laSalle
UNIVERSITAT RAMON LLULL

Escola Tècnica Superior d'Enginyeria
Electrònica i Informàtica La Salle

Treball Final de Màster

Màster Universitari en Data Science

Validació de certificats energètics d'edificis i habitatges

Alumne

Pau Durà Yvern

Professor Ponent

Dr. Álvaro Sicília Gómez

ACTA DE L'EXAMEN
DEL TREBALL FI DE MASTER

Reunit el Tribunal qualificador en el dia de la data, l'alumne

D. Pau Durà Yvern

va exposar el seu Treball de Fi de Màster, el qual va tractar sobre el tema següent:

Validació de certificats energètics d'edificis i habitatges

Acabada l'exposició i contestades per part de l'alumne les objeccions formulades pels Srs. membres del tribunal, aquest valorà l'esmentat Treball amb la qualificació de

Barcelona,

VOCAL DEL TRIBUNAL VOCAL DEL TRIBUNAL

PRESIDENT DEL TRIBUNAL

RESUM

Els certificats energètics són un component important en l'estudi i gestió dels habitatges i edificis dels municipis de Catalunya. Degut a la gran quantitat de certificadors que hi ha i al alt cost que representa un càlcul d'aquestes característiques, està sota sospita la correcta realització d'aquests documents.

L'òrgan responsable de la seva recol·lecció i estudi (Institut Català d'Energia, ICAEN), ens ha proporcionat una base de dades amb més de 500.000 certificats pendents de validació i una altra amb 1.000 certificats teòricament validats. Una vegada haguem contrastat que podem utilitzar la base validada per extrapolar resultats als 500.000 certificats, mitjançant tècniques de regressió clàssiques i de *machine learning,* construirem models que permetran determinar quins certificats són correctes i quins s'haurien de repetir a partir de la predicció de les emissions de CO2.

RESUMEN

Los certificados energéticos son un componente importante en el estudio y gestión de las viviendas y los edificios de los municipios de Cataluña. Debido a la gran cantidad de certificadores que hay y al alto coste que representa un cálculo de dichas características, está bajo sospecha la adecuada realización de estos documentos.

El órgano responsable de su recolección y estudio (Institut Català d'Energia, ICAEN), nos proporciona una base de datos con más de 500.000 certificados pendientes de validación y otra con 1.000 certificados teóricamente validados. Una vez hayamos validado que podemos utilizar a base validada para extrapolar resultados a los 500.000 certificados, mediante técnicas de regresión clásicas y de *Machine learning*, construiremos modelos que permitirán determinar qué certificados son correctos y qué certificados se deberían repetir a partir de la predicción de las emisiones de CO2.

ABSTRACT

Energy certificates are an important component in the study and management of housing and buildings in the municipalities of Catalonia. Due to the large number of certifiers that exist and at the high cost that represents a calculation of these characteristics, the correct execution of these documents is suspected.

The responsible institution of its collection and study (Institut Català d'Energia, ICAEN), has provided us with a database with more than 500.000 certificates pending validation and another with 1,000 certificates theoretically validated. Once we have contrasted that we can use the validated base to extrapolate results to 500.000 certificates, using classical regression techniques and *machine learning*, we will build models that will allow to determine which

certificates are correct and which ones should be repeated from the prediction of the CO2 emissions.

PARAULES CLAU

Certificat energètic, ICAEN, Emissions, CO2, *smart city*, Eficiència energètica, Regressió lineal, *Gradient Boosting*, *Machine learning*, Aprenentatge automàtic.

PALABRAS CLAVE

Certificado energético, ICAEN, Emisiones, CO2, *smart city*, Eficiencia energética, Regresión lineal, *Gradient Boosting*, *Machine learning*, Aprendizaje Automático.

KEYWORDS

Energy certificate, ICAEN, Emissions, CO2, smart city, Energy efficiency, Linear regression, Gradient Boosting, Machine learning, Automatic learning.

Índex

1. INTRODUCCIÓ

1.1 QUÈ ÉS UN CERTIFICAT ENERGÈTIC I PER QUÈ SERVEIX?

L'eficiència energètica dels edificis i habitatges juga un paper cada cop més important en la transició energètica cap a un nou model energètic net i renovable. En aquest procés, el paper del ciutadà és clau ja que passa de ser un consumidor passiu a ser un agent actiu productor i consumidor d'energia. Serà a l'edifici on, per exemple, s'integrarà la producció d'energia distribuïda amb energies renovables, i on es podrà fer efectiu l'autoconsum i la recàrrega del vehicle elèctric.

La certificació energètica dels edificis i habitatges va quedar definida a la Directiva Europea 2002/91/CE [1] relativa a l'eficiència energètica dels edificis. Aquesta directiva indica al seu article 3 que, "els Estats membres aplicaran, a escala nacional o regional, una metodologia de càlcul de l'eficiència energètica dels edificis". Així mateix, també estableix a l'article 7 que "els Estats membres vetllaran per a què, quan els edificis siguin construïts, venuts o llogats, es posi a disposició del propietari o, per part del propietari, a disposició del possible comprador o llogater, segons correspongui, un certificat d'eficiència energètica". Com a transposició d'aquesta directiva es va publicar a Espanya el Reial decret 47/2007 [2], que aprovava el Procediment bàsic per a la certificació d'eficiència energètica d'edificis de nova construcció. Segons aquest Reial decret, està en mans de les Comunitats Autònomes crear un registre d'aquests certificats, fer un control i inspecció dels mateixos i establir el procediment per a la seva renovació o actualització. Al 2010 es va publicar la Directiva 2010/31/UE [3] relativa a l'eficiència energètica dels edificis. Tres anys més tard, es va publicar el Reial decret 235/2013 [4] pel que s'aprova el procediment bàsic per a la certificació de l'eficiència energètica dels edificis. És amb aquesta legislació que esdevé obligatòria a Espanya la certificació energètica d'edificis tant en els edificis de nova construcció, com en els edificis existents que es venen o lloguen, així com en els edificis ocupats per una entitat pública.

Per entendre-ho a nivell pràctic, el propietari que posa a la venda o lloga un habitatge, un local o una oficina té l'obligació de posar a disposició del comprador o del llogater un certificat d'eficiència energètica. Aquest certificat ha d'incloure la informació relativa a l'eficiència energètica de l'edifici amb l'objectiu que els nous inquilins la puguin comparar i avaluar. Aquest eficiència energètica es comprova exposant el consum total anual d'energia necessari per satisfer la demanda d'un habitatge concret o de tot l'edifici en condicions normals.

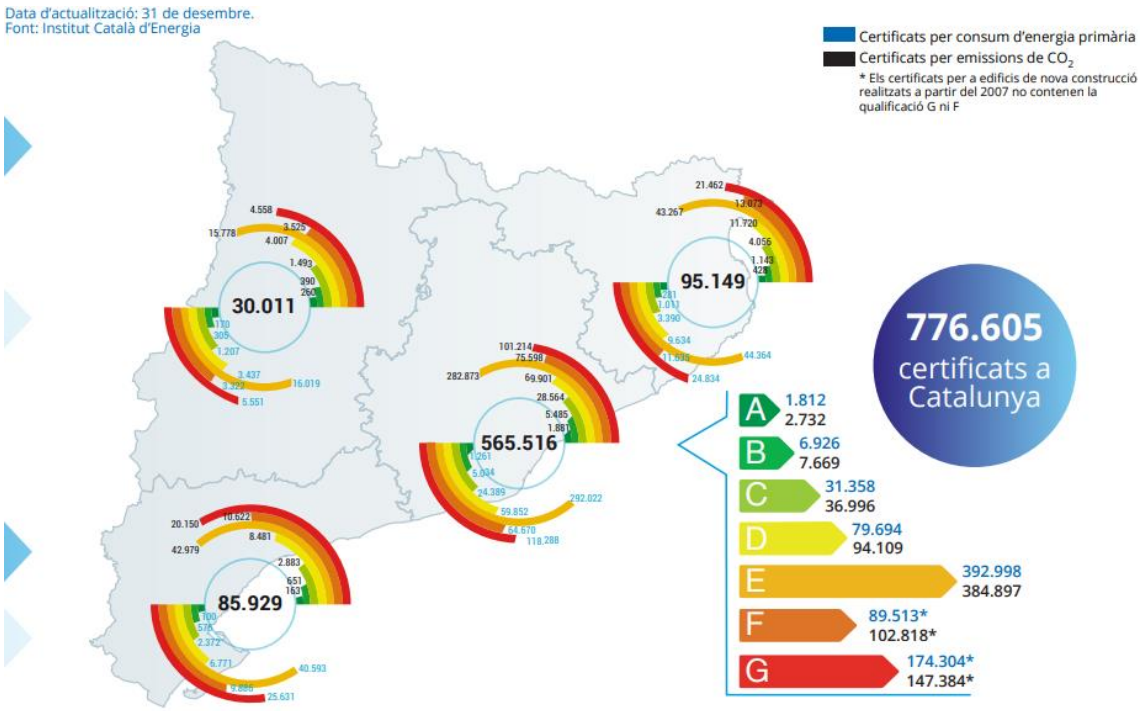
El certificat energètic ha d'estar fet per una persona amb un càrrec o estudis tècnics competents com per exemple arquitectes o enginyers. Aquest professional serà l'encarregat d'assignar a l'edifici o parts de l'edifici una etiqueta energètica amb una lletra, de la A a la G, de més a menys eficiència energètica.

Una vegada està redactat el certificat, aquest s'envia a l'Institut Català d'Energia [5] (ICAEN) que és l l'òrgan competent a Catalunya i aquest expedeix l'etiqueta d'eficiència energètica i fa la inscripció en el registre.

A Catalunya el 60% del parc d'habitatges és anterior al 1980, data en què es va iniciar el marc normatiu obligatori de l'aïllament tèrmic. Les Directives Europees d'Eficiència Energètica (2012/27/EU [6]) i d'Eficiència Energètica en Edificis (2010/31/EU) estableixen que tots els edificis (existents o d'obra nova) objecte de transaccions immobiliàries han d'obtenir un certificat d'eficiència energètica. A Catalunya, des de l'1 de novembre de 2007, els edificis de nova construcció han de disposar d'un certificat energètic, i des de l'1 de juny de 2013 també és obligatori per als edificis i habitatges ja existents que es lloguin o es venguin.

