03: Model predictiu de NO2 a Barcelona

Soulaiman el Hamri

2025-05-22

Contents

1. Context i motivació

En aquest Treball Final de Màster es construeix un model predictiu basat en Random Forest per estimar els nivells diaris de diòxid de nitrogen (NO2) a la ciutat de Barcelona, a partir de dades ambientals i meteorològiques.

Els objectius principals són:

- Avaluar la relació entre les variables meteorològiques i la concentració de NO2.
- Construir un model que expliqui les variacions de NO2 en funció de factors ambientals i temporals.
- Visualitzar els resultats mitjançant gràfics i anàlisis descriptius.

Aquest estudi no té com a finalitat predir el valor del dia següent (t+1), sinó explorar el potencial predictiu de les variables i identificar patrons de comportament.

2. Lectura i filtratge de dades

Primerament, carreguem els fitxers .csv de contaminants i meteorologia i filtrem el període d'estudi des de l'1 de gener de 2020 fins a l'actualitat. S'ha triat filtrar a partir del 2020, perque així ens permet captar patrons temporals actualitzats i reduir possibles distorsions associades a dades antiges o incompletes.

```
cont_diari <- read_csv("../data/processed/contaminants/contaminants_bcn_resum_diari.csv") %>%
  mutate(data = ymd(data)) %>%
  filter(data >= ymd("2020-01-01"))

est_cont <- read_csv("../data/processed/contaminants/estacions_contaminants_bcn_filtrat.csv")

meteo <- read_csv("../data/processed/meteocat/meteocat_1995_2025_bcn_processed.csv") %>%
  mutate(DATA_LECTURA = ymd(DATA_LECTURA)) %>%
  filter(DATA_LECTURA >= ymd("2020-01-01")) %>%
  mutate(ACRÒNIM = str_trim(str_to_upper(ACRÒNIM)))
```

3. Construcció del dataset de NO2

Un cop carregats els datasets, es filtra el conjunt de dades de contaminants per quedar-nos només amb els registres del contaminant NO2, ja que és un dels principals indicadors de la contaminació atmosfèrica urbana, i que esta estretament relacionat amb les emissions del trànsit i la combustió de combustibles fòssils.

El valor utilitzat com a variable target és la mitjana diària de NO2 (valor_mitja), s'ha triat aquesta variable per la seva estabilitat diària.

```
# Utilitzm el valor_mitja com a target
no2_diari <- cont_diari %>%
  filter(contaminant == "NO2") %>%
  select(codi_eoi, data, valor_mitja) %>%
  rename(
    station_code = codi_eoi,
    date         = data,
    NO2          = valor_mitja
  ) %>%
  left_join(
    est_cont %>% select(codi_eoi, latitud, longitud),
    by = c("station_code" = "codi_eoi")
  )
```

4. Transformació de la meteorologia

Per incorporar les variables meteorològiques al model predictiu, és necessari transformar el conjunt de dades de format llarg (on cada fila correspon a una mesura concreta) a format ample, de manera que cada fila representi un dia i una estació amb totes les variables meteorològiques rellevants com a columnes.

En aquesta transformació, es filtren només les variables meteorològiques d'interès per al model:

- **TM**: Temperatura mitjana diària
- **HRM**: Humitat relativa mitjana diària

- **PPT**: Precipitació acumulada diària
- **PM**: Pressió atmosfèrica mitjana diària
- **VVM10**: Velocitat mitjana del vent a 10 metres

Un cop seleccionades es reorganitzen les dades perquè cada combinació de data i estació meteorològica inclogui totes aquestes variables com a columnes independents. Per últim es fa un rename de les columnes per fer-les més interpretables.

```
meteo_filtered <- meteo %>%
  filter(ACRÒNIM %in% c("TM", "HRM", "PPT", "PM", "VVM10")) %>%
  select(DATA_LECTURA, CODI_ESTACIO, ACRÒNIM, VALOR) %>%
  pivot_wider(names_from = ACRÒNIM, values_from = VALOR) %>%
  rename(station_code = CODI_ESTACIO, date = DATA_LECTURA,
         Temp = TM, Hum = HRM, Prec = PPT, Press = PM, Wind = VVM10)
```

5. Fusió dades de NO2 amb les dades meteorològiques

Abans de fer la fusió entre el dataset de NO2 i les dades meteorològiques, cal tenir en compte que les estacions meteorològiques i les de contaminació no coincideixen. Per aquest motiu, el que farem serà associar a cada estació de contaminació la seva estació meteorològica més propera.

