TecnoCampus
Escola Superior Politècnica

Centre adscrit a la
upf. Universitat Pompeu Fabra Barcelona

Grau en Enginyeria Mecànica

DESENVOLUPAMENT D'UN MODEL DE DETECCIÓ D'ANOMALIES PER A INSTAL·LACIONS FOTOVOLTAIQUES

Memòria

RUT BERTRAN PIÑERO
PONENT: ARNAU GONZÁLEZ JUNCÀ

PRIMAVERA 2021

TecnoCampus
Mataró-Maresme

A totes les nenes que somnien ser grans dones, i a totes les dones que han fet que avui pugui ser enginyera.

Agraïments

Al Dr. Arnau Gonzalez, tutor del present treball, per confiar-me el projecte i posar-me en mans de professionals que m'han guiat durant el transcurs del projecte.

A l'Albert Garcia, per la seva incansable tasca de revisió i acompanyament en el desenvolupament del treball.

A la meva família, pel suport incondicional, les hores i els esforços que m'han dedicat des del primer dia.

I, finalment, a la Júlia i l'Eric, dues companyes vitals en la universitat, però sobretot en la vida.

Resum

El present treball pretén enfocar cap al manteniment predictiu de les instal·lacions fotovoltaiques, dins el marc de l'IIOT, creant un mòdul de detecció d'anomalies. Utilitzant un model predictiu de la producció d'energia solar fotovoltaica d'una instal·lació donada, el mòdul compara la producció real de la instal·lació i la producció predita per a detectar les anomalies que es generen i classificar-les.

Resumen

El presente proyecto pretende enfocar hacia el mantenimiento predictivo de las instalaciones fotovoltaicas, dentro del marco del IIOT, creando un módulo de detección de anomalías. Usando un modelo predictivo de la producción de energía solar fotovoltaica de una instalación dada, el módulo compara la producción real de la instalación y la producción predicha para detectar las anomalías que se generan y clasificarlas.

Abstract

This project aims to focus toward the predictive maintenance of photovoltaic systems, in IIOT frame, creating an anomaly detection module. This module uses a predictive model for photovoltaic energy production of a given installation. Subsequently, compares the real production of the installation. With this comparison, various anomalies may be detected and classified.

Índex

Índex de figures

Índex de taules

Glossari de termes

AC	Corrent altern
BOS	*Balance of Systems*
DC	Corrent continu
FV	Fotovoltaica
IIOT	*Industrial Internet of Things*
k-NN	*k Nearest Neighbours*
MAR	*Mean Absolute Error*
MAPE	*Mean Absolute Percentage Error*
NaN	*Not a Number*
RMSE	*Root Mean Square Error*

1 Objectius

1.1 Propòsit

Estudiar els diferents mètodes i models matemàtics de predicció dins del *Machine Learning* per comparar-ne l'eficàcia i eficiència en la predicció d'instal·lacions solars fotovoltaiques.

1.2 Finalitat

Aquest projecte neix com a part d'un projecte global (desenvolupament del producte "JoinEnergy" a càrrec de l'empresa Aiguasol) per a la impulsió, creació i gestió de comunitats energètiques ciutadanes. Dins d'aquest projecte es faciliten diferents mòduls i eines per a les comunitats, un d'ells consisteix en la detecció d'anomalies, enmarcat en el mòdul de gestió d'instal·lacions fotovoltaiques. L'objecte d'aquest Treball de Final de Grau és el desenvolupament del model de detecció d'anomalies.

1.3 Objecte del projecte

Elaboració d'un mòdul, pertanyent a una eina d'ús comercial, dissenyat per a la predicció horària de la generació d'energia elèctrica d'una instal·lació fotovoltaica i la detecció d'anomalies com a base per a establir un protocol de manteniment predictiu. Aquest mòdul està en sintonia amb els altres mòduls del producte, per tal de poder integrar-los, recopilar-ne dades i, si calgués, poder-ne fer una ampliació de manera optimitzada i entenedora.

1.4 Abast del projecte

Es contempla l'estudi de diferents models de regressió i l'elecció del més eficaç per a la realització de la tasca. Es tenen en compte els diferents factors físics, locals i generals, que poden afectar el rendiment de la instal·lació fotovoltaica essent o no anomalies tot diferenciant-les. Aquest model està plantejat de manera que la base de dades usada creix a mesura que s'usa la instal·lació. S'estima que el model es testegi i avaluï tant amb una base de dades de producció fotovoltaica amb entrades finites com amb una base de dades en continu creixement. A més a més, s'inclou la detecció d'anomalies en la producció fotovoltaica, comparant els resultats predits pel model i la producció actual. Tot i que l'estudi

i detecció d'anomalies és la base per al manteniment predictiu de la mateixa instal·lació, aquest no s'inclourà en aquest treball.

1.5 Context en les línies de recerca i transferència de coneixement del Tecnocampus

Aquest projecte busca introduir la detecció d'anomalies per a un futur manteniment predictiu dins de la generació d'energia fotovoltaica, com a tal, la seva finalitat màxima és emprendre noves vies d'investigació cap a la interconnexió d'instal·lacions, creant una xarxa, no només de generació d'energia, sinó també de generació i recull de dades per a l'optimització del seu funcionament i manteniment. Amb tot, és un projecte que cap dins l'àrea de recerca de la industria 4.0 de l'Escola Superior Politècnica del Tecnocampus.

2 Introducció

El present projecte té com a objectiu l'elaboració d'un mòdul dissenyat per a la predicció horària de la generació d'energia elèctrica d'una instal·lació fotovoltaica i la detecció d'anomalies. Com a tal, aquest projecte pretén enfocar una línia de recerca envers les instal·lacions fotovoltaiques i com se'n pot aconseguir fer un manteniment predictiu, per evitar la pèrdua d'energia produïda i poder obtenir un retorn de la inversió feta en la instal·lació en menys temps i de manera més satisfactòria per a l'usuari.

Mitjançant aquest projecte es vol aconseguir la promoció de la creació de comunitats energètiques per a l'apoderament de la ciutadania en la transició energètica cap a un consum d'energia més sostenible.

2.1 Abast de detall

En la present documentació es guia al lector per la metodologia del projecte, anant capítol a capítol s'aprofundeix en els coneixements necessaris i punts seguits en la implementació del projecte per a arribar als resultats obtinguts i que aquests siguin entenedors.

Per començar s'especifica si s'ha tingut en compte o no la perspectiva de gènere en aquest projecte i, si aplica, com afecta en el desenvolupament del treball.

Seguidament es descriu un marc teòric per a situar el lector en la matèria que es tracta i els fonaments bàsics per al seguiment del projecte.

En tercer lloc s'especifiquen els objectius a assolir en el llarg del projecte, tot incloent-hi les accions necessàries per assolir-los i indicadors per a la posterior avaluació dels mateixos objectius.

En quart lloc es descriu amb detall el funcionament del mòdul, des de l'elecció del model per a la predicció de la producció d'energia fins a l'algoritme de detecció i classificació d'anomalies.

En cinquè lloc s'estudien els resultats obtinguts del desenvolupament explicat en el punt anterior.

En sisè lloc es revisa la planificació feta per al desenvolupament del projecte, detallant les fases i tasques del projecte, així com la seva durada i les desviacions, en hores, de la planificació inicial, tot acompanyat d'un diagrama de Gantt.

En setè lloc es fa un estudi de l'impacte mediambiental que té un projecte que utilitza mínimament recursos externs i que es basa en la millora d'una energia renovable.

En vuitè lloc s'exposen les conclusions extretes del desenvolupament del projecte en el seu conjunt i se n'estudia l'assoliment dels objectius marcats en un inici.

Per últim s'inclou un estudi econòmic on s'exposa el pressupost econòmic del projecte i la seva viabilitat econòmica.

3 Perspectiva de gènere

Degut a la naturalesa pròpia del projecte, disseny d'un algorisme matemàtic, que serà integrat en la plataforma del producte final. Com ha tal, el resultat del projecte funciona de manera autònoma i, per tant, el client final del producte no hi té accés. Per tant, en aquest projecte no aplica la perspectiva de gènere.

4 Marc teòric

4.1 Recerca d'informació

4.1.1 Energia fotovoltaica

Des dels inicis de la Segona Revolució Industrial (s. XIX) fins a l'època actual, l'electricitat ha estat, i seguirà sent, l'eix motor de la societat. Cada nou producte es basa més que l'anterior en l'ús de l'electricitat i totes les facilitats, característiques, innovacions, etc. que l'ús d'aquesta aporta. D'aquesta manera, la necessitat de generar-ne augmenta, ja no només en la demanda industrial i per al sector terciari, també en les demandes residencials. És en aquest context, i en la cerca de nous models de generació d'energia elèctrica, en què l'energia fotovoltaica pren una gran importància.

Mètodes d'obtenció d'energia elèctrica i necessitat d'ús d'energies renovables

En funció de la font d'energia, podem diferenciar dues maneres d'obtenir l'electricitat, amb energies no renovables i amb energies renovables.

Entenem per energia no renovable aquella que s'obté a partir de fonts exhauribles i que, per tant, es consumeixen de manera més ràpida de la que es generen. Aquestes energies es coneixen per utilitzar combustibles fòssils i contribuir amb contaminants a l'entorn.

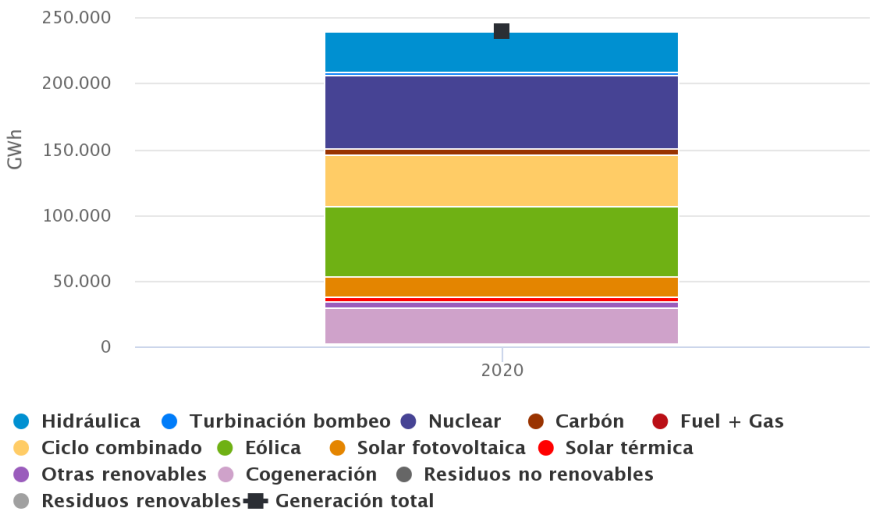

Figura 4.1 Repartició de la generació d'energia per tecnologia [1].

Les energies renovables són aquelles que, al contrari, utilitzen recursos naturals que es regeneren de manera immediata o quasi immediata, evitant així l'esgotament de les fonts. A més a més, aquestes energies no acostumen a deixar residus i, per tant, són més favorables per al medi ambient [1].

Durant el 2020, a la península espanyola, la generació d'aquest recurs s'ha repartit en un 45,5% renovable i un 54,5 % no renovable. En la figura 4.1 podem veure de quina manera s'han distribuït els diferents mètodes. Segons [2], això ha suposat un total de 29.538.819 tones de CO_2 emeses durant el darrer any amb una taxa de 0,12 tCO_2/MWh. Sense anar gaire més enrere, del 2018 al 2020, aquesta taxa ha minvant en gairebé la meitat. Aquesta baixada es deu al desús del carbó com a tecnologia per a generar electricitat. Cal notar que tot i que la reducció de l'ús del carbó és del 86%, la de la taxa és només del 46%, evidenciant que, encara que el consum baixi, la petjada mediambiental que deixa segueix sent gran. Per aquest motiu, és de gran importància reduir la despesa energètica de tecnologies no renovables, i passar a fer ús de les que si ho són.

Fonament físic i funcionament

En una instal·lació fotovoltaica es troben un seguit de panells solars fotovoltaics compostos, cada un, per diferents cel·les fotovoltaiques que, unides en sèrie, formen mòduls encapsulats. Alhora, cada cel·la està composta per dues capes de silici, una amb càrrega negativa (N) i una altra de positiva (P) formant un conjunt PN, que genera una diferència de potencial suficient perquè els electrons viatgin cap a una càrrega externa.

Quan un feix de llum, en aquest cas solar, "impacta" sobre la capa N de la cel·la, els fotons d'aquest feix provoquen que els electrons que es troben en la capa de valència dels àtoms del silici es desprenguin de l'àtom i puguin viatjar de manera lliure pel material, convertint així el silici en un semiconductor. En el cas de la banda P, en la capa de valència dels àtoms s'hi troben forats en els quals arriben els electrons que han perdut prou energia i tornen a formar part d'aquesta capa. És a dir, com es veu en la figura 4.2, quan un fotó arriba a la cel·la, impacta en un electró de la capa de valència d'un àtom, en fer això aquest electró se'n desprèn, deixant-hi un forat. Tant electró com forat són dominats pel camp que regenta la cel·la i són transportats en direccions contràries, on generen un corrent continu elèctric.

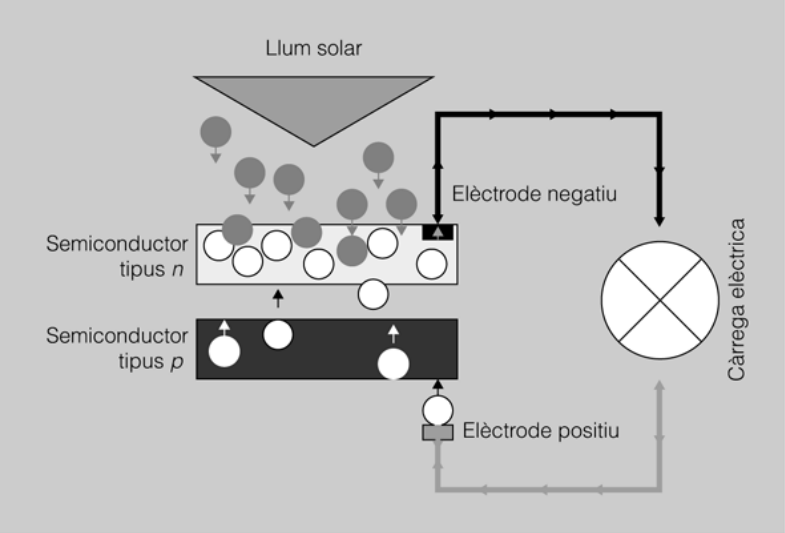

Figura 4.2 Funcionament d'una cel·la fotovoltaica [1].

Cal tenir en compte que, a major radiació solar, major el nombre de fotons i, per tant, major el nombre d'electrons que es desprenen i "viatgen" per generar electricitat. Tenint en compte aquest fet, i sabent que a major nombre d'electrons, major corrent, interessa que les plaques solars estiguin orientades de manera que puguin aprofitar la màxima radiació solar. És per això que les plaques s'orienten, a l'hemisferi nord, a sud, est o oest (en més o menys mesura) i amb una inclinació variant respecte el terra segons la latitud en la qual es troba la instal·lació, per a Catalunya, la inclinació òptima és de 37º.

A part dels mòduls, la instal·lació està formada pel BOS (*Balance of System*), que inclou la resta del sistema, està compost per l'inversor o alternador, encarregat de transformar el corrent DC que genera el mòdul fotovoltaic a corrent AC, útil per a l'ús domèstic, les bateries, en cas que fossin necessàries, la instal·lació elèctrica, suports per a la inclinació dels mòduls, etc. [3]

4.1.2 Teoria de models, predicció i detecció d'anomalies

Què és la modelització

La modalització és l'aplicació de models matemàtics numèrics a una representació analítica d'un sistema per tal de representar-lo teòricament, aconseguir-ne les característiques i poder ser estudiat sense tenir disponibilitat del sistema real. És a dir, tenint un sistema qualsevol, una biga en una casa per exemple, es pot crear un model matemàtic analític per tal de resoldre quines són les tensions màximes que pot suportar la biga. Si a aquest model analític, que

normalment serà prou complex per a poder-ne trobar una solució només analíticament, li apliquem un model matemàtic numèric, que en general és més complex, podem dir que estem fent la modelització del nostre sistema principal, la biga.

Què és una sèrie temporal?

Segons [4] les sèries temporals són un conjunt d'observacions x_t, que es graven en un instant de t concret. Es pot diferenciar entre sèries temporals discretes i contínues: en el primer tipus, que és el què es tracta en aquest projecte, les mostres estan agafades de manera discreta, com per exemple, amb una freqüència equivalent a un període de temps determinat. Les segones s'obtenen quan les observacions es registren durant un període de temps continu.

Una gran majoria de sèries temporals, es poden descompondre en tres factors (figura 4.3): tendència, estacionalitat i soroll.

$$X_t = T_t + E_t + S_t \tag{4.1}$$

La *tendència* (T_t) fa referència a com es mou el conjunt de dades en el temps, es poden donar tendències constants, creixents o decreixents. L'*estacionalitat* (E_t) descriu els moviments d'oscil·lació que es poden trobar en un mateix període de temps, per exemple les hores de sol durant una setmana. El *soroll* (S_t) es defineix com a punts aleatoris al voltant de les definicions anteriors. En els casos que la sèrie està definida per aquests factors és recomanable descompondre-la per a treballar-hi.

Algorismes *Machine learning*

A priori podem definir tres tipus d'algorismes diferents: aprenentatge supervisat, aprenentatge no supervisat i aprenentatge reforçat.

