Análisis Exploratorio y Preprocesamiento del Conjunto de Datos: Australia Weather Data.

Jonathan Andrés Jiménez Trujillo y Sebastian Prada Padilla, *Ests., Universidad Nacional de Colombia.*

***Abstract*—The present paper is about the study of a data set related to temperature and weather in different sites in Australia and the process that is needed to make use of the data set in order to later get information of the different records of it applying methods of classification, group and association. First is needed to know the different attributes the data set has and the range or type of data of it, also understanding the info and distribution it has making some statistical analysis such as mean, median, quantiles covariance between others important statistical measures and doing graphs to better understand the info acquired, after this being done it goes to the preprocessing of the data in order to discover information from the data set or this one being more accessible for further data mining methods. Some examples of tasks that need to be done in this phase are cleaning the data set, look for missing data, eliminating outliers, transformation of variables (standardization/normalization), sampling, discretization and feature selection.**

***Index Terms*—Míneria de datos, análisis exploratorio, preprocesamiento, predicción, clima**

I. Introducción

El conjunto de datos es obtenido de la pagina web https://www.kaggle.com/datasets, este conjunto describe mediciones climáticas durante 10 años en diferentes ciudades de Australia. EL objetivo principal para que el que se usara esta información es desarrollar un modelo para predecir si un día llueve basado en las mediciones del dia anterior, por lo tanto es un problema de clasificacíon, de igual forma también se pueden aplicar otros modelos de minería de datos como asociación y agrupación que se irán estudiando a lo largo del curso. En el presente documento se hace una descripción del conjunto de datos y un preprocemiento para limpiar el conjunto de datos y transformaciones para poder usar varios métodos de minería de datos para lograr la predicción de lluvia en un día dadas ciertas condiciones.

II. Descripción de los Datos

El conjunto de datos cuenta con 99516 registros con 23 variables que constan de 22 atributos: 16 numericos, 6 categoticos y una clase, cada registro hace referenia a datos de un día particular y la clase si el dia siguiente llovio o no, hay atributos que son de la misma variable física pero tomadas en diferente hora del día, entre estas estan la temperatura, humedad, presión atmosférica, viento y oscuridad de las nubes, por otro lado las categoricas cosntan de un identificador para cada registro, la ubicación e indican la dirección del viento y por ultimo hay un atributo binario junto a la clase, en la tabla 1 se detalla cada variable con una descripción, tipo de variable, rango y unidad de medida.

TABLE I: Descripción de los Datos.

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	-

III. Exploración de datos

A. Variables Numericas

1) Temperatura: Para la temperatura hay 4 variables en el conjunto de datos: minima temperatura, maxima temperatura,

temperatura a las 9 am y temperatura a las 3pm las medidas de centralidad y dispersion se muestran en la tabla II, la media se obaserva en el boxplot con una linea roja y es muy cercana a la mediana para las cuatro variables, en los histogrmas se puede observar que las cuatro medidas tienen un comportamiento muy similar, incluso entre la temperatura maxima y temperatura a las 3pm es casi el mismo histograma y sus medidas de centralidad y dispersion son cercanas.

TABLE II: Medidas de centralidad y dispersión de las variables de temperatura.

variable	Media	Desviación Estandar
MinTemp	12.17	6.3
MaxTemp	23.21	7.11
Temp9am	16.97	6.4
Temp3pm	21.68	6.9

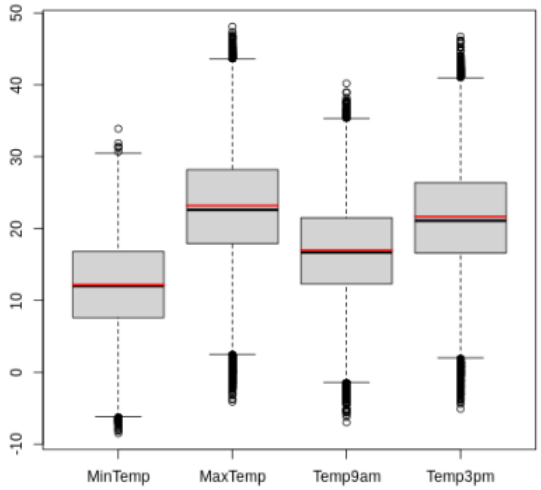

Fig. 1: Boxplot de las variables temperatura.