Per fer-ho, utilitzem les coordenades geogràfiques per calcular la distància entre cada parella d'estacions i assignem a cada estació de contaminació el codi de la meteorològica amb menor distància (precalculat i emmagatzemat a est_cont$meteo_code).

```
# Llegim ubicació de les estacions meteorològiques
est_meteo <- read_csv("../data/processed/meteocat/estacions_meteorologiques_bcn.csv") %>%
  select(codi_estacio, latitud, longitud) %>%
  rename(meteo_code = codi_estacio, meteo_lat = latitud, meteo_lon = longitud)

# Calculam distància entre cada estació de contaminació i totes les meteorològiques
matriu_dist <- distm(est_cont[, c("longitud", "latitud")],
                     est_meteo[, c("meteo_lon", "meteo_lat")], fun = distHaversine)

# Obtenim la index de l'estació meteo més propera per a cada estació de contaminació
index_meteo_propera <- apply(matriu_dist, 1, which.min)

# Assignem codi de l'estació meteo més propera
est_cont$meteo_code <- est_meteo$meteo_code[index_meteo_propera]
```

A continuació, unim les taules segons la data i el codi de la **estació meteorològica més propera**.

```
data_full <- no2_diari %>%
  left_join(est_cont %>% select(codi_eoi, meteo_code),
            by = c("station_code" = "codi_eoi")) %>%
  left_join(meteo_filtered,
            by = c("meteo_code" = "station_code", "date" = "date")) %>%
  drop_na(NO2, Temp, Hum, Prec, Press, Wind)
```

Per últim, construïm el dataset definitiu (data_model) incorporant variables temporals i variables de retard:

- **month**: captura l'estacionalitat anual (efectes estacionals del clima).
- **weekday**: permet detectar diferències entre dies laborables i caps de setmana.
- **lag1_NO2**: valor de NO2 del dia anterior, per capturar l'autocorrelació immediata.
- **lag7_NO2**: valor de NO2 d'una setmana abans, per captar repeticions setmanals.

Les variables de retard (lag) ajuden a capturar la dependència temporal pròpia dels contaminants atmosfèrics i milloren la capacitat predictiva del model, que pot tenir inèrcies i cicles relacionats amb la meteorologia i/o l'activitat humana.

```
data_model <- data_full %>%
  arrange(station_code, date) %>%
  group_by(station_code) %>%
  mutate(
    month   = factor(month(date), levels = 1:12),
    weekday = factor(wday(date), levels = 1:7),
    lag1_NO2 = lag(NO2, 1),
    lag7_NO2 = lag(NO2, 7)
  ) %>%
  ungroup() %>%
  drop_na()

summary(data_model)
```

```
##  station_code            date                 NO2            latitud
##  Length:11238       Min.   :2020-01-08   Min.   :  1.00   Min.   :41.38
##  Class :character   1st Qu.:2021-05-05   1st Qu.: 11.96   1st Qu.:41.39
##  Mode  :character   Median :2022-08-18   Median : 20.08   Median :41.39
##                     Mean   :2022-08-21   Mean   : 22.84   Mean   :41.40
##                     3rd Qu.:2023-12-10   3rd Qu.: 30.96   3rd Qu.:41.42
##                     Max.   :2025-04-07   Max.   :110.23   Max.   :41.43
##
##     longitud      meteo_code             Hum             Press
##  Min.   :2.115   Length:11238       Min.   : 24.00   Min.   : 942.5
##  1st Qu.:2.124   Class :character   1st Qu.: 59.00   1st Qu.: 973.0
##  Median :2.148   Mode  :character   Median : 68.00   Median :1006.1
##  Mean   :2.138                      Mean   : 67.58   Mean   : 996.8
##  3rd Qu.:2.153                      3rd Qu.: 76.00   3rd Qu.:1012.5
##  Max.   :2.154                      Max.   :100.00   Max.   :1033.7
##
##       Prec             Temp            Wind            month      weekday
##  Min.   : 0.000   Min.   : 1.70   Min.   : 0.500   3      :1076   1:1596
##  1st Qu.: 0.000   1st Qu.:12.50   1st Qu.: 1.700   1      :1061   2:1604
##  Median : 0.000   Median :16.50   Median : 2.300   2      :1006   3:1605
##  Mean   : 1.352   Mean   :17.32   Mean   : 2.696   10     : 918   4:1607
##  3rd Qu.: 0.000   3rd Qu.:22.50   3rd Qu.: 3.400   8      : 917   5:1606
##  Max.   :93.200   Max.   :33.40   Max.   :15.300   7      : 916   6:1616
##                                                    (Other):5344   7:1604
##     lag1_NO2         lag7_NO2
##  Min.   :  1.00   Min.   :  1.00
##  1st Qu.: 11.96   1st Qu.: 11.97
##  Median : 20.08   Median : 20.12
```

```
##  Mean   : 22.85   Mean   : 22.89
##  3rd Qu.: 30.96   3rd Qu.: 31.00
##  Max.   :110.23   Max.   :110.23
##
```

