# HarvardX PH125.9x

CYO Project

Miguel Angel Bustos Sáez

2022-10-13

## Contents

# 1 The introduction & Data

The following project it´s about a random forest algorithm, a technique that can transform a single tree model with high variance and predictive power into a fairly accurate prediction function.

Random Forest it´s a modification of bagging that builds a large collection of correlated trees and have become a very popular and has a good performance.

The objective of this project is to obtain different outputs and conclusions about predict methods, and at the finish a Random Forest output that predicts the categorical variable month between others but escencialy when the pm2.5 or air quality can be in a quality range, using a Random Forest algorithm.

The data was obtained from archive.ics.uci.edu and it´s the Bijing environmental environment information, has thirteen variables, each of them has a different meaning, that are the folliwing:

- No: Row number
- Year: The year of each row registration
- Month: The month of each row registration
- Day: The day of each row registration
- Hour: The hour of each row registration
- Pm2.5: Pm2.5 concentration (ug/mˆ3)
- Dewp: Dew Point (â„f)
- Temp: Temperature (â„f)
- Pres: Pressure (hPa)
- Cbwd: Combined wind direction
- Iws: Cumulated wind speed (m/s)
- Is: Cumulated hours of snow
- Ir: Cumulated hours of rain

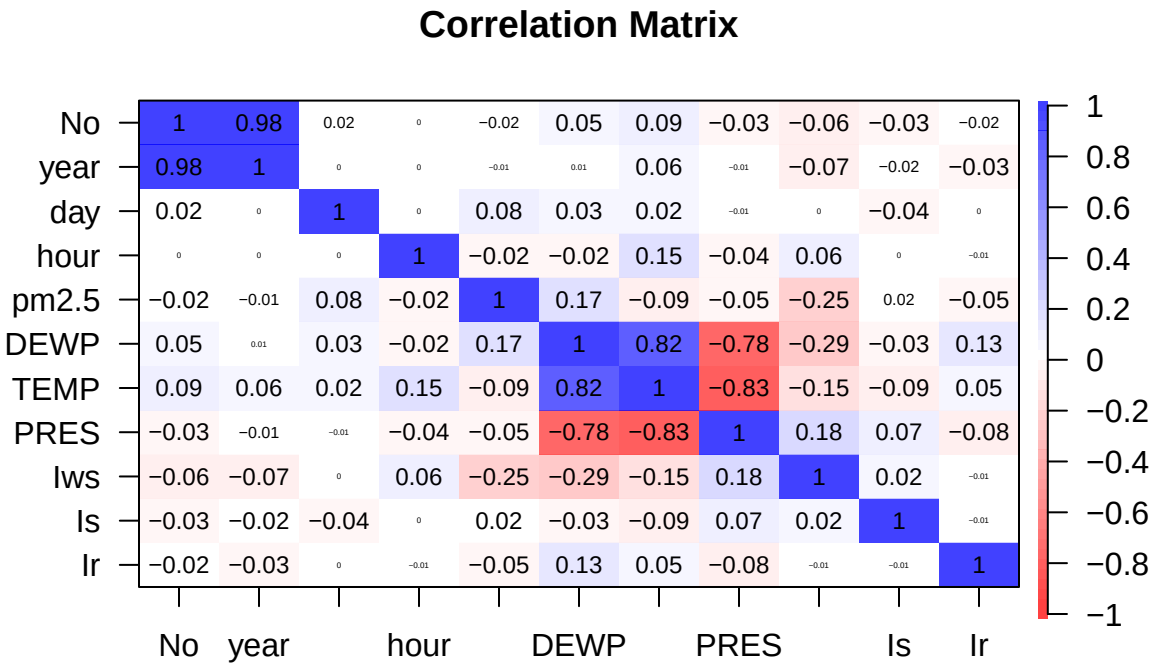
**Data Changes**

It´s recommended to remove NA´s rows, and we reduce 43748 rows to 41681. Now the data has 41681 rows and without NA´s. This summary shows the range but additionaly it´s necessary some comments:

- The cbwd range it´s 43824 (unique(cbwd) shows the cardinal points SE, cv, NW, NE)
- The pm2.5 range it´s 0 to 994 It is an air pollutant that is a concern for people's health when levels in air are high. The 35.4 it´s acceptable but 35.5 it´s unhealthy.
- The month it´s one to twelve, it´s necessary to change it to january to december, changing the class from numerica to character.
- The data has 43748 rows, but some had NA rows, that were removed.
- The month it´s the most critical variable to predict the Temperature in this project.

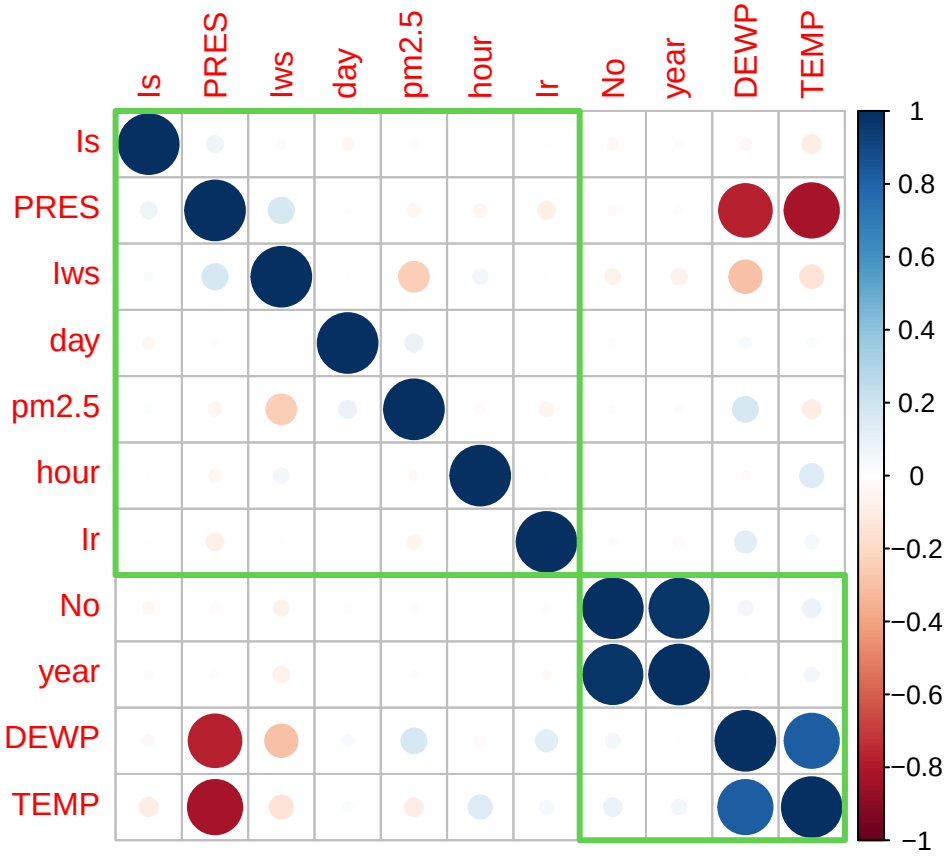
## 2 Correlation Matrix

There seems to be a slight negative correlation between the wind speed Iws and pm2.5 Also a positive or high correlation between the temperature TEMP and the dew point DEWP (Dew point) The correlation:




In this project, it´s necessary see these correlations of all the variables, to have a global vision and the importance of each variable between others.