1.2 PROBLEMÀTICA ACTUAL I OBJECTIU DEL TREBALL

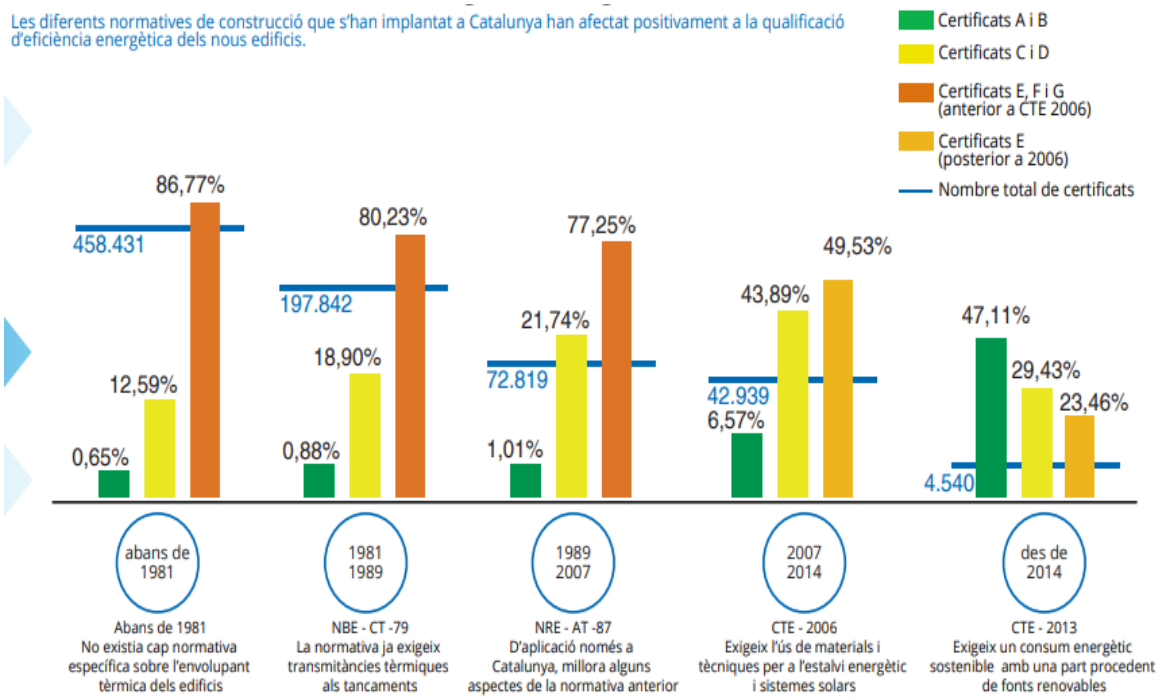
El total de certificats energètics a Catalunya el podem veure en el següent gràfic:



Imatge 1: Mapa de certificats. Imatge procedent del ICAEN [5]
http://icaen.gencat.cat/ca/energia/usos_energia/edificis/certificacio

Com podem veure, estem parlant de més de 776.000 certificats expedits. A nivell de distribució, els de mala qualificació (E, F, G) són encara més nombrosos que els de bona qualificació (A, B, C). Això és així perquè com hem dit abans hi ha molts edificis antics a Catalunya. Com era d'esperar, la província de Barcelona és la que té més certificats amb un 72% del total i en un principi sorprèn la poca representativitat de la província de Lleida amb només un 3% del total.

Una altra opció que tenim és veure aquesta distribució de certificats al llarg del temps per veure si les diferents normatives que s'han anat aplicant han tingut impacte i el resultat és el següent:



Imatge 2: Evolució dels certificats. Imatge procedent de l'ICAEN [5]
http://icaen.gencat.cat/ca/energia/usos_energia/edificis/certificacio

Com podem veure a la imatge 2, l'aplicació de normatives ha estat molt positiva en els edificis de nova construcció ja que ara mateix ja quasi la meitat dels edificis nous són de molt bona qualificació. És a dir, anem per bon camí però encara queda molta feina a fer ja que es segueixen construint edificis poc eficients energèticament parlant.

Fet aquest primer punt de situació, cal comentar que el procés de certificació és llarg i requereix d'un seguit de càlculs extensos. Això fa que sigui un servei costós a nivell econòmic ja que el tècnic que es dedica a fer les inspeccions s'hi ha de passar vàries hores al edifici. En un mercat on hi ha molts certificadors i no hi ha un control directe sobre la validesa final del certificat és racional pensar que es poden estar cometent infraccions en els protocols per tal de reduir el cost i ser més atractius davant la competència. És a dir, si ningú revisarà el certificat, pot ser que sigui una bona estratègia comercial reduir el temps de visita als edificis, "falsificar" els certificats i reduir el cost del servei.

És per aquestes sospites que té l'ICAEN que s'estan plantejant un seguit de possibles solucions per evitar aquestes conductes. La primera idea seria revisar un per un tots els certificats que arriben al registre però queda descartada directament per la volumetria que hem vist en la imatge 1 ja que no podem estar revisant milers de certificats 1 a 1. És per això que es planteja la idea de fer un controlador automàtic que faci saltar l'alarma per a aquells certificats que siguin més difícils de creure.

Per tant, l'objectiu d'aquest treball és desenvolupar un controlador automàtic que serà un model que farà prediccions de les emissions de CO2 del edificis mitjançant les dades del certificat. És a dir, si les emissions de CO2 que ha posat el tècnic són molt diferents a les esperades tenint en compte les condicions de l'edifici saltarà l'alarma i els tècnics del ICAEN sabran que aquell certificat és potencialment poc rigorós.

1.3 METODOLOGIA I PROCÉS SEGUIT

Una vegada vist l'enfoc del problema i la solució que plantegem cal veure amb quines eines podem treballar. Com hem dit anteriorment, la font d'informació és l'ICAEN. L'Institut Català de l'Energia és el responsable de col·leccionar tots els certificats i és qui ens ha donat accés a la informació que comentarem a continuació.

Partim de dues bases de dades amb les mateixes variables però de diferents mides i condicions. Tenim una base amb 1.281 registres de certificats validats que és amb la que farem el model. A més, tenim una segona base amb més de 500.000 certificats no validats que són els que hauran de rebre el vist i plau del model creat amb la base petita.

Dit això, en la primera fase del treball ens centrarem a validar les hipòtesis necessàries perquè la solució plantejada funcioni:

1) Existeix una relació entre les variables independents del certificat i les emissions de CO2 per a poder fer el model.

2) La base petita és representativa de la base gran:

    - La variable resposta (emissions de CO2) segueix la mateixa distribució a les dues bases.
    - Les variables independents del certificat segueixen la mateixa distribució a les dues bases.
    - La relació entre les variables independents del certificat i les emissions de CO2 és la mateixa a les dues bases.

3) Les dades tenen una qualitat suficient com per creure'ns-les i crear relacions.

4) La variable resposta (emissions de CO2) segueix una distribució coneguda i modelable.

Si qualsevol d'aquestes hipòtesis no es compleix no té sentit fer el model amb la base petita ni aplicar-lo a la base gran i per tan la solució plantejada en aquest treball al problema dels certificats no seria vàlida.

Un cop vistes les hipòtesis bàsiques que hem de complir, els passos que seguirem durant el present document seran els següents:

Primer de tot, presentació de la base de dades: estructura, neteja d'errors i explicació de les possibles variables d'interès així com la construcció de noves variables per a una millor modelització. Un cop tinguem la base de dades preparada, caldrà revisar les hipòtesis que hem plantejat en aquest mateix punt. És per això que compararem les dues bases (validada i no validada) i mirarem la relació entre la variable resposta i la resta d'atributs de la base.

Un cop tinguem aquests punts fets, passarem a la modelització de les emissions de CO2 on compararem dues tècniques: primer de tot utilitzarem la regressió simple i posteriorment la tècnica de *Machine Learning Gradient Boosting* per tal de veure les diferències entre un mètode i l'altre. En aquest punt, entendrem els models i interpretarem els resultats.

Tots els processos de càlcul i de visualització es portaran a terme amb el software estadístic R i els seus paquets. S'ha triat aquest software ja que ens és especialment vàlid per la creació de models amb *Machine learning* i a més és gratuït.

Per finalitzar, passarem a comentar les conclusions del treball i els potencials camins de futur.

2. BASES DE DADES DE L'ESTUDI

En aquest apartat farem un anàlisi bàsic de les bases de dades que treballarem i ens centrarem en la validació de les hipòtesis bàsiques que hem establert en la introducció.

2.1 DIMENSIONALITAT I VALORS ABSENTS DE LES BASES

Les dimensions de les bases són les següents:

- La base sense validar consta de 514.724 registres i 129 variables.
- La base validada consta de 1.281 registres i 129 variables.

Les variables de les dues bases són les mateixes i una breu descripció de cada una es pot veure a l'annex 1.

Pel que fa als valors faltants, hi ha un total de 78 variables amb valors absents a la base validada o a la base sense validar. És important tenir en compte la quantitat de *missings* que hi ha a les variables de la base per ser conseqüents després a l'hora d'aplicar les metodologies que plantegem. Per una banda, la regressió clàssica no funciona quan hi ha *missings* a les variable regressores mentre que el Xgboost sí que ho accepta però simplement el que fa és crear un valor mitjà per aquests valors. Concretament tenim:

Variable	Percentatge de *missings* a la base validada (%)	Percentatge de *missings* a la base no validada (%)
VENTILACIO_USO_RESIDENCIAL	5,87%	9,26%
DEMANDA_CONJUNTA	60,53%	0,80%
POTENCIA_TOTAL_INSTAL	38,91%	91,33%
INSTAL_TERM_CONSUM	99,12%	99,82%
CONSUMGASNATURAL_GLOBAL	21,7%	0,22%
CONSUMGASNATURAL_CALEF	21,7%	0,22%
CONSUMGASNATURAL_REFRIG	21,7%	0,22%
CONSUMGASNATURAL_ACS	21,7%	0,22%
CONSUMGASNATURAL_ILU	46,22%	0,32%
CONSUMELECT_GLOBAL	1,29%	0,03%
CONSUMELECT_CALEF	1,29%	0,03%
CONSUMELECT_REFRIG	1,29%	0,03%
CONSUMELECT_ACS	1,29%	0,03%
CONSUMELECT_ILU	31,99%	0,16%
CONSUMCARBON_GLOBAL	92,68%	0,97%
CONSUMCARBON_CALEF	92,68%	0,97%
CONSUMCARBON_REFRIG	92,68%	0,97%
CONSUMCARBON_ACS	92,68%	0,97%
CONSUMCARBON_ILU	92,68%	0,97%
CONSUMGASOLEO_GLOBAL	88,18%	0,93%