6. Divisió del conjunt d'entrenament i prova

Per construir i validar el model de manera robusta, es divideix el conjunt de dades en dos subconjunts: entrenament i prova. En aquest cas, donat que treballem amb una sèrie temporal, s'ha opta per una divisió cronològica i no pas aleatòria, amb l'objectiu d'imitar un escenari real de predicció futura.

- El **80% inicial** de les dates s'utilitza per entrenar el model (train_set).
- El **20% final**, que conté les observacions més recents, es reserva com a conjunt de prova (test_set).

Aquesta metodologia evita contaminacions d'informació entre el passat i el futur i permet avaluar el model en condicions properes a les reals.

```
set.seed(123)
all_dates <- sort(unique(data_model$date))
cutoff <- all_dates[floor(0.8 * length(all_dates))]
train_set <- filter(data_model, date <= cutoff)
test_set <- filter(data_model, date > cutoff)
```

7. Entrenament del model Random Forest

En aquesta fase es construeix el model predictiu utilitzant l'algoritme Random Forest, una tècnica basada en la combinació de múltiples arbres de decisió que ofereix bons resultats en contextos amb variables correlacionades i relacions no lineals.

Per optimitzar l'entrenament, es configura una validació creuada simple (cv) amb 3 particions. Això permet una avaluació eficient del model en un temps de càlcul raonable, especialment en conjunts de dades de mida considerable. En cada iteració, es reentrena el model amb diferents particions del conjunt d'entrenament per avaluar-ne l'estabilitat i evitar l'overfitting.

El model s'ajusta mitjançant el paquet ranger, una implementació eficient i ràpida de Random Forest. A més, es defineix el paràmetre importance = "impurity" per obtenir una estimació de la importància relativa de cada variable explicativa, la qual serà útil per interpretar el comportament del model.

Es fixa una llavor (set.seed(123)) per assegurar la reproduïbilitat dels resultats.

```
ctrl <- trainControl(
  method = "cv",
  number = 3
)

set.seed(123)
rf_model <- train(
  NO2 ~ Temp + Hum + Prec + Press + Wind +
    lag1_NO2 + lag7_NO2 + factor(month) + factor(weekday),
  data = train_set,
  method = "ranger",
```

```
  trControl = ctrl,
  importance = "impurity",
  tuneLength = 2,
  num.trees = 100
)
```

8. Avaluació i visualització de resultats

Un cop entrenat el model, es procedeix a avaluar-ne el rendiment sobre el conjunt de prova (test_set), que conté observacions no utilitzades durant l'entrenament. Això permet valorar la capacitat predictiva del model en un escenari realista.

Es generen les prediccions del NO2 i es comparen amb els valors reals mitjançant les mètriques.

```
# Prediccions sobre el conjunt de test
pred <- predict(rf_model, newdata = test_set)

# Afegim la columna de prediccions al conjunt de test
dfp <- test_set %>%
  mutate(Pred = pred,
         mes = month(date, label = TRUE, abbr = FALSE))

# Mostrem mètriques globals
metrics <- postResample(dfp$Pred, dfp$NO2)
print(metrics)
```

```
##      RMSE  Rsquared       MAE
## 6.3586905 0.7132801 4.7458369
```