Here it´s another visualization that confirms that the negative relationship between PRES and DEWP but a slighly negative correlation between the wind speed Iws and pm2.5. Also a positive or high correlation between the temperature TEMP and the dew point DEWP (Dew point)




**Correlation Importance**

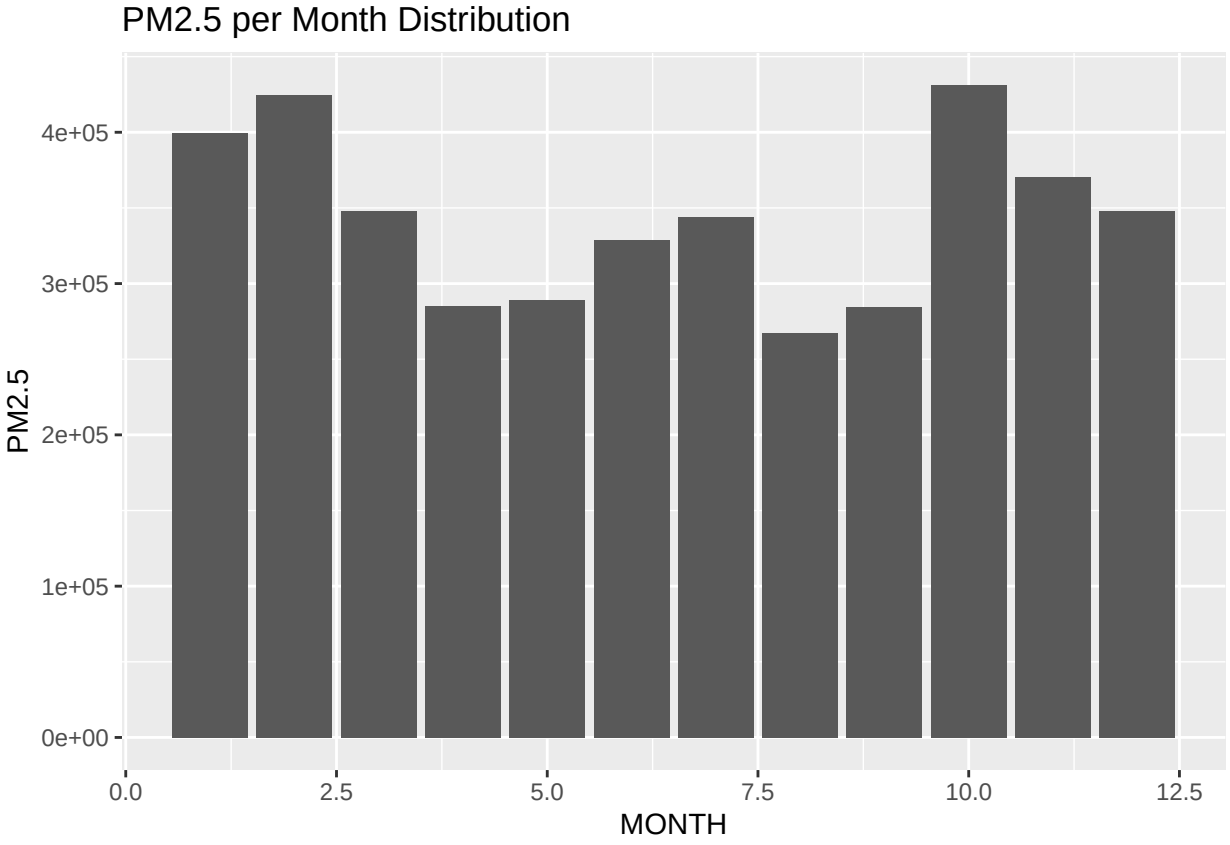
Anyway the main variable and critical point it´s the PM2.5 variable, some countries tries to reduce the exposure to fine particle pollution, and Smart Air www.smartairfilters.com reported that since 2014 to 2019, the PM2.5 in Beijing were reduce to 50%, that information it´s not available in the data set of this project, but in the Bibliography it´s available the source of that important information.

Additionaly the site informs thas despite the harms of PM2.5, studies have found that wearing masks prevents effects on blood pressure and heart rate variability and reducing particulate in the home prevents harm blood pressure, inflammation, and immune response. Exposure to PM2.5 has been linked to premature death, particularly in people who have chronic heart or lung diseases, and reduced lung function growth in children.

# 3 PM2.5 Distribution

### PM2.5 per Month Distribution

At first glance, it appears that the fine dust concentration is low in April and September and the fine dust concentration is high in October, January and February.

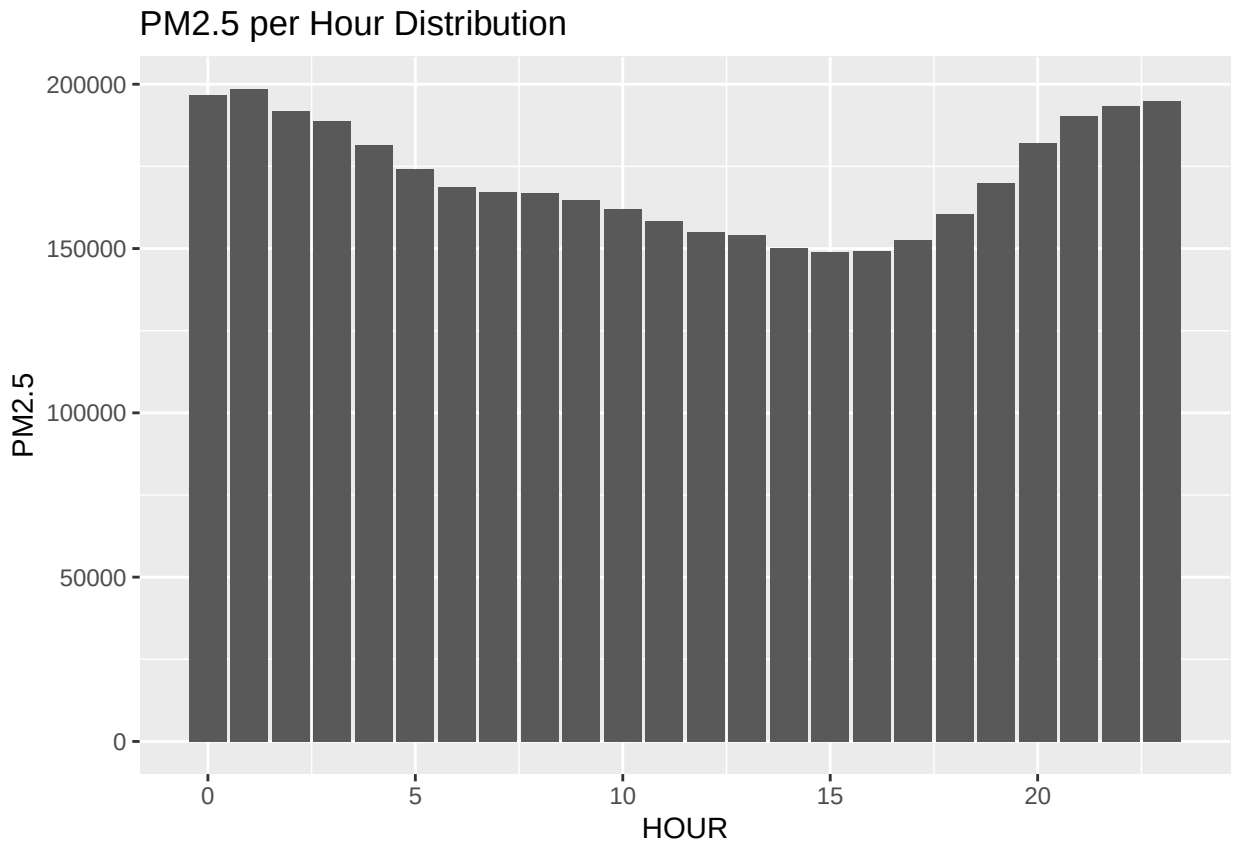



### Month

Month is of integer data type, it tells the month in which data is collected in all rows and in this column plot the month variable has different proportion of this dust or PM2.5. On february and October months are the more frequent observarions.

