Facultad de Ingeniería
Ingeniería Civil Informática

Predicción de Lluvia en Australia

Angelo Cancino, Felipe Ruiz
John Rios Griego, Mayo 2024

Facultad de Ingeniería
Ingeniería Civil Informática

Introducción

Australia es uno de los países más extensos del mundo y presenta una gran diversidad climática y geográfica. A pesar de ser el continente más seco y llano del planeta, este cuenta con regiones de clima tropical, desértico, semiárido y templado-continental. Las precipitaciones varían considerablemente en términos de distribución, esto debido a que estas precipitaciones son escasas en el interior y más abundantes en las zonas costeras.

Para el siguiente estudio se propone predecir los patrones de lluvia en Australia, y es para esto que se dispone de un dataset que recoge observaciones meteorológicas diarias de diversas ubicaciones en Australia. Estos datos han sido obtenidos de la Oficina de Meteorología de la Commonwealth de Australia.

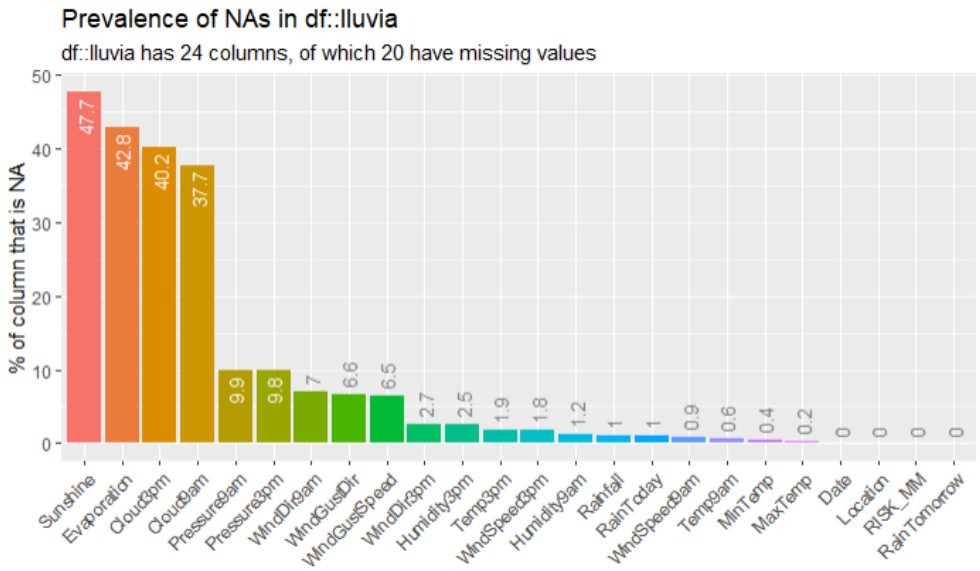
Modelo de Regresión

Para la predicción de cuánta lluvia cae en milímetros en Australia se decidió realizar un modelo de regresión lineal múltiple.
La elección de este modelo se basa en la naturaleza del conjunto de datos y el problema a resolver. En nuestro caso estamos intentando predecir la cantidad de lluvia en milímetros, lo cual es una variable contínua, esto a partir de variables independientes como lo son MinTemp, MaxTemp, Sol, Humedad, etc.

Primero se realizó la limpieza del dataset, para lo cual comenzamos buscando si existen datos NA dentro de las columnas.

```r
```{r}
#ver la cantidad de na
porcentaje=inspect_na(lluvia)
show_plot(porcentaje)

```
```