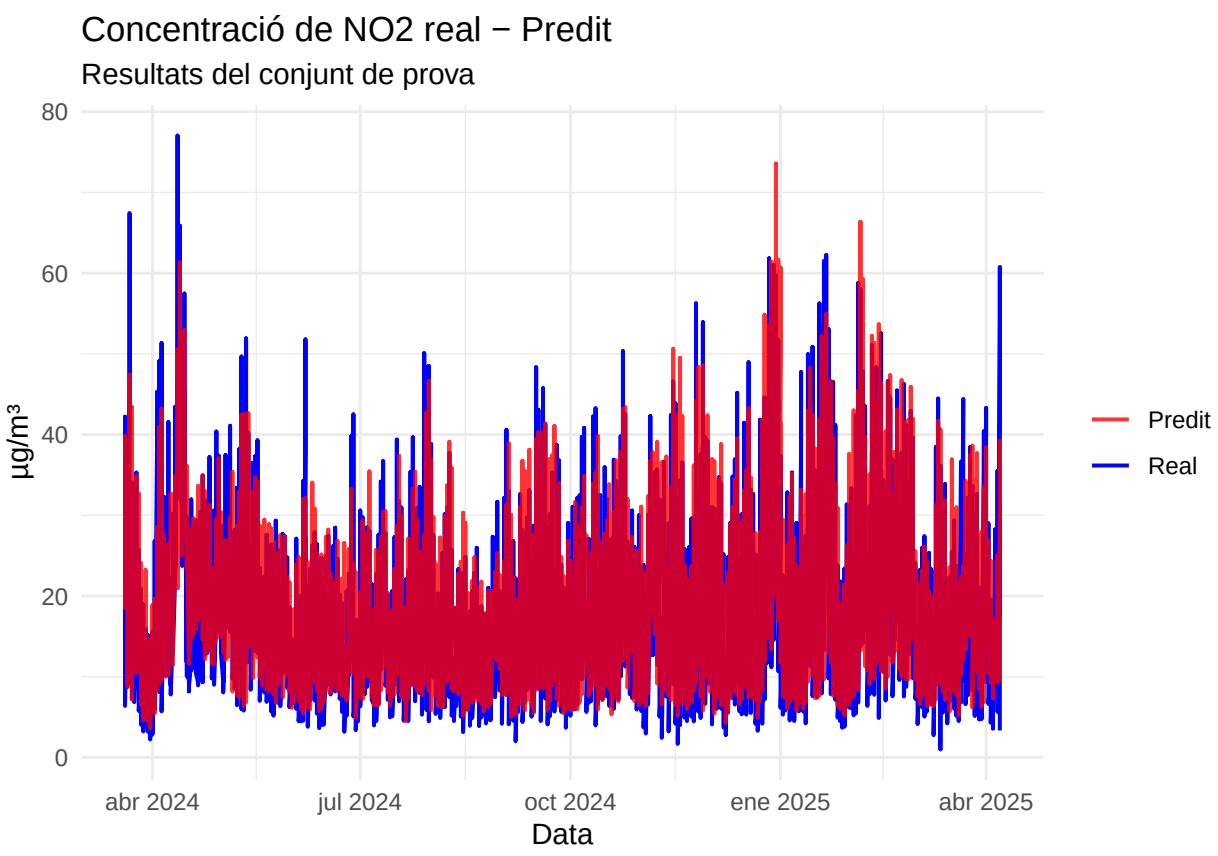
Observem que:

- **MSE (Root Mean Squared Error):** Reflecteix la mitjana de les diferències entre els valors reals i els predits. Amb un valor de 6.36 indica que el model s'equivoca uns 6.36 µg/m³ de mitjana.
- **R2 (Coeficient de determinació):** Mesura quina proporció de la variabilitat del NO2 es pot explicar pel model. Amb un resultat de 0.713,vol dir que el 71,3% de la variabilitat observada en NO2 és explicada pel model.
- **MAE (Mean Absolute Error):** És la mitjana de les diferències absolutes entre els valors reals i els predits. Ens dona un error mitjà absolut de 4.75 µg/m³, lo qual reforça que el model és bastant precís, tot i que pot ser millorable si calgués aplicacions més sensibles.