CONSUMGASOLEO_CALEF	88,18%	0,93%
CONSUMGASOLEO_REFRIG	88,18%	0,93%
CONSUMGASOLEO_ACS	88,18%	0,93%
CONSUMGASOLEO_ILU	90,43%	0,94%
CONSUMRENOVABLE_ILU	87,06%	0,41%
CONSUMGLP_GLOBAL	91,56%	0,93%
CONSUMGLP_CALEF	91,56%	0,93%
CONSUMGLP_REFRIG	91,56%	0,93%
CONSUMGLP_ACS	91,56%	0,93%
CONSUMGLP_ILU	92,36%	0,94%
CONSUMBIOMASSA_GLOBAL	90,19%	0,96%
CONSUMBIOMASSA_CALEF	90,19%	0,96%
CONSUMBIOMASSA_REFRIG	90,19%	0,96%
CONSUMBIOMASSA_ACS	90,19%	0,96%
CONSUMBIOMASSA_ILU	90,68%	0,96%
CONSUMBIOCARB_GLOBAL	92,68%	0,97%
CONSUMBIOCARB_CALEF	92,68%	0,97%
CONSUMBIOCARB_REFRIG	92,68%	0,97%
CONSUMBIOCARB_ACS	92,68%	0,97%
CONSUMBIOCARB_ILU	92,68%	0,97%
CONSUMPELLET_GLOBAL	87,30%	0,95%
CONSUMPELLET_CALEF	87,30%	0,95%
CONSUMPELLET_REFRIG	87,30%	0,95%
CONSUMPELLET_ACS	87,30%	0,95%
CONSUMPELLET_ILU	89,07%	0,96%
CERRAMIENTOSOPACOSSUP	0,08%	0,01%
HUECOSYLUCERNARIOSSUP	0,00%	0,07%
PUENTESTERMICOSTRANS	32,80%	3,12%
HUECOSYLUCERNARIOSTRANS	0,08%	0,07%
PUENTESTERMICOSLONG	32,80%	3,12%
CERRAMIENTOSOPACOSESPESSOR	90,68%	98,3%
PUENTESTERMICOSESPESSOR	94,13%	98,44%
HUECOSYLUCERNARIOSESPESSOR	90,68%	98,30%
CERRAMIENTOSOPACOSCOND	90,68%	98,31%
PUENTESTERMICOSCOND	94,21%	98,44%
HUECOSYLUCERNARIOSCOND	90,68%	98,31%
CERRAMIENTOSOPACOSRESIS	94,94%	99,31%
PUENTESTERMICOSRESIS	97,75%	98,95%
HUECOSYLUCERNARIOSRESIS	95,58%	99,27%
CERRAMIENTOSOPACOSDENS	90,68%	98,30%
PUENTESTERMICOSDENS	94,13%	98,44%
HUECOSYLUCERNARIOSDENS	90,68%	98,30%
CERRAMIENTOSOPACOSVAPO	90,68%	98,30%
PUENTESTERMICOSVAPO	94,13%	98,44%
HUECOSYLUCERNARIOSVAPO	90,68%	98,30%

CERRAMIENTOSOPACOSCALOR	90,68%	98,30%
PUENTESTERMICOSCALOR	94,13%	98,44%
HUECOSYLUCERNARIOSCALOR	90,68%	98,30%
INSTCALEF_POTENCIA	61,50%	33,45%
INSTCALEF_RENDIM	3,05%	33,45%
INSTACS_POTENCIA	64,63%	4,60%
INSTACS_RENDIM	61,09%	4,60%
INSTREFRIG_POTENCIA	79,90%	64,82%
INSTREFRIG_RENDIM	21,95%	64,82%
EMISSIONS_ILLUMINACIO	85,56%	0,89%
NOMB_PLANT_BAJORASANT	8,20%	0,01%
ANY_CONSTRUCCIO	7,88%	0,00%

Com podem veure, algunes variables tenen en ambdues bases més del 90% dels valors absents. En el nostre cas, decidim tractar-les eliminant-les del nostre estudi en comptes d'imputar els valors ja que considerem que són uns percentatges massa grans com per imputar i correríem el risc de desvirtuar la distribució original de la variable reduint dràsticament la seva variabilitat original. Aquest fet comportaria un possible mal càlcul d'estimació dels seus coeficients i de la seva variabilitat i a més estaríem tractant aquesta variable de forma errònia al considerar que coneixem la seva distribució original quan en realitat no ho fem.

2.2 TRACTAMENT DE LA BASE DE DADES VALIDADA

Com a primer pas de l'estudi, cal que portem a terme una valoració de la qualitat de les dades i una possible modificació/eliminació d'alguns valors que considerem erronis (per exemple, una edat o un temps negatiu).

Per fer això, ens valdre'm d'anàlisis descriptius bàsics de totes les variables de les quals disposem i anirem modificant allò que calgui modificar. Inicialment, portem a terme:

1) Uniformització de la variable MUNICIPI que indica el municipi al que pertany l'edifici o la llar. En aquest cas, passem totes les lletres a minúscules per evitar tenir en grups diferents els mateixos municipis escrits amb mida de lletra diferent.

2) Canvi de format de totes les variables de la base: passem a caràcter totes aquelles que són categòriques i passem a numèriques totes aquelles que són ordinals/numèriques.

3) Eliminació de tres variables sense informació que només contenen *missings* [1]:
   a. CERRAMIENTOSOPACOSLONG
   b. HUECOSLUCERNARIOSLONG
   c. PUENTESTERMICOSSUP

4) Modificació de valors de la variable CODIGO_POSTAL: Es substitueix el valor "código postal" per valor *missing.*

[1] Podem veure les descripcions de les variables a l'annex 1

5) Eliminació de registres on la variable EMISSIONS_GLOBAL no està informada o on el seu valor és 0.

6) Modificació de valors de les variables SUPERF_HABIT_CALEF i SUPERF_HABIT_REFRIG: Es canvia els valors superior al 100% per un valor *missing.*

7) Modificació de valor a la variable CERRAMIENTOSOPACOSSUP: Es canvia un valor de 45.405 per un valor *missing* ja que és massa elevat per creure-se'l.

2.3 SELECCIÓ DE VARIABLES D'INTERÈS

El mètode de selecció de variables d'interès ha estat el numèric o estadístic. És a dir, a falta de coneixement expert sobre el domini sobre el que treballem, hem decidit seleccionar aquelles variables que tinguessin una relació positiva o negativa amb la nostra variable objectiu (les emissions de CO2). Aquest procés ha estat fet seguint la següent metodologia:

Pel que fa a les variables numèriques hem utilitzat el coeficient de correlació de Spearman [7] [2]. Hem seleccionat aquelles variables numèriques que tenien una correlació amb la variable resposta superior a 0,3. Tot i no ser un valor elevat, considerem que a l'estar parlant del "món real", qualsevol efecte, tot i que sigui petit, s'ha de tenir en compte i podria ser d'interès pel nostre model.

Pel que fa a les variables categòriques, hem utilitzat el mètode *apply* de R per tal de crear les mitjanes de les emissions de CO2 segons categories de forma ràpida. Si el valor de la mitjana és diferent segons la categoria que mirem llavors aquella variable és considerada d'interès.

Una vegada portat a terme aquest procés, ens quedem amb un total de 24 variables que són les següents:

- ZONA_CLIMATICA
- ANY_CONSTRUCCIO
- NORMATIVA_CONSTRUCCIO
- US_EDIFICI
- PROCEDIMENT_UTILITZAT
- ID_TIPUS_TRAMIT
- PROVINCIA
- PERC_SUPERF_HABIT_CALEFA
- PERC_SUPERF_HABIT_REFRIG
- PERC_SUPERF_ACRIST_NORDEST
- PERC_SUPERF_ACRIST_EST
- PERC_SUPERF_ACRIST_SUROEST

[2] La gran diferència entre el coeficient de Pearson i el de Spearman és que aquest últim és considerat com a un estadístic no paramètric. Això fa que puguem calcular la correlació entre dues variables sense tenir en compte la hipòtesis de que ambdues segueixen una distribució normal. La interpretació del coeficient de Spearman és igual que la de l'coeficient de correlació de Pearson. Oscil·la entre -1 i +1, indicant-nos associacions negatives o positives respectivament.

- PERC_SUPERF_ACRIST_SUREST
- PERC_SUPERF_ACRIST_OEST
- PERC_SUPERF_ACRIST_NORDOEST
- DENSITAT_FONTS_INTERNAS
- CERRAMIENTOSOPACOSSUP
- HUECOSYLUCERNARIOSSUP
- CERRAMIENTOSOPACOSTRANS
- HUECOSYLUCERNARIOSTRANS
- INSTCALEF_RENDIM
- NOMB_PLANT_BAJORASANT
- NOMB_PLANT_SOBRERASANT
- DEMANDA_GLOBAL

2.4 CREACIÓ DE NOVES VARIABLES

Una vegada tenim seleccionades les 24 variables que volem utilitzar a l'estudi, les modificarem per tal d'ajudar als models (o complir hipòtesis bàsiques en el cas de la regressió lineal simple), per crear relacions lineals més evidents en les variables numèriques, i per agrupar categories dins de les variables categòriques per tal de que no hi hagi grups amb una representació ínfima respecte al total de la població de la base. Aquest procés és bàsic perquè les metodologies que aplicarem vegin suficient gruix de població com per identificar canvis en la variabilitat entre grups. A més, també ens serveix per fer més robustos els models ja que si deixéssim les subpoblacions molt petites podríem estar parlant d'errors mostrals i llavors els models sortirien totalment esbiaixats a la realitat

Com a límit, hem considerat que cada grup ha de tenir com a mínim el 5% de la població total[3]. Tot i que sabem que és molt poc, seran categories que s'entendran com alertes: no es detecta quasi mai però quan es detecta vol dir que alguna cosa important està passant.

Aquest procés, com els anteriors, també serà variable a variable i seguint criteri numèric i no d'expert de domini.

Els resums dels canvis realitzats són els següents:

1) Creació de ZONA_CLIMATICA_2: S'agrupen algunes zones climàtiques per fer grups consistents.

2) Creació de ANY_CONSTRUCCIO_2: Dividim entre els edificis construïts abans del 2013 i després del 2013.

3) Creació de NORMATIVA_CONSTRUCCIO_2: Agrupem normatives segons espais temporals per fer grups consistents.

4) Creació de US_EDIFICI_2: Agrupem usos d'edificis per tal de tenir grups consistents.

[3] Criteri establert per la consultoría Boston Consulting Group pels models relacionats amb consum energètic. Com que la nostra variable està molt relacionada amb el consum energètic considerem que és un bon criteri de selecció del límit inferior de la mida de les categories.

5) Creació de PROCEDIMENT_UTILITZAT_2: Agrupem procediments per tal de tenir grups consistents.

6) Creació de ID_TIPUS_TRAMIT_2: Agrupem tipus de tràmits per tal de tenir grups consistents.