Un *aprenentatge supervisat* es dóna quan, a l'entrada del model, s'especifiquen exemples de dades d'entrada i els seus resultats. D'aquesta manera s'ensenya al model quina ruta ha de seguir per a trobar els resultats esperats donats un seguit d'entrades, i es segueix fent fins que el nivell d'*accuracy* (precisió) és acceptable per a les dades treballades. En aquests tipus de models es separa la base de dades entre dades d'entrenament i dades de validació [5].

L'*aprenentatge no supervisat* no té com a objectiu trobar un resultat a partir d'unes dades d'entrada, sinó clusteritzar, agrupar les dades en diferents grups de característiques semblants [5].

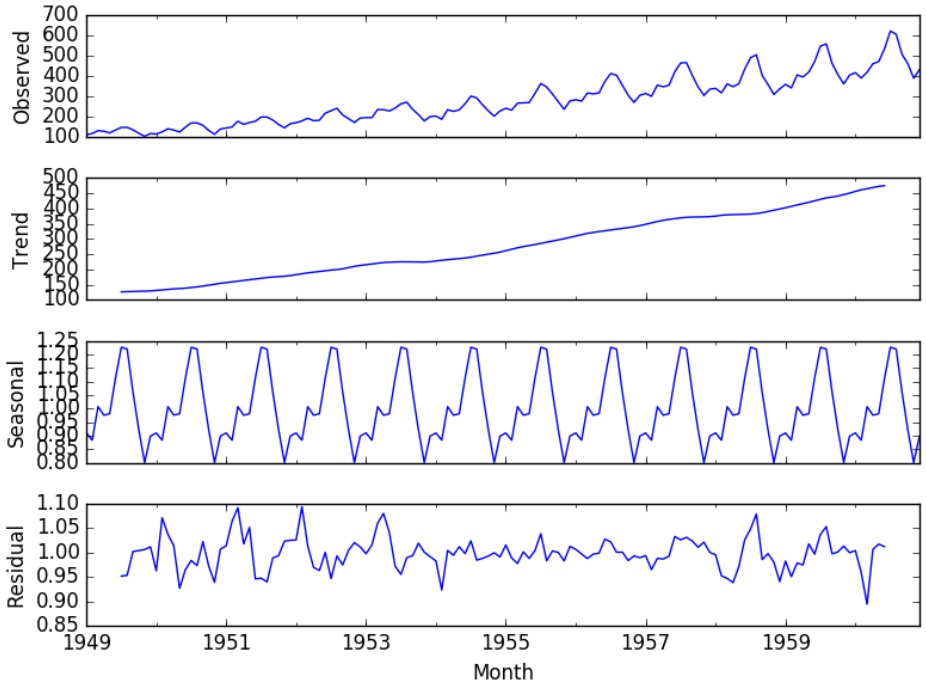

Figura 4.3 Descomposició d'una mostra observada en tendència, estacionalitat i soroll [6].

En l'*aprenentatge reforçat* el model es desenvolupa en un entorn en el qual ha d'obtenir un objectiu global. Aquest model va aprenent de situacions anteriors i és capaç de prendre decisions que milloraran el seu comportament davant un nou problema. Un exemple d'aquest tipus d'aprenentatge són els cotxes de conducció autònoma, que han de conduir sense xocar contra altres vehicles, persones o objectes [5].

L'objecte d'aquest projecte defineix un model del primer tipus i, com a tal, només s'exposen alguns dels models dins l'aprenentatge supervisat. Aquests són:

- **Arbre de decisions:** en anglès *Decision Tree*, és un algorisme que classifica les dades d'una part de la base de dades, les d'entrenament, seguin una classificació en arbre. Tal com es pot veure en la figura 4.4, a cada nivell de l'arbre es pren una decisió que classifica i divideix les dades fins a obtenir-ne els grups finals [5].
- **Random Forest:** es tracta d'un pas més enllà dels arbres de decisions. En aquest "bosc" es generen diferents arbres de decisions. Per tal de classificar les dades

rebudes, aquestes passen per tots els arbres i obtenen, en cada un, una puntuació. Aquella puntuació que sigui més alta és la que classificarà la dada segons en quin arbre hagi estat obtinguda [5].

- **k-Veïns Propers (kNN):** o en anglès *k-Nearest Neighbours*. Té en compte totes les classes de dades (per exemple valors) que es poden donar en la base de dades i, segons els k veïns més propers al cas ha estudiat, decideix a quina classe li correspon. Si el nombre k és igual a 1, aleshores al cas d'estudi se li assigna la mateixa classe que a aquest. La distància entre el cas a estudiar i els veïns es pot estudiar amb diferents funcions matemàtiques, Euclidiana, Manhattan, Minkowski i Hamming [5].

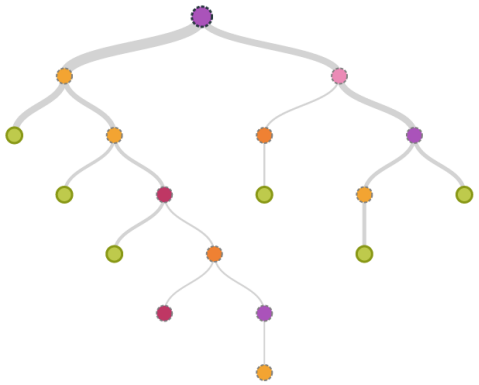

Figura 4.4 Dibuix esquemàtic d'un arbre de decisions.

- **Xarxa Neuronal (NN):** són un tipus de model basat en la interconnexió de les neurones del cervell humà. Es divideixen diferents nodes en als quals els arriben uns inputs, aleshores, fent algun tipus de funció matemàtica, s'operen tots aquests inputs i se n'extreu una sortida. S'estructuren en tres capes, una d'inputs, una capa amagada, o *hidden layer*, i una d'output i un seguit de pesos, *weights*, entre capa i capa (veure figura 4.5) [6]:

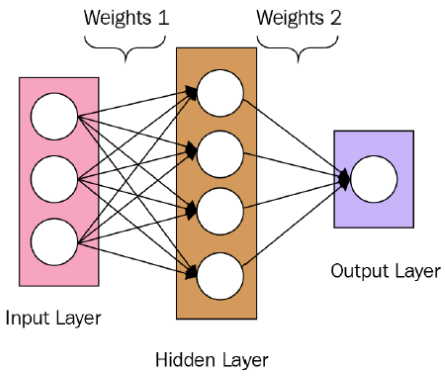

Figura 4.5 Estructura interna d'una xarxa neuronal [6]

Per poder triar quin és el mètode més efectiu i el que prediu millor, en aquest cas, la producció fotovoltaica, cal tenir un indicador capaç de puntuar com s'ajusta el model a la realitat. Per fer això podem implementar el càlcul de diferents errors matemàtics:

- **MAE:** És calcula fent la mitja de tots els errors absoluts. [6]

$$MAE = 1/n \sum_{t=1}^{n} |x_t - x| \tag{4.2}$$

On *n* són el nombre d'errors i $|x_t - x|$ és l'error absolut.

- **MAPE:** És la suma dels errors absoluts individuals entre la mitja de tots els valors. Es calcula de la següent manera [7]:

$$MAPE = 100/n \sum_{t=1}^{n} |y_t - p_t/\bar{y}_t| \tag{4.3}$$

On y_t és el valor actual p_t és el valor predit i $\hat{y}_t$ és la mitja de tots els valors.

- **R^2 *score*:** També conegut com a coeficient de determinació, defineix segons Walker Rowe, [8], la proporció de la variància de la variable dependent que es pot predir a partir de la variable independent.

$$R^2 = 1 - SS_{RES}/SS_{TOT} \tag{4.4}$$

On SS_{RES} és la suma dels quadrats dels residuals de i SS_{TOT} és la suma total dels quadrats.

- **RMSE:** Calcula la desviació estàndard dels residuals, aquests mesuren com de lluny es troba la mostra respecte a la línia de regressió del model [11]:

$$RMSE = \left[\sum_{i=1}^{N} \left(z_{fi} - z_{oi}\right)^2 / N\right]^{1/2} \tag{4.5}$$

On (z_{fi} - z_{oi}) són les diferències i *N* és el nombre de mostres.

Detecció d'anomalies

Per a decidir si una dada és o no, una anomalia, no n'hi ha prou en assenyalar aquells valors que s'allunyen de la tendència de les mostres, sinó que cal quantificar aquesta distància. Aquest càlcul acostuma a ser un múltiple de la desviació estàndard de la diferència entre el valor real i el predit de les dades, en escènica, es calcula la diferència de cada valor predit

contra el real i, d'aquest conjunt de mostres, se'n mira la desviació estàndard i es multiplica per un factor. El resultat de fer aquesta operativa marcarà el límit el qual si és superat per un valor, es considerarà que aquest és una anomalia.

Cal notar, però, la diferència entre els tipus d'anomalies, que poden ser puntuals, col·lectives o contextuals.

Les *anomalies puntuals* són aquells punts, és a dir, punts aïllats del conjunt de dades, que es desvien de manera significant de la resta [10].

És defineixen *anomalies col·lectives* a aquelles que, per si soles no són una anomalia, però la repetició consecutiva del mateix valor sí ho és. Per exemple, si un client d'un banc retira 600€ del banc en un caixer, pot ser una transacció corrent, però, si repeteix aquesta mateixa acció en diferents caixer durant el mateix dia, es considera una anomalia [10].

Per últim, les *anomalies contextuals* fan referència a aquells punts que, depenent de la situació i context de les dades en general, es consideraran anomalies o no [10].

5 Especificacions tècniques

5.1 Objectius del projecte

A continuació s'especifiquen els objectius del projecte, les accions que comporten i els seus indicadors per a poder ser avaluats posteriorment.

1) Desenvolupar un model matemàtic de la producció d'una instal·lació fotovoltaica.
   a) Condicionar la base de dades.
      i) La base de dades no té buits, totes les dades necessàries són completes i les no necessàries o incompletes no hi són.
   b) Entrenar el model de manera correcta.
      i) S'utilitzen les dades mínimes necessàries per entrenar el model
      ii) Les dades tenen en compte la importància de l'ordre i del mostreig en una sèrie temporal
   c) Incrementar la precisió del model.
      i) El model té en compte més paràmetres que sols la producció, contempla potència de pic, inclinació, orientació, tecnologia i radiació.
2) Detectar les diverses anomalies que es produeixen en la producció respecte al model.
   a) Implementar un algoritme predictiu capaç de distingir entre anomalies i hores de no producció.
      i) El model diferencia entre les hores en les quals és natural que no hi hagi producció, és a dir, des del capvespre fins a l'albada, i les hores en les quals no es produeix per algun motiu diferent.
3) Classificar les anomalies en diferents grups per a un possible manteniment preventiu futur.
   a) Diferenciar entre anomalies puntuals i temporals, i dins d'aquestes segones, degradacions i pèrdues de producció sostingudes.
4) El codi ha de ser compatible amb altres mòduls de l'eina global.
   a) Escriptura de variables, dades, funcions, etc. comú entre els diferents mòduls.
      i) Les dades, funcions i variables, tant d'entrada com de sortida, són admeses i cridades per altres mòduls.

6 Mòdul de detecció d'anomalies

L'entorn de test del mòdul es dividirà en les següents etapes:

- **Comprovació de les dades d'entrada:** Cal que, abans de fer els models, es comprovi que la informació que s'està donant d'entrada al model està ordenada d'una manera adequada per al seu processat. Com que es treballarà amb més d'una base de dades, cal que les dades i com estan etiquetades sigui igual per a tots els sets de dades.
- **Testeig del model:** Es comprova, utilitzant les dades d'entrada, configuracions diferents dels paràmetres intrínsecs de cada model per a trobar-ne el més eficient.
- **Calibració del model:** Utilitzant els paràmetres òptims trobats anteriorment, es calibra un model i, si s'obté un resultat adequat en totes les instal·lacions, es guarda aquell model com a vàlid.
- **Detecció d'anomalies:** Es carrega el model guardat en el pas anterior, es prediu la producció de la instal·lació i se'n detecten les anomalies.
- **Comprovació dels resultats obtinguts:** S'analitzen les anomalies que s'han detectat, si són o no anomalies, i si s'han classificat correctament. Si els resultats no són satisfactoris es revisaran les etapes anteriors i es comprovarà que les especificacions adjudicades al model i l'algorisme de detecció d'anomalies són correctes.

6.1 Dades d'entrada

L'entrada de dades del mòdul està basada en bases de dades on consta, entre altra informació, la generació d'energia solar fotovoltaica.

6.1.1 Reestructuració de les bases de dades

Les bases de dades són pertanyents a quatre localitzacions, Austin, ciutat de Nova York i ciutat de Barcelona. Les dades dels Estats Units d'Amèrica s'han obtingut a través del portal *Pecan Street* [12] i les dades de Barcelona han estat obtingudes a través del Consorci de l'Agència d'Energia de Barcelona (AEB), departament de l'Ajuntament de Barcelona.

	dataid	local_15min	air1	...	winecooler1	leg1v	leg2v
0	661	21/11/2018 15:15	0.0	...	NaN	123.915	124.277
1	661	21/11/2018 15:30	0.0	...	NaN	123.959	124.293
2	661	21/11/2018 15:45	0.0	...	NaN	123.886	124.240
3	661	21/11/2018 16:00	0.0	...	NaN	123.880	124.175
4	661	21/11/2018 16:15	0.0	...	NaN	123.633	124.226

Taula 6.1 Primeres cinc files de la base de dades i algunes columnes de la base de dades d'Austin.

	Source	Location ID	City	...	Fill Flag 5	Surface Albedo Units	Version
0	NSRDB	399337	-	...	Rayleigh Violation	NaN	unknown
1	Year	Month	Day	...	NaN	NaN	NaN
2	2018	1	1	...	NaN	NaN	NaN
3	2018	1	1	...	NaN	NaN	NaN
4	2018	1	1	...	NaN	NaN	NaN

Taula 6.2 Primeres cinc files de la base de dades i algunes columnes de la base de dades de Barcelona.

Com que les dades d'Amèrica i les de Barcelona no estan extretes de la mateixa font, és d'esperar que les dades obtingudes, la manera d'anomenar-les i la seva ordenació no coincideixin. En les taules 6.1 i 6.2 es pot veure una mostra, les primeres cinc files i algunes columnes, de la base de dades d'Austin i Barcelona respectivament. La base de dades que es modifica és la catalana on les columnes es defineixen en funció de les columnes definides a les bases de dades americanes.

No totes les dades que es faciliten en la base de dades són estrictament necessàries, per exemple, no és interessant saber si hi ha més d'un cotxe aparcat al recinte o si existeix un o dos banys, per això cal tractar aquestes dades i seleccionar-ne només les que siguin de vital importància per a la tasca que es vol fer. En aquest cas interessa saber quina és la generació de la instal·lació en cada hora disponible, per tant, es modifica l'estructura de les dades per a que, en un índex temporal es mostri la generació elèctrica de cada instal·lació.

	661	1642	2335	2361	...	8156	9019	9160	9278
1/1/2018 0:00	NaN	-0.005	-0.006	-0.008	...	-0.006	-0.008	NaN	-0.004
1/1/2018 0:15	NaN	-0.005	-0.007	-0.008	...	-0.006	-0.008	NaN	-0.004
1/1/2018 0:30	NaN	-0.003	-0.007	-0.009	...	-0.006	-0.008	NaN	-0.004
1/1/2018 0:45	NaN	-0.005	-0.007	-0.008	...	-0.006	-0.007	NaN	-0.004

Taula 6.3 Capçalera de les dades reordenades.

Com es pot observar en la taula 6.3 l'índex ha estat modificat respecte la taula 6.1 a la data i hora en qüestió (iniciant-se en el primer dia de l'any) i les columnes han passat a ser l'identificador numeral de la instal·lació en concret.

6.1.2 Filtratge de les dades

En la taula 6.4 es mostra un resum de la quantitat d'instal·lacions que es tenen de cada ciutat, la quantitat de punts que contenen entre totes les ciutats, el nombre de valors en blanc (NaN) i el seu percentatge i el nombre de valors anòmals o *outliers*.

	Núm. inst	Núm. punts	Núm. NaN	% NaN	Núm. *Outliers*	% *Outliers*
Barcelona	32	3363840	963252	28,64	5776	0,17
Austin	20	700720	40731	5,81	1	0,00
Nova York	14	247296	1	0,00	0	0

Taula 6.4 Resum del nombre d'instal·lacions i les seves característiques per ciutat.

Tal i com es presenten les dades, amb els buits i *outliers* que hi ha presents, la base de dades és inservible i, per tant, s'hi ha d'aplicar canvis per a que el resultat sigui més adequat.

Com es veu en la taula 6.3, hi ha valors negatius que, per a generació d'energia, no tenen sentit, per tant, mitjançant una simple manipulació s'indica que tots els valors negatius es marquin com a 0 (notis la diferència entre la taula 6.3 i la taula 6.5).

	661	1642	2335	2361	...	8156	9019	9160	9278
1/1/2018 0:00	NaN	0.0	0.0	0.0	...	0.0	0.0	NaN	0.0
1/1/2018 0:15	NaN	0.0	0.0	0.0	...	0.0	0.0	NaN	0.0
1/1/2018 0:30	NaN	0.0	0.0	0.0	...	0.0	0.0	NaN	0.0
1/1/2018 0:45	NaN	0.0	0.0	0.0	...	0.0	0.0	NaN	0.0

Taula 6.5 Capçalera un cop els negatius han estat canviats a 0.

En segon lloc s'eliminen els punts que superin tres vegades la desviació estàndard de les dades de la mateixa instal·lació a la que pertanyen, d'aquesta manera els valors que, de per si són anòmals, és a dir, que s'allunyen en excés de la tendència de les dades, són eliminats.

Seguidament però, s'eliminaran aquelles instal·lacions en les quals el valor de buits sigui major o igual a un 10% de les dades totals. S'ha decidit que el tall sigui un 10% degut a la composició de les instal·lacions, en la taula 6.6 es pot veure el percentatge de NaN per les 13 primeres instal·lacions per ciutat.