Luego se cambiaron los valores de tipo char a tipo numérico para poder ser procesadas por la red neuronal.

```r
#transformar una variable cuantitativa a numerica

lluvia$RainToday<- ifelse(lluvia$RainToday   == "Yes", 1, 0)
lluvia$RainTomorrow<- ifelse(lluvia$RainTomorrow   == "Yes", 1, 0)

lluvia$WindGustDir  <- ifelse(lluvia$WindGustDir== "W", 1, ifelse(lluvia$WindGustDir == "SE", 2,  ifelse(lluvia$WindGustDir == "E", 3,
ifelse(lluvia$WindGustDir == "N",4,ifelse(lluvia$WindGustDir=="SSE",5,ifelse(lluvia$WindGustDir=="S",6,ifelse(lluvia$WindGustDir=="WSW",7,ifelse(ll
uvia$WindGustDir=="SW",8,ifelse(lluvia$WindGustDir=="SSW",9,ifelse(lluvia$WindGustDir=="WNW",10,ifelse(lluvia$WindGustDir=="NW",11,ifelse(lluvia$Wi
ndGustDir=="ENE",12,ifelse(lluvia$WindGustDir == "ESE",13,ifelse(lluvia$WindGustDir=="NE", 14,ifelse(lluvia$WindGustDir=="NNW",15,16)))))))))))))))

lluvia$WindDir9am  <- ifelse(lluvia$WindDir9am== "W", 1, ifelse(lluvia$WindDir9am == "SE", 2,  ifelse(lluvia$WindDir9am == "E", 3,
ifelse(lluvia$WindDir9am == "N",4,ifelse(lluvia$WindDir9am=="SSE",5,ifelse(lluvia$WindDir9am=="S",6,ifelse(lluvia$WindDir9am=="WSW",7,ifelse(lluvia
$WindDir9am=="SW",8,ifelse(lluvia$WindDir9am=="SSW",9,ifelse(lluvia$WindDir9am=="WNW",10,ifelse(lluvia$WindDir9am=="NW",11,ifelse(lluvia$WindDir9am
=="ENE",12,ifelse(lluvia$WindDir9am == "ESE",13,ifelse(lluvia$WindDir9am=="NE", 14,ifelse(lluvia$WindDir9am=="NNW",15,16)))))))))))))))

lluvia$WindDir3pm  <- ifelse(lluvia$WindDir3pm== "W", 1, ifelse(lluvia$WindDir3pm == "SE", 2,  ifelse(lluvia$WindDir3pm == "E", 3,
ifelse(lluvia$WindDir3pm == "N",4,ifelse(lluvia$WindDir3pm=="SSE",5,ifelse(lluvia$WindDir3pm=="S",6,ifelse(lluvia$WindDir3pm=="WSW",7,ifelse(lluvia
$WindDir3pm=="SW",8,ifelse(lluvia$WindDir3pm=="SSW",9,ifelse(lluvia$WindDir3pm=="WNW",10,ifelse(lluvia$WindDir3pm=="NW",11,ifelse(lluvia$WindDir3pm
=="ENE",12,ifelse(lluvia$WindDir3pm == "ESE",13,ifelse(lluvia$WindDir3pm=="NE",14,ifelse(lluvia$WindDir3pm=="NNW",15,16)))))))))))))))
```

```r
#lluvia$WindGustDir[is.na(lluvia$WindGustDir)] <- 0

summary(lluvia)
head(lluvia)
```

A tibble: 6 × 24

Date <date>	Location <chr>	MinTemp <dbl>	MaxTemp <dbl>	Rainfall <dbl>	Evaporation <dbl>	Sunshine <dbl>	WindGustDir <dbl>	WindGustSpeed <dbl>	WindDir9am <dbl>
2008-12-01	Albury	13.4	22.9	0.6	NA	NA	1	44	1
2008-12-02	Albury	7.4	25.1	0.0	NA	NA	10	44	15
2008-12-03	Albury	12.9	25.7	0.0	NA	NA	7	46	1
2008-12-04	Albury	9.2	28.0	0.0	NA	NA	14	24	2
2008-12-05	Albury	17.5	32.3	1.0	NA	NA	1	41	12
2008-12-06	Albury	14.6	29.7	0.2	NA	NA	10	56	1