7) Creació de NOMB_PLANT_BAJORASANT_2: Dividim entre els que en tenen i els que no.

8) Creació de NOMB_PLANT_SOBRERASANT_2: Dividim entre els que en tenen 0 o 1 i la resta.

9) Creació de PERC_SUPERF_HABIT_CALEFA_2: Dividim entre el 100% calefactat, el 0% o un percentatge intermedi.

10) Creació de PERC_SUPERF_HABIT_REFRIG_2: Dividim entre el 100% refrigerat, el 0% o un percentatge intermedi.

11) Creació de PERC_SUPERF_ACRIST_NORDEST_2: Dividim entre els que tenen un 0% i més d'un 0%.

12) Creació de PERC_SUPERF_ACRIST_EST_2: Dividim entre els que tenen un 0% i més d'un 0%.

13) Creació de PERC_SUPERF_ACRIST_SUREST_2: Dividim entre els que tenen un 0% i més d'un 0%.

14) Creació de PERC_SUPERF_ACRIST_SUROEST_2: Dividim entre els que tenen un 0% i més d'un 0%.

15) Creació de PERC_SUPERF_ACRIST_OEST_2: Dividim entre els que tenen un 0% i més d'un 0%.

16) Creació de PERC_SUPERF_ACRIST_NORDOEST_2: Dividim entre els que tenen un 0% i més d'un 0%.

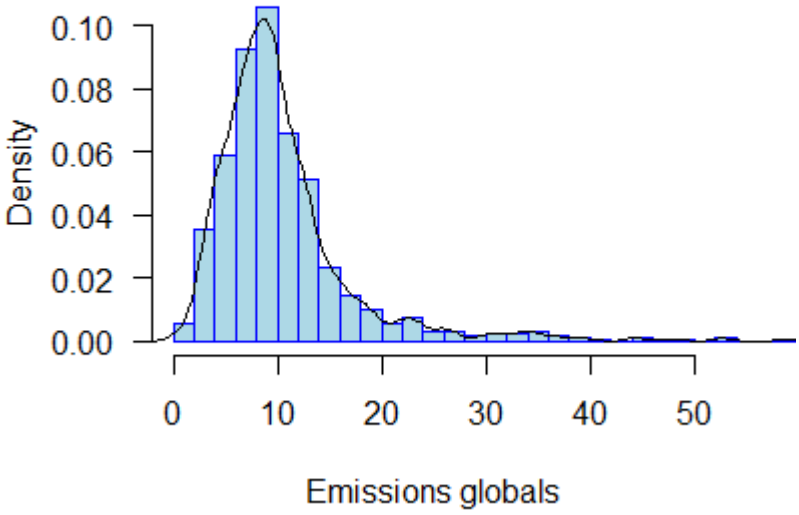
17) Creació de DENSITAT_FONTS_INTERNAS_2: Dividim entre els que tenen 0 i més de 0.

Una vegada construïdes aquestes noves variables, les canviarem per les originals a l'hora de modelar.

2.5 VARIABLE RESPOSTA: EMISSIONS DE CO2, ESTUDI I TRACTAMENT

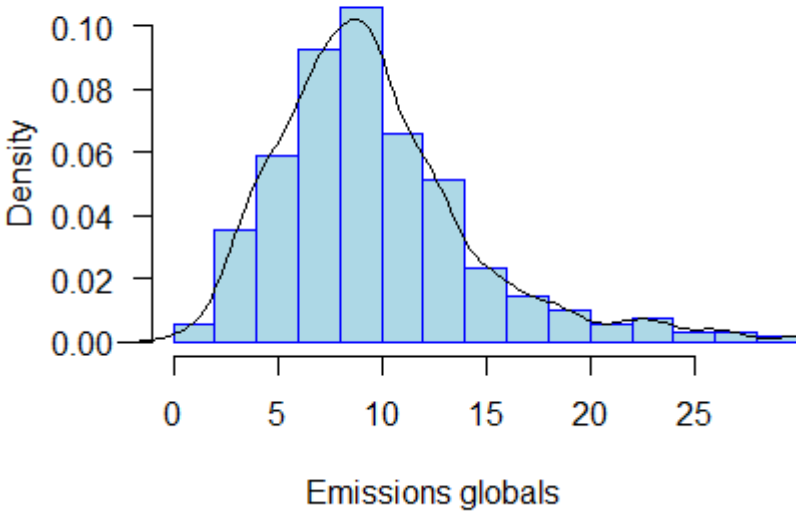
Una de les hipòtesis que hem de validar per tal que la idea que teníem (validar la base gran amb la petita) pugui funcionar, és comprovar que la variable d'emissions de CO2 de la base validada sigui representativa (mateixa distribució) a la de la base no validada.

Primer de tot, però, podem veure la distribució de les emissions de CO2 de la base validada a nivell numèric i gràfic.



Imatge 3: histograma de les emissions globals de la base validada

Si ampliem una mica més veiem:



Imatge 4: Histograma de les emissions globals de la base validada

Numèricament, ho veiem de la següent manera:

Mínim	1r quartil	Mediana	Mitjana	3r quartil	Màxim	Desviació
0	6,65	9,08	10,68	12,25	120,42	7,81

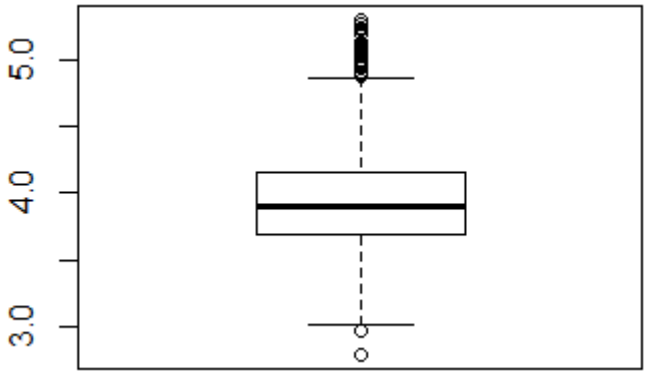
Com podem veure, la mitjana de les emissions de CO2 es situa al voltant del 10 i mig mentre que la mediana ho fa sobre el 9. Això ens indica que estem davant d'una distribució asimètrica i possiblement amb apuntament.

Pels següents apartats ens interessa saber si la variable es distribueix segons una distribució normal o no. Per comprovar-ho, utilitzem el test de Shapiro [8][4] amb els següents resultats:

Un estadístic de 0,689 i un p-valor de 2^-16. Recordem que el p-valor indica la probabilitat d'obtenir un valor igual o més extrem que l'observat, suposant que sigui veritat la hipòtesis nul·la.

És per això que rebutgem la hipòtesis nul·la i concloem que la distribució de les emissions globals no podem dir que sigui una distribució normal. Això ens portarà a transformar la variable com veurem més endavant.

Vista la distribució que segueixen les emissions, és possible que tinguem molts casos *outliers*. Considerem com a *outlier* totes aquelles observacions que estiguin fora de l'interval format per la mediana i dues vegades el rang interquartílic. Després d'eliminar els registres on les emissions són igual a 0, detectem un total de 91 casos outliers. D'aquests 91 outliers, decidim eliminar els més extrems al ser possibles dades errònies i quedar-nos amb la resta (76) ja que al cap i a la fi, representen el comportament de la variable resposta. Gràficament, ho podem veure amb un Boxplot:



Imatge 5: Boxplot outliers

Una alternativa per a un futur estudi seria imputar aquests valors *outliers* amb uns de nous mitjançant alguna tècnica d'imputació que generi variabilitat com per exemple la *Predictive Mean Matching.*

Una vegada vista la distribució inicial de la variable i retocats/eliminats alguns registres, podem veure si realment la hipòtesi en la que es sustenta el treball es compleix o no:

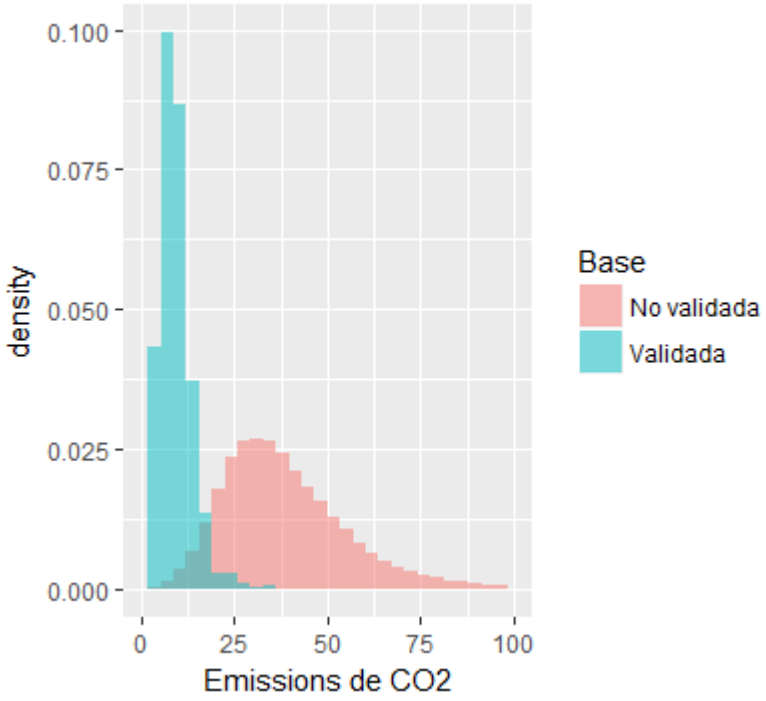
[4] Test que s'utilitza per contrastar la normalitat de les dades. Publicat el 1965 per Samuel Shapiro.

H0: Les dades de la base validada són representatives de la base per validar.

H1: Les dades de la base validada no són representativa de la base per validar.