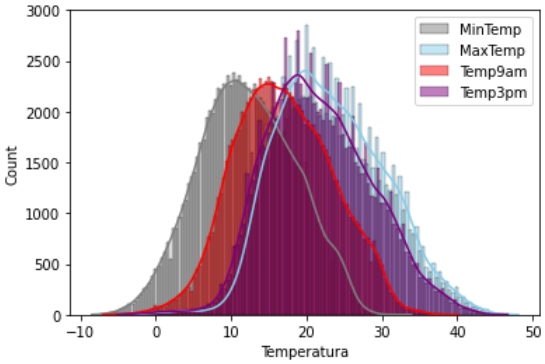

Fig. 2: histograma de las variables temperatura.

2) Viento: Hay tres variables relacionadas con el viento, tanto en los boxplot como en el histograma se puede obervar que la velocidad del viento a las 9 am y 3 pm tienen un comportamient similiar, sus desviaciones estandar son casi iguales, tabla III, para las tres variables su media es muy cercana a su mediana (linea roja en el boxplot) y también se observa que la velocidad de la rafaga de viento mas rapida tiene un comportamiento similar a las otras dos variables pero con valores más altos.

TABLE III: Medidas de centralidad y dispersión de las variables del viento.

variable	Media	Desviación Estandar
WindGustSpeed	39.9	13.58
WindSpeed9am	14	8.9
WindSpeed3pm	18.65	8.8

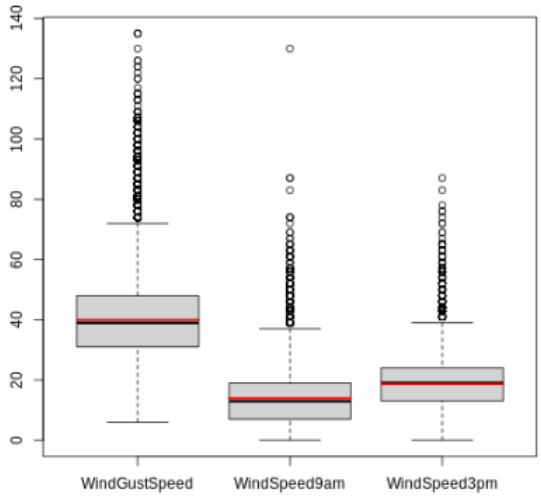

Fig. 3: Boxplot de las variables viento.

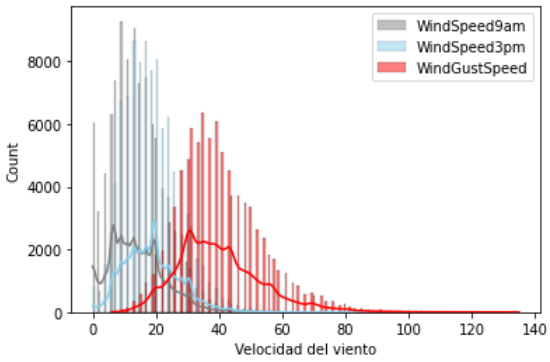

Fig. 4: Histograma de las variables de viento.

3) Presión atmosférica: Hay dos variables asociadas a la presión atmosférica y ambas son similares, tienen la misma distribución como se osbserva en el histograma, los boxplots son parecidos sus medidas de centralidad y dispersion son muy cercanos, tabla IV, sus medias y sus medianas son practicamente iguales para cada variable.

TABLE IV: Medidas de centralidad y dispersión de las variables de la Presión atmosférca.

variable	Media	Desviación Estandar
Pressure9am	1017.68	7.11
Pressure3pm	1015.28	7

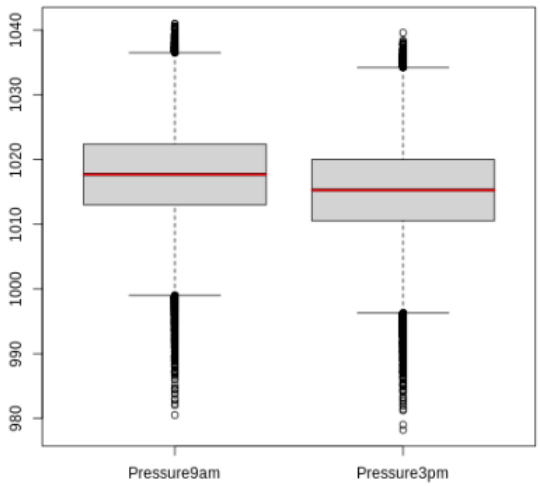

Fig. 5: Boxplot de las variables presión atmosférica.

