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: aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa asdasd kagdskdas asdas

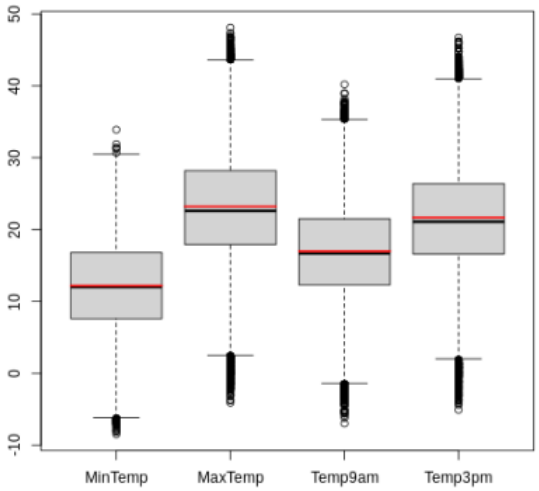

Fig. 1: Boxplot de las variables temperatura.

2) Viento: aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa asdasd kagdskdas asdas

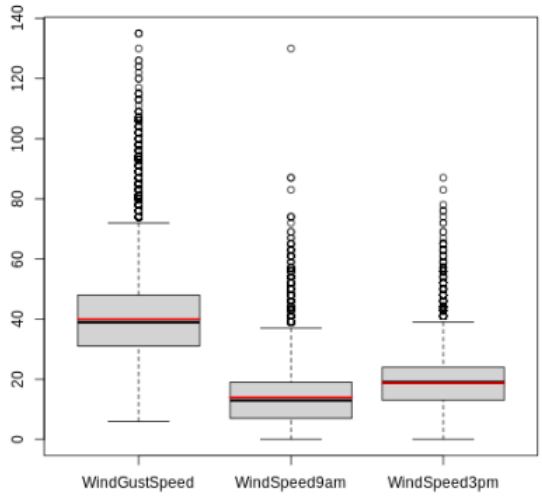

Fig. 2: Boxplot de las variables viento.

3) presión atmosférica: aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa asdasd kagdskdas asdas

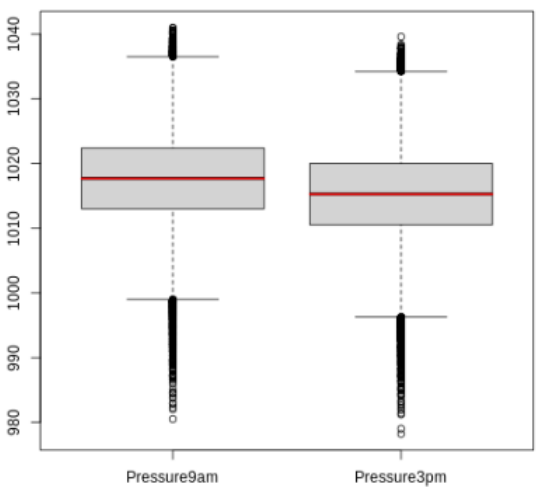

Fig. 3: Boxplot de las variables presión atmosférica.

4) Humedad: aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa asdasd kagdskdas asdas

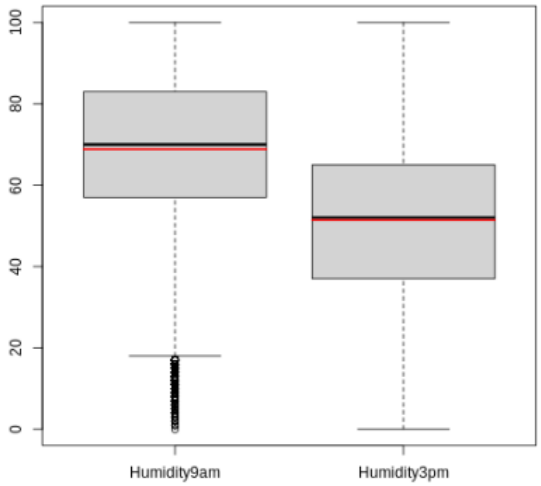

Fig. 4: Boxplot de las variables humedad.

5) Sunshine: aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa asdasd kagdskdas asdas

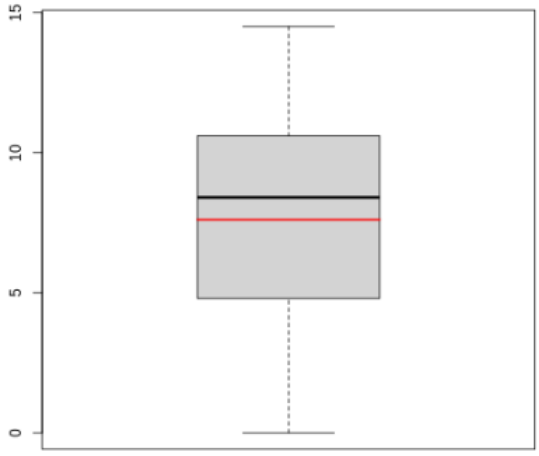

Fig. 5: Boxplot de la variable Shunshine.