Ciutat / Instal·lació	0	1	2	3	4	5	6	7	8	9	10	11	12
Barcelona	72	0	0	0	61	0	43	0	17	39	22	19	61
Austin	1	2	2	0	0	1	0	0	0	1	0	100	0
Nova York	0	0	0	0	0	0	0	0	0	0	0	0	0

Taula 6.6 Percentatge de NaN per les primeres tretze instal·lacions de cada ciutat.

Un cop s'han modificat totes les dades, el nombre d'instal·lacions per ciutat, i el nombre de punts, es redueix (notis la diferència entre la taula 6.4 i 6.7).

	Núm. inst	Núm. punts	Núm. NaN	% NaN	Núm. *Outliers*	% *Outliers*
Barcelona	8	840960	11884	1,41	600	0,07
Austin	18	630648	30073	4,77	0	0
Nova York	14	247296	11087	4,48	0	0

Taula 6.7 Resum del nombre d'instal·lacions i les seves característiques per ciutat un cop fets els canvis adients.

6.1.3 Imputació de dades

Els models de predicció que s'utilitzaran més endavant no poden treballar amb valors en blanc, és a dir, cal substituir-los o eliminar-los depenent de les característiques de les dades. A fer aquesta substitució dels valors NaN amb un valor conegut se'n diu imputar un valor. Aquesta imputació pot ser de diferents tipus, entre els quals, imputació d'una constant (un 0 en aquest cas) o de la mitjana. Aquestes dues imputacions s'han cregut adients per a les dades que es tenen ja que si el forat d'informació es troba durant les hores que no hi ha sol, imputar-hi un 0 és el valor més adequat, per altra banda, si el forat es troba durant el dia, imputar-hi la mitjana de la generació propera a aquell buit pot ser un bon indicador. Tot i que s'ha valorat personalitzar la imputació, en certes hores del dia utilitzar la mitjana i en altres una constant, la diferència de l'error, tal i com es veu en el capítol 5. Resultats i Anàlisi, entre ambdós *imputers* no és prou significativa per justificar aquesta personalització, ja que no la incorpora el mateix software i s'hauria d'adequar a aquest cas concret.

6.1.4 Escalat de les dades

Quan en aquest projecte s'usen xarxes neuronals per a la predicció de dades, les variables d'entrada estan presentades en diferents unitats que, en si, pot suposar una escala diferent per a cada variable [13].

Si s'utilitzen aquestes dades amb diferents escales, la dificultat del problema augmenta, portant cap a un major temps de predicció, a una inestabilitat del model i a una predicció poc acurada. Per tal d'evitar aquest error les dades s'escalen, de manera que totes es moguin dins un mateix rang [13].

Es poden utilitzar diferents escaladors o *scalers*, en el cas que ocupa se'n compararan quatre de diferents:

- ***Standard Scaler*:** Es basa en la desviació estàndard segueix la formula bàsica (6.1) de l'escalat de dades. No té perquè situar les dades escalades entre 0 i 1 [13].

$$y = \frac{x - mitjana}{desviació\ estàndard} \tag{6.1}$$

On x són els valors d'entrada, la mitjana és la mitjana de les dades d'entrada i la desviació estàndard també és respecte les dades d'entrada.

- ***Min Max Scaler*:** Posiciona les dades escalades entre 0 i 1, o entre -1 i 1 si en les dades d'entrada hi ha algun nombre negatiu [13].
- ***Max Abs Scaler*:** En lloc d'utilitzar la desviació estàndard, fa us del valor absolut màxim [13]:

$$y = \frac{x - mitjana}{valor\ absolut\ màxim} \tag{6.2}$$

- ***Robust Scaler*:** Aquest tipus d'*scaler* opera en dos passos, en un primer elimina els *outliers* que es puguin trobar entre les dades i, seguidament, utilitza un dels escaladors anteriors. És lògic esperar que aquesta sigui la millor opció per a usar en l'algorisme [13].

6.2 Predicció de la generació d'energia

La predicció de la generació s'ha fet amb dos mètodes diferents.

El primer (M1) es basa en l'article [8], el qual defensa que el comportament de tota instal·lació solar fotovoltaica pot estar predit per a les cinc instal·lacions, de característiques semblants, més properes a la primera. És a dir, que coneixen la producció d'energia de cinc instal·lacions fotovoltaiques, se'n pot predir la generació d'una sisena de característiques

semblants si aquestes no estan gaire allunyades entre si. Per a aquest mètode es comparen dos models de regressió, el *Random Forest* i el *k-NN* per triar-ne el més eficaç.

El segon mètode (M2) n'és un de més convencional on, usant les dades meteorològiques d'una localització i la potència màxima de la instal·lació a predir, es defineix el model i, posteriorment, es prediu la producció en la instal·lació desitjada. Per al desenvolupament d'aquest mètode s'utilitzarà una xarxa neuronal amb diferents *solvers*, triant-ne a posterior el més adequat per a la tasca a realitzar.

6.2.1 Models de predicció

La diferència de tots tres models recau en la manera de calcular la predicció, però es basen en la mateixa metodologia: la base de dades es divideix entre *X*, dades punter i *y*, dades objectiu. Les dades punter són les que, quan són entrades en el model, s'analitzen per donar una sortida de dades el més semblant a les dades objectiu possible. Seguidament, s'entrena el model amb una selecció de les dades, tant *X* com *y* (*X_train* i *y_train*), per a que creï una relació entre aquestes. I, finalment, es valida, amb la resta de dades que no s'han usat per a l'entrenament (*X_valid* i *y_valid*), que el model predigui unes dades de sortida a partir d'unes dades d'entrada que mai ha vist i se'n calcula l'error comès en la predicció respecte les dades reals de validació. En la figura 6.1 es mostra un esquema del funcionament explicat.

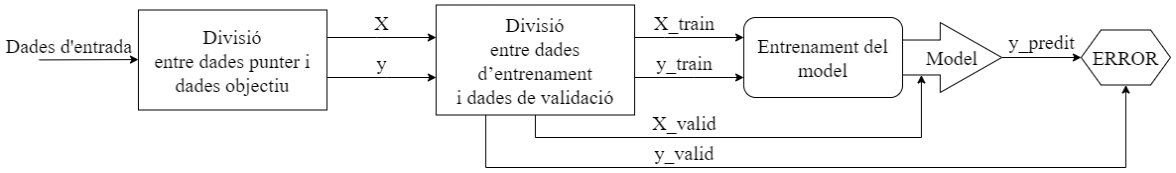

Figura 6.1 Diagrama del funcionament dels models de predicció utilitzats.

- **Divisió de les dades**

Com s'ha dit, la base de dades de la qual es parteix, s'ha de dividir en dues mostres diferents: les dades punter i les dades objectiu. En la figura 6.2 es pot veure un exemple de com es divideixen aquestes dades.

Les **dades objectiu (*y*)** són aquelles que es volen predir en el model, és a dir, a les quals les dades de sortida del model s'han d'assemblar el màxim possible. En ambdós mètodes descrits, M1 i M2, les dades objectiu són les quals ofereixen la informació sobre generació

elèctrica fotovoltaica d'una instal·lació en concret. Per tant, per triar-les de la resta de la base de dades, tan sols cal indicar sobre quina instal·lació es vol treballar.

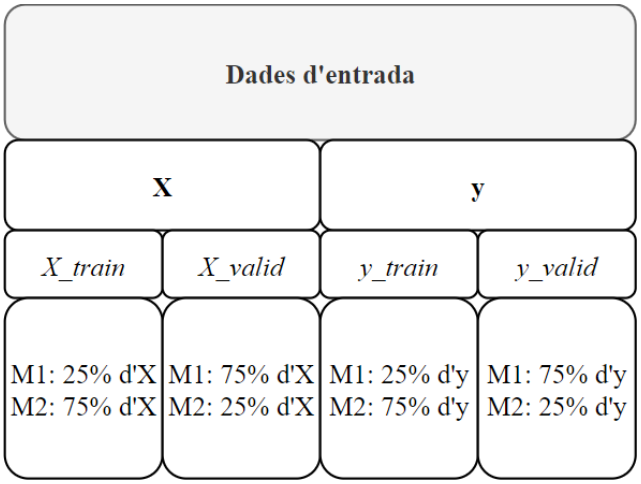

Figura 6.2 Divisió de les dades d'entrada.

Les **dades punter (*X*)** són aquelles que al ser donades d'entrada al model, s'utilitzen per a fer les prediccions, per tant, es pot dir que són les dades d'entrada que assenyalen quina és la dada de sortida que pertoca. En el cas de l'M1, les dades punter són totes aquelles que no són objectiu, és a dir, totes les dades de generació elèctrica fotovoltaica de la resta d'instal·lacions que no es volen predir. En el cas de l'M2 les dades punter són aquelles que fan referència a la meteorologia de la localització de la qual s'estan usant les dades objectiu.

Un cop s'han separat les *X* i les *y*, cal dividir aquestes, un cop més, entre dades d'entrenament i dades de validació cada grup, tant *X* com *y*.

Les **dades d'entrenament (_train)** són aquelles que permeten entrenar el model per a que, després, sigui capaç de predir un resultat des de qualsevols dades d'entrada. Són un extracte d'*X* i d'*y* que creen una relació entre les dades punter i objectiu. En el mètode 1, com que la quantitat de dades és gran i la seva correlació és alta s'usen un 25% de les dades per a entrenar el model, en canvi, per al mètode de la xarxa neuronal, com que les dades que s'utilitzen són menys i amb una correlació més baixa, s'utilitzen un 75% de les dades totals per a l'entrenament.

Les **dades de validació (_valid)** són les que s'utilitzen per a comprovar la fiabilitat del model generat. S'utilitza la par d'*X* i d'*y* que no s'ha destinat a les dades d'entrenament i, quan el model ja ha estat entrenat, se li donen d'entrada aquestes dades *X* de validació per a

que en faci una predicció. Aquesta, aleshores, es compararà amb les dades *y* de validació per observar la resposta i la precisió del model.

En aquest punt de divisió de les dades, i degut a la component temporal cíclica que caracteritza les dades, cal pensar que, al final, el que s'analitza és la radiació solar entorn l'any i, com és sabut, la radiació del mes de juliol no és la mateixa que la del mes de gener, per tant, és imprescindible decidir si les dades d'entrenament comprendran les dades de tres mesos seguits, per exemple, o diferents mostres repartides al llarg de l'any i si aquestes dades es presentaran ordenades cronològicament o no.

També cal fer mencionar que, com s'està tractant amb una sèrie temporal, l'ordre en el qual estan presentades les dades i en el qual es separaran és vinculant als resultats que es puguin obtenir un cop el model s'hagi entrenat i validat ja que, com s'ha dit, cada valor en la producció fotovoltaica va estretament lligat al moment de l'any en que es pren.

- **Creació i entrenament del model**

Per a que un model predigui no només calen unes dades d'entrenament, també cal definir diferents paràmetres, no iguals per tots els models, per a que aquest pugui obtenir una bona resposta sense sobreajustar-lo.

El sobreajustament, o *overfitting* en anglès, es dona quan un model ha estat entrenat amb massa dades i, com a resultat, ofereix una predicció molt semblant a les dades objectiu de validació. Un model sobreajustat és excel·lent en reproduir la resposta per a les dades amb les quals ha estat entrenat però, en usar-lo amb altres dades, falla, degut a que té en compte, no només la tendència general de les dades, sinó que també el soroll o les fluctuacions aleatòries que té les dades d'entrenament [12].

L'*overfitting* es pot controlar de diverses maneres, entre elles, utilitzant uns bons paràmetres en la creació del model i triant un percentatge prou baix de dades d'entrenament que sigui, a la vegada, prou alt per fer una bona predicció. La tria d'aquests paràmetres i percentatges no està lligada a cap norma escrita i, per tant, cal experimentar amb cada model per a trobar quin és el més adequat.

Tant el *Random Forest*, el *kNN* i la xarxa neuronal són models els quals ja existeix una llibreria Python on estan descrites les funcions i, per tant, s'usaran aquestes per al

procediment. Cal notar que totes tres funcions contenen diferents paràmetres que ja estan definits amb anterioritat, la majoria d'aquests paràmetres no es modificaran.

- **Validació del model**

Un cop el model ja ha estat entrenat, tan sols falta fer la predicció de les dades punter de validació i comparar el resultat amb les dades objectiu de validació.

Per fer la validació i tenir un criteri objectiu amb el qual mesurar la precisió dels models, s'ha utilitzat diferents càlculs d'error en cada model, on es compararà cada valor predit amb el corresponent real. Aquest valor predit no només es calcularà d'una, sinó de totes les instal·lacions de les quals es disposa, degut a que el model escollit per a la predicció podria ser molt favorable en una de les instal·lacions i no ser adequat en cap de les altres, evitant així l'*overfitting* del model.

6.2.2 Pipeline

Un *pipeline*, o canonada en català, és un sistema de programació que permet organitzar en compartiments el preprocessat de les dades i la modelització d'aquestes. Tot i que és possible generar models sense *pipelines*, aquests aporten diferents avantatges: [14]

- **Codi més net:** aplicant aquest sistema de canonades es redueix la quantitat de línies que conté el codi i, per tant, es veu un codi més net i organitzat. [14]
- **Menys *bugs*:** partint de l'avantatge anterior, com que el codi té menys línies i està millor organitzat, els possibles errors que s'hagin comés durant la programació són més fàcils de trobar i, a la vegada, més difícils de que succeeixin. [14]
- **Facilitat de producció:** tenir un sistema *pipeline* facilitarà la posada en marxa en un entorn de producció on la quantitat de les dades obtingudes sigui d'una magnitud major a la de l'entorn de proves, facilitant també l'intercanvi de mètodes de predicció. [14]

Pel que fa la implementació, es parteix de les dades dividides entre dades d'entrenament i de validació, tant d'X com d'y. Al contrari del que es faria en altres casos, aquestes dades no es processen en quan a omplir forats, es farà dins de la mateixa *pipeline*. El que si cal, és definir una sèrie de variables en les quals es notifiquin les transformacions que s'han d'aplicar i el model que es vol utilitzar:

La **transformació numèrica** fa referència a la imputació de valors als espais buits de la base de dades. S'assigna a una variable el tipus de transformació que es vol fer (*numerical_transformer*) i, en una altra, a quines columnes de les dades es vol aplicar (*numerical_cols*), d'aquesta manera es pot definir un tipus de transformació, per exemple una constant igual a 0, a un seguit de columnes i, a unes altres, una transformació diferent, com una mitjana de les dades properes. Com s'ha dit hi ha diferents tipus d'imputacions però, per al cas que ocupa, no són vàlides i, per tant, no es mencionen.

En les següents línies de codi es defineix la variable *numerical_transformer* amb un *imputer* d'una constant igual a zero i la variable *numerical_cols* amb les columnes que conté el fitxer de dades punter X.

```
numerical_trasformer = SimpleImputer(strategy="constant", fill_value=0)
numerical_cols = [j for j in range(X_train.columns.size)]
```

Seguidament cal muntar el **preprocessador**. Aquest fa ús de la funció *ColumnTransformer* en la qual s'estableix cada transformació que s'ha configurat amb anterioritat. És en aquest punt que es defineix quina transformació aplicarà a cada columna en concret. És defineix:

```
preprocessor = ColumnTransformer(
            transformers = [("num", numerical_transformer, numerical_cols)])
```

Si hi hagués altres transformadors a aplicar, per exemple, que una columna en específic en lloc d'aplicar-hi un *imputer* constant hi imputéssim la mitjana, caldria crear una altra funció com la primera i afegir-la al preprocessador, per exemple:

```
# Imputer de la constant
constant_trasformer = SimpleImputer(strategy="constant", fill_value=0)
constant_cols = [0, 1, 3, 4, 5]
# Imputer de la mitjana
mean_trasformer = SimpleImputer(strategy="mean")
mean_cols = [2]
# Preprocessador amb els imputers de la constant i la mitjana
preprocessor = ColumnTransformer(
            transformers = [("const", constant_transformer, constant_cols),
                            ("mean", mean_transformer, mean_cols)])
```

La definició del **model** segueix la mateixa lògica que la transformació numèrica. Es crea una variable a la qual s'assigna un model de regressió amb les seves característiques.

```
model = RandomForestRegressor()
```

Finalment es fa **creació** del *pipeline* i la seva **predicció**. Utilitzant la funció *Pipeline*, es defineixen els passos a seguir amb les dades d'entrada, és a dir, s'indica l'ordre dels passos, primer el preprocessador i, seguidament el model. Definida la funció, tan sols cal entrenar-la amb les dades d'entrenament *X_train* i *y_train* (fer els processos previs a la predicció) i predir les dades amb els valors de validació *X_valid*.

```
my_pipeline = Pipeline(steps=[("preprocessor", preprocessor),
                              ("model", model)])
my_pipeline.fit(X_train, y_train)
prediction = my_pipeline.predict(X_valid)
```

Com es pot comprovar, és un codi molt compacte, s'han d'escriure poques línies, amb un funcionament bastant senzill i fàcil d'aplicar. Si aquesta mateixa operativa es fes sense el *pipeline*, no només suposaria més línies de codi, també més variables intermèdies, on guardar, per exemple, les diferents columnes on imputar amb diferents mètodes, aquestes columnes ja imputades, una altra variable amb tot el conjunt de dades un cop s'ha fet la imputació, etc.

6.3 Detecció d'anomalies

6.3.1 Factors transitoris i anomalies

Els diferents factors que afecten a la generació d'energia fotovoltaica es poden resumir en la quantitat de radiació solar que rep cada placa fotovoltaica i defectes en el *hardware*. El segon inclou, entre altres, defectes en els mòduls solars, en inversors, en connexions cablejat, etc. Pel que fa el primer factor, la radiació solar, es pot veure afectada per una gran quantitat de variables com ara un cel ennuvolat, pols, el dia de l'any o ombres d'edificis propers, arbres, faroles, etc [8]. Tots aquests factors es poden dividir en dues classes: factors transitoris i anomalies.