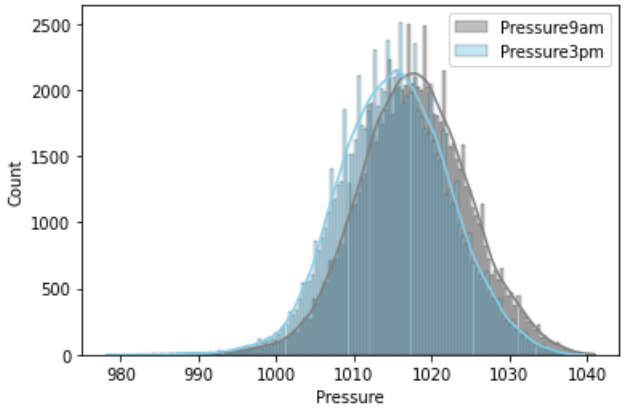

Fig. 6: histograma de las variables presión atmosférica.

4) Humedad: Hay dos variables asociadas a la huemdad y tienen un comportamiento similiar aunque no tanto como las variables anteriores, aunque en el histograma se oberva que hay valores que son más frecuentes que otros en el boxplot para la humedad a las 3 pm no se observan outlaiers y en la tabla V se observa que las medias si tienen cierta diferencia entre las variables y como en las anteriores varibales sus medias y sus medianas son cercanas para cada variable.

TABLE V: Medidas de centralidad y dispersión de las variables de la humedad.

variable	Media	Desviación Estandar
Pressure9am	68.8	19.1
Pressure3pm	51.43	20.77

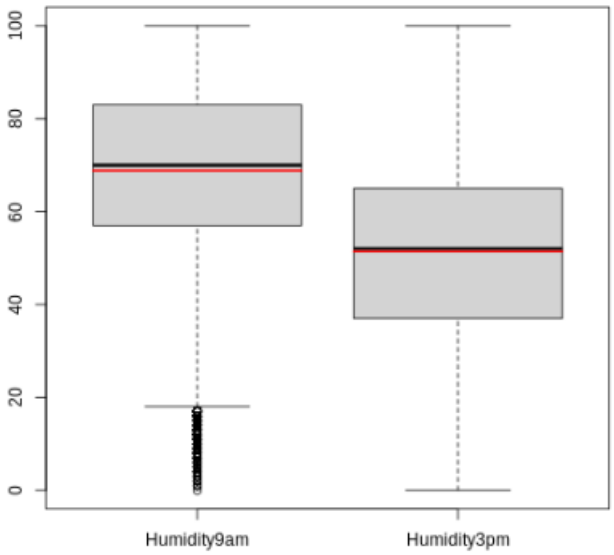

Fig. 7: Boxplot de las variables humedad.

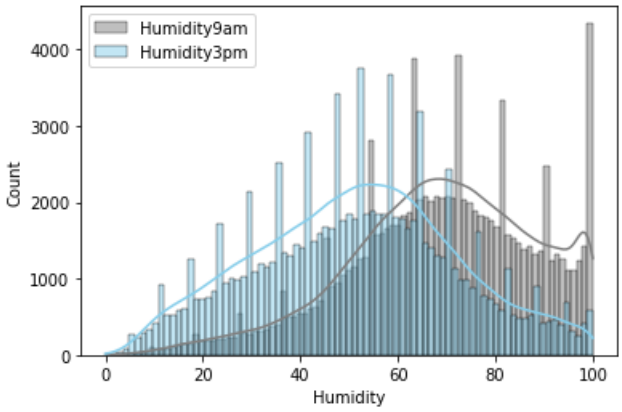

Fig. 8: histograma de las variables humedad.

5) Nubosidad oscuca: la nubosidad oscura se mide con valores del 0 a al 9 y estan medidas a las 9 am y a las 3 pm sus medidas de centralidad y dispersion, tabla VI reflejan que ocurre lo mismo que las otras variables, sus medias y desviaciones estandar son cercanas pero en los bloxplots se observa que las medias son diferentes con sus respectivas medianas y no presentan outliers.

TABLE VI: Medidas de centralidad y dispersión de las variables de la nubosida oscura.

variable	Media	Desviación Estandar
Cloud9am	4.44	2.88
Cloud3pm	4.51	2.71

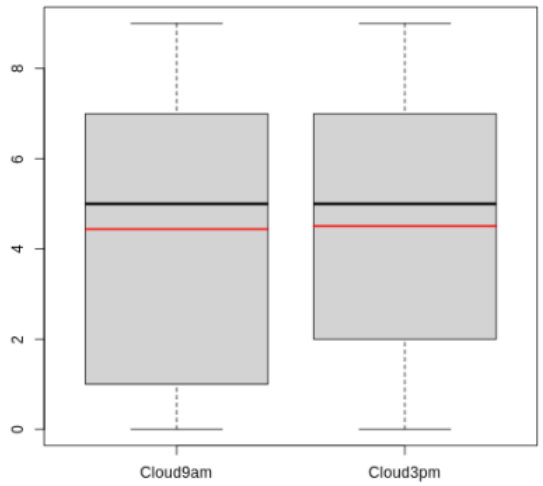

Fig. 9: Boxplot de las variables nubosidad oscura.