6 rows | 1-10 of 24 columns

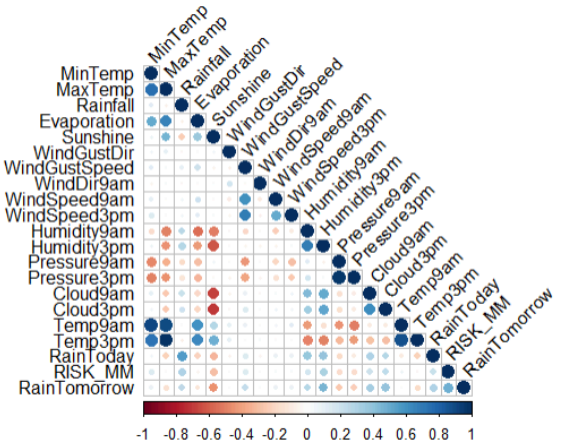
y para concluir con el proceso de limpieza se realiza un .omit a los valores NA.

```r
#data sin na

lluvia_sin_na<-na.omit(lluvia)
head(lluvia_sin_na)

summary(lluvia_sin_na)
```

```
      Date              Location            MinTemp          MaxTemp         Rainfall        Evaporation        Sunshine       WindGustDir
 Min.   :2007-11-01   Length:56420       Min.   :-6.70   Min.   : 4.10   Min.   :  0.00   Min.   : 0.000   Min.   : 0.000   Min.   : 1.000
 1st Qu.:2010-07-19   Class :character   1st Qu.: 8.60   1st Qu.:18.70   1st Qu.:  0.00   1st Qu.: 2.800   1st Qu.: 5.000   1st Qu.: 4.000
 Median :2012-07-28   Mode  :character   Median :13.20   Median :23.90   Median :  0.00   Median : 5.000   Median : 8.600   Median : 8.000
 Mean   :2012-09-17                      Mean   :13.46   Mean   :24.22   Mean   :  2.13   Mean   : 5.503   Mean   : 7.736   Mean   : 7.868
 3rd Qu.:2014-10-10                      3rd Qu.:18.40   3rd Qu.:29.70   3rd Qu.:  0.60   3rd Qu.: 7.400   3rd Qu.:10.700   3rd Qu.:12.000
 Max.   :2017-06-25                      Max.   :31.40   Max.   :48.10   Max.   :206.20   Max.   :81.200   Max.   :14.500   Max.   :16.000
 WindGustSpeed     WindDir9am      WindDir3pm        WindSpeed9am    WindSpeed3pm    Humidity9am      Humidity3pm      Pressure9am
 Min.   :  9.00   Min.   : 1.000   Length:56420       Min.   : 2.00   Min.   : 2.00   Min.   :  0.00   Min.   :  0.0   Min.   : 980.5
 1st Qu.: 31.00   1st Qu.: 4.000   Class :character   1st Qu.: 9.00   1st Qu.:13.00   1st Qu.: 55.00   1st Qu.: 35.0   1st Qu.:1012.7
 Median : 39.00   Median : 8.000   Mode  :character   Median :15.00   Median :19.00   Median : 67.00   Median : 50.0   Median :1017.2
 Mean   : 40.88   Mean   : 8.083                      Mean   :15.67   Mean   :19.79   Mean   : 65.87   Mean   : 49.6   Mean   :1017.2
 3rd Qu.: 48.00   3rd Qu.:12.000                      3rd Qu.:20.00   3rd Qu.:26.00   3rd Qu.: 79.00   3rd Qu.: 63.0   3rd Qu.:1021.8
 Max.   :124.00   Max.   :16.000                      Max.   :67.00   Max.   :76.00   Max.   :100.00   Max.   :100.0   Max.   :1040.4
  Pressure3pm       Cloud9am        Cloud3pm        Temp9am         Temp3pm        RainToday         RISK_MM          RainTomorrow
 Min.   : 977.1   Min.   :0.000   Min.   :0.000   Min.   :-0.7   Min.   : 3.70   Min.   :0.0000   Min.   :  0.000   Min.   :0.0000
 1st Qu.:1010.1   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:13.1   1st Qu.:17.40   1st Qu.:0.0000   1st Qu.:  0.000   1st Qu.:0.0000
 Median :1014.7   Median :5.000   Median :5.000   Median :17.8   Median :22.40   Median :0.0000   Median :  0.000   Median :0.0000
 Mean   :1014.8   Mean   :4.242   Mean   :4.327   Mean   :18.2   Mean   :22.71   Mean   :0.2209   Mean   :  2.347   Mean   :0.2203
 3rd Qu.:1019.4   3rd Qu.:7.000   3rd Qu.:7.000   3rd Qu.:23.3   3rd Qu.:27.90   3rd Qu.:0.0000   3rd Qu.:  0.600   3rd Qu.:0.0000
 Max.   :1038.9   Max.   :8.000   Max.   :9.000   Max.   :39.4   Max.   :46.10   Max.   :1.0000   Max.   :367.600   Max.   :1.0000
```

Luego se realizó una matriz de correlación para así poder ver cuáles variables eran

```r
#Gráfico de correlación

corrplot(mcor, method = "circle", type = "lower", tl.col = "black", tl.srt = 45)
```



A continuación se muestran los 3 modelos realizados.

