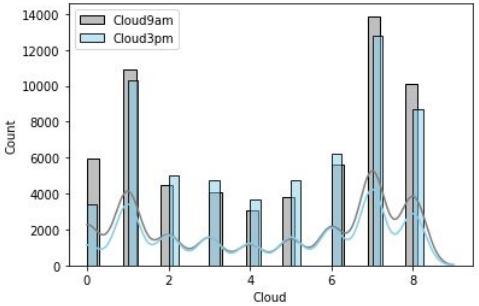


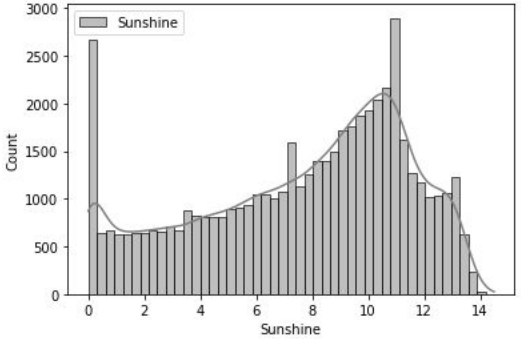
Exploración de variables categóricas

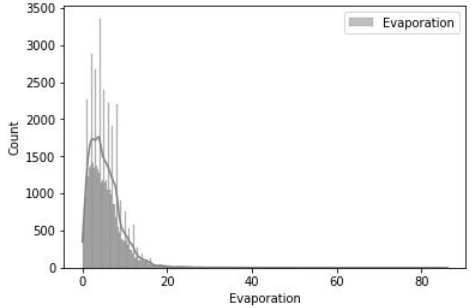


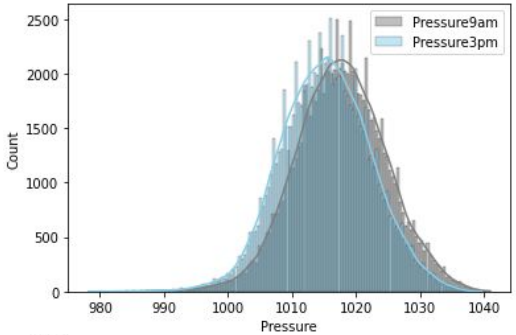


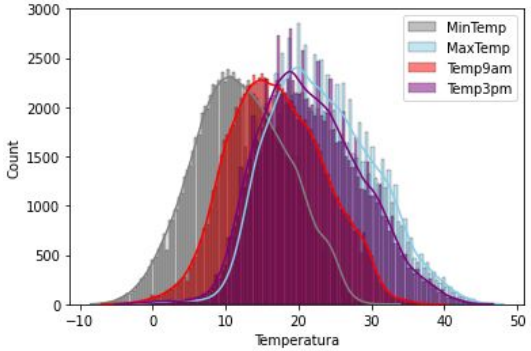
Histogramas

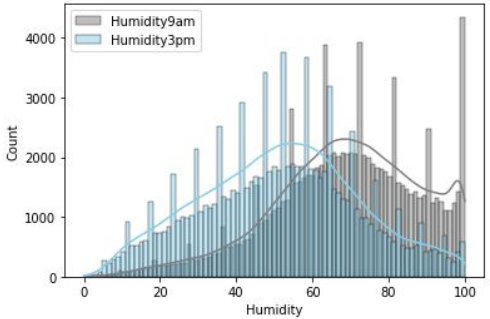


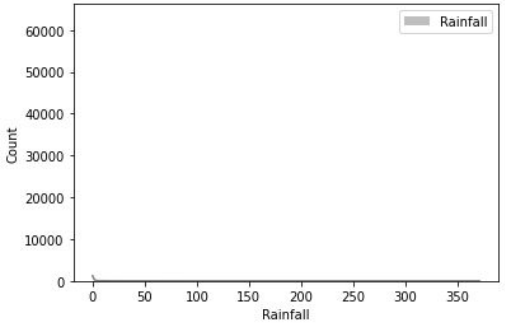


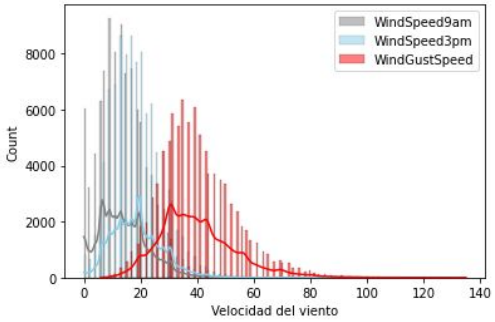




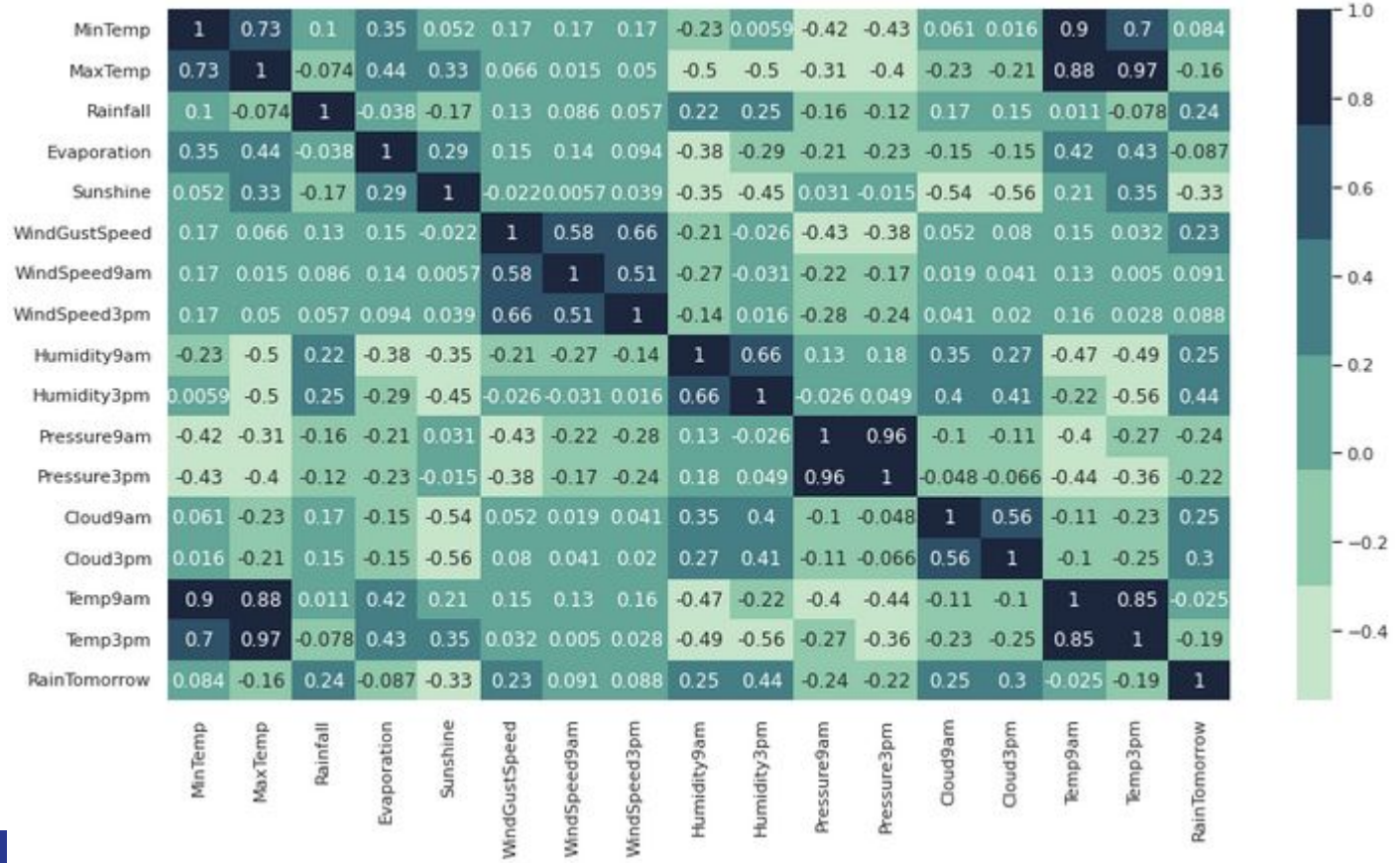








Matriz de Correlación



Preprocesamiento

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Limpieza de datos

UNIVERSIDAD NACIONAL DE COLOMBIA

- No se encuentran registros duplicados
- Hay variables con casi la mitad de registros faltantes
- Valores numéricos se reemplazan por la media y valores categóricos se reemplazan por el valor con más frecuencia.