**PM2.5 per Hour Distribution**

During the day, the concentration of fine dust appears to be high in the early morning hours:




**Hour**

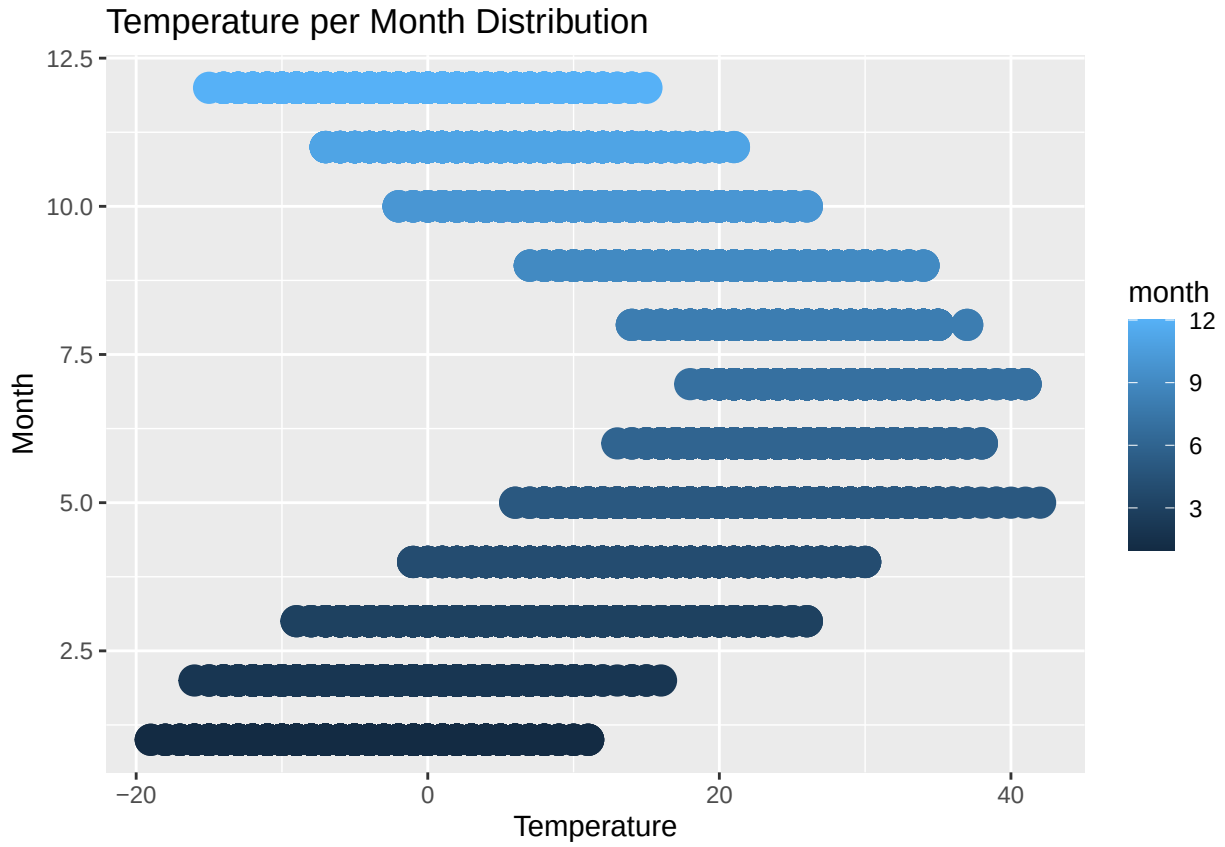
Hour and rush hour it´s the critical point that must be reduced to have a better air. This plot shows the concentration intervals during all day, and mornings it´s the more concentrated PM2.5 dust.

**Hour Importance**

Identifying the PM2.5 concentration, the decision must be focusing the decision making in that timetables, is for that reason that governments impart the vehicular restriction or choose electrical cars, but they are expensive and perhpas in the future will be more accessible. But the importance of identify the hour it´s very important and focusing in that can help to take some decisions or discussions about how to reduce the carbon footprint.

# 4 Temperature based on month

Each month has different temperature and here we have a temperature per month distribution:




```
w <- summary(e$TEMP)
print(paste("TEMP:", w))
```

```
## [1] "TEMP: -19"              "TEMP: 2"               "TEMP: 14"
## [4] "TEMP: 12.4015614148527" "TEMP: 23"              "TEMP: 42"
```