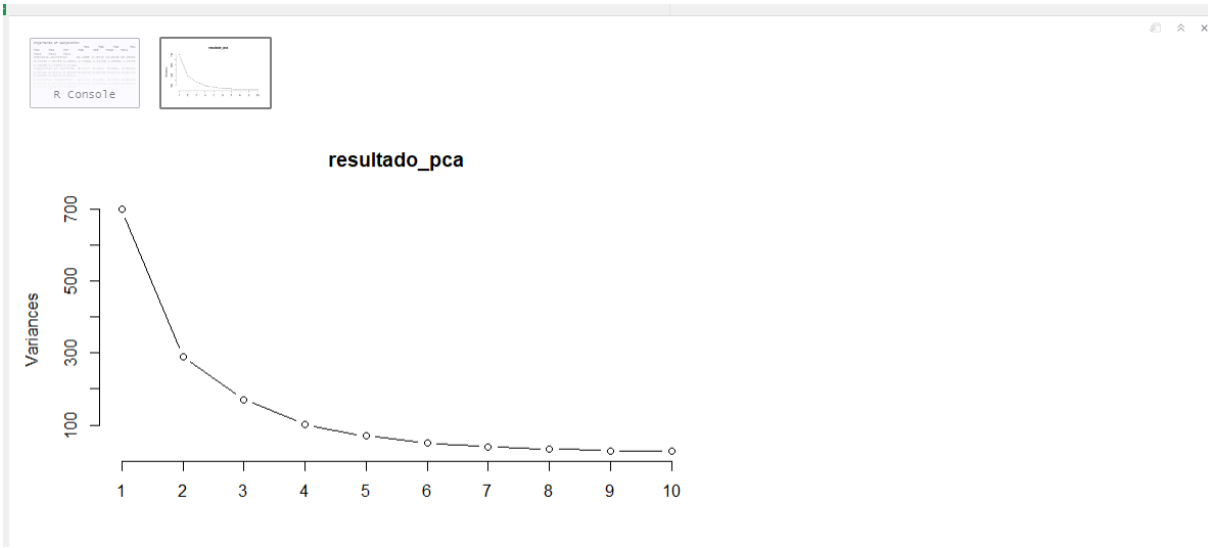
```r
#regresion lineal multiple

modelolm <- lm(RISK_MM ~ MinTemp + MaxTemp+ Rainfall + Evaporation + Sunshine + WindGustDir +WindDir9am + WindDir3pm +WindSpeed9am +
               WindSpeed3pm +Humidity9am + Humidity3pm + Pressure9am +Pressure3pm + Cloud9am +
               Cloud3pm + Temp9am + Temp3pm + RainToday + RainTomorrow -1, data = lluvia_sin_na)

summary(modelolm)
```

```r
#regresion lineal multiple 2 limpio

modelolm2 <- lm(RISK_MM ~  MaxTemp+ Rainfall + Evaporation + Sunshine + WindGustDir +WindSpeed9am +
                WindSpeed3pm + Humidity3pm + Pressure9am + Cloud3pm+RainToday + RainTomorrow -1, data = lluvia_sin_na)

summary(modelolm2)
```

```r
#regresion lineal multiple 3 limpio

modelolm3 <- lm(RISK_MM ~  MaxTemp+ Rainfall + Sunshine +WindSpeed9am +
                WindSpeed3pm + Humidity3pm + Pressure9am + Cloud3pm+RainToday + RainTomorrow - 1, data = lluvia_sin_na)

summary(modelolm3)
```