```
Number of instances = 99516
Number of attributes = 23
Number of missing values:
        row ID: 0
        Location: 0
        MinTemp: 443
        MaxTemp: 230
        Rainfall: 979
        Evaporation: 42531
        Sunshine: 47317
        WindGustDir: 6521
        WindGustSpeed: 6480
        WindDir9am: 7006
        WindDir3pm: 2648
        WindSpeed9am: 935
        WindSpeed3pm: 1835
        Humidity9am: 1233
        Humidity3pm: 2506
        Pressure9am: 9748
        Pressure3pm: 9736
        Cloud9am: 37572
        Cloud3pm: 40002
        Temp9am: 614
        Temp3pm: 1904
        RainToday: 979
        RainTomorrow: 0
```

Outliers

Todas las variables numéricas que quedan en el conjunto de datos menos la humedad a las 3 pm presentan outliers por lo que se realiza una normalización Z-score y de ahí se eliminan los registros que en alguno de sus atributos este fuera del umbral (Z > 3 or Z <= -3), los registros de la variable rainfall con outliers no se eliminaron.

Con los registros eliminados nos queda un total de 96225 de 99516 del conjunto original

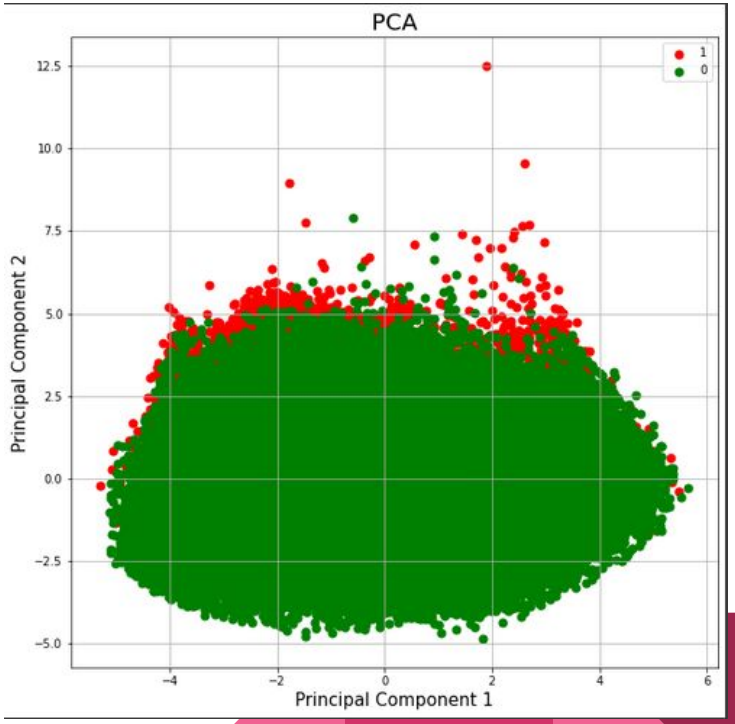
Ranking por Entropía.

Atributo	Ganancia	Entropía
WindSpeed3pm	0.227	780.381
WindGustSpeed	1.597	780.154
WindSpeed9am	1.925	781.75
Pressure9am	4.016	783.677
Pressure3pm	3.641	787.694
MinTemp	6.368	791.335
Humidity9am	9.767	797.703
Humidity3pm	11.587	807.471
Temp9am	16.178	819.058

Location	WindSpeed9am
WindGustDir	WindSpeed3pm
WindDir9am	Humidity9am
WindDir3pm	Humidity3pm
RainToday	Pressure9am
MinTemp	Pressure3pm
MaxTemp	Temp9am
Rainfall	Temp3pm
WindGustSpeed	