Els **factors transitoris** es caracteritzen per tenir un efecte temporal més aviat curt (per exemple el cel tapat per núvols). Aquests es poden dividir, un cop més, entre factors *comuns* i factors *locals*. Els comuns afecten a tota una mateixa regió a la vegada i, per tant, totes les instal·lacions en aquesta àrea veuran reduïda la seva generació, un exemple és, altre cop, un dia tapat. L'efecte que tindrà aquesta condició meteorològica farà variar la generació d'energia a mesura que vagi canviant. En canvi, els factors locals tan sols afecten a una

instal·lació en concret, com pot ser l'ombra d'un edifici proper que farà més o menys ombra depenent de l'hora del dia. Aquests factors transitoris no es poden modificar, al menys no fàcilment, en contraposició a les **anomalies**, com poden ser pols o deposicions d'ocells, que es poden solucionar [8].

6.3.2 Algorisme de detecció

Obtinguda la predicció d'energia per a cada instal·lació, es segueix el mètode que es veu en la figura 6.3, ones poden observar els diferents passos per a la detecció d'anomalies. Primer s'aconsegueix la diferència entre el valor real (y) i predit (y_t), d'aquesta resta se n'obté una sèrie temporal en la qual coexisteixen anomalies i factors transitoris (L_t).

$$L_t = y - y_t \tag{6.3}$$

Com que tan sols interessa tenir les anòmales, els factors transitoris s'han d'eliminar. Per a fer-ho, tenint en compte que aquests factors no canvien ràpidament en el temps, es descompon temporalment la sèrie de valors reals amb una freqüència d'una setmana, d'aquesta manera, tots aquells valors que es repeteixen en el temps són els que s'eliminen del residu L_t només quan el valor predit supera quatre vegades la desviació estàndard del mateix residu.

$$A_t = \begin{cases} si\ L_t < 4\sigma_t & L_t \\ si\ L_t \geq 4\sigma_t & L_t - D_e \end{cases} \tag{6.4}$$

És a dir, aquells punts que poden ser anòmals són els que es troben fora dels límits de quatre vegades la desviació estàndard (σ_t), aleshores, per descartar que es tracti d'un factor transitori, si el valor predit supera aquesta desviació, se li resta el component estacional de la sèrie de valors reals. Finalment, tots els valors que segueixen sent superiors a $4\sigma_t$ es definiran com a anomalia.

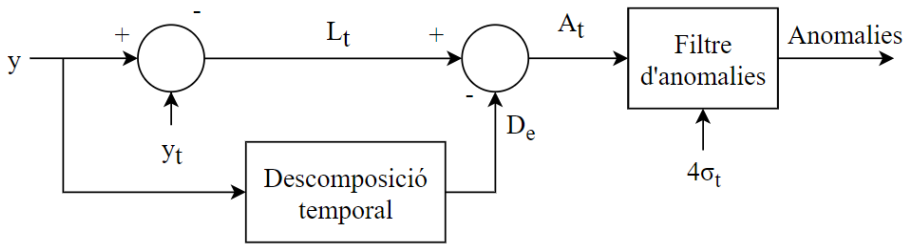

Figura 6.3 Diagrama del funcionament de l'algorisme de detecció d'anomalies.

7 Resultats i anàlisis

En aquest capítol es pretén analitzar els resultats obtinguts de la implementació del mòdul de detecció d'anomalies. Per tal revisar tots els casos estudiats, es compararà la resposta del mòdul en les tres localitzacions de les quals es tenen dades per als dos mètodes que s'han descrit, predicció a base de la producció d'altres instal·lacions (mètode regressiu) i predicció a base de la meteorologia local (mètode neuronal).

7.1 Testeig

Per tal de decidir quin és el model més adequat, és a dir, que prediu millor la producció, per a després fer la detecció d'anomalies del sistema, cal comprovar quins paràmetres s'adapten millor a cada mètode. Per decidir-ho, es fan córrer diferents models amb combinacions diferents d'aquests paràmetres per tal de trobar-ne l'idoni.

7.1.1 Mètode regressiu

A continuació es mostren els resultats obtinguts de comparar els dos models de regressió, *Random Forest* (rf) i *k-NN* (knn) i la imputació d'un valor constant igual a zero (constant) o de la mitjana dels valors més propers (mean). En les figures 7.2, 7.3 i 7.4 es pot veure una representació gràfica dels errors que es cometen en cada una de les combinacions model/valor imputat. En aquest tipus de gràfics és necessari tenir en ment que, com menys dispersió hi hagi en la figura, més precís és el model al qual es fa referència.

Tal i com es pot apreciar en els errors de totes tres localitzacions (figures 7.2, 7.3 i 7.4), el model que n'aconsegueix un de més petit, la dispersió és molt menor, i una puntuació major en l'R^2, és el *Random Forest*. Per tant, per a la calibració del model de predicció s'usarà aquest i no el model *k-NN*.

A simple vista, sembla que tant imputant un zero com la mitjana l'error que es comet és el mateix. Si ampliem en l'R^2 en les tres ciutats (veure figura 7.1), es pot veure com fent ús de la mitjana, la puntuació que obté el model és més gran (fixant-nos en la línia mitja de les caixes) i, tot i que subtilment, és menys dispers, per tant, es pot considerar que el model prediu millor la generació fotovoltaica de les instal·lacions en general. Tal i com s'ha

comentat amb anterioritat, la diferència entre els errors d'utilitzar la mitjana o utilitzar zeros per a omplir forats buits és mínima i, per tant, no es contempla l'opció d'omplir els forats de la nit amb zeros i els del dia amb la mitjana, perquè el resultat obtingut no variaria en gran mesura dels que ja s'han obtingut.

Per tant, per a la calibració del model i la detecció d'anomalies s'utilitzarà la mitjana dels punts propers com a *imputer* i el model *Random Forest*.

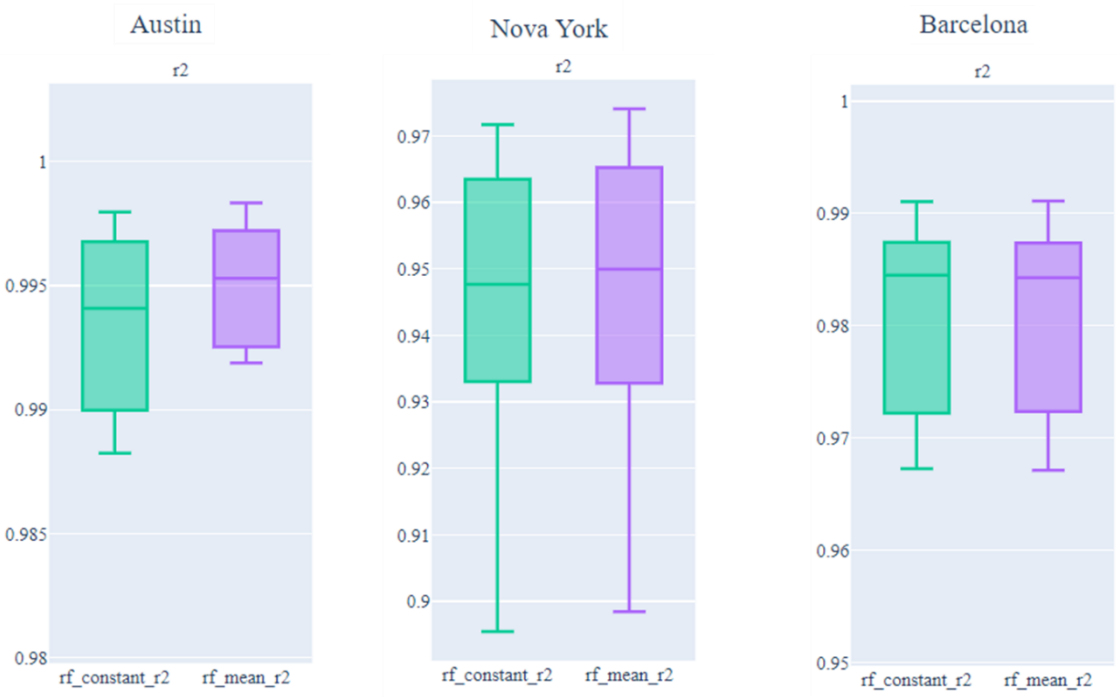

Figura 7.1 Comparativa dels errors comesos en les diferents ciutats entre imputar una constant igual a zero o la mitjana.

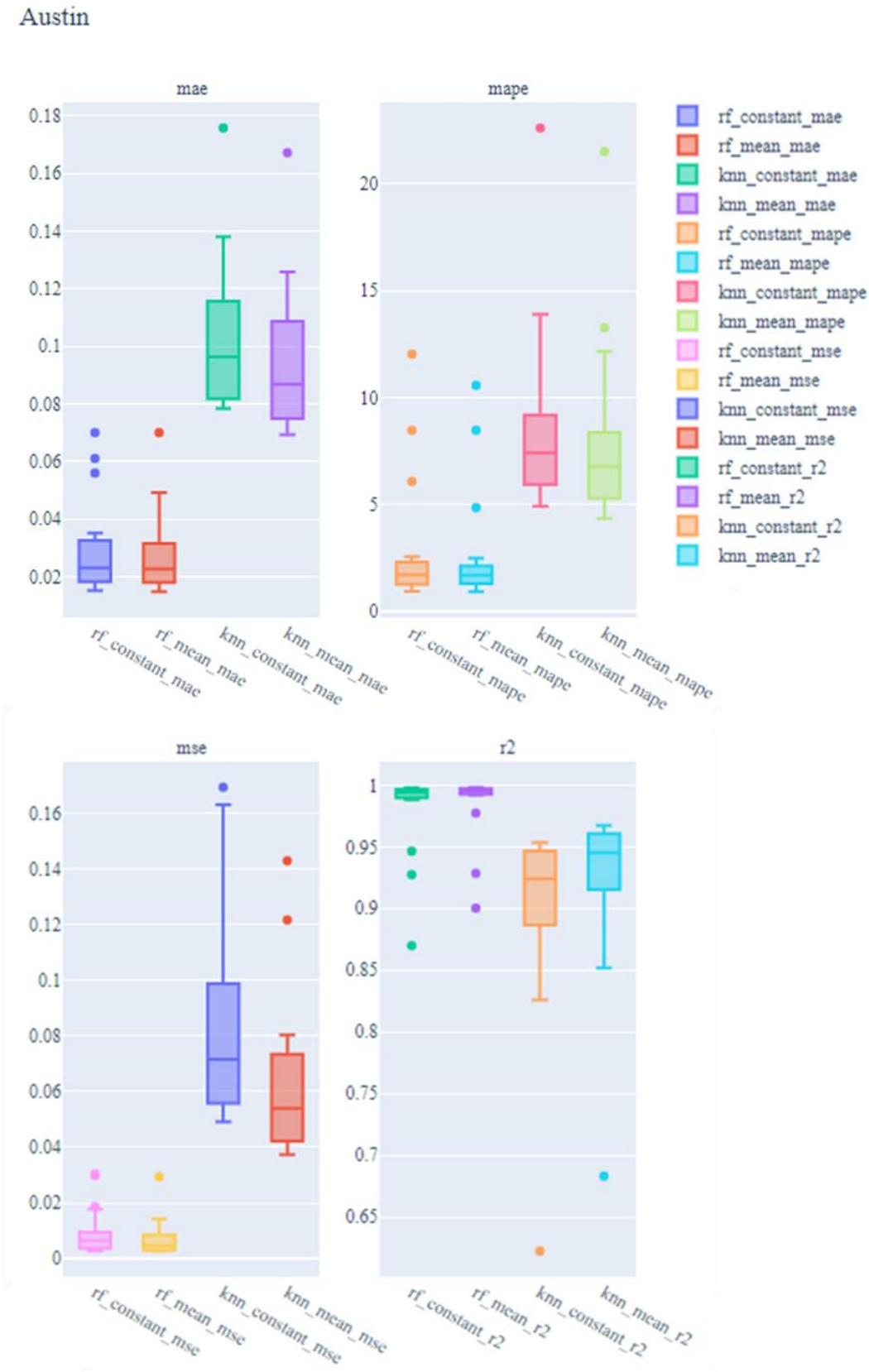

Figura 7.2 Representació gràfica dels diferents errors comesos en la predicció de les instal·lacions de la ciutat d'Austin.

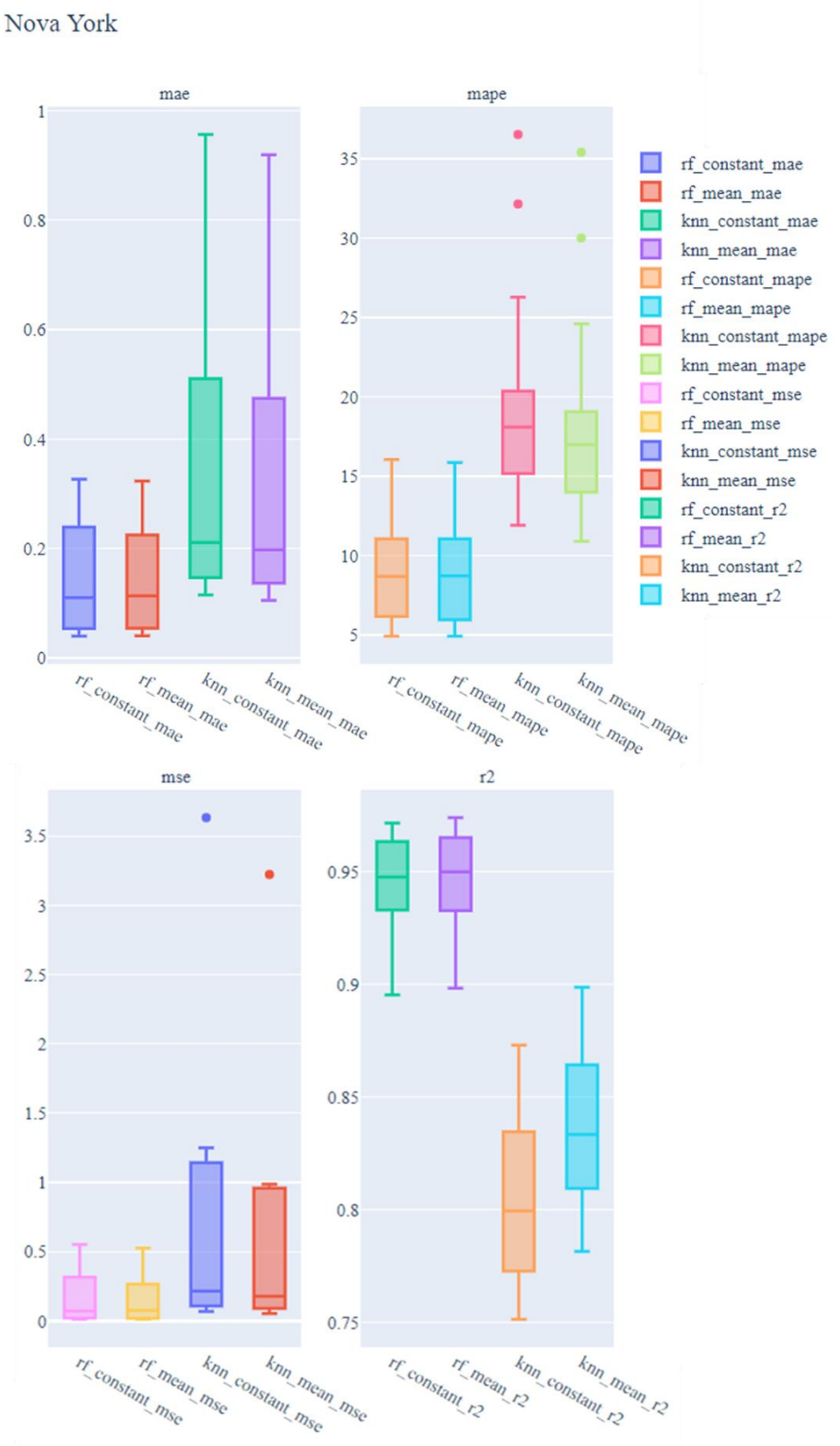

Figura 7.3 Representació gràfica dels diferents errors comesos en la predicció de les instal·lacions de la ciutat de Nova York.