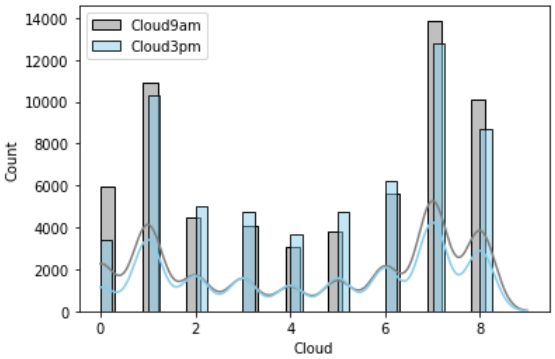

Fig. 10: Histograma de las variables nubosidad oscura.

6) Sunshine: La variable Sunshine tiene dos valores frecuentes incluyendo el cero como se observa en su histograma, su media es 7.6 con desviacón estandar de 3.8 para la variable su media si es diferente a su media (linea roja y negra en el boxplot). También se puede observar que no hay outlaiers.

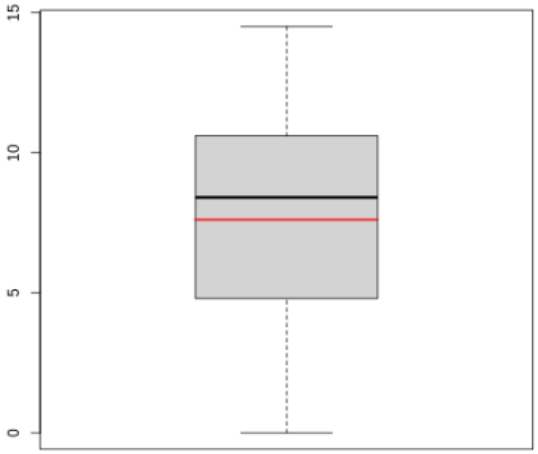

Fig. 11: Boxplot de la variable Shunshine.

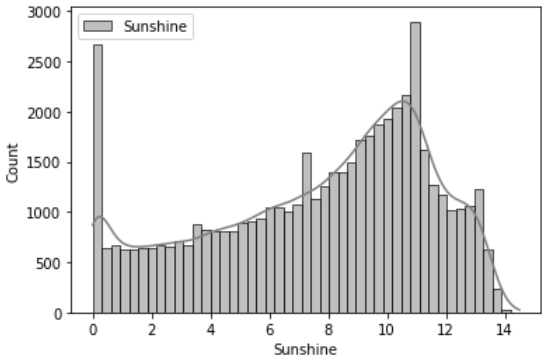

Fig. 12: histograma de la variable Shunshine.

7) Rainfall: la variable de presipitación presenta una media de 2.35 y una desviación estandar de 8.4, en el boxplot se puede observar muchos valores atipicos tantos que no se puede apreciar la caja y esto se puede explicar con el histograma donde se obserba que hay la mayoria de valores es cercano a cero o cero, esto queire decir que casi no llueve en austraia y cuando llueve se registran valores altos.

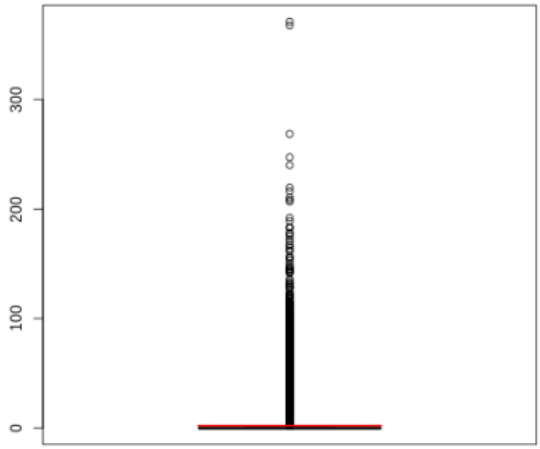

Fig. 13: Boxplot de la variable Rainfall.

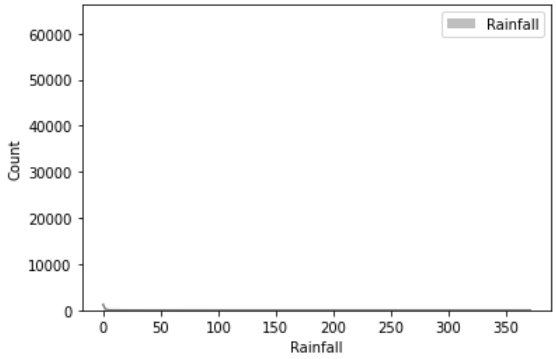

Fig. 14: histograma de la variable Rainfall.