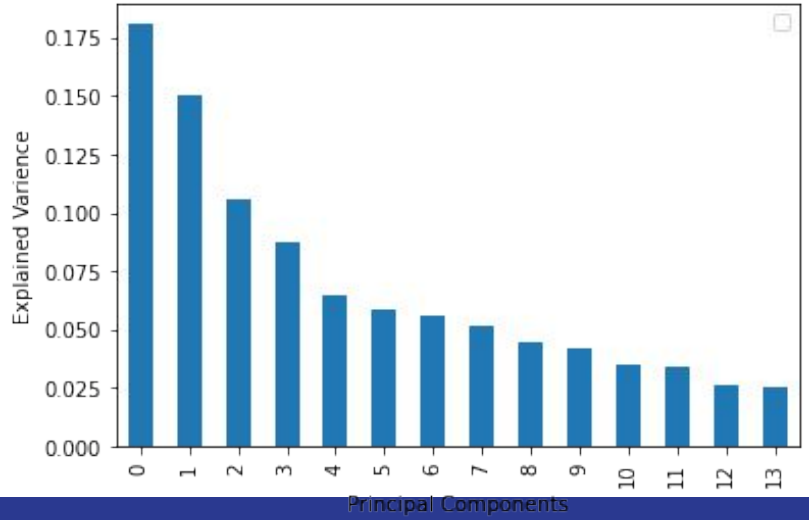
PCA

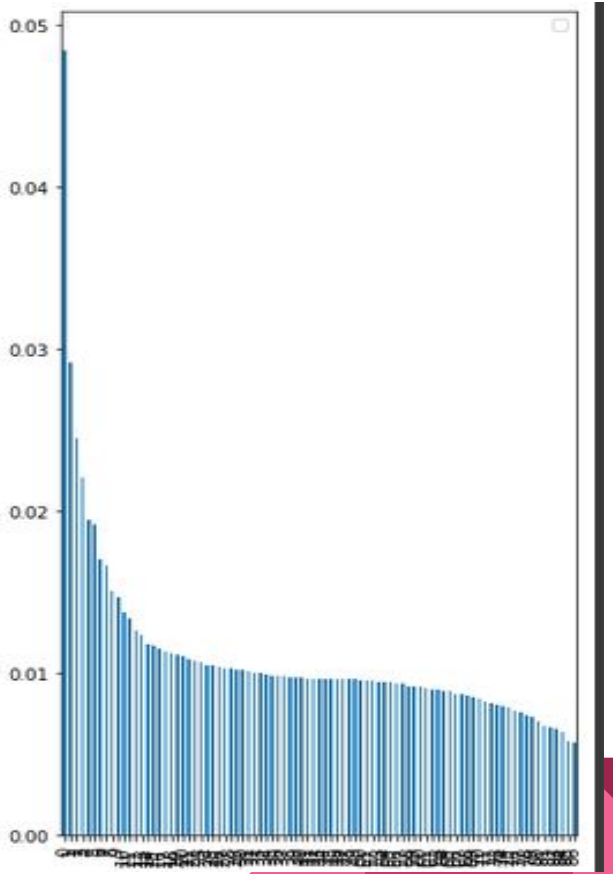
- Convertir los datos categóricos a numéricos, estandarizarlos y aplicar objeto PCA de sklearn
- No se logra evidenciar una categoría simple para separación de los atributos basada en la clase



Varianza explicada (PCA)

Se logra notar que a partir de 4 componentes principales la varianza explicada empieza a converger y se supone que a partir de este número de componentes se logra evidenciar los dos grupos (Si llueve o no el dia de mañana)





Asociación

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Discretización

- Libreria: KBinsDiscretizer.
- N. Bins: 5
- Estrategia: Igual Ancho.
- Cada intervalo se asigna un número Ordinal.