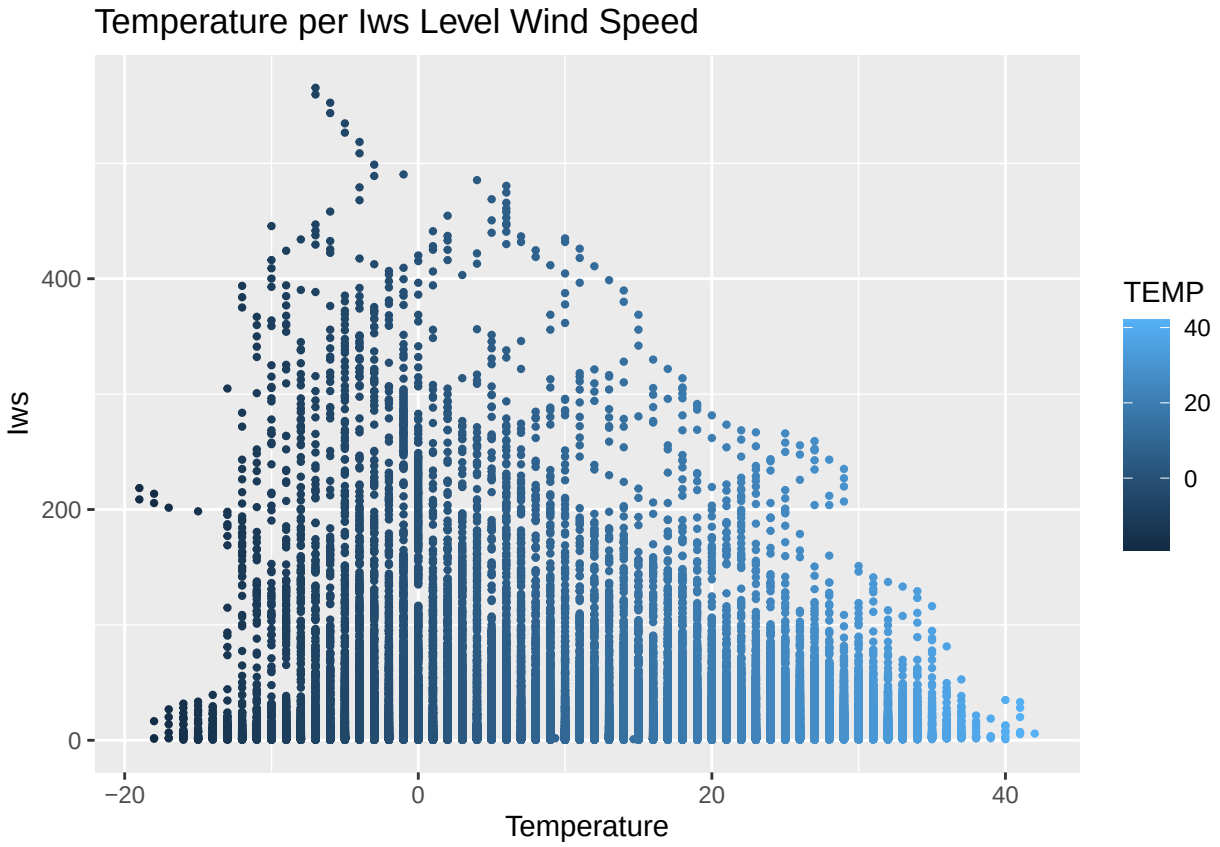
### Monthly Temperature

Where the minimum temperature it´s minus nineteeen and maximun it´s forty two. May it´s the warmest month and July it´s the second warmest.

In the next page, it´s presented some temperature and wind speed means, where these highest warmest months.

# 5 Temperature based on Iws

Temperature is higher in the afternoons and cooler in the twighlight time in the morning.




**Cumulated wind speed & Temperature**

Between january to march the mean TEMP was 2.767 with a Iws mean 17.30 and Between april to july the mean TEMP was 21.85 with a Iws mean 21.59. Iws, Cumulated wind speed it´s higher in april to july:

```
##       Iws              TEMP             month           year
##  Min.   :  0.45   Min.   :-13.000   Min.   :1.000   Min.   :2014
##  1st Qu.:  1.79   1st Qu.: -2.000   1st Qu.:1.000   1st Qu.:2014
##  Median :  4.92   Median :  1.000   Median :2.000   Median :2014
##  Mean   : 17.30   Mean   :  2.767   Mean   :2.003   Mean   :2014
##  3rd Qu.: 18.77   3rd Qu.:  7.000   3rd Qu.:3.000   3rd Qu.:2014
##  Max.   :328.55   Max.   : 26.000   Max.   :3.000   Max.   :2014
```

```
##       Iws              TEMP           month           year
##  Min.   :  0.45   Min.   : 3.00   Min.   :4.000   Min.   :2014
##  1st Qu.:  1.79   1st Qu.:18.00   1st Qu.:5.000   1st Qu.:2014
##  Median :  6.26   Median :23.00   Median :5.000   Median :2014
##  Mean   : 18.39   Mean   :22.85   Mean   :5.504   Mean   :2014
##  3rd Qu.: 21.45   3rd Qu.:28.00   3rd Qu.:7.000   3rd Qu.:2014
##  Max.   :310.22   Max.   :42.00   Max.   :7.000   Max.   :2014
```

# 6 Temperature TEMP Linear Model, MSE and Correlation

**TEMP Linear Model**

```
##
## Call:
## lm(formula = TEMP ~ ., data = e)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.0745  -3.3445  -0.2435   3.0941  21.5120
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -1.380e+06  2.909e+04  -47.454   <2e-16 ***
## No          -7.830e-02  1.650e-03  -47.445   <2e-16 ***
## year         6.869e+02  1.447e+01   47.470   <2e-16 ***
## month        5.727e+01  1.206e+00   47.491   <2e-16 ***
## day          1.894e+00  3.960e-02   47.832   <2e-16 ***
## hour         3.072e-01  3.955e-03   77.669   <2e-16 ***
## pm2.5       -2.458e-02  2.758e-04  -89.143   <2e-16 ***
## DEWP         4.545e-01  2.987e-03  152.185   <2e-16 ***
## PRES        -4.178e-01  3.960e-03 -105.511   <2e-16 ***
## cbwdNE       1.829e-01  8.735e-02    2.094   0.0362 *
## cbwdNW       7.145e-02  7.228e-02    0.989   0.3229
## cbwdSE       1.281e+00  6.688e-02   19.157   <2e-16 ***
## Iws          6.458e-03  5.386e-04   11.990   <2e-16 ***
## Is          -6.283e-01  3.054e-02  -20.572   <2e-16 ***
## Ir          -4.749e-01  1.690e-02  -28.102   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.816 on 41742 degrees of freedom
## Multiple R-squared:  0.8436, Adjusted R-squared:  0.8435
## F-statistic: 1.608e+04 on 14 and 41742 DF,  p-value: < 2.2e-16
```

This linear model shows the temperature output between all others variables. The R-squared it´s a correlation, knowning as "goodness of fit", is represented as a value between 0.0 and 1.0 and in this lineal model reached 0.8435 or 0.8. It means that temperature it´s the **80%** of the variance between others are very strong. The Std. Error can be used to calculate confidence interval and others in the next page has the R-quared explanation Linear Model result. Additionally MSE and Correlation.

- **TEMP Residuals Standard Error**

**Residuals Standard Error* or *Residual Standard Deviation **is a measure used to assess how well a Linear Regression model fits the data. The Regression Model predicts the TEMP or temperature has an average error about** 4.8**. (Even the lower this value is, that means the better the model will be)

**TEMP F-Statistic & P-value**

- If F-statistic is little bit larger than 1 is already sufficient to reject the null hypothesis, in this case it´s **1.6**. The P-value it´s very small **2.2** so lower than 5 and this is enough to reject the null hypothesis. We reject the null hypothesis and conclude that there is strong evidence that a relationship does exist between TEMP Temperature and others variables.

**TEMP R-squared**

- When **R-squared** is small, the **Adjusted R-squared** will become negative and it is not in this case.
- An adjusted r-squared is a more accurate measure than r-squared about how much variance in the response or dependent variable (Y)

| **Concept** | **Result** |
|---|---|
| *R-squared* | 0.8436 |
| *Adjusted R-squared* | 0.8435 |

- It´s necessary to get some conclusions about the variables, before to start with Random Forest and to analize the data set, the **Adjusted R-squared** it´s the more important measure and it means that temperature has a strong relationship with all others variables.

**TEMP MSE**

- The residual standard error is an estimate of the standard deviation of the response relative to the population regression line.

$$MSE = (1/n) * sum(actual \breve{} forecast)2$$

- MSE it´s a risk function that measures the square of errors and in the case of Temperature between others variables, the MSE result it´s **23.18**.

```
## [1] "MSE: 23.1834365640924"
```

**TEMP Correlation**

```
## [1] "Correlation: 0.918477359195698"
```

- The Temperature variable between others variables has a strong correlation.

# 7 Random Forest Temperature in different Months

**Random Forest**

Now it´s time to make some predictions in different months january to december or 1 to 12. **¿How we can predict the *air quality* in different month?** It´s neccesary to create:

- The Data
- The ThePM (PM2.5)
- The Training Model
- The Testing Model
- The Training
- The Model and Model OOB
- Categorize Forecasts
- Some PM2.5 testings in Months Forecasts
- The Prediction PM2.5 air in months
- Air Predictions in different Months.