8) Evaporation: la variable evaporación tiene una media de 5.46 y una desviación estandar de 4.16 y observando el boxplot y el histograma ocurre lo mismo que con la variable

presipitación pero menos critica y tiene sentido que cuando llueve es cuando ocurre más evaporación.

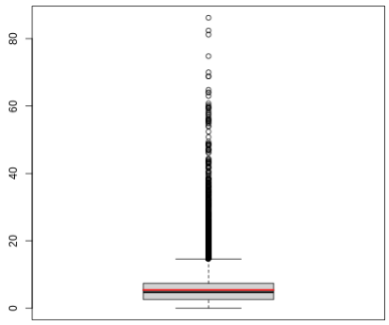

Fig. 15: Boxplot de la variable Evaporation.

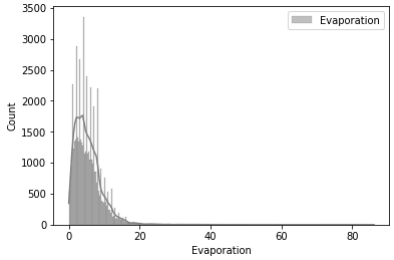

Fig. 16: histograma de la variable Evaporation.

B. *Variables categoricas*

La exploración de datos para las variables categoricas fue observar las frecuencias de sus valores, se decidio juntar en la misma grafica las variables de las direcciones del viento ya que tienen los mismos valores, 16 puntos cardinales y para la ubicación 46 valores distintos, en todaslas variables se observa que las frecuencias son cercanas excepto para tres ubicacinoes que hay una menor cantidad de registros con esa locación.

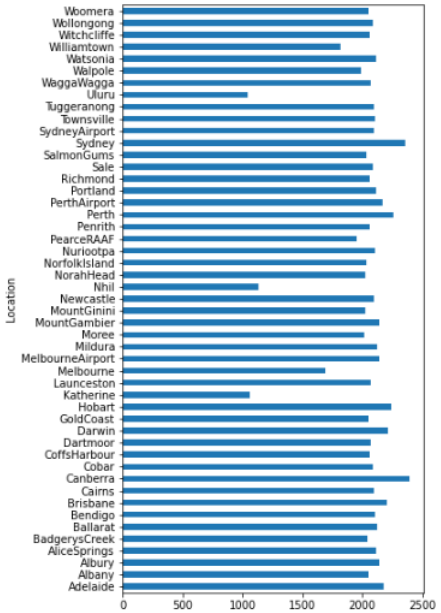

Fig. 17: Frecuencias de los valores de Location.

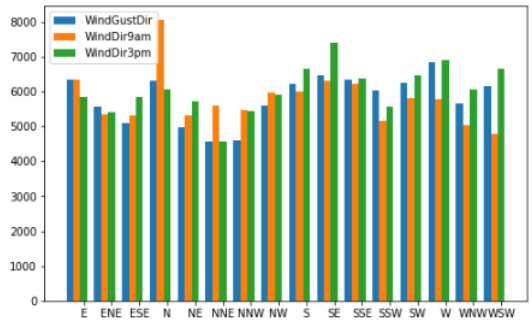

Fig. 18: Frecuencias de los valores de las variables de la dirección del viento.

C. *Correlación*

Se calculo la correlación de Pearson y se observo una alta correlación, mayor al 0.90 entre las variables de maxima temperatura y la temperatura a las 3 pm y entre las varibales de presión medidas a las a las 9 am y 3 pm, figura **xx**, resultado que sugeria los histogramas y boxplots de dichas variables.

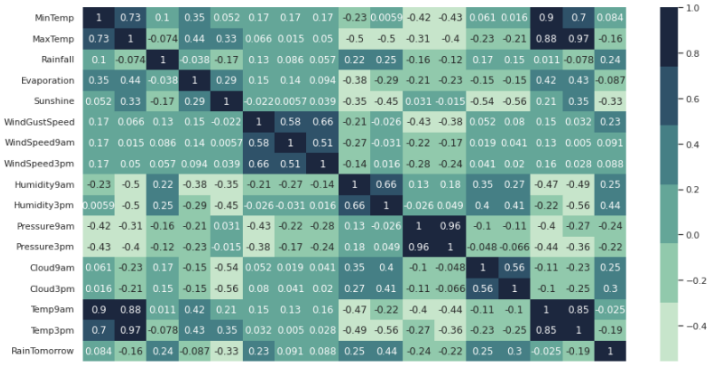

Fig. 19: Mapa de calor de la matriz de correlación.