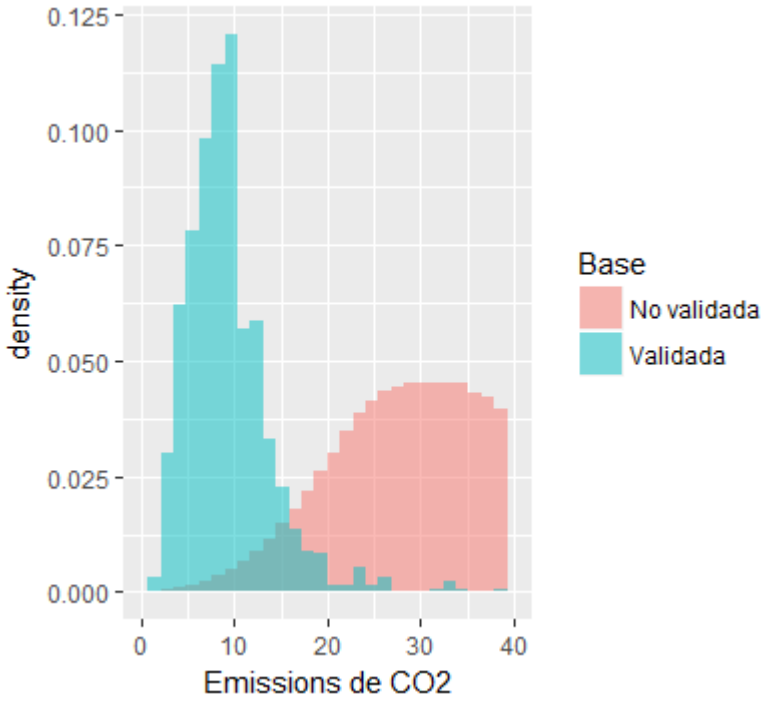
Per comprovar aquesta hipòtesi, comparem les dues distribucions (emissions globals de la base validada contra les emissions globals de la base per validar) gràfica i numèricament.

Gràficament, el resultat és:



Imatge 6: Comparació de distribucions 1

Fent una mica de zoom veiem més clarament que no s'assemblen gens:



Imatge 7: Comparació de distribucions 2

Numèricament, el canvi és molt substancial també:

	Mínim	1r quartil	Mediana	Mitjana	3r quartil	Màxim	Desviació
Sense validar	-3,01	26,82	36,27	40,2	48,6	21929,28	48,86
validada	1,55	6,36	8,67	9,28	11,28	39,14	4,51

A més, hi podem aplicar un test de comparació de distribucions no paramètric com el de Kolmogorov-smirnov [9][5]. El resultat és:

Un estadístic de 0,87 i un p-valor de 2.2e-16.

És a dir, rebutgem la hipòtesi nul·la i concloem que no tenim suficients indicis com per dir que estem parlant de la mateixa distribució. Com podem veure, les diferències entre les dues distribucions són abismals.

Veient els resums numèrics i gràfics, la conclusió és clara: **la hipòtesis nul·la sobre la que residia l'estudi queda rebutjada i per tan no té sentit validar la base sense validar amb models creats amb la base validada.**

A partir d'aquí, doncs, la resta del treball és a mode explicatiu però no es podrà posar en pràctica.

[5] És una prova no paramètrica que s'utilitza per determinar la bondat d'ajust de dues distribucions de probabilitat entre si.

3. METODOLOGIES DE MODELATGE

En els següents punts del treball utilitzarem dues metodologies per modelitzar les emissions globals de C02 utilitzant la base validada que ens ofereix l'ICAEN. Aquestes dues tècniques, tal com hem comentat a la introducció, són la regressió lineal simple i la tècnica de *Machine learning* anomenada *Gradient Boosting*. És necessari, doncs, fer com a mínim un breu esment sobre les tècniques que utilitzarem.

3.1 REGRESSIÓ LINEAL CLÀSSICA

Es tracta del model més senzill que podem trobar a la literatura. És a dir, és el model més bàsic però per això no ha de deixar de ser útil. El problema d'aquesta metodologia és que s'han de complir moltes hipòtesis perquè el model sigui vàlid, eficient i interpretable.

3.1.1 DEFINICIÓ

Podem considerar la regressió lineal com un model matemàtic que es fa servir per aproximar la relació de dependència entre una variable dependent Y, les variables independents X i un terme aleatori. Ho podem expressar com:

$$Y_t = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

Per calcular els coeficients de les variables independents, tenim el següent procés:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{21} & x_{31} & \cdots & x_{k1} \\ 1 & x_{22} & x_{32} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{2n} & x_{3n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

A partir d'aquí, estimem segons mínims quadrats. Tenint com a S a la suma dels quadrats dels residus:

$$S = \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} \left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki} \right]^2$$

Per aplicar el criteri de mínims quadrats en el model de regressió múltiple, calculem la primera derivada de S respecte a cada beta a l'expressió anterior i ens queda:

$$\frac{\partial S}{\partial \hat{\beta}_1} = 2\sum_{i=1}^{n}\left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki}\right][-1]$$

$$\frac{\partial S}{\partial \hat{\beta}_2} = 2\sum_{i=1}^{n}\left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki}\right][-x_{2i}]$$

$$\frac{\partial S}{\partial \hat{\beta}_3} = 2\sum_{i=1}^{n}\left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki}\right][-x_{3i}]$$

$$\cdots \qquad \cdots \qquad \cdots \qquad \cdots$$

$$\frac{\partial S}{\partial \hat{\beta}_k} = 2\sum_{i=1}^{n}\left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki}\right][-x_{ki}]$$

Els estimadors per mínims quadrats s'obtenen al igualar a 0 les derivades anteriors:

$$\sum_{i=1}^{n}\left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki}\right] = 0$$

$$\sum_{i=1}^{n}\left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki}\right] x_{2i} = 0$$

$$\sum_{i=1}^{n}\left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki}\right] x_{3i} = 0$$

$$\cdots \qquad \cdots \qquad \cdots \qquad \cdots$$

$$\sum_{i=1}^{n}\left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki}\right] x_{ki} = 0$$

O amb notació matricial:

$$\mathbf{X'X\hat{\beta}} = \mathbf{X'y}$$

A aquest sistema anterior se l'anomena genèricament sistema d'equacions normals de l'hiperplà. En notació matricial ampliada, el sistema d'equacions normals és el següent:

$$\begin{bmatrix} n & \sum x_{2i} & \cdots & \sum x_{ki} \\ \sum x_{2i} & \sum x_{2i}^2 & \cdots & \sum x_{2i}x_{ki} \\ \vdots & \vdots & \ddots & \vdots \\ \sum x_{ki} & \sum x_{ki}x_{2i} & \cdots & \sum x_{ki}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_{2i}y_i \\ \vdots \\ \sum x_{ki}y_i \end{bmatrix}$$

Com podem veure, en aquest sistema tenim *k* equacions i *k* incògnites (les betes). Aquest sistema es pot resoldre fàcilment utilitzant àlgebra matricial. Amb l'objectiu final de resoldre el sistema respecte a les betes, és necessari que el rang de la matriu X'X sigui igual a *k.* Si això es compleix, podem invertir:

$$\left[\mathbf{X'X}\right]^{-1}\mathbf{X'X\hat{\beta}} = \left[\mathbf{X'X}\right]^{-1}\mathbf{X'y}$$

Obtenint l'expressió del vector d'estimadors per mínims quadrats, o més exactament, el vector d'estimadors per mínims quadrats ordinaris. Per tant, la solució és la següent:

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \hat{\boldsymbol{\beta}} = [\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{y}$$

Tenint en compte que la matriu de segones derivades 2X'X és una matriu definida positiva, la conclusió és que S presenta un mínim en les betes estimades.

3.1.2 HIPÒTESIS I SUPÒSITS

Com hem comentat anteriorment, el problema principal d'aquesta metodologia és que conté un seguit d'hipòtesis i supòsits molt restrictius/ves.

HIPÒTESIS

1) Esperança matemàtica nul·la. Per a cada valor de X, la pertorbació prendrà diferents valors de forma aleatòria, però no prendrà sistemàticament valors positius o negatius, sinó que es suposa que prendrà alguns valors majors que zero i alguns menors que zero de tal manera que el seu valor esperat sigui 0.

2) Homoscedasticitat: Tots els termes de la pertorbació tenen la mateixa variància que és desconeguda. La dispersió de cada pertorbació en torn al seu valor esperat és sempre la mateixa.

3) Incorrelació o independència: Les covariàncies entre les diferents pertorbacions són nul·les (no correlacionades). Això implica que el valor de la pertorbació per a qualsevol observació mostral no ve influenciada pels valors de les pertorbacions corresponents a altres observacions mostrals.

4) Regressors estocàstics: si repetíssim el model amb una nova mostra, els resultats haurien de ser els mateixos.

5) Independència lineal: No existeixen relacions lineals exactes entre els regressors.

6) No existeixen errors d'especificació en el model, ni errors de mesura en les variables explicatives.

7) Les distribucions de les pertorbacions segueixen una normal amb mitjana 0 i variància sigma al quadrat.

SUPÒSITS

1) La relació entre les variables és lineal.
2) Els errors de mesura de les variables explicatives són independents entre si.
3) Els errors han de tenir variància constant (homoscedasticitat una altra vegada).
4) Els errors han de tenir una esperança matemàtica igual a 0.
5) L'error total ha de ser la suma de tots els errors.

3.2 GRADIENT BOOSTING

El *Gradient Boosting* és una tècnica d'aprenentatge automàtic per resoldre problemes de regressió i classificació que construeix el model de forma escalonada, tal i com fan altres mètodes de *boosting*, però permetent la optimització d'una funció de pèrdua arbitrària i diferenciable. La idea del *Gradient Boosting* va sorgir a partir de la interpretació que va fer Leo Breiman[6] sobre les metodologies *boosting* "el *boosting* pot ser interpretat com un algorisme d'optimització en una funció de pèrdua adequada".

Posteriorment, va ser Jerohme H.Friedman [10] [7] qui va desenvolupar l'algorisme del *Gradient Boosting,* introduint la idea de veure els algorismes de *boosting* com algorismes iteratius del mètode d'optimització *Gradient Descent.*

Així, per entendre el funcionament del *Gradient Boosting* és necessari entendre primerament el funcionament general de *boosting*, així com conèixer l'algorisme del *Gradient Descent*.

3.2.1 BOOSTING

El *boosting* és una tècnica d'aprenentatge automàtic basada en la construcció d'un model complex a partir d'una seqüència additiva de models simples als que anomenarem deixebles base.

És un procés iteratiu, en el que cada nova iteració consisteix en la creació d'un nou deixeble. El nou deixeble depèn de la seqüència de deixebles creada fins el moment i es crea en funció de l'error observat. Cal remarcar que una nova iteració no canvia els deixebles de la seqüència obtinguda fins el moment.

[6] Leo Breiman (gener de 1928-juliol de 2005) va ser un estadístic de la Universitat de California, Berkeley. La seva feina va contribuir a tancar el forat entre la estadística i el *machine learning*. Les seves contribucions més importants van ser en el camp dels arbres de classificació i de regressió.
[7] Jerome H. Friedman, estadístic de la Universitat de Stanford, conegut per les seves aportacions en el camp de l'estadística i de la mineria de dades.