Puntos de Corte.

```
MinTemp [-7.    1.18  9.36 17.54 25.72 33.9 ]
MaxTemp [ 2.6 11.7 20.8 29.9 39.  48.1]
WindSpeed9am [ 2. 19. 36. 53. 70. 87.]
WindSpeed3pm [ 2. 19. 36. 53. 70. 87.]
WindGustSpeed [  7.   32.6  58.2  83.8 109.4 135. ]
Humidity9am [  0.  20.  40.  60.  80. 100.]
Humidity3pm [  0.  20.  40.  60.  80. 100.]
Pressure9am [ 980.5  992.6 1004.7 1016.8 1028.9 1041. ]
Pressure3pm [ 978.2   990.48 1002.76 1015.04 1027.32 1039.6 ]
Temp9am [-3.1   5.56 14.22 22.88 31.54 40.2 ]
Temp3pm [ 1.7 10.7 19.7 28.7 37.7 46.7]
```

Algoritmo Apriori

Confianza mínima = 0.7 y soporte mínimo = 0.4

```
Confianza: 0.7
{Rainfall--0.0, WindGustSpeed--1.0} -> {RainToday--No} (conf: 1.000, supp: 0.403, lift: 1.291, conv: 225347485.469)
{RainToday--No, WindGustSpeed--1.0} -> {Rainfall--0.0} (conf: 0.837, supp: 0.403, lift: 1.302, conv: 2.192)
{WindSpeed3pm--1.0} -> {WindGustSpeed--1.0} (conf: 0.808, supp: 0.406, lift: 1.297, conv: 1.966)
{RainTomorrow--0, WindGustSpeed--1.0} -> {RainToday--No} (conf: 0.841, supp: 0.409, lift: 1.086, conv: 1.419)
{RainToday--No, WindGustSpeed--1.0} -> {RainTomorrow--0} (conf: 0.849, supp: 0.409, lift: 1.091, conv: 1.467)
{Pressure9am--3.0} -> {RainTomorrow--0} (conf: 0.846, supp: 0.410, lift: 1.087, conv: 1.441)
{Pressure9am--3.0} -> {Pressure3pm--3.0} (conf: 0.855, supp: 0.415, lift: 1.860, conv: 3.736)
{Pressure3pm--3.0} -> {Pressure9am--3.0} (conf: 0.902, supp: 0.415, lift: 1.860, conv: 5.257)
{Rainfall--0.0, WindSpeed9am--0.0} -> {RainToday--No} (conf: 1.000, supp: 0.441, lift: 1.291, conv: 225347485.469)
{RainToday--No, WindSpeed9am--0.0} -> {Rainfall--0.0} (conf: 0.830, supp: 0.441, lift: 1.291, conv: 2.097)
{RainTomorrow--0, WindSpeed9am--0.0} -> {RainToday--No} (conf: 0.857, supp: 0.456, lift: 1.106, conv: 1.572)
{RainToday--No, WindSpeed9am--0.0} -> {RainTomorrow--0} (conf: 0.859, supp: 0.456, lift: 1.105, conv: 1.580)
{WindGustSpeed--1.0} -> {RainToday--No} (conf: 0.772, supp: 0.481, lift: 0.997, conv: 0.990)
{WindGustSpeed--1.0} -> {RainTomorrow--0} (conf: 0.779, supp: 0.486, lift: 1.002, conv: 1.006)
{WindSpeed9am--0.0} -> {RainToday--No} (conf: 0.797, supp: 0.531, lift: 1.028, conv: 1.108)
{WindSpeed9am--0.0} -> {RainTomorrow--0} (conf: 0.799, supp: 0.533, lift: 1.027, conv: 1.106)
{Rainfall--0.0} -> {RainTomorrow--0} (conf: 0.872, supp: 0.560, lift: 1.121, conv: 1.735)
{RainTomorrow--0} -> {Rainfall--0.0} (conf: 0.721, supp: 0.560, lift: 1.121, conv: 1.278)
{RainTomorrow--0, Rainfall--0.0} -> {RainToday--No} (conf: 1.000, supp: 0.560, lift: 1.291, conv: 225347485.469)
{RainToday--No, Rainfall--0.0} -> {RainTomorrow--0} (conf: 0.872, supp: 0.560, lift: 1.121, conv: 1.735)
{RainToday--No, RainTomorrow--0} -> {Rainfall--0.0} (conf: 0.853, supp: 0.560, lift: 1.326, conv: 2.423)
{Rainfall--0.0} -> {RainToday--No, RainTomorrow--0} (conf: 0.872, supp: 0.560, lift: 1.326, conv: 2.675)
{RainTomorrow--0} -> {RainToday--No, Rainfall--0.0} (conf: 0.721, supp: 0.560, lift: 1.121, conv: 1.278)
{RainToday--No} -> {RainTomorrow--0, Rainfall--0.0} (conf: 0.723, supp: 0.560, lift: 1.291, conv: 1.589)
{Rainfall--0.0} -> {RainToday--No} (conf: 1.000, supp: 0.643, lift: 1.291, conv: 225347485.469)
{RainToday--No} -> {Rainfall--0.0} (conf: 0.830, supp: 0.643, lift: 1.291, conv: 2.098)
{RainTomorrow--0} -> {RainToday--No} (conf: 0.845, supp: 0.657, lift: 1.091, conv: 1.455)
{RainToday--No} -> {RainTomorrow--0} (conf: 0.849, supp: 0.657, lift: 1.091, conv: 1.467)
```