## 7.1 The Data

Month is character variable in Data, having 29637 rows and for doing any Random Forest it´s necessary to transform the class of month, from character to factor.

```
## 'data.frame':    29637 obs. of  13 variables:
##  $ No   : int  25 26 27 28 29 30 31 32 33 34 ...
##  $ year : int  2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 ...
##  $ month: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ day  : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ hour : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ pm2.5: int  129 148 159 181 138 109 105 124 120 132 ...
##  $ DEWP : int  -16 -15 -11 -7 -7 -7 -7 -7 -8 -7 ...
##  $ TEMP : num  -4 -4 -5 -5 -5 -6 -6 -5 -6 -5 ...
##  $ PRES : num  1020 1020 1021 1022 1022 ...
##  $ cbwd : chr  "SE" "SE" "SE" "SE" ...
##  $ Iws  : num  1.79 2.68 3.57 5.36 6.25 ...
##  $ Is   : int  0 0 0 1 2 3 4 0 0 0 ...
##  $ Ir   : int  0 0 0 0 0 0 0 0 0 0 ...
```

## 7.2 The ThePM (PM2.5)

The PM2.5 it´s the contamination dust, we´re going to call it just thePM to make it simple and we have 994 PM2.5 values registrations:

- Summary

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0     145     290     301     435     994
```

- Structure

```
##  int [1:581] 129 148 159 181 138 109 105 124 120 132 ...
```

**Concern:** The PM2.5 basically it´s the main *concern* about the air quality, **¿How we can predict the *air quality* in different month?**

## 7.3 The Training Model

training_u it´s the sample of 70% that helps to train the predict model, 70% because it´s a few data and the 70% it´s 406 values in training_u:

```
training_u = sample(thePM, length(thePM) * 0.70, replace = FALSE)
```

training_u

- Summary

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   141.2   286.5   298.6   434.8   994.0
```

- Structure

```
##  int [1:406] 144 184 618 101 886 4 23 277 570 310 ...
```

## 7.4 The Testing Model

A test set is reserved for future evaluations of predictive power, this is called The Testing Model or testing_u Where it´s de difference of training_u

```
##   [1] 129 159 124 140 152  91  78  98  95  70  76  73  58  26  28  30  33  34
##  [19]  36  60  84  44  42 131  43  24  67 198  81 135 200 212 227 225 119  22
##  [37]  12 257 174 242 261 269 208 182 230 191 282 349 146 188 203 233 403 360
##  [55] 358 297 310 305 426 364  38 340 298 299 300 213 143 104  99  74 185 163
##  [73] 258 176   8 351 980 599 115 265 215 206 220 246 353 309 284 248 202 224
##  [91] 252 239 232 473 784 761 194 286 385 238 281 318 334 293   1 320 503 405
## [109] 359 355 348 279 287 327 391 436 431 387 350 329 335 397 392 428 420 395
## [127] 352 419 466 467 504 569 488 501 559 557 540 454 363 308 404 427 474 500
## [145] 548 607 573 510 492 527 563 452 494 469 451 375 356 338 370 456 484 450
## [163] 994 529 802 845 776 824 886 852 858 744 731 722 684 651 475 523 447 448
## [181] 603 512 532 516 541 539 567 659 671 487 579 507 545 577 519 551 542 580
```

- Summary

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0   198.5   344.0   352.7   498.5   994.0
```

- Structure

```
##  int [1:198] 129 159 124 140 152 91 78 98 95 70 ...
```

Now it´s necessary to create a training model.

## 7.5 Training

Here it´s the creation of the data for training and this training itps the diference from training_u. And it´s a filter of the *concern*, and that *concern*, it´s the PM2.5 air quality variable. For randomForest, the variable you use in the Model must be factor, here in training it is necessary to fix it to factor.

Training: Selecting all variables and mutate the month as a factor, this because for Random Forest any categorical variable must pass from character to a factor.

- Summary

```
##        No              year          month           day              hour
##  Min.   :   25   Min.   :2010   7      :2020   Min.   : 1.00   Min.   : 0.0
##  1st Qu.:11618   1st Qu.:2011   6      :1973   1st Qu.: 8.00   1st Qu.: 5.0
##  Median :22118   Median :2012   5      :1926   Median :16.00   Median :11.0
##  Mean   :22098   Mean   :2012   8      :1782   Mean   :15.95   Mean   :11.4
##  3rd Qu.:32915   3rd Qu.:2013   10     :1745   3rd Qu.:23.00   3rd Qu.:18.0
##  Max.   :43795   Max.   :2014   2      :1679   Max.   :31.00   Max.   :23.0
##                                 (Other):9611
##      pm2.5            DEWP              TEMP             PRES
##  Min.   : 36.0   Min.   :-29.000   Min.   :-18.00   Min.   : 992
##  1st Qu.: 70.0   1st Qu.: -6.000   1st Qu.:  2.00   1st Qu.:1007
##  Median :111.0   Median :  7.000   Median : 15.00   Median :1014
##  Mean   :134.3   Mean   :  5.166   Mean   : 13.24   Mean   :1015
##  3rd Qu.:166.0   3rd Qu.: 17.000   3rd Qu.: 24.00   3rd Qu.:1023
##  Max.   :994.0   Max.   : 28.000   Max.   : 38.00   Max.   :1042
##
##       Iws               Is                Ir              cbwd
##  Min.   :  0.45   Min.   : 0.00000   Min.   : 0.0000   Length:20736
##  1st Qu.:  1.78   1st Qu.: 0.00000   1st Qu.: 0.0000   Class :character
##  Median :  3.58   Median : 0.00000   Median : 0.0000   Mode  :character
##  Mean   : 12.76   Mean   : 0.08054   Mean   : 0.1579
##  3rd Qu.: 13.86   3rd Qu.: 0.00000   3rd Qu.: 0.0000
##  Max.   :485.46   Max.   :26.00000   Max.   :25.0000
##
```

- Structure

```
## 'data.frame':    20736 obs. of  13 variables:
##  $ No   : int  25 27 28 29 30 31 33 34 35 36 ...
##  $ year : int  2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 ...
##  $ month: Factor w/ 12 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ day  : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ hour : int  0 2 3 4 5 6 8 9 10 11 ...
##  $ pm2.5: int  129 159 181 138 109 105 120 132 140 152 ...
##  $ DEWP : int  -16 -11 -7 -7 -7 -7 -8 -7 -7 -8 ...
##  $ TEMP : num  -4 -5 -5 -5 -6 -6 -6 -5 -5 -5 ...
##  $ PRES : num  1020 1021 1022 1022 1022 ...
##  $ Iws  : num  1.79 3.57 5.36 6.25 7.14 ...
##  $ Is   : int  0 0 1 2 3 4 0 0 1 0 ...
##  $ Ir   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ cbwd : chr  "SE" "SE" "SE" "SE" ...
```

**As you can see here in this structure, now month variable it´s factor**

## 7.6 Model

First, let's proceed by setting a seed to run a model that makes tests and that tests starts in a set.seed zero and the forecast must be the same zero seed. Additionally this Random Forest algorithm needs an estimator or the number of trees, in this case it´s 10.

Numerically where to start testing from, hence the seed. You can improve the quality of the Model, giving it different seeds. If the Model is not good, another seed is given. For this project a set.seed is zero.

Random Forest Model the data comes from the training, month based on PM2.5 air quality that in the *concern*.

- Summary

```
##                 Length Class  Mode
## call                 9 -none- call
## type                 1 -none- character
## predicted        20736 factor numeric
## err.rate           130 -none- numeric
## confusion          156 -none- numeric
## votes           248832 matrix numeric
## oob.times        20736 -none- numeric
## classes             12 -none- character
## importance         168 -none- numeric
## importanceSD       156 -none- numeric
## localImportance      0 -none- NULL
## proximity            0 -none- NULL
## ntree                1 -none- numeric
## mtry                 1 -none- numeric
## forest              14 -none- list
## y                20736 factor numeric
## test                 0 -none- NULL
## inbag                0 -none- NULL
## terms                3 terms  call
```

- Structure