3.2.2 GRADIENT DESCENT

El *Gradient Descent* és un algorisme d'optimització per trobar un mínim en una funció. Sent *F* una funció definida i diferenciable al voltant d'un punt *a*, llavors la funció decreix el més ràpid possible si un va, des de a en la direcció del gradient negatiu de *F* en *a*. És a dir, per trobar un mínim local a *F*, fem passos proporcionals al gradient negatiu de *F* en *a*.

3.2.3 ALGORISME GRADIENT BOOSTING

En un problema de predicció es té un sistema format per una variable resposta y i un conjunt de variables explicatives x = (x1, ..., xn).

Resoldre el problema és trobar la funció que s'aproxima millor a la funció f(x) que posa sobre un mapa x i y, és a dir, és trobar la funció que minimitza l'esperança en una funció de pèrdua donada. L'algorisme del *gradient boosting* fa servir la connexió entre el *boosting* i la optimització per trobar aquesta funció.

Inicialització: S'inicialitza la funció com una constant = C.

Per a t en 1, ..., T:

Es calcula el gradient:

$$z_i = -\frac{\partial}{\partial f(\mathbf{x}_i)}\Psi(y_i, f(\mathbf{x}_i))\Big|_{f(x_i)=\hat{f}(x_i)} \quad (1)$$

S'ajusta el model g(x) que prediu Z_i a partir de les variables explicatives x_i i s'escull el pas en direcció al gradient:

$$\rho = \arg\min_{\rho} \Psi(y_i, \hat{f}(\mathbf{x}_i) + \rho g(\mathbf{x}_i)) \quad (2)$$

S'actualitza l'estimació de f(x):

$$\hat{f}(\mathbf{x}) = \hat{f}(\mathbf{x}) + g(\mathbf{x}) \quad (3)$$

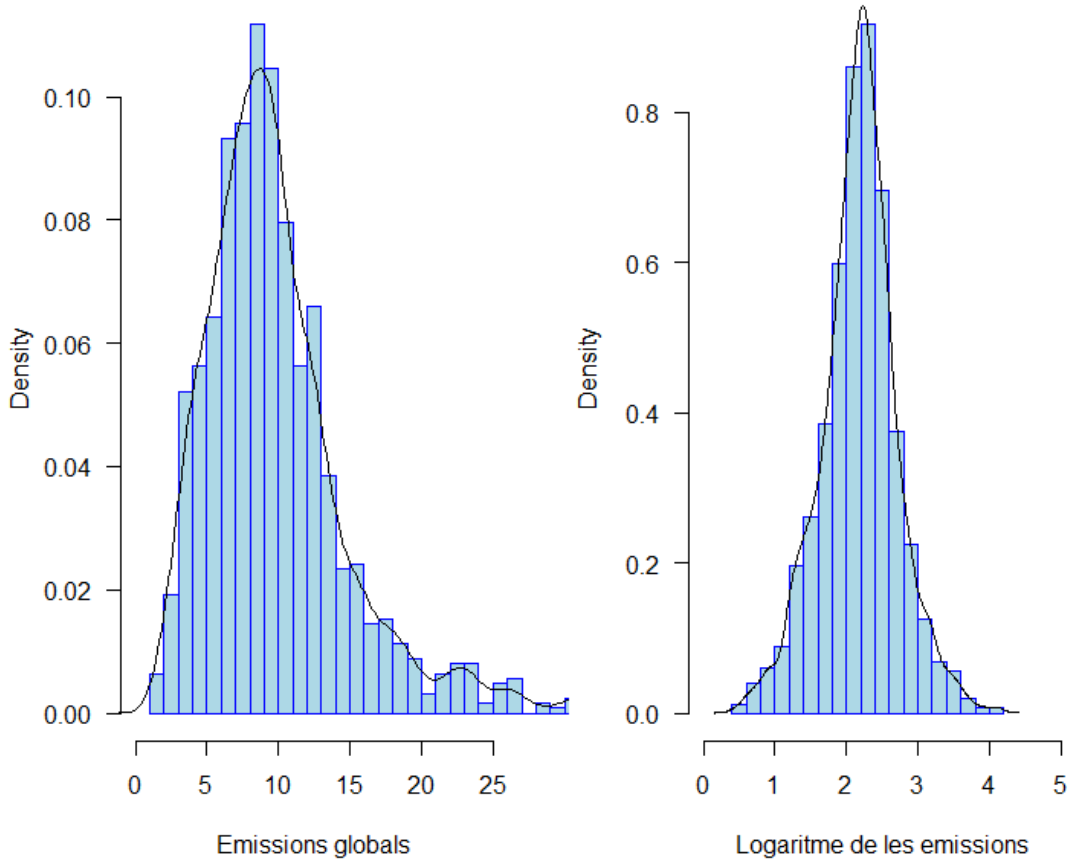
En aquesta treball, s'ha utilitzat el paquet *xgboost* disponible en el software lliure R. Aquest paquet es denomina *Extreme Gradient Boosting* (XGBoost) i és una implementació eficient del *Gradient Boosting*. El que fa únic al XGBoost és que utilitza una formalització més regularitzada del model per poder controlar millor el sobre ajustament i que realitza, de forma automàtica, la

paral·lelització del procés en una sola màquina, la qual cosa suposa que treballa al voltant de 10 vegades més ràpid que altres implementacions.

3.3 TRANSFORMACIÓ DE LA VARIABLE RESPOSTA

Com hem vist en un capítol anterior, les emissions de CO2 no segueixen una distribució normal. Això fa que no hi puguem aplicar els models plantejats ja que en les seves hipòtesis queda remarcat que la variable resposta ha de ser una distribució normal. Si no ho és, els principis bàsics queden afectats i el model passa a no ser vàlid.

És per això que hem de transformar la variable resposta per tal de que segueixi la distribució normal desitjada i la transformació escollida és la logarítmica. És a dir, modelitzarem el logaritme de les emissions de CO2 ja que com podem veure a continuació segueix una distribució més semblant a la normal. Comparant-ho amb les emissions sense logaritme:



Imatge 8: Transformació de la variable resposta

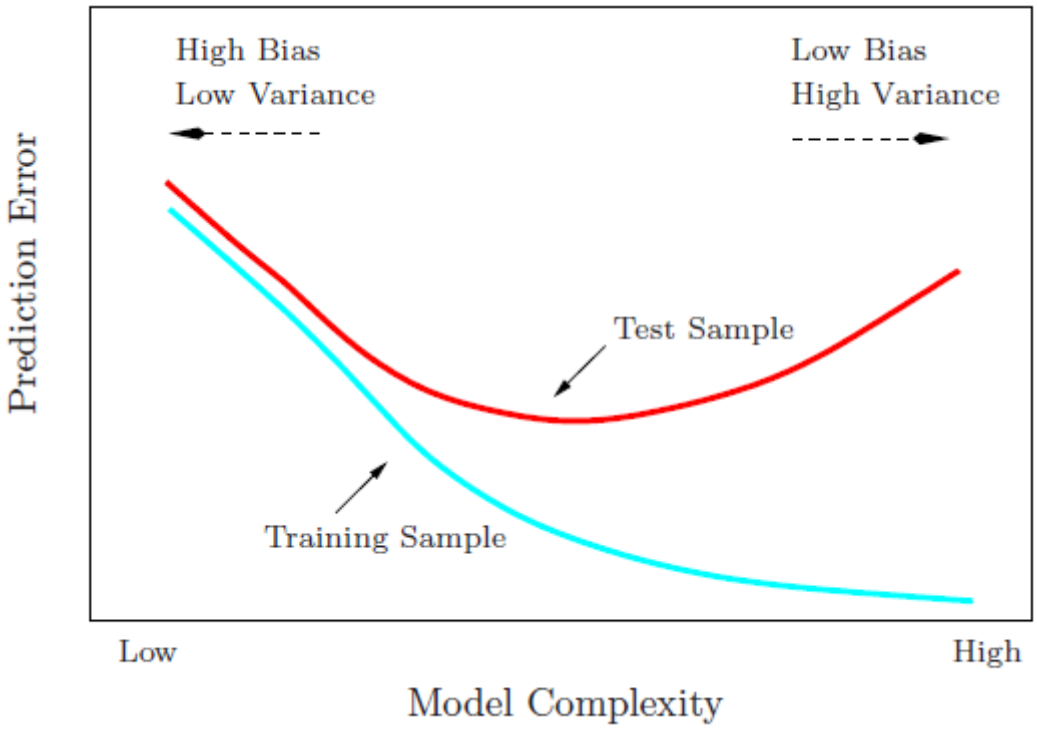
Com podem veure, la distribució del logaritme de les emissions s'acosta molt més a una normal. Tot i que sembla que tingui un apuntament massa pronunciat i que no passi els tests de normalitat com el de Shapiro per molt poc, podem donar per vàlida la distribució del logaritme de les emissions com l'escollida per modelitzar.

3.4 PROCÉS DE MODELATGE

El procés o els passos que seguirem a l'hora de fer els models serà el mateix independentment de la metodologia de creació del model que fem servir.

El primer que farem serà dividir les dades en un conjunt d'entrenament i un conjunt de test. Això ho fem per evitar el sobre ajustament i per veure quina és la capacitat predictiva del model real al enfrontar-lo davant unes dades que no ha vist mai. Aquest procés, que es pot anomenar amb angles *one leaf out* és una de les dues alternatives que tenim (l'altra seria el *cross-validation*).

En el cas del *gradient boosting,* l'error d'entrenament només farà que reduir-se com més iteracions fem ja que a cada iteració s'anirà ajustant més a les dades d'entrenament. És per això que haurem de mirar l'error del model a les dades de test per tal de triar el millor número d'iteracions. Gràficament, podem veure això de la següent manera:



Imatge 9: Mostra teòrica dels errors

Una vegada comença el sobre ajustament podem veure com l'error de predicció a les dades de test comença a augmentar. Per tan, ens quedarem el nombre d'iteracions que minimitzi l'error de predicció a les dades de test.

En el cas de la regressió simple, no hem de triar iteracions però sí les variables d'entrada del model. El que s'ha fet ha estat entrar-li totes les variables de les que disposàvem al model i aplicar-hi un algorisme de *stepwise* basat amb el criteri d'informació AIC[8]. Aquest mètode ens permetrà escollir el millor model possible donades totes les variables d'entrada.

Una vegada tinguem els models construïts, passarem a analitzar els resultats. Primer de tot, mirarem la qualitat de l'ajust del model mitjançant diferents tècniques com el R quadrat ajustat, la mitjana dels errors al quadrat i un mètode alternatiu de classificació i error. Per últim, interpretarem el model i les seves variables/coeficients/importància relativa i a més compararem els resultats entre les dues metodologies que aplicarem.