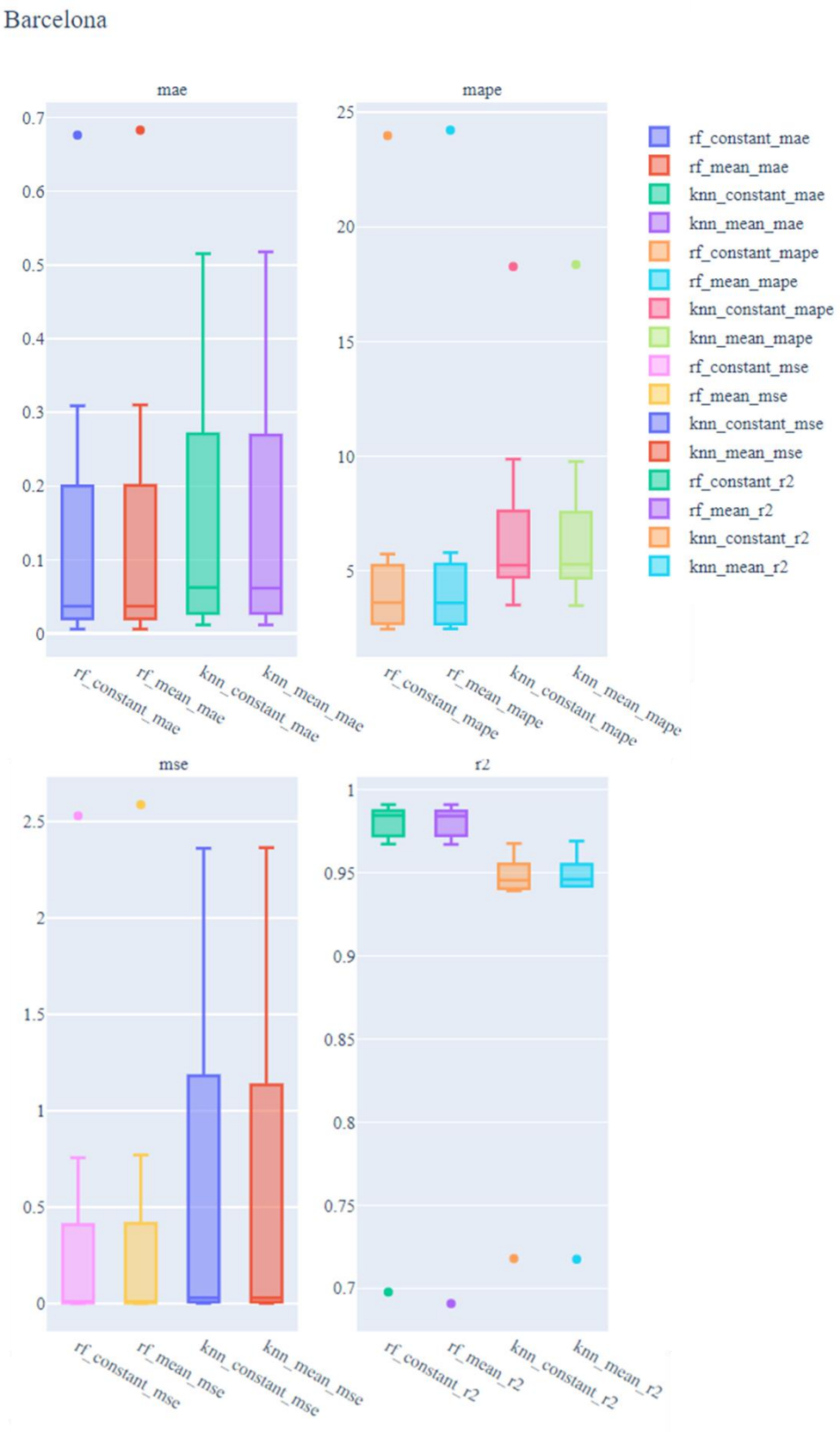

Figura 7.4 Representació gràfica dels diferents errors comesos en la predicció de les instal·lacions de la ciutat de Barcelona.

7.1.2 Mètode neuronal

Per a aquest segon mètode, implementat amb una xarxa neuronal, no s'ha usat més d'un model diferent, però si que s'ha variat la manera d'escalar els valors d'entrada. S'han testejat quatre escaladors diferents (*standard, minmax, maxabs, robust* explicats en el capítol 6.1.4) i, de la mateixa manera que en el mètode anterior, s'ha utilitzat dues imputacions per als valors amb blanc, constant (*constant*) i mitjana (*mean*).

Observant les figures 7.5, 7.6 i 7.7, es pot destacar el millor rendiment dels escaladors *standard* i *robust* per sobre dels altres dos. Per a poder fer un anàlisi més concret, es fan els gràfics dels dos escaladors per a cada ciutat i s'observa, tan sols, el coeficient de determinació R^2, que puntua el nivell de precisió de les dades predites envers les reals, veure figura 7.8. En el mètode anterior, la diferència entre els models i els valors imputats era clara, en les tres localitats sobresortia, clarament, la combinació *constant* i *Random Forest*, però en aquest cas, per a cada ciutat hi ha una combinació més adient. Fent una simple classificació, que es pot veure en la taula 7.1, els models on s'usa la mitjana obtenen una millor qualificació, també es veu com els models que usen l'*scaler* estàndard obtenen una millor puntuació respecte els *robust*.

	Mean/Standrad	*Mean/Robust*	*Constant/Standard*	*Constant/Robust*
Austin	2	3	1	4
Nova York	1	4	2	3
Barcelona	3	1	4	2
TOTAL	**6**	**8**	**7**	**9**

Taula 7.1 Classificació dels models utilitzats en el segon mètode

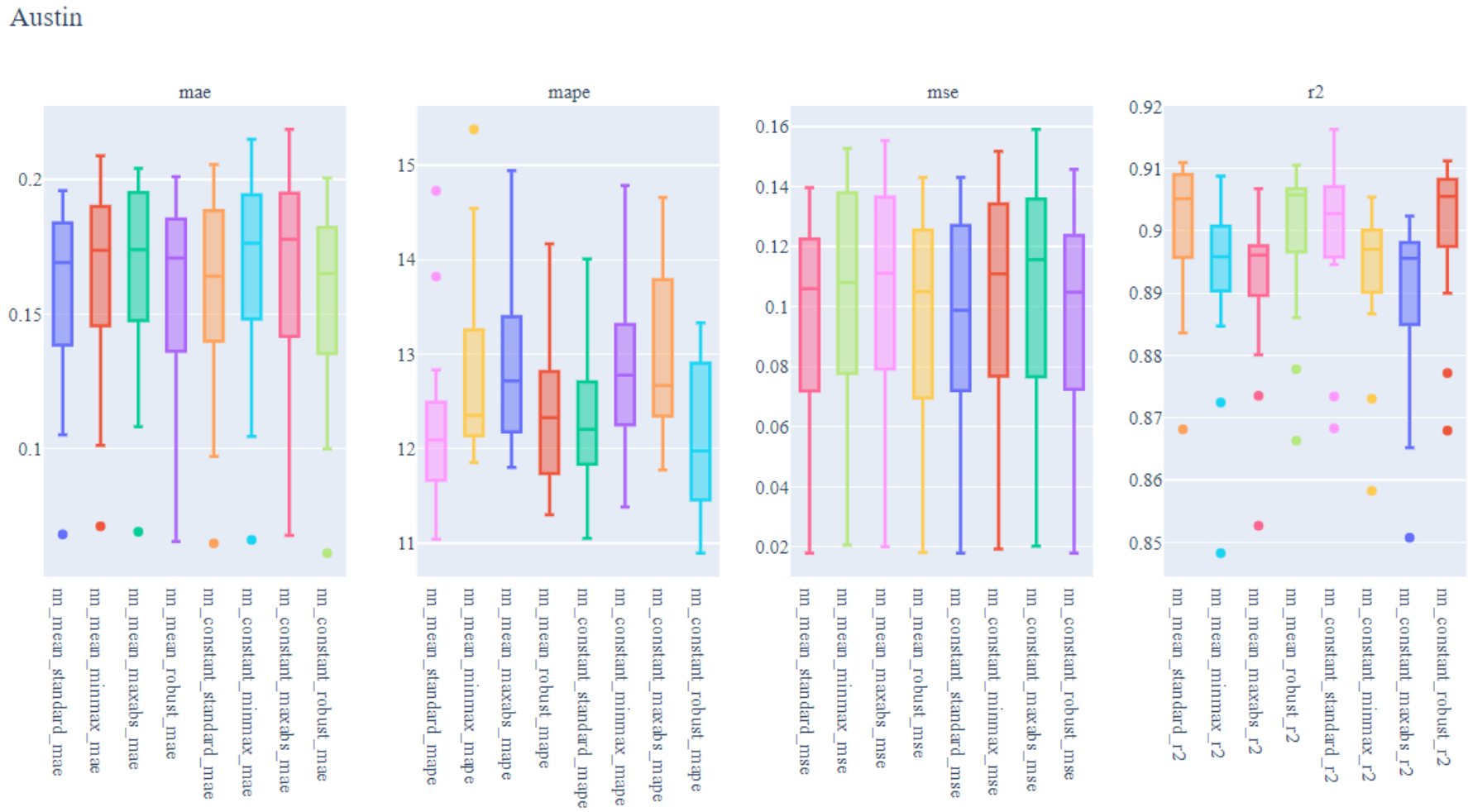

Figura 7.5 Representació gràfica dels diferents errors comesos en la predicció de les instal·lacions de la ciutat d'Austin.

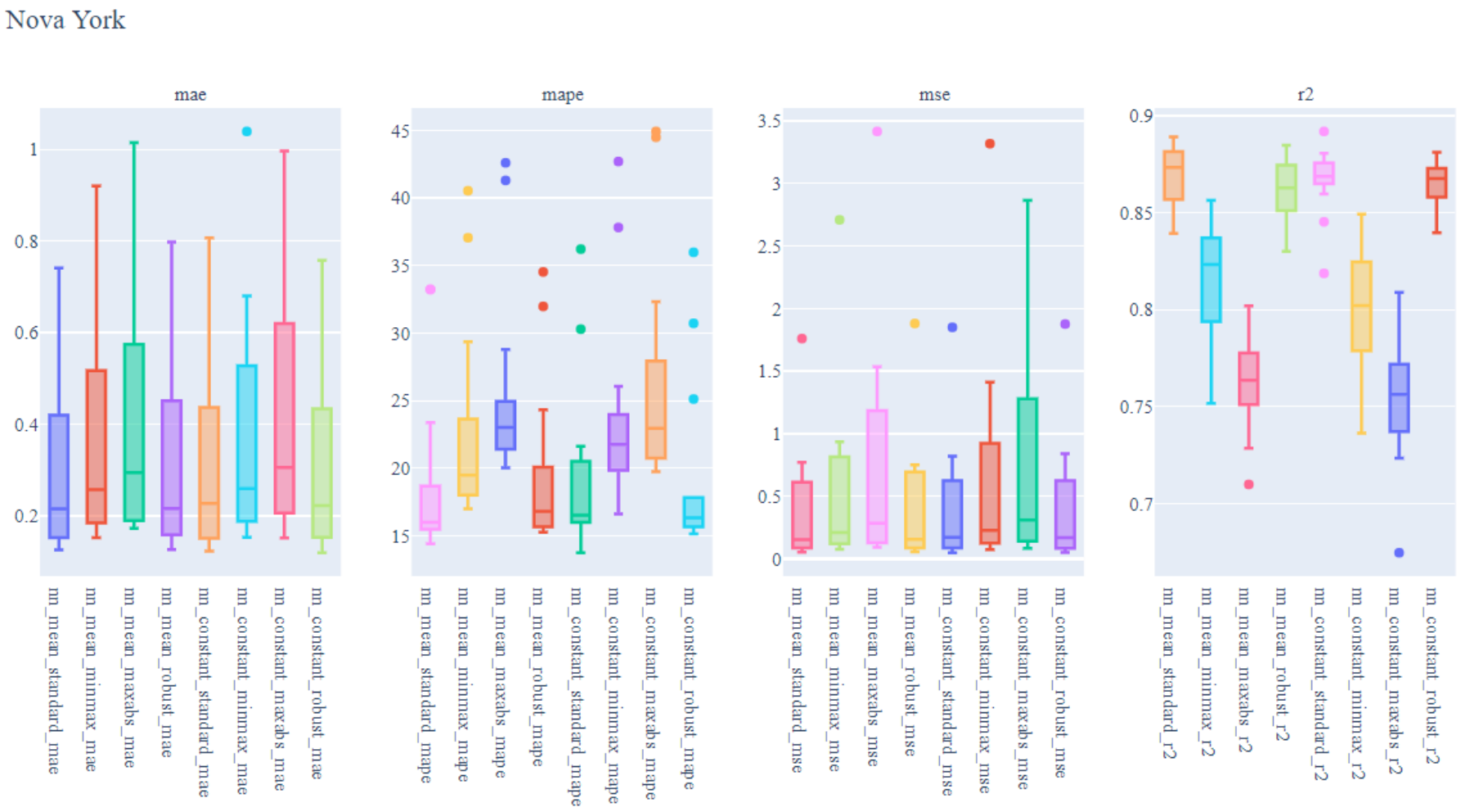

Figura 7.6 Representació gràfica dels diferents errors comesos en la predicció de les instal·lacions de la ciutat de Nova York.

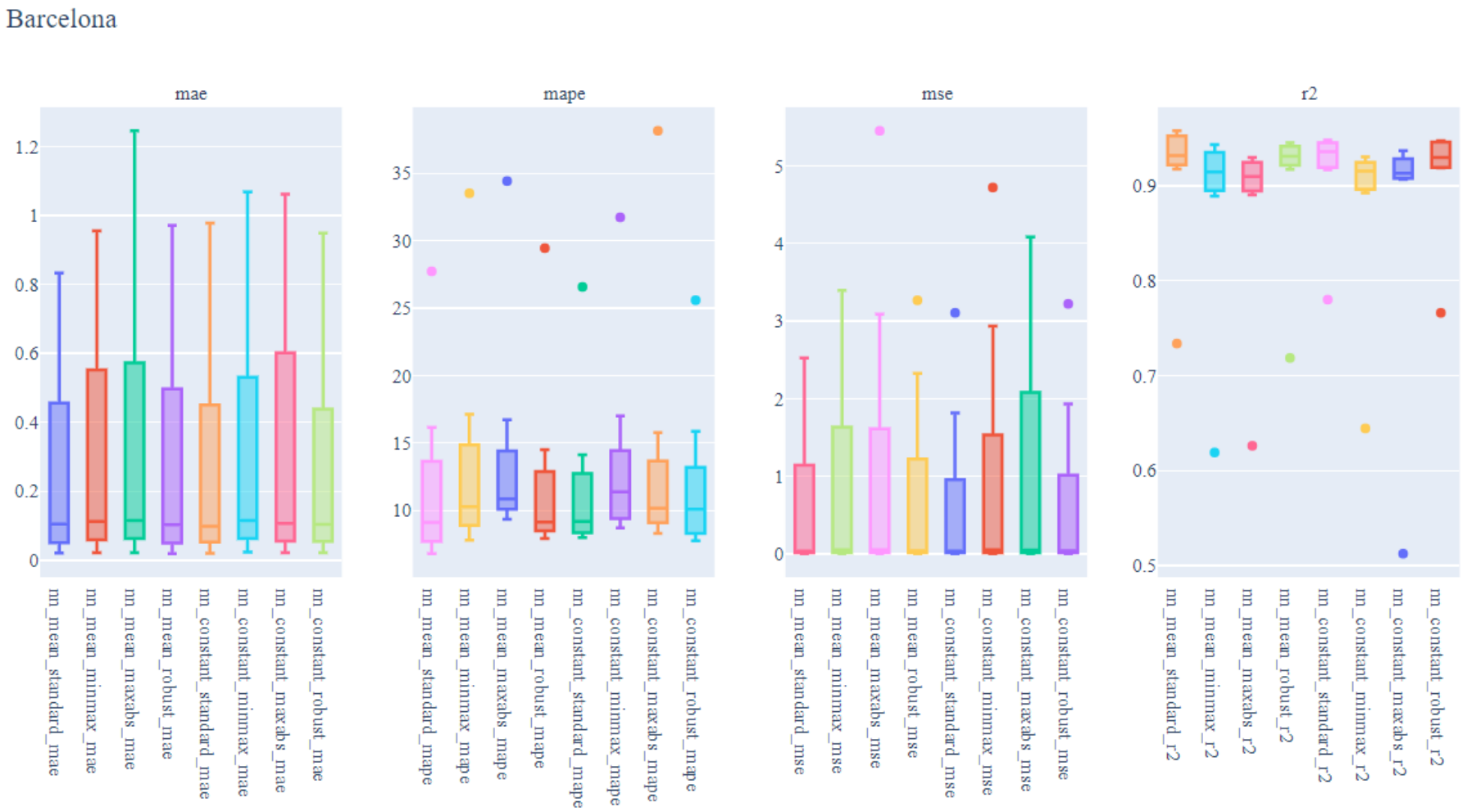

Figura 7.7 Representació gràfica dels diferents errors comesos en la predicció de les instal·lacions de la ciutat de Barcelona.