```
## List of 19
##  $ call           : language randomForest(formula = month ~ ., data = training, ntree = 10, method =
##  $ type           : chr "classification"
##  $ predicted      : Factor w/ 12 levels "1","2","3","4",..: 2 2 1 1 1 1 1 1 1 1 ...
##   ..- attr(*, "names")= chr [1:20736] "1" "2" "3" "4" ...
##  $ err.rate       : num [1:10, 1:13] 0.187 0.189 0.171 0.168 0.165 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : NULL
##   .. ..$ : chr [1:13] "OOB" "1" "2" "3" ...
##  $ confusion      : num [1:12, 1:13] 1410 93 27 2 0 0 0 0 0 1 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:12] "1" "2" "3" "4" ...
##   .. ..$ : chr [1:13] "1" "2" "3" "4" ...
##  $ votes          : 'matrix' int [1:20736, 1:12] 1 1 1 5 7 3 4 3 4 5 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:20736] "1" "2" "3" "4" ...
##   .. ..$ : chr [1:12] "1" "2" "3" "4" ...
```

```
##  $ oob.times      : num [1:20736] 2 2 2 5 7 3 4 5 4 5 ...
##  $ classes        : chr [1:12] "1" "2" "3" "4" ...
##  $ importance     : num [1:12, 1:14] 0.352 0.2148 0.1341 0.0526 0.0777 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:12] "No" "year" "day" "hour" ...
##   .. ..$ : chr [1:14] "1" "2" "3" "4" ...
##  $ importanceSD   : num [1:12, 1:13] 0.02953 0.01518 0.01336 0.00661 0.01273 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:12] "No" "year" "day" "hour" ...
##   .. ..$ : chr [1:13] "1" "2" "3" "4" ...
##  $ localImportance: NULL
##  $ proximity      : NULL
##  $ ntree          : num 10
##  $ mtry           : num 3
##  $ forest         :List of 14
##   ..$ ndbigtree : int [1:10] 5337 5723 4725 6005 6255 5731 5283 6005 5873 6599
##   ..$ nodestatus: int [1:6599, 1:10] 1 1 1 1 1 1 1 1 1 -1 ...
##   ..$ bestvar   : int [1:6599, 1:10] 12 8 5 9 1 8 6 5 3 0 ...
##   ..$ treemap   : int [1:6599, 1:2, 1:10] 2 4 6 8 10 12 14 16 18 0 ...
##   ..$ nodepred  : int [1:6599, 1:10] 0 0 0 0 0 0 0 0 0 1 ...
##   ..$ xbestsplit: num [1:6599, 1:10] 3.5 1010.5 281.5 27.3 743.5 ...
##   ..$ pid       : num [1:12] 1 1 1 1 1 1 1 1 1 1 ...
##   ..$ cutoff    : num [1:12] 0.0833 0.0833 0.0833 0.0833 0.0833 ...
##   ..$ ncat      : Named int [1:12] 1 1 1 1 1 1 1 1 1 1 ...
##   .. ..- attr(*, "names")= chr [1:12] "No" "year" "day" "hour" ...
##   ..$ maxcat    : int 1
##   ..$ nrnodes   : int 6599
##   ..$ ntree     : num 10
##   ..$ nclass    : int 12
##   ..$ xlevels   :List of 12
##   .. ..$ No   : num 0
##   .. ..$ year : num 0
##   .. ..$ day  : num 0
##   .. ..$ hour : num 0
##   .. ..$ pm2.5: num 0
##   .. ..$ DEWP : num 0
##   .. ..$ TEMP : num 0
##   .. ..$ PRES : num 0
##   .. ..$ Iws  : num 0
##   .. ..$ Is   : num 0
##   .. ..$ Ir   : num 0
##   .. ..$ cbwd : num 0
##  $ y              : Factor w/ 12 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
##   ..- attr(*, "names")= chr [1:20736] "1" "2" "3" "4" ...
##  $ test           : NULL
##  $ inbag          : NULL
##  $ terms          :Classes 'terms', 'formula'  language month ~ No + year + day + hour + pm2.5 + DEW
##   .. ..- attr(*, "variables")= language list(month, No, year, day, hour, pm2.5, DEWP, TEMP, PRES, Iws
##   .. ..- attr(*, "factors")= int [1:13, 1:12] 0 1 0 0 0 0 0 0 0 0 ...
##   .. .. ..- attr(*, "dimnames")=List of 2
##   .. .. .. ..$ : chr [1:13] "month" "No" "year" "day" ...
##   .. .. .. ..$ : chr [1:12] "No" "year" "day" "hour" ...
##   .. ..- attr(*, "term.labels")= chr [1:12] "No" "year" "day" "hour" ...
##   .. ..- attr(*, "order")= int [1:12] 1 1 1 1 1 1 1 1 1 1 ...
```

```
##   .. ..- attr(*, "intercept")= num 0
##   .. ..- attr(*, "response")= int 1
##   .. ..- attr(*, ".Environment")=<environment: R_GlobalEnv>
##   .. ..- attr(*, "predvars")= language list(month, No, year, day, hour, pm2.5, DEWP, TEMP, PRES, Iws
##   .. ..- attr(*, "dataClasses")= Named chr [1:13] "factor" "numeric" "numeric" "numeric" ...
##   .. .. ..- attr(*, "names")= chr [1:13] "month" "No" "year" "day" ...
##  - attr(*, "class")= chr [1:2] "randomForest.formula" "randomForest"
```

*The model and the interpretation in the next page*

# 8 Categorize Forecasts

### The OOB muestra

```
Model
```

```
##
## Call:
##  randomForest(formula = month ~ ., data = training, ntree = 10,      method = "class", norm.votes = 
##                Type of random forest: classification
##                      Number of trees: 10
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 11.01%
## Confusion matrix:
##       1    2    3    4    5    6    7    8    9   10   11   12 class.error
## 1  1410   73   18    0    0    0    0    0    0    0   10   31  0.08560311
## 2    93 1449   40    4    0    0    0    0    0    1   35   35  0.12552806
## 3    27   55 1425   56    1    0    0    0    0   12   52   20  0.13531553
## 4     2    5   60 1418   82    5    0    1   15   35   23    3  0.14008490
## 5     0    0    4  107 1655   63    4   16   25   28    3    0  0.13123360
## 6     0    0    0    4   55 1789   36   36   25    3    0    0  0.08162218
## 7     0    0    0    1    2   69 1876   55    1    0    0    0  0.06387226
## 8     0    0    1    1   18   57   84 1574   23    1    0    0  0.10517339
## 9     0    1    2   27   37   32    9   29 1483   35    0    0  0.10392749
## 10    1    3   24   45   35    5    0    0   38 1561   17    1  0.09768786
## 11   12   44   61   31    1    0    0    0    0   37 1375   32  0.13684871
## 12   60   49   26    5    0    0    0    0    0    0   46 1257  0.12889813
```

OOB Out of Bag error or estimate of error rate: 9.86%, **81%** of accuracy. The data set model accuracy is approximately 90% (100 minus OBB Out of Bag error)

### Interpretation

- 1434values were correctly classified as january or 1
- 1456 values were correctly classified as february or 2
- 31 values that should have been classified as february were classified as january.
- So on (If you run the code many times, many times will change these numbers, because it´s a matrix)

### Overall Accuracy

The overall accuracy is calculated by summing the number of correctly classified values and dividing by the total number of values:

- Total of number of values: To sum all rows and the vertical total
- Correctly classified values: **17803** (It´s necessary to sum the correctly classified values: 1383 + 1420 + 1351 + 1427 + 1667 + 1712 + 1822 + 1499 + 1409 + 1527 + 1330 + 1256 = 17803)
- Overall Accuracy : 1256/Total of number of values = **Accuracy**.

It can takes a considerable time, for that reason it´s the importance of **OOB estimate of error rate**, in this case it´s the **10%** approximately. So the accuracy it´s the difference a **90%**.

# 9 Testing PM2.5

If you think to be necessary, you can do some testing choosing any pm2.5, any of these values:

```
unique(testing$pm2.5)
```

```
##   [1] 129 159 124 140 152  91  78  98  95  70  76  73  58  36  60  84  44  42
##  [19] 131  43  67 198  81 135 200 212 227 225 119 257 174 242 261 269 208 182
##  [37] 230 191 282 349 146 188 203 233 403 360 358 297 310 305 426 364  38 340
##  [55] 298 299 300 213 143 104  99  74 185 163 258 176 351 980 599 115 265 215
##  [73] 206 220 246 353 309 284 248 202 224 252 239 232 473 784 761 194 286 385
##  [91] 238 281 318 334 293 320 503 405 359 355 348 279 287 327 391 436 431 387
## [109] 350 329 335 397 392 428 420 395 352 419 466 467 504 569 488 501 559 557
## [127] 540 454 363 308 404 427 474 500 548 607 573 510 492 527 563 452 494 469
## [145] 451 375 356 338 370 456 484 450 994 529 802 845 776 824 886 852 858 744
## [163] 731 722 684 651 475 523 447 448 603 512 532 516 541 539 567 659 671 487
## [181] 579 507 545 577 519 551 542 580
```

To set a **Forecast**. In my case, I´m going to set a PM2.5 40 to 60 forecast range:

```
testingair = testing %>% filter(pm2.5 %in% (40:60)) %>% select(No, year, month, day, hour, pm2.5, DEWP,
```

Now as the page 14 makes mention, in the point 7.6, the set.seed must be the same as used in the model, here it is a prediction air, in a vector called predictionair:

```
  set.seed(0)
  predictionair <- predict(Model, newdata = testingair, type = "class")
```

Now month as a factor and we can see the first prediction according to the last set PM2.5 in the range of 40 to 60.

```
  predictionair = as.factor(predictionair)
  testingair$month = as.factor(testingair$month)
  table(predictionair)
```

```
## predictionair
##   1   2   3   4   5   6   7   8   9  10  11  12
##  78  58  78  95 147  95 111 112 101  91  87  76
```

### Predictions Matrix

In the next page it´s presented a confusion matrix about the predict Model, and next: Random Forest Error Rate Plot and Importance of Variable Plot.

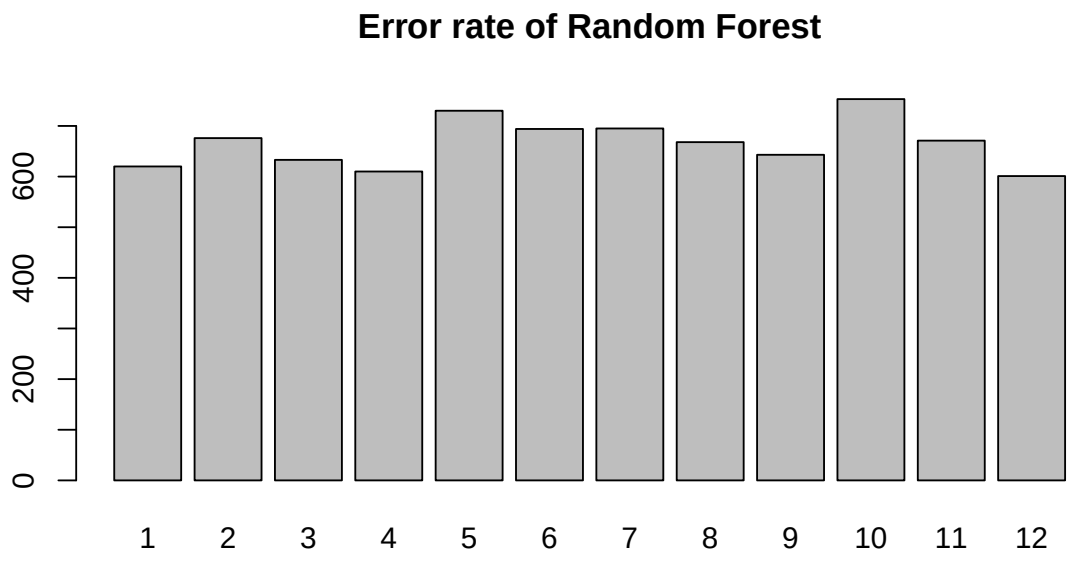
```
predictions.rf <- predict(Model, newdata = testing)
```

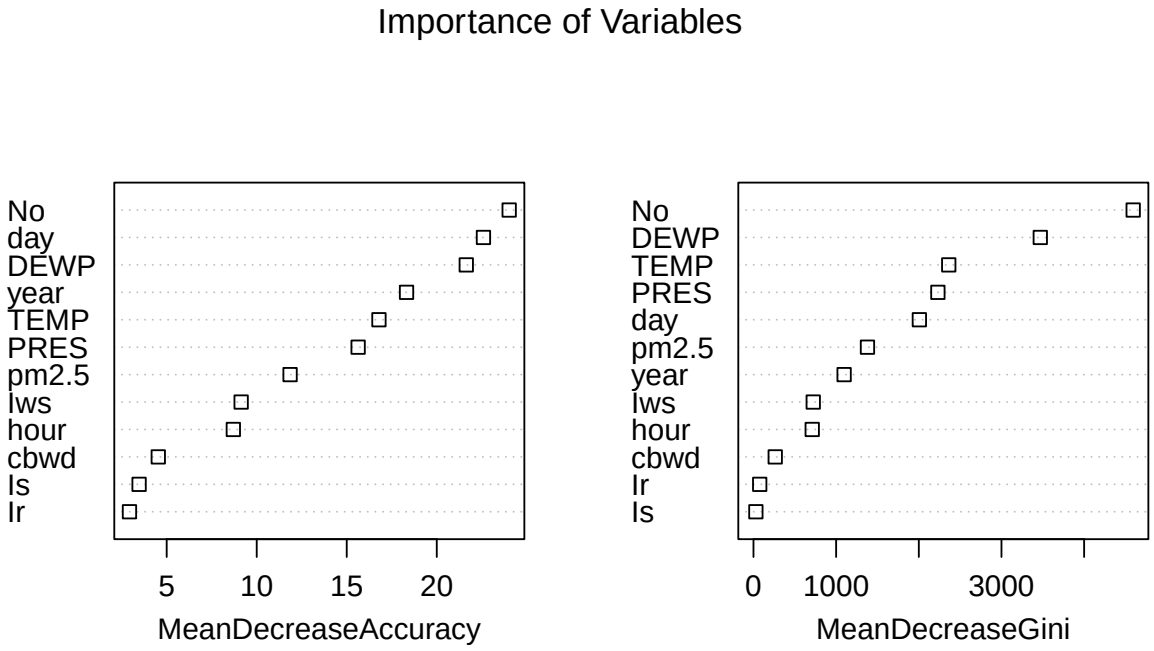
**Confusion Matrix**