[8] Akaike Information Criterion, és una mesura de la qualitat relativa d'un model estadístic per a un conjunt de dades. Utilitza un trade-off entre la bondat de l'ajust del model i la seva complexitat basant-se en l'entropia de la informació: S'ofereix una estimació relativa de la informació perduda quan s'utilitza un model determinat per representar el procés que genera les dades.

4. PREDICCIÓ DE LES EMISSIONS DE CO2

En aquest apartat construirem els models que s'utilitzaran per fer prediccions de les emissions de CO2 modelitzant el logaritme de les emissions en funció de les variables independents plantejades en els anteriors capítols. Primer ho farem amb la regressió clàssica i després amb el XGboost.

4.1 REGRESSIÓ CLÀSSICA

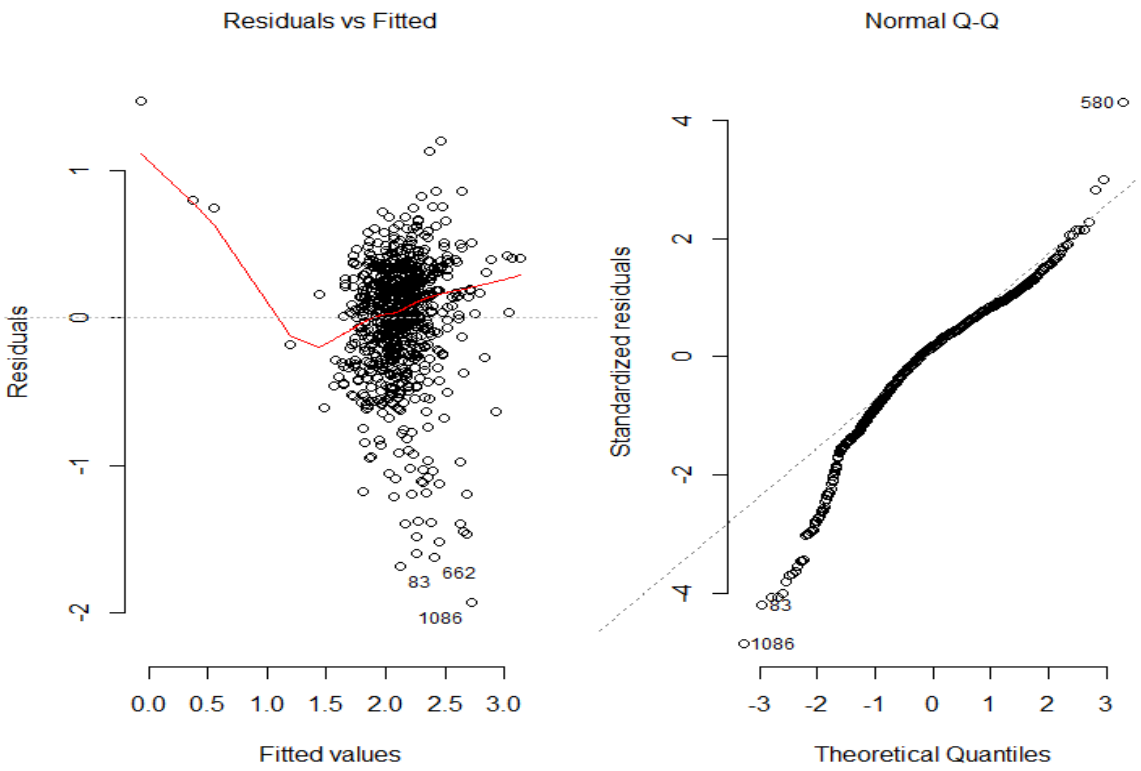
Com hem dit anteriorment, utilitzem el mètode de *stepwise* amb totes les variables de les que disposem i el model resultant és el següent:

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.93430 -0.18819  0.07291  0.25634  1.47106