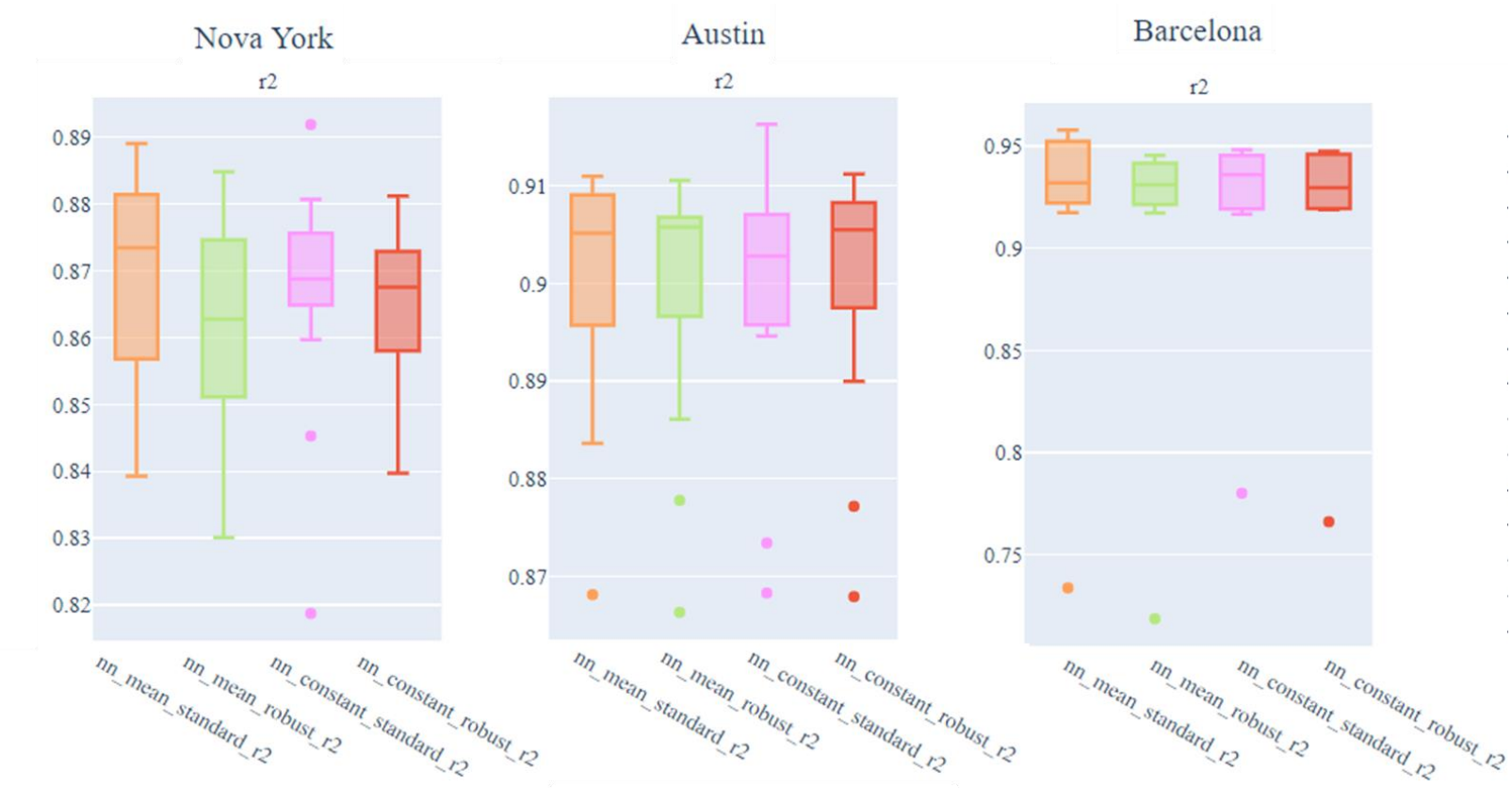

Figura 7.8 Comparativa dels errors comesos en les diferents ciutats segons *imputer* i *scaler*.

7.2 Calibració

Un cop s'ha decidit quin model i quins paràmetres s'utilitzaran en aquest, cal entrenar el model. Tot i que aquest punt no és crític i, en si, no presenta resultats més que els models entrenats, s'ha volgut extreure una comparativa de l'eficàcia dels diferents models en les diferents ciutats.

7.2.1 Mètode regressiu

Com es pot observar en la figura 7.9, l'error en les diferents localitzacions és dispar. Si bé és un error acceptable per als tres casos, un coeficient de determinació (R^2) de més de 0.8 és més que suficient per a la predicció amb models de regressió, la teoria diria que tots tres models haurien de tenir un error semblant. Aleshores, quin és el motiu d'aquesta disparitat?

Tot i que el model i mètode per predir la producció d'energia ha estat el mateix, les bases de dades de les dades d'entrada no són coincidents en totes tres ciutats. Com es veu en la taula 7.2, cap de les tres localitats conegudes té el mateix nombre d'instal·lacions i, molt menys, el mateix nombre de punts, per tant, és adequat pensar que, a major nombre de punts i instal·lacions, la quantitat de dades per a la calibració del model és major i, per tant, la predicció és més ajustada a la producció real. És d'esperar doncs que la ciutat de Nova York tingui una pitjor puntuació i que es dubti dels resultats d'Austin i Barcelona.

	Número instal·lacions	Número punts
Austin	18	630648
Nova York	14	247296
Barcelona	8	840960

Taula 7.2 Nombre real d'instal·lacions i quantitat de dades usades per ciutat.

Cal tenir en compte, però, que la diferència no només la marca la quantitat de les dades obtingudes, sinó la qualitat d'aquestes. En les taules 7.3 i 7.4 es pot veure la correlació de les dades de les diferents instal·lacions per a les ciutats d'Austin i Barcelona respectivament.

En aquesta correlació, s'analitzen les dades, comparant instal·lació a instal·lació, per estudiar quina relació tenen les unes amb les altres. Com més proper a u (1) sigui el nombre de la correlació entre dues instal·lacions, més tenen a veure, i més poden ajudar a predir el

comportament de les altres instal·lacions, unes dades amb les altres. El codi de colors també va relacionat amb el mateix fet, a major correlació, més color vermell.

Tot i que de Barcelona es tenen més punts, en la correlació de les dades es pot observar com la instal·lació nº 31 té una correlació baixa amb la resta d'instal·lacions, en la ciutat d'Austin també hi ha un cas, instal·lació nº 17, en que la correlació és més baixa, però, ni és tant baixa com la correlació de la 31 de Barcelona, ni suposa el mateix impacte, en el cas d'Austin és la correlació baixa és d'una instal·lació de divuit, en el cas de Barcelona, una de vuit.

Per aquests motius, es pot concloure que la ciutat amb una millor predicció és Austin, seguida de Barcelona i, deixant en últim lloc Nova York.

	1	3	5	7	15	16	27	31
1	1	0,92	0,90	0,86	0,91	0,91	0,82	0,40
3	0,92	1	0,92	0,88	0,90	0,90	0,84	0,38
5	0,90	0,92	1	0,90	0,89	0,92	0,85	0,40
7	0,86	0,88	0,90	1	0,83	0,93	0,80	0,42
15	0,91	0,90	0,89	0,83	1	0,89	0,84	0,39
16	0,91	0,90	0,92	0,93	0,89	1	0,81	0,45
27	0,82	0,84	0,85	0,80	0,84	0,81	1	0,37
31	0,40	0,38	0,40	0,42	0,39	0,45	0,37	1

Taula 7.3 Correlació entre les dades de les instal·lacions d'Austin.

	0	1	2	3	4	5	6	7	8	9	10	12	13	14	15	16	17	18
0	1	0,98	0,99	0,99	0,97	0,84	0,98	0,94	0,92	0,92	0,97	0,92	0,98	0,99	0,99	0,97	0,83	0,94
1	0,98	1	0,99	0,98	0,99	0,88	0,99	0,97	0,94	0,95	0,99	0,95	0,99	0,99	0,97	0,99	0,80	0,95
2	0,99	0,99	1	0,99	0,99	0,89	0,99	0,98	0,95	0,96	0,99	0,94	1,00	1,00	0,98	0,99	0,80	0,97
3	0,99	0,98	0,99	1	0,97	0,86	0,99	0,95	0,93	0,93	0,98	0,93	0,99	0,99	0,99	0,98	0,83	0,95
4	0,97	0,99	0,99	0,97	1	0,93	0,99	0,99	0,96	0,98	0,99	0,94	0,99	0,98	0,96	0,99	0,75	0,98
5	0,84	0,88	0,89	0,86	0,93	1	0,89	0,96	0,92	0,97	0,93	0,83	0,92	0,88	0,84	0,94	0,62	0,96
6	0,98	0,99	0,99	0,99	0,99	0,89	1	0,97	0,95	0,96	0,99	0,94	1,00	0,99	0,98	0,99	0,78	0,97
7	0,94	0,97	0,98	0,95	0,99	0,96	0,97	1	0,96	0,99	0,99	0,93	0,98	0,97	0,94	0,99	0,72	0,99
8	0,92	0,94	0,95	0,93	0,96	0,92	0,95	0,96	1	0,96	0,96	0,88	0,96	0,93	0,91	0,96	0,69	0,96
9	0,92	0,95	0,96	0,93	0,98	0,97	0,96	0,99	0,96	1	0,98	0,90	0,97	0,95	0,92	0,98	0,69	1,00
10	0,97	0,99	0,99	0,98	0,99	0,93	0,99	0,99	0,96	0,98	1	0,94	1,00	0,99	0,97	1,00	0,77	0,98
12	0,92	0,95	0,94	0,93	0,94	0,83	0,94	0,93	0,88	0,90	0,94	1	0,94	0,94	0,93	0,93	0,77	0,90
13	0,98	0,99	1,00	0,99	0,99	0,92	1,00	0,98	0,96	0,97	1,00	0,94	1	0,99	0,98	1,00	0,78	0,98
14	0,99	0,99	1,00	0,99	0,98	0,88	0,99	0,97	0,93	0,95	0,99	0,94	0,99	1	0,99	0,99	0,82	0,96
15	0,99	0,97	0,98	0,99	0,96	0,84	0,98	0,94	0,91	0,92	0,97	0,93	0,98	0,99	1	0,97	0,82	0,93
16	0,97	0,99	0,99	0,98	0,99	0,94	0,99	0,99	0,96	0,98	1,00	0,93	1,00	0,99	0,97	1	0,77	0,99
17	0,83	0,80	0,80	0,83	0,75	0,62	0,78	0,72	0,69	0,69	0,77	0,77	0,78	0,82	0,82	0,77	1	0,70
18	0,94	0,95	0,97	0,95	0,98	0,96	0,97	0,99	0,96	1,00	0,98	0,90	0,98	0,96	0,93	0,99	0,70	1

Taula 7.4 Correlació entre les dades de les instal·lacions de Barcelona.

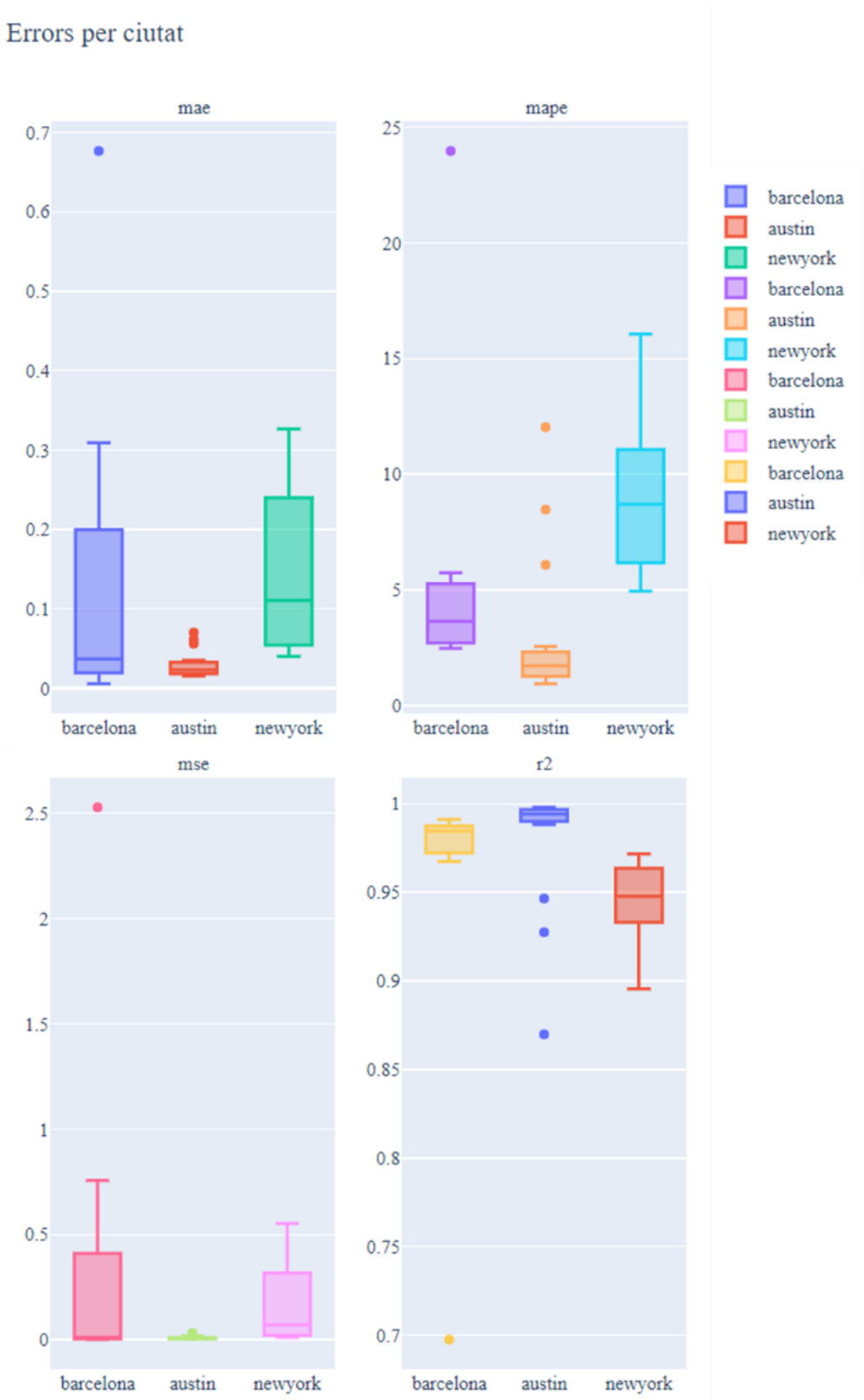

Figura 7.9 Comparativa dels errors comesos en les diferents ciutats en la calibració dels models.

7.2.2 Mètode neuronal

Tal i com s'ha discutit en el punt anterior, i com es pot veure en la figura 7.10, aquella ciutat que té més punts, és aquella que obté una millor resposta i, per tant, un error més baix. Si bé és cert que en el primer mètode s'ha reflexionat sobre la diferència entre Barcelona i Austin, en aquest cas no és vàlid el mateix argument. Quan s'ha predit la producció en el mètode regressiu s'ha usat diferents dades reals de producció d'instal·lacions properes, per tant, és lògic pensar que, a major quantitat d'instal·lacions (suposant que aquestes tenen el mateix nombre de valors), millor serà la predicció. En la predicció d'aquest segon mètode tan sols s'utilitza la producció de la instal·lació a predir i la meteorologia de la localització en el dia i hora assenyalats, per tant, el volum d'instal·lacions no és tant important com el nombre de valors per instal·lació. Si es fa un cop d'ull a la taula 7.5, la localització que té més punts per instal·lació és Barcelona, seguida d'Austin i, finalment, Nova York. Per tant, es coherent que els errors calculats vagin de menys a més en el mateix ordre.

	Número punts per intsal·lació
Barcelona	105.120
Austin	35.036
Nova York	17.664

Taula 7.5 Nombre de valors cada quart horari per instal·lació en les diferents localitzacions.

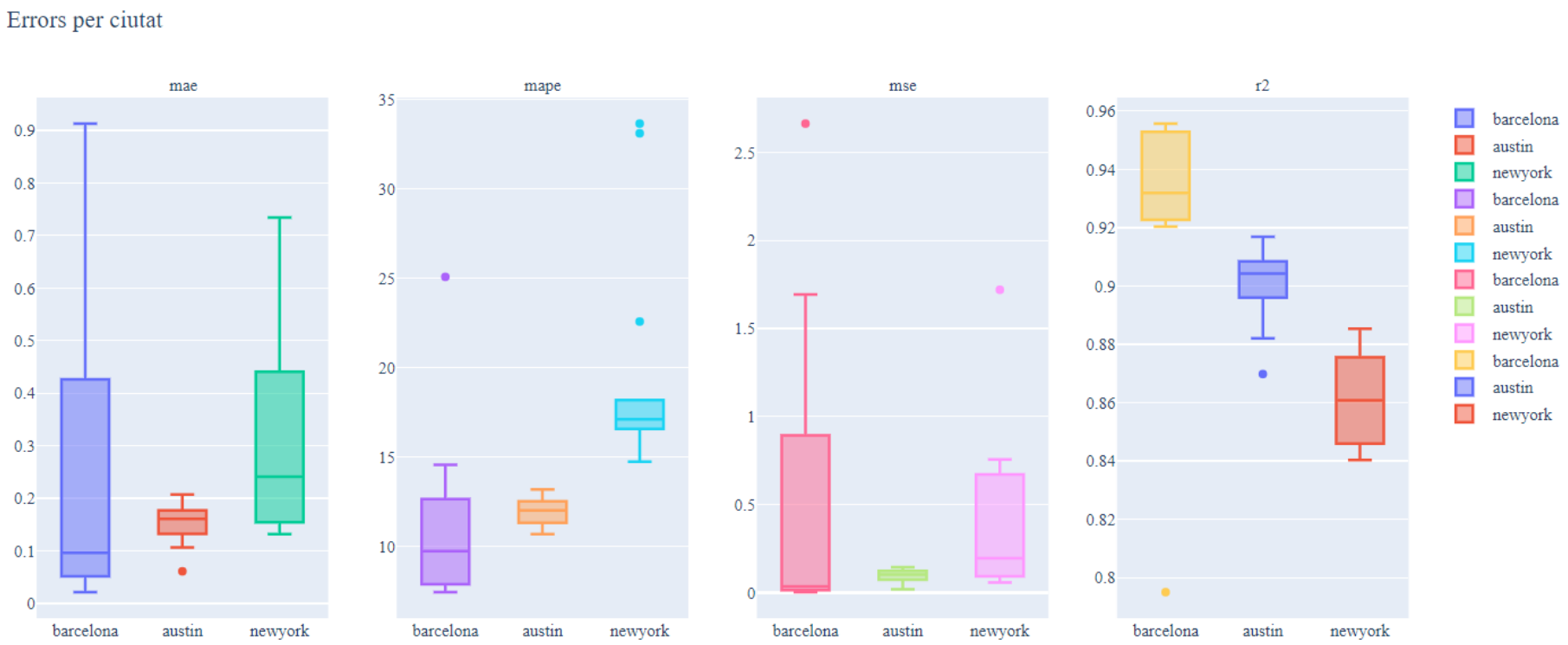

Figura 7.10 Comparativa dels errors comesos en les diferents ciutats en la calibració dels models.

7.3 Detecció d'anomalies

Discutida la precisió dels models, i acceptant que tots ells són adequats per al procediment a seguir, es procedeix a detectar les anomalies sorgides en cada una de les instal·lacions per a cada ciutat estudiada. Com que es compta amb un total de 40 instal·lacions, es repassaran aquelles que tinguin un resultat més interessant de cara a l'estudi de quin dels dos mètodes és el més adequat per a la tasca a realitzar.