```
confusionMatrix(predictions.rf, as.factor(testing$month))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   1   2   3   4   5   6   7   8   9  10  11  12
##         1  608   7   0   0   0   0   0   0   0   0   1   4
##         2    1 667   3   0   0   0   0   0   0   0   4   1
##         3    0   1 625   3   0   0   0   0   0   0   3   1
##         4    0   0   4 595   6   0   0   0   0   1   4   0
##         5    0   0   0   6 715   4   0   0   2   2   1   0
##         6    0   0   0   0   4 681   5   1   2   1   0   0
##         7    0   0   0   0   0   1 692   2   0   0   0   0
##         8    0   0   0   0   0   0   4 663   1   0   0   0
##         9    0   0   0   0   1   3   0   3 634   2   0   0
##         10   0   1   0   2   4   0   0   0   5 737   4   0
##         11   0   0   2   3   1   0   0   0   0   5 654   6
##         12   0   3   0   0   0   0   0   0   0   0   3 595
##
## Overall Statistics
##
##                Accuracy : 0.984
##                  95% CI : (0.981, 0.9866)
##     No Information Rate : 0.0936
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9825
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 1 Class: 2 Class: 3 Class: 4 Class: 5 Class: 6
## Sensitivity           0.99836  0.98233  0.98580  0.97701  0.97811  0.98839
## Specificity           0.99838  0.99877  0.99891  0.99797  0.99793  0.99822
## Pos Pred Value        0.98065  0.98669  0.98736  0.97541  0.97945  0.98127
## Neg Pred Value        0.99986  0.99836  0.99878  0.99810  0.99780  0.99890
## Prevalence            0.07618  0.08494  0.07931  0.07618  0.09144  0.08619
## Detection Rate        0.07606  0.08344  0.07818  0.07443  0.08944  0.08519
## Detection Prevalence  0.07756  0.08456  0.07918  0.07631  0.09132  0.08682
## Balanced Accuracy     0.99837  0.99055  0.99236  0.98749  0.98802  0.99330
##                      Class: 7 Class: 8 Class: 9 Class: 10 Class: 11 Class: 12
## Sensitivity           0.98716  0.99103  0.98447   0.98529   0.97033   0.98023
## Specificity           0.99959  0.99932  0.99878   0.99779   0.99768   0.99919
## Pos Pred Value        0.99568  0.99251  0.98600   0.97875   0.97466   0.99002
## Neg Pred Value        0.99877  0.99918  0.99864   0.99848   0.99727   0.99838
## Prevalence            0.08769  0.08369  0.08056   0.09357   0.08431   0.07593
## Detection Rate        0.08656  0.08294  0.07931   0.09219   0.08181   0.07443
## Detection Prevalence  0.08694  0.08356  0.08044   0.09420   0.08394   0.07518
## Balanced Accuracy     0.99337  0.99517  0.99162   0.99154   0.98400   0.98971
```

## 10 Random Forest Error Rates and Importance of variables




Plot with varImpPlot() function from Random Forest library

# 11 The Prediction PM2.5 air in months

We can contiuous with the prediction:

```
table(predictionair)
```

```
## predictionair
##   1   2   3   4   5   6   7   8   9  10  11  12
##  78  58  78  95 147  95 111 112 101  91  87  76
```

And aleatory suggestions of months:

# 12 Air Predictions in different Months

```
table(predictionair)
```

```
## predictionair
##   1   2   3   4   5   6   7   8   9  10  11  12
##  78  58  78  95 147  95 111 112 101  91  87  76
```

Recommendation, even can be more, it depends of the choosen pm2.5 In this case where three the prediction and these are the predicted months that the air quality will be between *40 and 42*, taking this code:

**Months prediction in Air quality PM2.5**

```
  suggestion = names(sort(table(predictionair), decreasing = T))
  suggestion = names(sort(table(predictionair), decreasing = T)[1:3])
  suggestion
```

```
## [1] "5" "8" "7"
```

So these are the months predicted.

And here it´s the head of the prediction, in a filter taking the suggestion:

```
  head(testingair %>% filter(month %in% suggestion))
```

```
##     No year month day hour pm2.5 DEWP TEMP PRES   Iws Is Ir cbwd
## 1 2931 2010     5   3    2    43   -1   15 1007  2.23  0  0   cv
## 2 2934 2010     5   3    5    44   -2   13 1008  3.58  0  0   NW
## 3 3044 2010     5   7   19    58    1   23 1004 76.88  0  0   NW
## 4 3103 2010     5  10    6    58    6    8 1005  1.78  0  0   cv
## 5 3108 2010     5  10   11    43   -8   23 1003 20.11  0  0   NW
## 6 3111 2010     5  10   14    58   -5   22 1004 61.23  0  0   NW
```

And finish the suggestions have many rows

```
## # A tibble: 374 x 13
## # Groups:   month [3]
##       No  year month   day  hour pm2.5  DEWP  TEMP  PRES   Iws    Is    Ir cbwd
##    <int> <int> <fct> <int> <int> <int> <int> <dbl> <dbl> <dbl> <int> <int> <chr>
##  1  2931  2010 5         3     2    43    -1    15  1007  2.23     0     0 cv
##  2  2934  2010 5         3     5    44    -2    13  1008  3.58     0     0 NW
##  3  3044  2010 5         7    19    58     1    23  1004 76.9      0     0 NW
##  4  3103  2010 5        10     6    58     6     8  1005  1.78     0     0 cv
##  5  3108  2010 5        10    11    43    -8    23  1003 20.1      0     0 NW
##  6  3111  2010 5        10    14    58    -5    22  1004 61.2      0     0 NW
##  7  3146  2010 5        12     1    43     6    14  1019 17.4      0     0 SE
##  8  3156  2010 5        12    11    42     6    20  1019  1.79     0     0 cv
##  9  3157  2010 5        12    12    44     6    21  1018  4.92     0     0 SE
## 10  3158  2010 5        12    13    44     5    22  1018  1.79     0     0 cv
## # ... with 364 more rows
```

# 13 Conclusions

The air quality depends on multi-dimensional factors including location, time, weather parameters, such as temperature, humidity, wind direction and force, air pressure, carbon dioxide, relative humidity, sulfur dioxide, wind speed, etc.

In this project were presented jsome variables: Temperature, Iws, Month, PM2.5, and Hour. Aditionally the correlation of each of tese variables and others and was finally to obtain the minimum prediction and in the testing with a pm2.5 40.0 to 60.0 obtaining a prediction month values.

Random Forest depends on the data, and it´s critical to have categorical data, in this data set the only categorical data was month and cbwd. Month with 1 to 12 months and cbwd with cardinal points. And for that reason the decision was take the more appropiate for this algorithm and was month. Additionally one of the objectives of the course was a simple code and for that reason was this data choosen.

To obtain a better result in this data or similar projects, it´s necessary to obtain more variables, and ideal:

- Categorical data
- More variables, to have a better accuracy.
- Use the same seed.set in the Model and in the Testing
- Make more than one testing with different set.seed and see which set.seed has a better accuracy in the model.
- Change the class of the predicted Random Forest variable from character to factor.

This project it´s simple and easy to read and thanks for the opportunity to present it.