FP Growth

Confianza mínima = 0.2 y soporte mínimo = 0.4

```
{('Pressure3pm--3.0',): (('Pressure9am--3.0',), 0.9020464221947535),
 ('Pressure9am--3.0',): (('RainTomorrow--0',), 0.8458071743038008),
 ('RainTomorrow--0',): (('RainToday--No',), 0.8451390694047309),
 ('WindSpeed3pm--1.0',): (('WindGustSpeed--1.0',), 0.8083869993962568),
 ('RainToday--No',): (('RainTomorrow--0',), 0.8485303232962516),
 ('RainToday--No', 'Rainfall--0.0'): (('RainTomorrow--0',),
  0.8719133375275244),
 ('RainToday--No', 'WindGustSpeed--1.0'): (('RainTomorrow--0',),
  0.8485198404366996),
 ('Rainfall--0.0', 'WindGustSpeed--1.0'): (('RainToday--No',), 1.0),
 ('RainToday--No', 'RainTomorrow--0'): (('WindSpeed9am--0.0',),
  0.6944252210688197),
 ('RainTomorrow--0', 'WindGustSpeed--1.0'): (('RainToday--No',),
  0.841212436581241),
 ('RainToday--No', 'WindSpeed9am--0.0'): (('RainTomorrow--0',),
  0.859379088854526),
 ('Rainfall--0.0', 'WindSpeed9am--0.0'): (('RainToday--No',), 1.0),
 ('RainTomorrow--0', 'Rainfall--0.0'): (('RainToday--No',), 1.0),
 ('RainTomorrow--0', 'WindSpeed9am--0.0'): (('RainToday--No',),
  0.8566279345506284)}
```

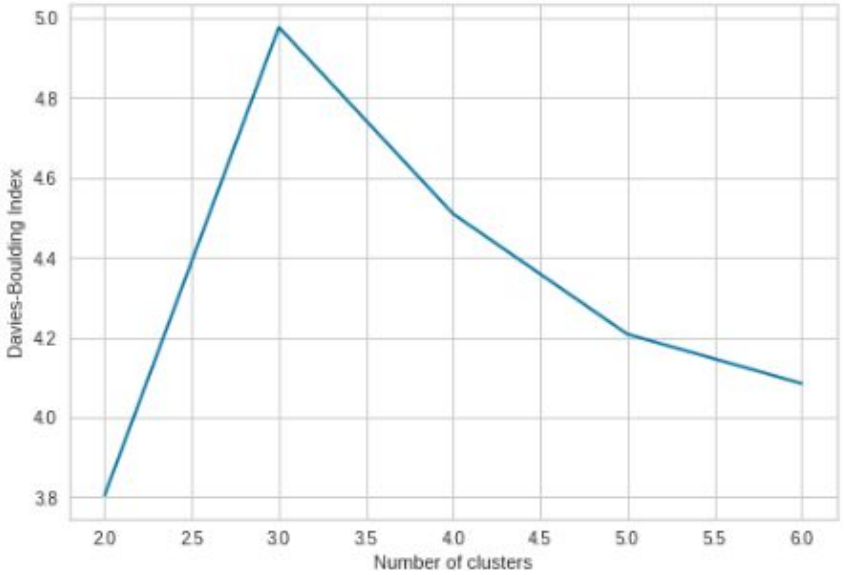
Agrupación

UNIVERSIDAD
NACIONAL
DE COLOMBIA

KMeans

KMeans usando distancia euclidiana y tc

Índice Davis-Bouldin para clustering con k = [2, 3, 4, 5 , 6].