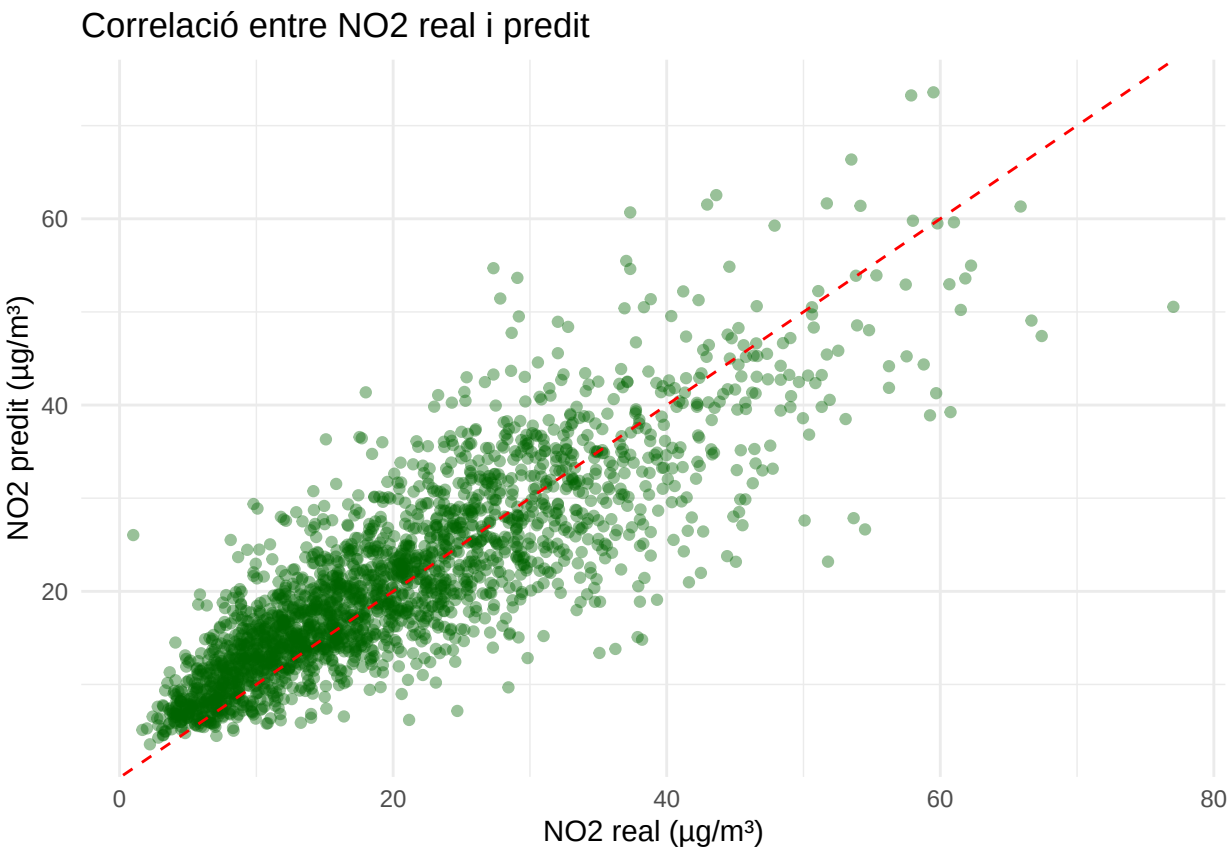
A continuació, representem les prediccions realitzades pel model en comparació amb els valors reals de NO2 al llarg del temps. Aquesta visualització facilita la detecció de patrons temporals, així com possibles desajustos del model.

```
ggplot(dfp, aes(x = date)) +
  geom_line(aes(y = NO2, colour = "Real"), linewidth = 0.7) +
  geom_line(aes(y = Pred, colour = "Predit"), linewidth = 0.7, alpha = 0.8) +
  labs(
      title = "Concentració de NO2 real - Predit",
```

```
    subtitle = "Resultats del conjunt de prova",
    x = "Data", y = "µg/m³"
  ) +
  scale_color_manual(name = NULL, values = c("Real" = "blue", "Predit" = "red")) +
  theme_minimal()
```




```
ggplot(dfp, aes(x = NO2, y = Pred)) +
  geom_point(alpha = 0.4, color = "darkgreen") +
  geom_abline(slope = 1, intercept = 0, color = "red", linetype = "dashed") +
  labs(
    title = "Correlació entre NO2 real i predit",
    x = "NO2 real (µg/m³)",
    y = "NO2 predit (µg/m³)"
  ) +
  theme_minimal()
```



9. Conclusions i línies futures

Aquest document ha desenvolupat un model predictiu de la concentració diària de NO2 a la ciutat de Barcelona utilitzant l'algoritme Random Forest. El model es basa en variables meteorològiques (temperatura, humitat, precipitació, pressió atmosfèrica i vent) així com en variables temporals i de retard (lags) que permeten capturar dinàmiques temporals pròpies de la qualitat de l'aire.

Els resultats obtinguts mostren una capacitat predictiva sòlida, amb valors de R2 superiors al 0.7, indicant que el model captura una part significativa de la variabilitat del NO2. Tot i això, s'observen certs desajustos en períodes concrets, especialment quan hi ha canvis sobtats en les condicions meteorològiques.

Liníes de millora:

- **Afegir noves fonts de dades**: com dades de trànsit, ocupació urbana o emissions industrials, que podrien millorar la precisió del model.
- **Explorar altres algoritmes**: com XGBoost, regressió GAM o xarxes neuronals per comparar-ne el rendiment.
- **Modelització per estació**: crear models específics per a cada estació de mesura per capturar patrons locals més detallats.
- **Representació espacial i interpolació**: una línia de futur podria ser aplicar tècniques d'interpolació per generar mapes continus de NO2 a partir de les prediccions.