7.3.1 Mètode regressiu

En les figures 7.11, 7.12 i 7.13 es poden veure les anomalies detectades per a cada instal·lació en cada una de les localitzacions estudiades. En conjunt, s'han detectat un total de 36496 anomalies. En la taula 7.6 es mostra el detall de quantes anomalies s'han detectat per ciutat i el nombre de dades que es tenen.

	Número d'anomalies	Número punts
Barcelona	8.848	840.960
Austin	5.683	630.648
Nova York	4.147	247.296

Taula 7.6 Nombre d'anomalies i dades conegudes per ciutat.

Aquestes anomalies s'han dividit, posteriorment, en anomalies puntuals, si tan sols es donen en un moment puntual, i en anomalies sostingudes, si l'anomalia es manté durant un cert temps.

Com a exemple d'anomalies sostingudes en el temps, en la figura 7.14 es pot veure la detecció d'anomalies d'una instal·lació de la ciutat de Barcelona en la segona meitat de l'any. En blau es pot veure la producció real de la instal·lació, i en verd, es veu com en aquells períodes que, per exemple, no hi ha producció, es marquen com a anomalies sostingudes. És necessari notar que, com l'anomalia es produeix durant el dia, i no la nit, en les hores de no sol no es detecta anomalia, veure ampliació, on es veu, també, la predicció en vermell.

Per altra banda, com a anomalies puntual es por observar la figura 7.15, on s'observen tres dies seguits amb la mateixa anomalia puntual. Aquest fet es podria explicar com a una ombra puntual d'un arbre proper a una instal·lació que, al cap de les setmanes, desapareix.

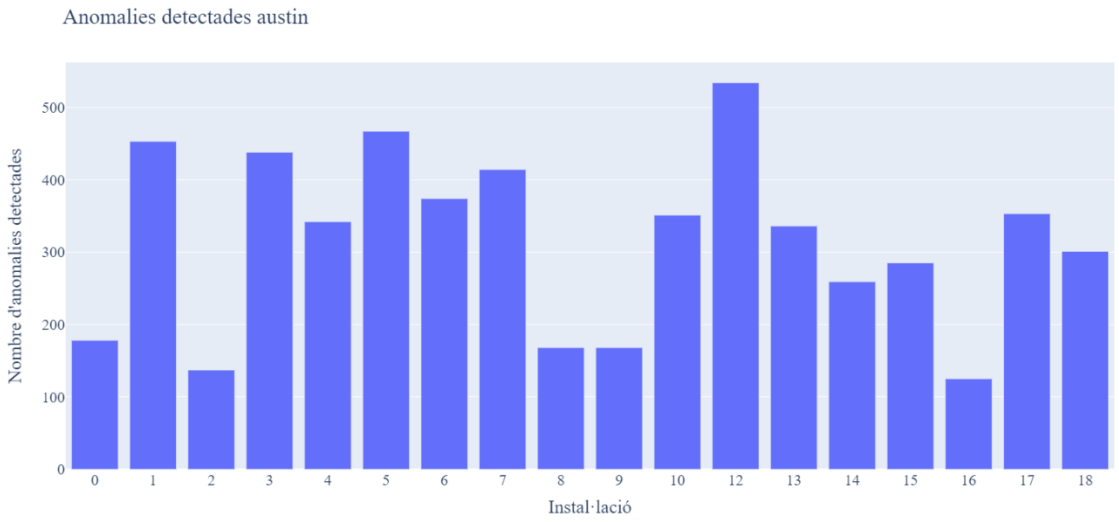

Figura 7.11 Nombre d'anomalies detectades per instal·lació en la ciutat d'Austin.

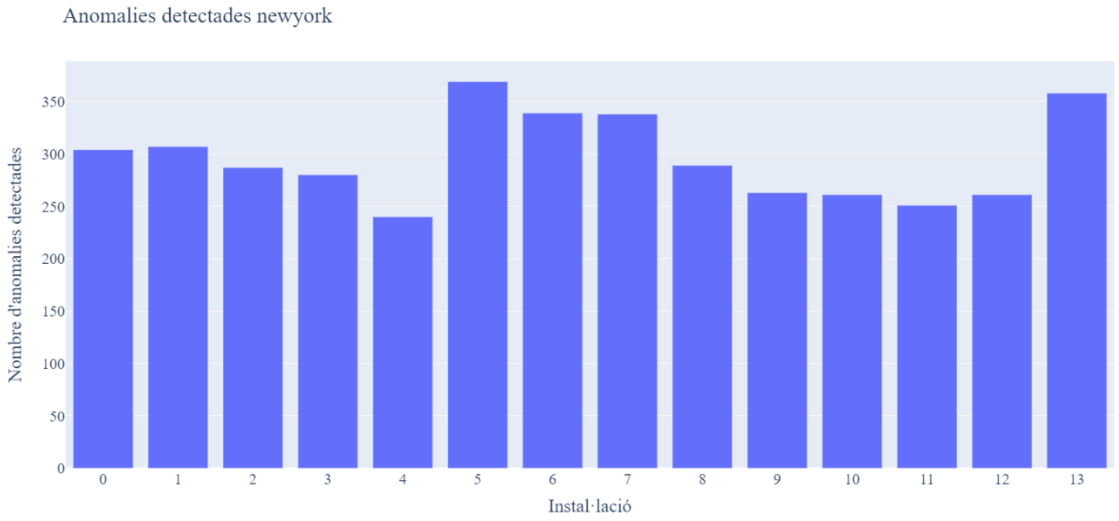

Figura 7.12 Nombre d'anomalies detectades per instal·lació en la ciutat de Nova York.

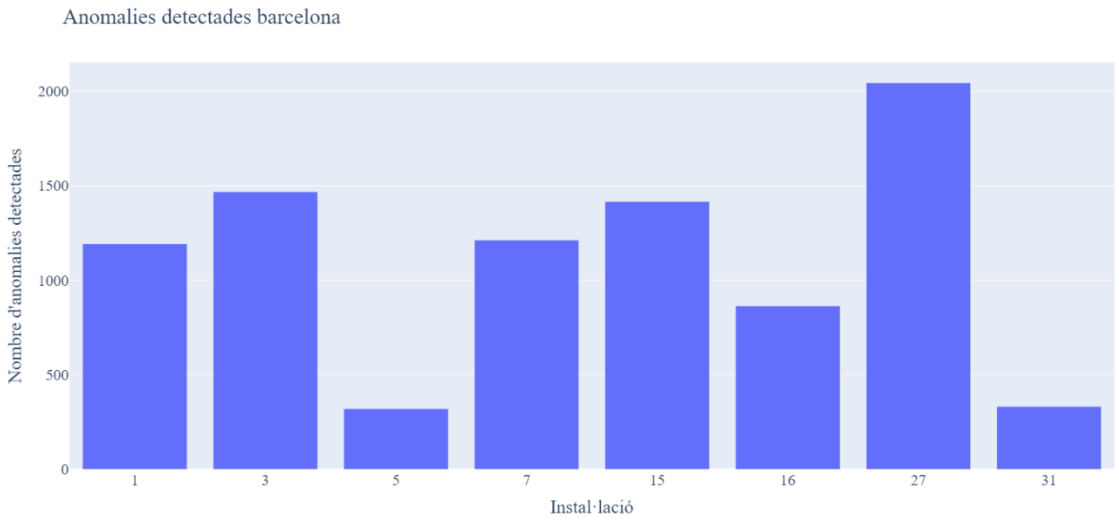

Figura 7.13 Nombre d'anomalies detectades per instal·lació en la ciutat de Barcelona.

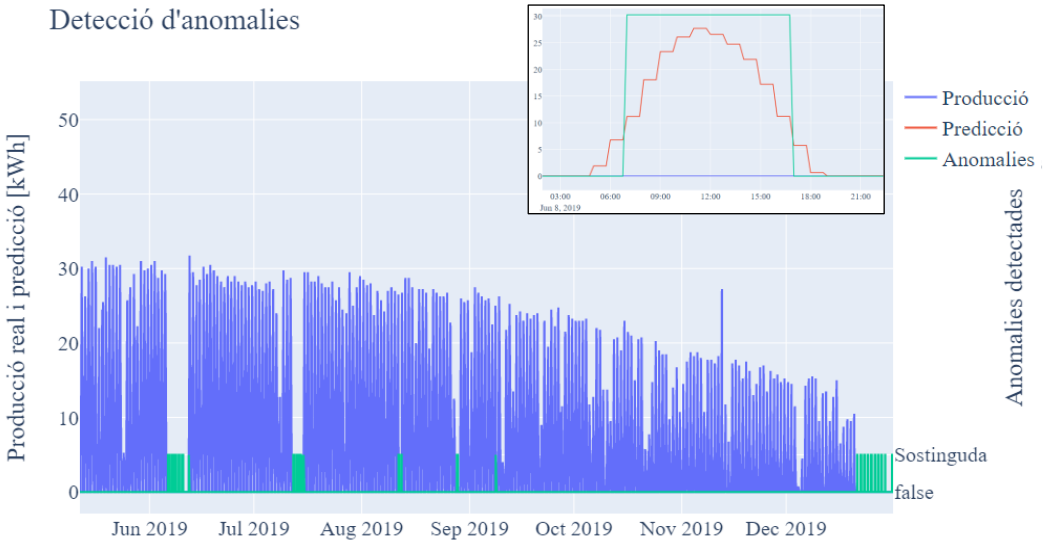

Figura 7.14 Detecció d'anomalies d'una instal·lació de la ciutat de Barcelona.

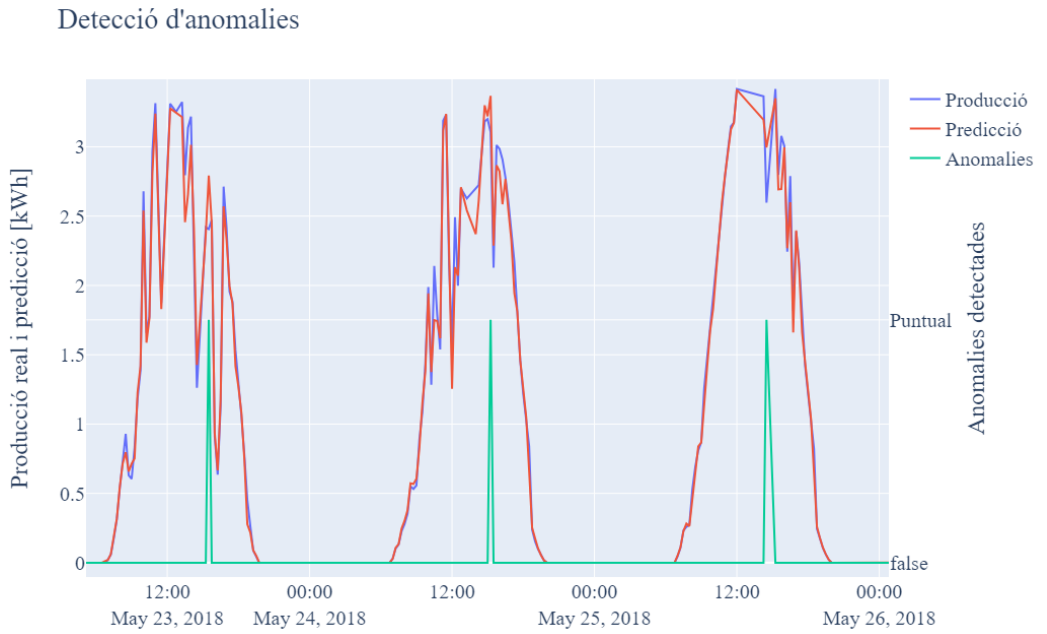

Figura 7.15 Detall de tres dies consecutius amb anomalia puntual en una instal·lació de la ciutat d'Austin.

7.3.2 Mètode Neuronal

De la mateixa manera que en el mètode anterior, s'han comptat el total d'anomalies per ciutat, es poden veure recollides en la taula 7.7, i les anomalies per instal·lació, que es representen en les figures 7.16, 7.17 i 7.18. Tot i que en la ciutat de Barcelona sembla que

detecti menys anomalies, en les altres dues ciutats el nombre d'anomalies detectades és molt més superior a l'obtingut anteriorment. Aquest fet, es pot deure a dues raons, o l'algorisme utilitzar és més sensible a les anomalies, o la predicció d'aquest segon mètode no és prou eficient per a fer una predicció d'anomalies amb els resultats obtinguts. Cal fer menció que, tal i com s'ha vist en el punt 7.2.2, els models obtinguts amb aquest segon mètode no són tan acurats, és a dir cometen més error que els models definits amb el primer mètode regressiu i, per tant, és raonable assumir que els resultats obtinguts amb aquest segon mètode neuronal poden no ser tant precisos com en el cas anterior.

	Número d'anomalies	Número punts
Barcelona	2.152	840.960
Austin	24.755	630.648
Nova York	9.589	247.296

Taula 7.7 Nombre d'anomalies i dades conegudes per ciutat.

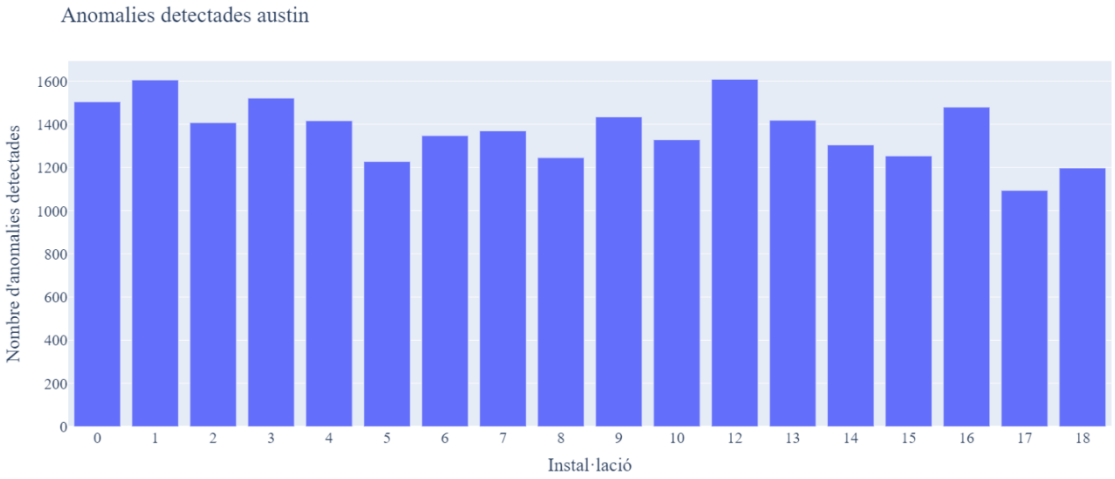

Figura 7.16 Nombre d'anomalies detectades per instal·lació en la ciutat d'Austin.

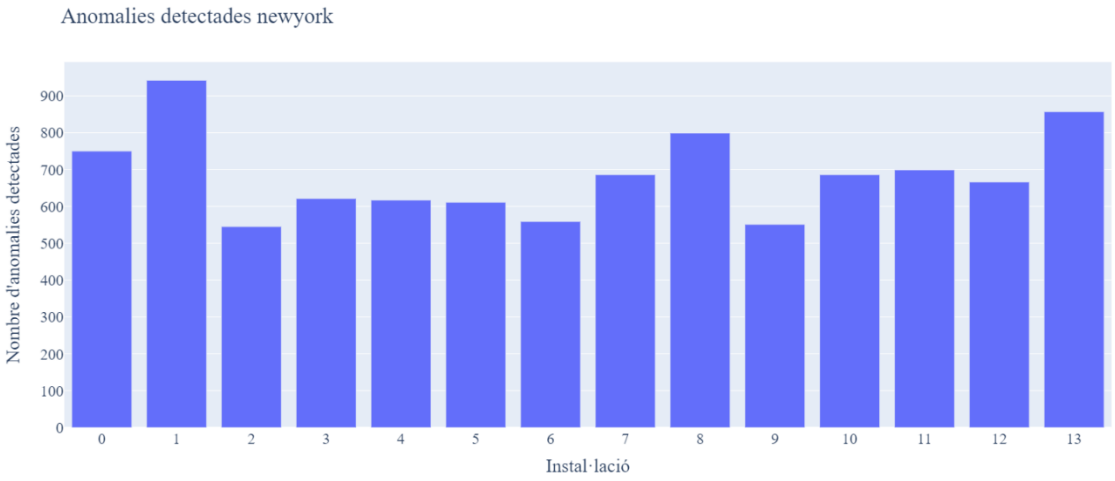

Figura 7.17 Nombre d'anomalies detectades per instal·lació en la ciutat de Nova York.

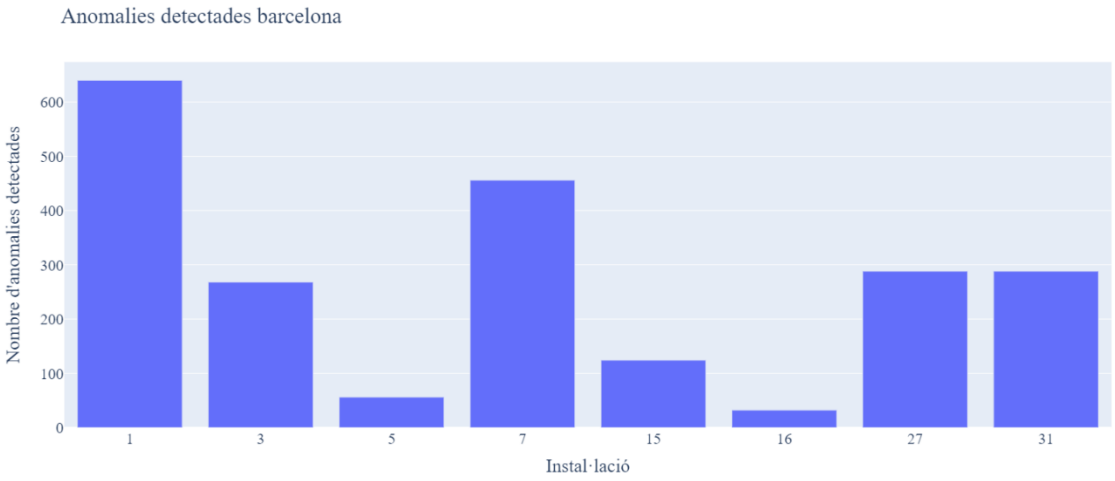

Figura 7.18 Nombre d'anomalies detectades per instal·lació en la ciutat de Barcelona.