Modelo con KMeans; k=2

- Precision: 0.47
- recall: 0.46
- F-score: 0.43

KNearestNeighbors

Distancia euclidiana.

Como parámetro número de vecinos = [5, 6, 7, 8].

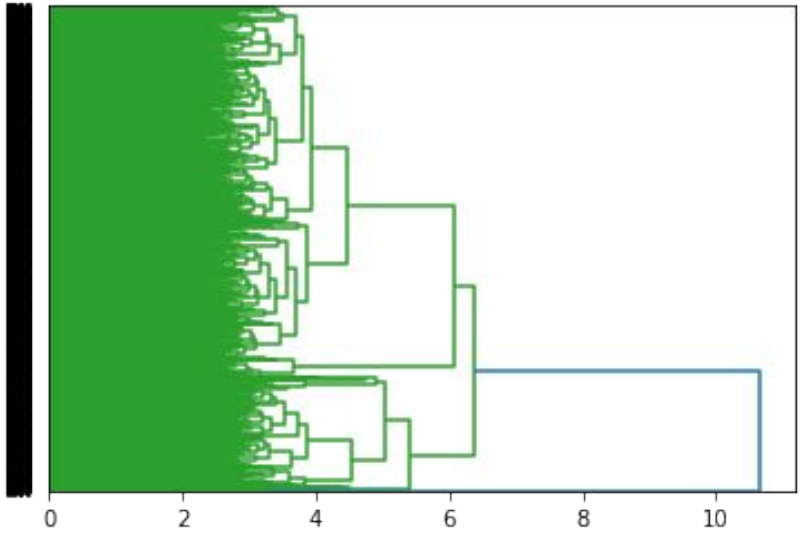
	precision	recall	F_score	support
n_5	0.730295	0.637369	0.658301	None
n_6	0.756298	0.603827	0.619479	None
n_7	0.745837	0.634712	0.65683	None
n_8	0.758171	0.602571	0.61781	None

Agrupación Jerárquica

Haciendo un muestreo del dataset con 4000 datos se logra evidenciar el dendograma respectivo por registros y se diferencian las dos clases

El dendograma es realizado a partir de los siguientes criterios:

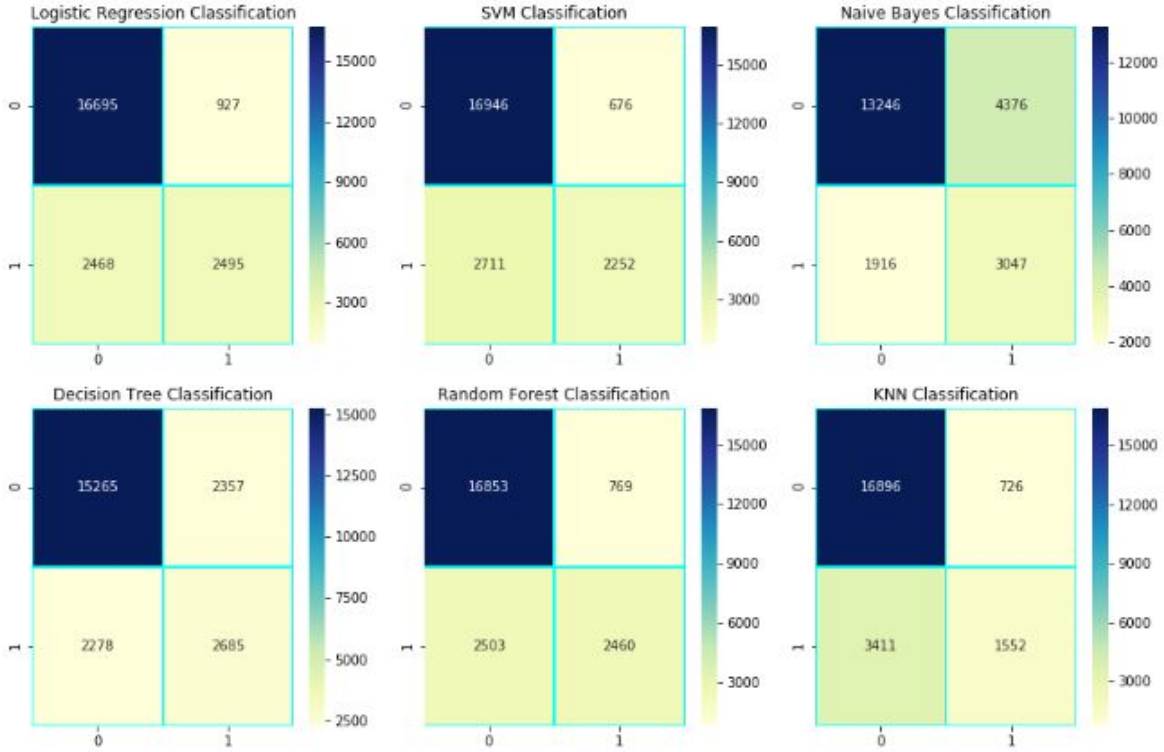
- Distancia de Gower.
- Proximidad Intercluster el promedio.