IV. Preprocesamiento

Al empezar con el preprocesamiento de datos para conseguir información de un conjunto de datos se deben pasar por varias fases las cuales serán nombradas y desarrolladas a continuación.

A. *Limpieza de datos*

Teniendo en cuenta que son 99516 registros en total en el data set se debe conocer cuantos registros estan faltando por atributo lo que nos da la tabla **??**.

Se observa que en todos los atributos, menos 3, se encuentra en alguna proporción una cantidad de datos faltantes. Teniendo en cuenta la gran cantidad de datos faltantes en los atributos Evaporation, Sunshine, Cloud9am y Cloud3pm se proceden a eliminar estas, porque al hacer imputación en los registros de estas variables se estaría sesgando el modelo de una u otra forma.

Teniendo el nuevo data set se procede a realizar imputación según la media de cada atributo, en el caso de los atributos numericos, mientras que para los datos categoricos se hara imputación por el valor con más frecuencia.

TABLE VII: Cantidad de datos faltantes por atributo.

Variable	Cantidad de datos faltantes
Row ID	0
Location	0
MinTemp	443
MaxTemp	230
Rainfall	979
Evaporation	42531
Sunshine	47317
WindGusDir	6521
WindGuSpeed	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

B. Eliminación de outliers

Una vez imputados los valores, para reducir los valores faltantes en los registros, se procede a mirar los outliers que se presentan haciendo uso de la función boxplot, con esto se puede observar que todas las variables númericas menos Humidity3pm cuenta con outliers. Para descartar estos outliers del conjunto de datos se aplica Z-score a cada atributo y descartando las instancias que tienen atributos anormales altos o bajos, es decir si tiene un valor de Z ¿ 3 o Z ¡= -3. Tras aplicar la eliminación de outliers el data set quedaría con un total de 94698 registros comparados a los 99516 del conjunto de datos original.

C. Normalización

Para las secciones que van a ser desarrolladas a continuación como parte de la sección de preprocesamiento, esto con el objetivo de permitir que el conjunto de datos tenga una propiedad particular, ya que cuando se mezclan variables es necesario evitar tener una variable con una escala diferente y valores grandes dominantes en el resultado de los distintos cálculos en el preprocesamiento y en métodos que se aplicarán al conjunto de datos de manera posteriori. Ejemplo de ello se observo en la sección anterior donde se aplico la normalización z-score para eliminar outliers del conjunto de datos y de esta manera cada variable quedé con una escala propia para lograr hacer comparaciones despúes. La normalización z-score esta dada por la siguiente fórmula:

$$V' = \frac{v - \overline{v}}{\sqrt{\sigma^2}}, dondé \;\; \sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

D. Discretización

La tarea consiste en repartir una variable numérica y sus datos a una variable categorica esto repartiendo los datos en un número X de bins y dependiendo de si se quiere aplicar el parametro de que cada bin quede con igual frecuencia u otro tipo de parametro, tras repartir los bins cada intervalo quedará representado por una variable ordinal. Para este caso se escogio el nuemro de bins igual a 10 para cada variable y la estrategia de igual frecuencia. Los puntos de corte de los bins estan en la tabla 3.

TABLE VIII: puntos de corte de los bins para cada variable.

Variable	Cantidad de datos faltantes
MinTemp	[-8.5, 4,6.6, 8.6, 10.3, 12, 13.7, 15.7, 18, 20.8, 33.9]
MaxTemp	[-4.1, 14.5, 16.9, 18.9, 20.7, 22.7, 24.7, 26.9, 29.6, 32.9, 48.1]
WindGuSpeed	[6, 24, 30, 33, 35, 39, 41, 44, 50, 57, 135,]
WindSpeed9am	[0, 4, 6, 9, 11, 13, 15, 19, 20, 26, 130]
WindSpeed3pm	[0, 9, 11, 13, 15, 18.65, 20, 22, 26, 30, 87]
Humidity9am	[0, 44, 54, 60, 65, 70, 75, 80, 86, 94, 100]
Humidity3pm	[0, 23, 33, 41, 47, 51.43, 57, 62, 69, 79, 100]
Pressure9am	[980.5, 1009.2, 1012.4, 1014.7, 1016.7, 1017.68, 1018.7, 1020.7, 1023.1, 1026.31]
Pressure3pm	[978.2, 1006.8, 1009.8, 1012.2, 1014.2, 1015.28, 1016.2, 1018.3, 1020.7, 1023.9, 1039.6]
Temp9am	[-7, 8.9, 11.3, 13.2, 15, 16.7, 18.4, 20.4, 22.7, 25.8, 40.2]
Temp3pm	[-5.1, 13.3, 15.7, 17.6, 19.4, 21.3, 22.9, 25.1, 27.7, 31.1, 46.7]