8 Planificació del projecte

Per tal de fer la planificació del projecte s'ha utilitzat el programari Microsoft Project Professional, cedit per l'Escola Superior Politècnica del TecnoCampus. S'ha comprés la programació de les tasques designades entre el 13 de febrer de 2021, dia següent a l'entrega de l'avantprojecte, i el 16 de juny de 2021, dia anterior a l'entrega de la memòria final del projecte.

8.1 Planificació inicial

És necessari contemplar les dues seccions diferenciades en la planificació del projecte que es veuen en la Figura 8.1: La primera descriu els dos períodes de feina, marcant les dates d'entrega tant de la memòria intermèdia com de la memòria final, durant els quals es redactarà la documentació a mesura que es vagi assolint els objectius i finalitzant les tasques. La segona fa referència a les tasques pròpiament dites del projecte, deixant de banda la redacció, més enfocada al procés acadèmic, i enfocant-se en aquells passos pràctics per completar l'objectiu final.

S'ha dividit l'estructura i les tasques a desenvolupar en set fases diferents explicades i detallades a continuació:

F1: Presa de contacte

Llegir *paper*: lectura de l'article científic en el qual es basa l'algoritme utilitzat al projecte. **Preparació del *dataset*:** Avaluació i quantificació de NaNs (valors no existents) i metodologia d'emplenar de forats. Obtenir les dades necessàries i descartar aquelles que no són útils. **Descompondre sèries temporals:** la generació FV ve donada per una sèrie temporal, aquesta ha d'estar descomposta en tendència, estacionalitat i soroll. **M1:** Construcció del primer model (RF - RandomForest) **M2:** Construcció del segon model (KNN - K Nearest Neighbor) **Càlcul i representació gràfica d'errors:** MAPE, R^2 score, RMSE.

F2: Detecció d'anomalies

Implementació mètrica de detecció d'anomalies: seguint l'estudi fet en l'article, implementar la metodologia per a la detecció d'anomalies sense estudiar-ne la causa i/o efecte. **Classificació d'anomalies:** en puntuals i temporals, dins d'aquestes segones entre degradacions o pèrdua de capacitat de producció.

F3 i F4: Creació de *pipelines* i ordenació del codi

F3: Creació del *pipeline* calibració del model: Càrrega i comprovació del format dataset, selecció i calibració del model adequat i càlcul d'error.
F4: Creació del *pipeline* d'implementació del model: Lectura del *dataset*, càrrega del model, detecció d'anomalies, classificació d'anomalies.

F5: Introducció de noves dades

Preparar *dataset* BCN: per a què s'edacui a les necessitats de les dades d'entrada. **Testatge del codi en *datasets* complementaris:** en les ciutats de Nova York, California i Barcelona.

F6: Augment de la sensibilitat del model

M1: Inclusió de nous paràmetres (kWp, Orientació, Inclinació, Material/Eficiència del panell, etc.) **Testatge i comparació model:** comparació dels models primer abans i després de l'augment de paràmetres d'entrada en tots els *datasets*.

F7: Xarxes Neuronals

M3: Construcció d'un tercer model (NN - Neural Network) **Testeig i comparació model:** comparació dels models primer amb o sense els nous paràmetres d'entrada (depenent de quin sigui més rigorós) i tercer en tots els *datasets*.

Al ser un projecte en un marc empresarial, s'acorda amb el cap de projecte una reunió de seguiment setmanal per comprovar l'estat de les tasques actives i, si fos necessari, la creació de noves tasques i ampliació del temps de dedicació a aquestes.

A data de 22 d'abril de 2021 la programació s'està seguint tal i com va estar planificada en un inici i, per tant, no s'ha hagut de modificar.

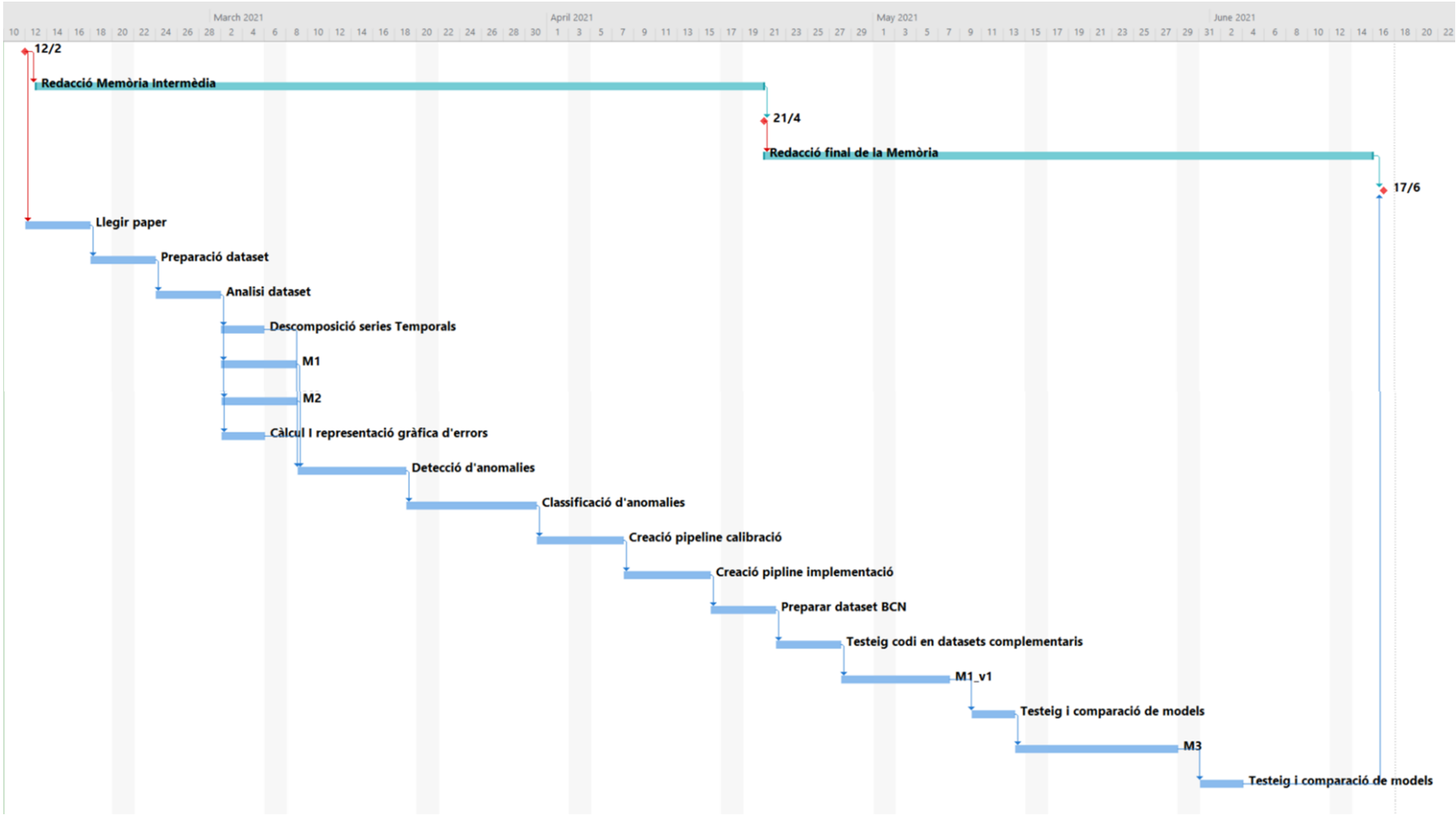

Figura 8.1 Diagrama de Gantt de la planificació del projecte.

8.2 Desviacions respecte la planificació

En aquesta secció es mencionen les desviacions respecte la planificació inicial. Per fer-ho, es mostren les tasques en la taula 8.1. Cal notar que a les dues primeres tasques de la primera fase s'han reduït en el temps, així com les tasques de la fase sis, que s'han eliminat del projecte, deixant més temps a les etapes intermèdies, que han augmentat en algunes hores. Aquest augment d'hores es deu a que, s'ha hagut de fer més recerca, envers les tasques que s'han allargat, sobre el funcionament i l'aplicació en programació.

Fase	*Tasca*	*Hores Planificació*	*Hores Afegides*	*Hores Totals*
Fase 1	Llegir paper	16	-12	4
	Preparació dataset	16	-12	4
	Analisi dataset	16	8	24
	Descomposició series Temporals	16	10	26
	M1	20	10	30
	M2	20	0	20
	Càlcul I representació gràfica d'errors	16	4	20
Fase 2	Detecció d'anomalies	32	8	40
	Classificació d'anomalies	32	8	40
Fase 3	Creació pipeline calibració	24	8	32
Fase 4	Creació pipline implementació	24	16	40
Fase 5	Preparar dataset BCN	16	0	16
	Testeig codi en datasets complementaris	16	0	16
Fase 6	M1_v1	32	-32	0
	Testeig i comparació de models	16	-16	0
Fase 7	M3	44	0	44
	Testeig i comparació de models	16	0	16
				372

Taula 8.1 Correlació de les tasques a fer per fase i el temps total dedicat.

9 Impacte mediambiental

Els efectes mediambientals que pugui ocasionar el desenvolupament del projecte són gairebé nuls. Cal tenir en compte que més enllà d'un ordinador i l'electricitat que aquest pugui necessitar, juntament amb les bases de dades necessàries, el projecte no es val de cap altre material/procés perillós per al medi ambient. Per altra banda, el projecte pretén aportar un producte per millorar la capacitat productiva d'energia solar, incentivant, doncs, l'ús d'energia renovable per davant de la no renovable i aportant un impacte positiu al medi.

10 Conclusions

A mode de conclusions, s'ha volgut fer una avaluació exhaustiva dels objectius marcats en l'inici del projecte, marcant el seu grau d'assoliment i valorant la tasca realitzada. A més a més, s'inclou un apartat amb millores addicionals que s'aplicaran al projecte que, degut a la falta de temps, no s'han pogut implementar.

10.1 Revisió dels objectius

Del primer objectiu, *Desenvolupar un model matemàtic de la producció d'una instal·lació fotovoltaica*, s'ha pogut extreure de les bases de dades aquelles dades no necessàries, no existents o incompletes, deixant un set de dades òptim per a la modelització de la producció elèctrica d'una instal·lació fotovoltaica. S'ha usat el mínim de dades necessàries per a entrenar el model amb un 25% de les dades totals, una dada poc habitual en la predicció amb models de regressió lineal. S'ha tingut en compte tant l'ordre de les dades com la periodicitat d'aquestes. Finalment no s'ha pogut disposar ni de la orientació, ni tecnologia de les instal·lacions, però si s'ha tingut en compte la potència de pic i la radiació per a la predicció en el mètode neuronal. Tot i així, aquest afegit no ha pogut incrementar l'eficàcia del model. Es conclou que l'objectiu ha estat assolit satisfactòriament.

Del segon objectiu, *Detectar les diverses anomalies que es produeixen en la producció respecte al model*, s'ha implementat un algoritme (explicat en l'apartat 6.3.2) el qual ha detectat anomalies en les hores de producció i no ho ha fet en les hores de no producció. Es conclou que l'objectiu ha estat assolit satisfactòriament.

Del tercer objectiu, *Classificar les anomalies en diferents grups per a un possible manteniment preventiu futur*, si bé s'ha pogut diferenciar entre anomalies puntuals i temporals, no s'ha pogut arribar a fer la diferència entre les degradacions i les pèrdues de producció sostingudes tal i com s'havia volgut en un inici. Es conclou que l'objectiu ha estat assolit però caldrà seguir-hi treballant.

Del quart objectiu, *El codi ha de ser compatible amb altres mòduls de l'eina global*, s'ha adequat el codi per a que funcioni de manera independent, amb una estructura robusta i compacta, tot i així encara no s'ha implementat en el producte final pel propi

desenvolupament del projecte, ja que s'han desenvolupat en línies de temps diferents. Per tant, no es pot valorar l'assoliment d'aquest objectiu.

10.2 Millores i properes passes

En aquest punt del projecte, cal seguir treballant, primer, en els objectius que no s'han assolit satisfactòriament o que no es poden valorar i, seguidament, en l'optimització de la modelització de la predicció.

Tot i que la detecció d'anomalies es fa, i detecta, sens dubte, anomalies d'una manera satisfactòria, cal millorar la classificació d'aquestes entre puntuals, temporals de degradació i temporals sostingudes. Cal, doncs, repensar que és el que fa que una anomalia caigui dins d'un d'aquests tres grups i desenvolupar-ho, matemàticament, per a una millor resposta del mòdul de treball.

Per altra banda, cal seguir millorant l'optimització del codi, la seva estructura interna i les diferents variables que s'usen. Si bé és cert que, en un principi, aquest projecte funciona com a mòdul independent dins del projecte global (desenvolupament del producte "JoinEnergy" a càrrec de l'empresa Aiguasol), és necessari que s'adeqüi al funcionament global d'aquest. Com que el projecte d'aquest treball encara està en desenvolupament i no s'ha tancat, no s'ha buscat aquest encaix, però serà necessari de cara al tancament del producte final.

Finalment, cal afegir una etapa de recerca en el qual es millorin els paràmetres intrínsecs de cada model. Com s'ha explicat en el punt 6.2.1, els models *Random Forest*, *k-NN* i les xarxes neuronals disposen d'una sèrie de paràmetres que, per a aquest treball, no s'han modificat, però que es poden modificar a gust de l'usuari. Per tant, caldria aprofundir en la tria d'aquests paràmetres. Per a fer-ho es recomana l'ús de la híperparametrització, una eina on, donada una matriu de paràmetres, es generen diferents models amb diferents configuracions d'aquests paràmetres i se n'obtenen els més idonis, aquells que aconsegueixen un error menor de les dades predites en comparació a les dades reals [16].

Tot i que de moment el mètode que resol millor la predicció i detecció d'anomalies és el regressiu, es vol continuar treballant el mètode neuronal, que treballa amb les xarxes neuronals, per a aconseguir una major precisió i una detecció d'anomalies més fiable.

11 Referències

[1] Factorenergia, «Energies renovables: característiques, tipus i nous reptes,» 30 Agost 2018. [En línia]. Available: https://www.factorenergia.com/ca/blog/noticies/energies-renovables-caracteristiques-tipus-i-nous-reptes/. [Últim accés: 31 Gener 2021].

[2] Red Eléctrica de España, «Generación,» 31 Gener 2021. [En línia]. Available: https://www.ree.es/es/datos/generacion. [Últim accés: 31 Gener 2021].

[3] Grupo de Nuevas Actividades Profesionales, Energía Solar Fotovoltaiva, Madrid: Colegio Oficial de Ingenieros de Telecomunicación, 2002.

[4] P. J. Brockwell i R. A. Davis, Introduction to Time Series and Forecasting, Nova York: Springer, 2002.

[5] S. Ray, «Commonly used Machine Learning Algorithms,» vol. I, 2015.

[6] «Anomaly Detection in Univariate Time-Series: A Survey on the State-of-the-Art,» 2020.

[7] J. Loy, Neural Network Projects with Python, Birmingham: Packt Publishing Ltd, 2019.

[8] Statistics How To, «Statistics How TO,» 2021. [En línia]. Available: https://www.statisticshowto.com/absolute-error/. [Últim accés: 10 Juny 2021].

[9] S. Iyengar, S. Lee, D. Sheldon i P. Shenoy, «SolarClique: Detecting Anomalies in Residential Solar Arrays,» *Proceedings of ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS),* p. 10, 20-22 Juny 2018.

[10] W. Rowe, «Mean Square Error & R2 Score,» BMC Software, 5 Juliol 2018. [En línia]. Available: https://www.bmc.com/blogs/mean-squared-error-r2-and-variance-in-regression-analysis/. [Últim accés: 5 Febrer 2021].

[11] S. Glen, «RMSE: Root Mean Square Error,» [En línia]. Available: https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error. [Últim accés: 5 Febrer 2021].

[12] Pecan Street, «Pecan Street Dataport,» 2019. [En línia]. Available: https://dataport.pecanstreet.org/. [Últim accés: 2 Febrer 2021].

[13] J. Brownlee, «Machine Learning Mastery,» 4 Febrer 2019. [En línia]. Available: https://machinelearningmastery.com/how-to-improve-neural-network-stability-and-modeling-performance-with-data-scaling/. [Últim accés: 3 Juny 2021].

[14] J. Brownlee, «Machine Learning Algorithms: Overfitting and Underfitting With Machine Learning Algorithms,» Machine Learning Algorithms, 21 Març 2016. [En línia]. Available: https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/. [Últim accés: 19 Abril 2021].

[15] A. Cook, «Intermediate Machine Learning: Pipelines,» kaggle, [En línia]. Available: https://www.kaggle.com/alexisbcook/pipelines. [Últim accés: 10 Febrer 2021].

[16] W. Koehrsen, «Towards Data Sience,» 10 Gener 2018. [En línia]. Available: https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74. [Últim accés: 10 Juny 2021].