Clasificación

UNIVERSIDAD
NACIONAL
DE COLOMBIA

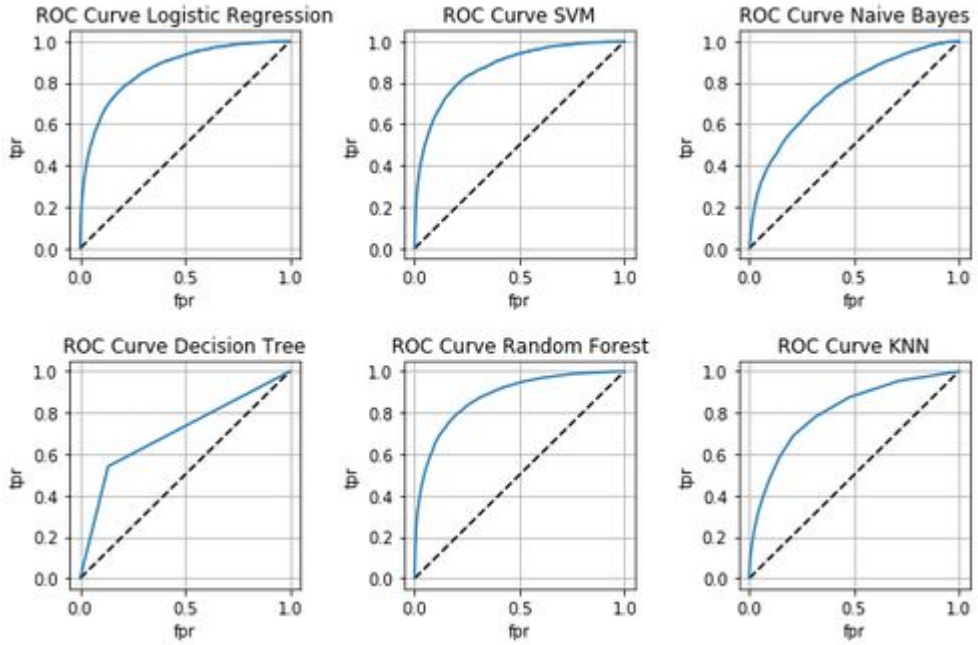
Comparación Modelos



Logistic Regression Score	0.849679
Support Vector Machine Score	0.850033
Naive Bayes Score	0.721408
Decision Tree Score	0.794775
Random Forest Score	0.855125
K-Nearest Neighbour Score	0.816825

dtype: float64

Curva ROC entre los modelos



Conclusiones

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Conclusiones

- La técnica de random forest presenta los mejores resultados en términos de predicción seguido por SVM y regresión logística con puntajes de prueba cercanos al 85%.
- Los peores resultados fueron presentados por la técnica Naive Bayes por lo que puede ser asociado a la suposición del modelo.
- Los modelos paramétricos tienen una mayor velocidad de aprendizaje a comparación de los modelos no paramétricos.
- Regresión logística nos brinda los mejores resultados teniendo en cuenta los tiempos de cómputo entre los distintos métodos.