E. Selección de características

La selección de características se realizo por medio del ranking usando la entropia, primero se calcula la entropia del conjunto total de los datos y luego se calcula la entropia elimiando cada una de las variables y se calcula la ganancia para cada varible eliminada y se elimina la varible con menor ganancia y se repite el proceso, el la tabla IX se observa el ranking que desde la primera variable que se elimina hasta la ultima con sus ganancias y la entropia total del conjunto de datos. La función de entropia utilisada hace uso de la función de similitud de hamming por lo tanto para las variables continuas se uso la discretización anterior.

Se puede observar que las varibles que se eliminan con las variables continuas discretizadas, esto posiblemente se debe a que no hay una optima discretización para mantener la información de los datos.

TABLE IX: Puntos de corte de los bins para cada variable.

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

F. Reducción de la dimensionalidad

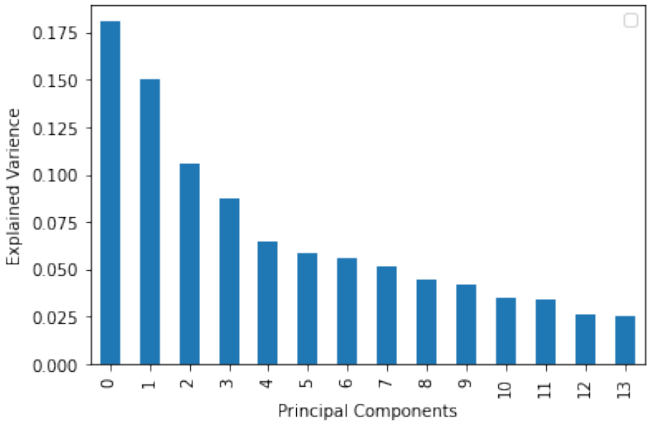

Fig. 20: Varianza explicada por componentes tras aplicar PCA.

Esta tarea consiste en reducir el data set original creando nuevos n atributos dependiendo de como se quiera interpretar el conjunto de datos, para la siguiente gráfica se aplica un ánalis PCA teniendo como base dos componentes y el objetivo que se tiene para clasificar el conjunto de datos. De igual forma depende sobre que atributos se quiera aplicar la reducción de la dimensionalidad, en el caso del conjunto de datos que se esta estudiando, se decidió hacer uso de todos los atributos, por lo que fue necesario convertir las variables categóricas a numéricas haciendo uso de la función get dummies, perteneciente a la biblioteca pandas, tras tener el conjunto de datos convertido a variables numéricas todos sus atributos es necesario estandarizar la información y de manera posterior se crea el objeto PCA haciendo uso de dos componentes para obtener la figura 20 donde se observa la varianza explicada por cada uno de los componentes seleccionados y se logra observar que cuando se llega a un número de componentes entre 4 y 6 esta varianza empieza a converger a un valor cercano a 0.062 por lo cual al hacer una selección de 4 componentes principales se tendría un número optimo de componentes para hacer uso en futuros métodos a aplicar dentro del conjunto de datos.

V. Asociación

A partir de este método se pueden descubrir reglas de asociación las cuales descubren hechos que ocurren en común dentro de un determinado conjunto de datos. Para encontrar estas reglas se encuentran distintos algoritmos, para el conjunto de datos se usaron los siguientes algoritmos

A. Principio Apriori

Este algoritmo es usado comunmente sobre bases de datos transaccionales dondé permite encontrar de manera eficiente conjuntos de items frecuentes que sirven de base para generar reglas de asociación. Manteniendo una confianza de 1 se logran obtener 4 reglas en las que se mantiene como atributo principal 'RainFall' con un soporte empezando desde 0.4 y un lift de 1.291 en las 4 reglas. Mientras se baja la confianza más reglas van apareciendo apareciendo nuevos atributos principales como la presión y la velocidad del viento, de igual forma se observa la aparición de la clase dentro de las reglas, donde estas tienen un soporte promedio de 0.5 y un lift mayor a 1 que nos indica que el conjunto aparece una cantidad de veces superior a lo esperado.