6) Rainfall: aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa asdasd kagdskdas asdas

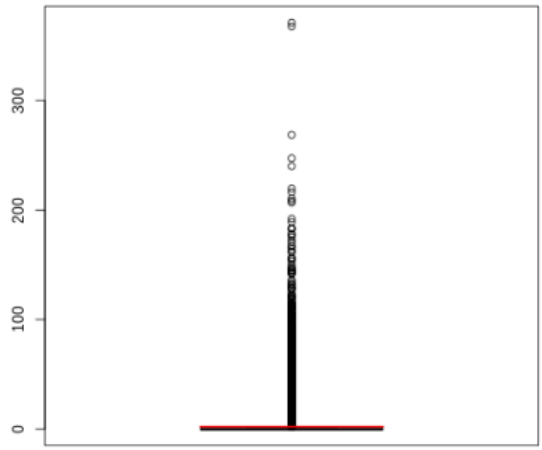

Fig. 6: Boxplot de la variable Rainfall.

7) Evaporation: aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa asdasd kagdskdas asdas

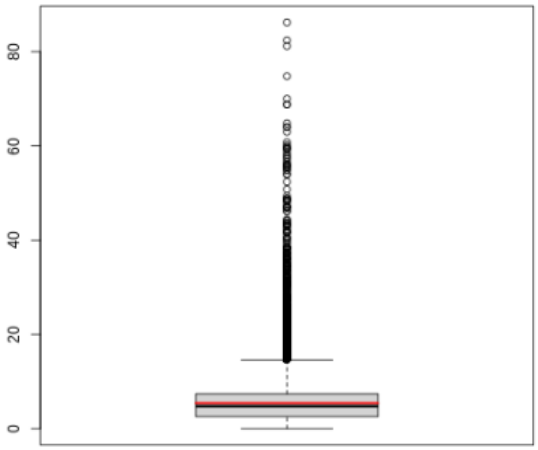

Fig. 7: Boxplot de la variable Evaporation.

B. Variables categoricas

1) Location:

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 II.

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

TABLE II: 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

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.

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:

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

D. Muestreo

Esta tarea se hace con el fin de seleccionar un subconjunto de datos del conjunto de datos a analizar, esto buscando reducir costos y tiempo en el procesamiento del conjunto de datos, lo ideal es que la muestra sea representativa es decir aplicar un muestreo estratificado donde los ejemplos de la muestra y su proporción entre clases sea una copia de la proporción entre clases del conjunto de datos original. Para el ejercicio fue necesario hacer uso de muestreo en dos casos para que las funciones de binarización y selección de características no tenga que explorar todo el conjunto de datos para dar un valor mas certero dado el costo de tiempo.

E. 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.

F. Binarización

G. Selección de características

H. Reducción de la dimensionalidad

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 siguiente gráfica:

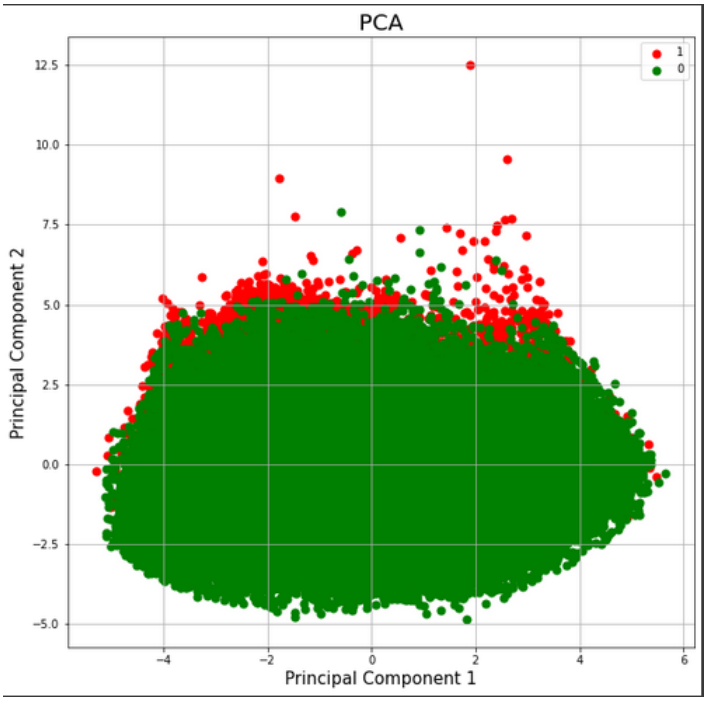

Fig. 8: Grafica por componentes tras aplicar PCA.

V. Conclusión