Procedimiento de Componentes Principales

```r
#Analisis de componentes principales


# Seleccionar solo las variables numéricas
lluvia_principal <- subset(lluvia_sin_na, select = -RainTomorrow)
lluvia_PCA <- lluvia_principal[, sapply(lluvia_principal, is.numeric)]

# Realizar el análisis de componentes principales
resultado_pca <- prcomp(lluvia_PCA, center = TRUE)

#Mostrar el resumen del PCA
summary(resultado_pca)
resultado_pca

plot(resultado_pca, type="l")
```



A continuación elegimos nuestros componentes principales.

```r
#tomamos el PC con una proporcion acumulada hasta el 80%

resultado_pca$rotation[,1:3]

lluvia_PCA$ValPC1 <- apply(resultado_pca$rotation[,1]*t(lluvia_PCA),2,sum)
lluvia_PCA$ValPC2 <- apply(resultado_pca$rotation[,2]*t(lluvia_PCA),2,sum)
lluvia_PCA$ValPC3 <- apply(resultado_pca$rotation[,3]*t(lluvia_PCA),2,sum)
```

Luego hacemos el modelo logístico.

```{r}
#modelo logistico

lluvia_PCA$RainTomorrow <- lluvia_sin_na$RainTomorrow
llueve_o_no <- glm(RainTomorrow ~ ValPC1 + ValPC2,"binomial", lluvia_PCA)
summary(llueve_o_no)

```

```
Call:
glm(formula = RainTomorrow ~ ValPC1 + ValPC2, family = "binomial",
    data = lluvia_PCA)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.064e+00  8.399e-02  -96.01  < 2e-16 ***
ValPC1       4.800e-02  5.644e-04   85.05  < 2e-16 ***
ValPC2      -1.535e-04  4.313e-05   -3.56 0.000371 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 59493  on 56419  degrees of freedom
Residual deviance: 49383  on 56417  degrees of freedom
AIC: 49389

Number of Fisher Scoring iterations: 5
```

Predicciones

Análisis de Componentes Principales.

Primero separamos los datos en un conjunto de entrenamiento que abarca un 80% de los datos y uno de testeo que abarca el 20% restante.

```{r}
#data de logist
set.seed(1234)
indextrain2 = createDataPartition(lluvia_PCA$RainTomorrow,p=0.8,list=F)

dat2.train=lluvia_PCA[indextrain2,]
dat2.test=lluvia_PCA[-indextrain2,]

```

```{r}

pred_log <- predict(llueve_o_no, newdata = dat2.train, type = "response")
plot(pred_log)

```

Luego se realizó la predicción, la cual tiene un Accuracy del 81%, valor bastante alto que indica que el modelo predice de manera correcta si lloverá o no la mayoría de las veces.

```{r}
#Prediccion modelo logistico

predict <- ifelse(pred_log > 0.5, 1, 0)
matriz_confusion <- confusionMatrix(factor(predict), factor(dat2.train$RainTomorrow))

exactitud <- matriz_confusion$overall["Accuracy"]

```

```{r}

exactitud

```

```
  Accuracy
 0.8104396
```

Finalmente se realizó una predicción con los datos que nos entrega el dataset y se obtuvo que hay un 90% de probabilidad de que no llueva “mañana”.

```{r}

Predictmañana <- sapply(predict, function(x) ifelse(x > 0.5, 1, 0))

conteo_predicciones <- table(Predictmañana)

porcentaje_predicciones <- conteo_predicciones / sum(conteo_predicciones) * 100

print(porcentaje_predicciones)


if (porcentaje_predicciones["0"] > porcentaje_predicciones["1"]) {
  print(" Según el modelo, es más probable que mañana no lloverá.")
} else {
  print(" Según el modelo, es más probable que mañana sí lloverá.")
}




```

```
Predictmañana
        0         1
90.860954  9.139046
[1] " Según el modelo, es más probable que mañana no lloverá."
```

La predicción del Modelo de Regresion lineal multiple con el 20% de los datos nos da:

```{r}
#prediccion 20 de los datos de RLM

rmse <- sqrt(mean((prediccionestest - dat.test$RISK_MM)^2))
rmse


```

```
[1] 7.018176
```