B. FP Growth

Es un algoritmo que se deriva del principio anterior y este se caracteriza por ser muy eficiente y escalable lo cual nos permite un gran procesamiento de datos en un tiempo relativamente bajo. Este algoritmo permite encontrar conjuntos de patrones frecuentes sin generar candidatos. A partir de esta definición y manteniendo un soporte del 0.4 junto a una confianza minima de 0.5 se llegaron a obtener 14 reglas donde se manejan atributos como la clase 'RainTomorrow' así como los atributos presión

VI. Agrupación

Al empezar con el preprocesamiento de datos para conseguir información de un conjunto de datos se deben pasar por varias fases las cuales serán nombradas y desarrolladas a continuación.

A. KMeans

El algoritmo de KMeans es un metodo de agrupamiento particional donde se define un numero de grupos y busca particionar los datos en es cantidad de grupos a partir de su distancia al centroide o punto medio del grupo. A partir de la exploración de los datos se decide eliminar los registros con algun dato perdido, las columnas que tienen la mayor cantidad de valores nulos (Evaporation, Sunshine, Cloud9am, Cloud3pm), la columna de identificación por no aportar información y la clumna de interes, posteriormente se normalizaron los atributos continuos para usar la distancia euclidiana y se codificaron los datos categoricos a datos numericos con una binarización, a este conjunto de datos transformado se aplica el algotirmo de KMeans con diferentes numero de grupos y se valida con el indice de Davies-Bouldin (figura 21), el cual indica que un numero de grupo igual a 2 los datos en los grupos son más compactos.

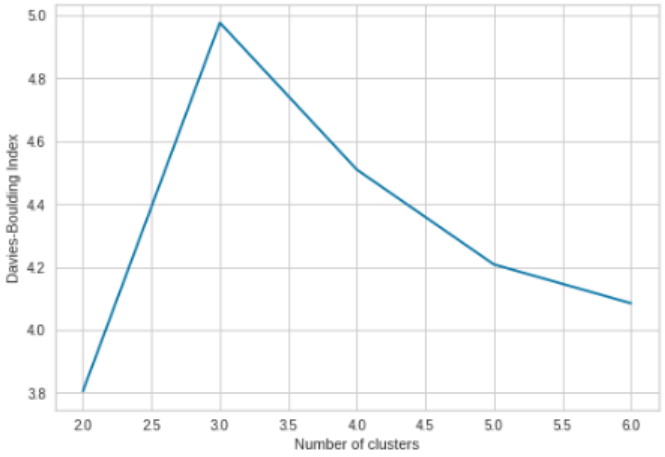

Fig. 21: Indice Davies-Bouldin para diferentes k.

B. KNearestNeighbors

El algoritmo de los k vecinos más cercanos es un metodo supervisado, el conjunto de entrada utilizado es el mismo tratado en la sección de KMeans ya que para KNearest-Neighbors se utiliza tambien la distancia euclidiana como distancia para el agrupamiento, se evaluaron 4 modelos con el parametro de n vecinos igual a 5, 6, 7 y 8, el tabla X se puede observar que las medidas de precision, recall y f1 no son diferentes significativamente entre los cuatro modelos pero si es relativamente baja alrrededor de 0.7.

TABLE X: Evaluación modelos KNearestNeighbors.

Modelo	**Precision**	**Recall**	**F1**
n = 5	0.73	0.63	0.63
n = 6	0.76	0.60	0.62
n = 7	0.77	0.61	0.64
n = 8	0.78	0.60	0.62

C. Agrupación jerárquica

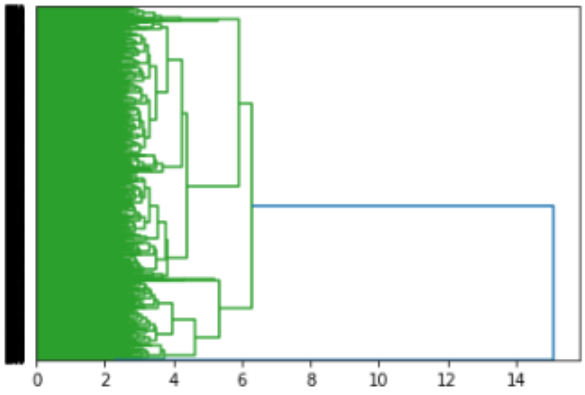

Fig. 22: Dendograma de agrupación jerarquica.

Se utilizo un algoritmo jerárquico alglomerativo con el promedio como medida de proximidad intercluster y como medida de distancia se uso Gower, por lo tanto los atributos continuos se normalizaron y los categóricos se dejaron igual.Hubo necesidad de hacer un muestreo del conjunto de aproximadamente 4000 datos manteniendo la proporción de la clase (RainTomorrow) debido a que la RAM del ordenador no era suficiente. En el dendograma generado por el algoritmo (figura 22) se puede observar dos grupos uno con mucha más cantidad de puntos que el otro.