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      1.925091   0.157896  12.192  < 2e-16 ***
ZONA_CLIMATICA_2C                0.255441   0.080380   3.178 0.001531 **
ZONA_CLIMATICA_2DE               0.276349   0.090337   3.059 0.002283 **
ANY_CONSTRUCCIO_2>2013          -0.360814   0.127359  -2.833 0.004708 **
PROCEDIMENT_UTILITZAT_2CTE-HE CEE -0.368295   0.137804  -2.673 0.007656 **
PROVINCIAGirona                  0.116922   0.033955   3.443 0.000600 ***
PROVINCIALleida                  0.076809   0.058948   1.303 0.192891
PROVINCIATarragona               0.098186   0.064345   1.526 0.127361
PERC_SUPERF_HABIT_REFRIG_20.5    0.108746   0.051128   2.127 0.033683 *
PERC_SUPERF_HABIT_REFRIG_21     -0.063818   0.057740  -1.105 0.269324
PERC_SUPERF_ACRIST_SUROEST_21   -0.074190   0.027477  -2.700 0.007056 **
DENSITAT_FONTS_INTERNAS_21       0.288110   0.080092   3.597 0.000338 ***
CERRAMIENTOSOPACOSTRANS          0.006823   0.001910   3.573 0.000371 ***
HUECOSYLUCERNARIOSTRANS          -0.004080   0.001079  -3.781 0.000166 ***
INSTCALEF_RENDIM                -0.046117   0.004739  -9.730  < 2e-16 ***
NOMB_PLANT_BAJORASANT_20         0.073085   0.031887   2.292 0.022126 *
NOMB_PLANT_SOBRERASANT_20       -0.094533   0.047218  -2.002 0.045565 *
NOMB_PLANT_SOBRERASANT_21       -0.047749   0.034409  -1.388 0.165554
DEMANDA_GLOBAL                   0.007846   0.000957   8.199 7.82e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.404 on 948 degrees of freedom
Multiple R-squared:  0.2859,    Adjusted R-squared:  0.2724
F-statistic: 21.09 on 18 and 948 DF,  p-value: < 2.2e-16
```

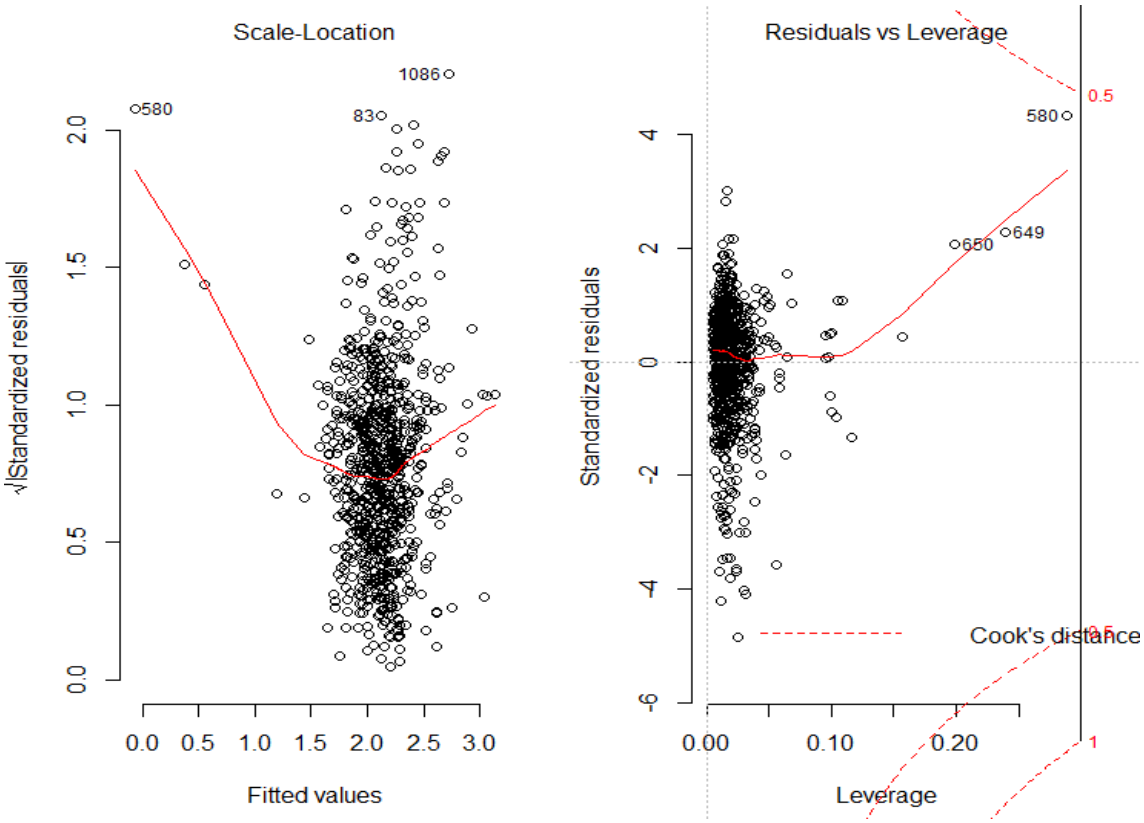
Ens queda un model amb 13 variables, 9 categòriques i 4 numèriques.

4.1.1 VALIDACIÓ DEL MODEL

El primer que hem de dir és que el model en conjunt apareix com a significatiu i les variables que el representen també (mirem p-valors). Una vegada dit això, la regressió clàssica s'acostuma a validar de forma gràfica mirant els errors. Els gràfics que podem mirar són els següents:



Imatge 10: Validació dels residus



Imatge 11: Validació dels residus 2

Com podem veure en els gràfics, tot i que tenim algun *outliers* que ens desvirtua una mica les rectes ajustades, la conclusió general és que no hi ha cap patró marcat en els residus així que podem donar per bo el model a nivell teòric.

Una altra de les hipòtesis de la regressió lineal clàssica que podem comprovar ràpidament és la de la absència de multicol·linealitat. Recordem que una presència de multicol·lieanitat afecta directament a la variància dels coeficients (sobre estimant-la) fent que passin a ser considerats com a no significatius. A més, podria donar-se el cas de coeficients canviats de signe o de compensació d'efectes entre dues variables correlacionades. La multicol·linealitat la podem comprovar mirant els vif[9] del model generat:

	GVIF	Df	GVIF^(1/(2*Df))
ZONA_CLIMATICA_2	3.594597	2	1.376932
ANY_CONSTRUCCIO_2	20.898340	1	4.571470
PROCEDIMENT_UTILITZAT_2	24.914262	1	4.991419
PROVINCIA	3.271799	3	1.218422
PERC_SUPERF_HABIT_REFRIG_2	4.155857	2	1.427793
PERC_SUPERF_ACRIST_SUROEST_2	1.069136	1	1.033990
DENSITAT_FONTS_INTERNAS_2	1.828490	1	1.352217
CERRAMIENTOSOPACOSTRANS	1.434953	1	1.197895
HUECOSYLUCERNARIOSTRANS	1.213873	1	1.101759
INSTCALEF_RENDIM	1.109299	1	1.053233
NOMB_PLANT_BAJORASANT_2	1.329783	1	1.153162
NOMB_PLANT_SOBRERASANT_2	1.410596	2	1.089810
DEMANDA_GLOBAL	1.500804	1	1.225073

Com podem veure, només tenim dues variables amb una alta correlació com són l'any de construcció i el procediment utilitzat. Té sentit que aquestes dues variables estiguin correlacionades i com a solució es planteja eliminar-ne una de les dues i veure que el resultat del model quasi no variarà.

4.1.2 BONDAT DE L'AJUST

La primera mesura que hem de mirar com a bondat de l'ajust en una regressió clàssica és el R quadrat ajustat[10]. En aquest cas, el valor del R quadrat ajustat és de 0,27. És un valor molt baix i es podria interpretar com que el model representa correctament el 27% de la variabilitat total de les emissions de CO2.

Una altra mesura que podem mirar és el MSE a les dades de test i és de 0,177. Aquesta dada ens servirà per poder comparar el model clàssic amb l'altra metodologia.

Per últim, podríem crear una mesura pròpia de capacitat de predicció i classificació. Creem un interval de confiança del valor real +/- el 10% del valor. Ara podem veure quin percentatge dels valors predits cauen dins d'aquest interval artificial i quins no. En el nostre cas, tenim una tassa

[9] Factor d'inflació de la variància, quantifica la intensitat de la multicol·linealitat en un anàlisis de regressió per mínims quadrats. A més, proporciona un índex que mesura fins a quin punt la variància d'un coeficient de regressió estimat s'incrementa a causa de la col·linealitat.

[10] És la mesura que defineix el percentatge explicat per la variància de la regressió en relació amb la variància de la variable explicada. A diferencia del R quadrat, penalitza la inclusió de variables.

d'èxit d'un 39%. És a dir, un 39% de les nostres prediccions es troben dins de l'interval format pel valor real i un +/- 10%. Aquesta capacitat predictiva és molt baixa i fa que el model no sigui explotable.

4.1.3 INTERPRETACIÓ

La interpretació dels coeficients del model variable a variable és la següent:

1) ZONA_CLIMATICA_2: Tenint com a base la zona climàtica B, podem dir que els edificis de la zona climàtica C emeten més CO2 i la zona climàtica D i E encara més.

2) ANT_CONSTRUCCIO_2: Els edificis construïts a partir del 2013 emeten menys CO2. Té molt sentit degut a que es construïen sota una altra llei i per tan l'eficiència energètica ja estava més regulada.

3) PROCEDIMENT_UTILITZAT_2: El procediment CTE-HE CEE porta a que s'emetin menys C02. El procediment en sí no hauria d'afectar així que podem estar davant d'un factor temporal. Si aquest procediment es fa servir més ara que abans llavors té sentit.
4) PROVINCIA: Tenint com a base la província de Barcelona, veiem que a les altres 3 s'emet més CO2.

5) PERC_SUPERF_HABIT_REFRIG_2: Tenint com a base els edificis que no tenen superfície refrigerada, veiem com els que tenen un percentatge entre el 0% i el 100% emeten més CO2 però els que tenen l'edifici 100% refrigerat emeten menys CO2. Això podria ser degut a que els edificis 100% refrigerats són més moderns.

6) PERC_SUPERF_ACRIST_SUROEST_2: Tenir l'edifici orientat al sud oest permet emetre menys CO2 al augmentar l'eficiència energètica.

7) DENSITAT_FONTS_INTERNAS_2: Tenir densitat de fonts internes suposa un increment de les emissions de CO2.

8) CERRAMIENTOSOPACOSTRANS[11]: A més transparència dels tancaments, més emissions de CO2.

9) HUECOSYLUCERNARIOSTRANS: A més transparència dels enllumenats, menys emissions de CO2.

10) INSTCALEF_RENDIM: A més rendiment de la instal·lació de la calefacció menys emissions de CO2.

11) NOMB_PLANT_BAJORASANT_2: Els edificis amb 0 plantes sota rasant emeten més CO2. Pot ser un tema d'any d'edificació també.

[11] La interpretació numèrica és complicada ja que estem modelitzant el logaritme de les emissions. Dit això, si fossin percentatges, ho podríem interpretar com una elasticitat.

12) NOMB_PLANT_SOBRERASANT_2: Tenint com a base els que tenen més de 1 planta sobre rasant, podem veure com tan els que no en tenen cap com els que només en tenen 1 emeten menys CO2.

13) DEMANDA_GLOBAL: A més demanda d'energia, més emissions de CO2.

4.2 REGRESSIÓ PER GRADIENT BOOSTING

Com hem dit anteriorment, utilitzem l'algorisme XGboost [11] per fer una aproximació per gradient boosting. Els paràmetres òptims després de provar diferents combinacions d'opcions són els següents:

Shrinkage	Iteracions	Profunditat
0.01	351	8

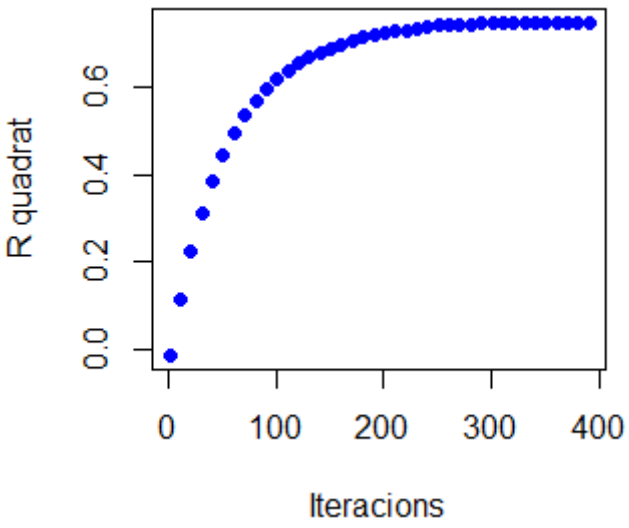
I les influències relatives de les variables són:

	var	rel.inf
INSTCALEF_RENDIM	INSTCALEF_RENDIM	49.53566151
DEMANDA_GLOBAL	DEMANDA_GLOBAL	32.67247260
CERRAMIENTOSOPACOSSUP	CERRAMIENTOSOPACOSSUP	3.61060878
CERRAMIENTOSOPACOSTRANS	CERRAMIENTOSOPACOSTRANS	2.54573151
DENSITAT_FONTS_INTERNAS_3	DENSITAT_FONTS_INTERNAS_3	2.20376273
HUECOSYLUCERNARIOSSUP	HUECOSYLUCERNARIOSSUP	1.98759365
ZONA_CLIMATICA_3	ZONA_CLIMATICA_3	1.90816285
HUECOSYLUCERNARIOSTRANS	HUECOSYLUCERNARIOSTRANS	1.43587876
PROVINCIA	PROVINCIA	0.77984751
PERC_SUPERF_HABIT_CALEFA_3	PERC_SUPERF_HABIT_CALEFA_3	0.64523827
US_EDIFICI_3	US_EDIFICI_3	0.50289902
ANY_CONSTRUCCIO_3	ANY_CONSTRUCCIO_3	0.44368570
NORMATIVA_CONSTRUCCIO_3	NORMATIVA_CONSTRUCCIO_3	0.43262587
PERC_SUPERF_HABIT_REFRIG_3	PERC_SUPERF_HABIT_REFRIG_3	0.28695204
PERC_SUPERF_ACRIST_OEST_3	PERC_SUPERF_ACRIST_OEST_3	0.17338741
NOMB_PLANT_SOBRERASANT_3	NOMB_PLANT_SOBRERASANT_3	0.15751126
PROCEDIMENT_UTILITZAT_3	PROCEDIMENT_UTILITZAT_3	0.12029588
PERC_SUPERF_ACRIST_SUREST_3	PERC_SUPERF_ACRIST_SUREST_3	0.11824581
PERC_SUPERF_ACRIST_SUROEST_3	PERC_SUPERF_ACRIST_SUROEST_3	0.09841735
ID_TIPUS_TRAMIT_3	ID_TIPUS_TRAMIT_3	0.08468027
NOMB_PLANT_BAJORASANT_3	NOMB_PLANT_BAJORASANT_3	0.07479700
PERC_SUPERF_ACRIST_EST_3	PERC_SUPERF_ACRIST_EST_3	0.07421844
PERC_SUPERF_ACRIST_NORDEST_3	PERC_SUPERF_ACRIST_NORDEST_3	0.05584537
PERC_SUPERF_ACRIST_NORDOEST_3	PERC_SUPERF_ACRIST_NORDOEST_3	0.05148041

4.2.1 BONDAT DE L'AJUST

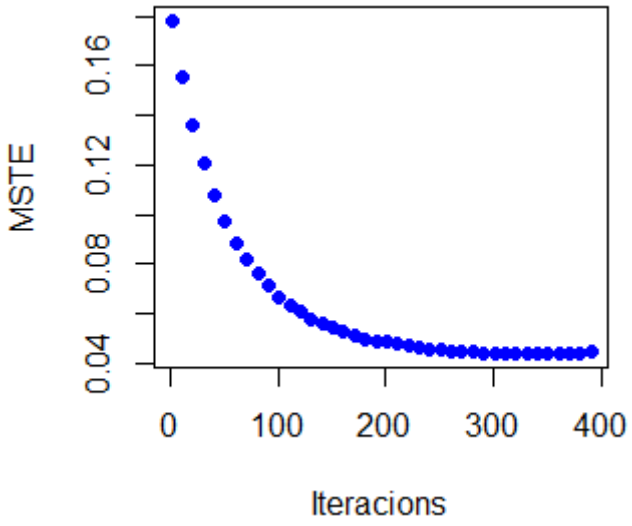
Podem mirar la bondat de l'ajust del *gradient boosting* recreant les mateixes mesures que hem triat per la regressió clàssica. Recordem que com en el cas de la regressió clàssica, totes les mesures de bondat d'ajust s'han construït a partir de les dades de test que no han participat en la construcció del model.

Pel que fa al R quadrat, podem veure gràficament com va augmentant en funció del nombre d'iteracions que apliquem:



En la iteració optima, el resultat és que el R quadrat és del 75%. És a dir, el model explica el 75% de la variància de les emissions de CO2.

També podem fer el mateix gràfic pel MSTE, i queda de la següent manera:



Com podem veure, com més iteracions fem, més petit és el MSTE (en algun moment tornarà a augmentar a causa del sobre ajustament). En la iteració òptima, el MSTE és de 0,0439.

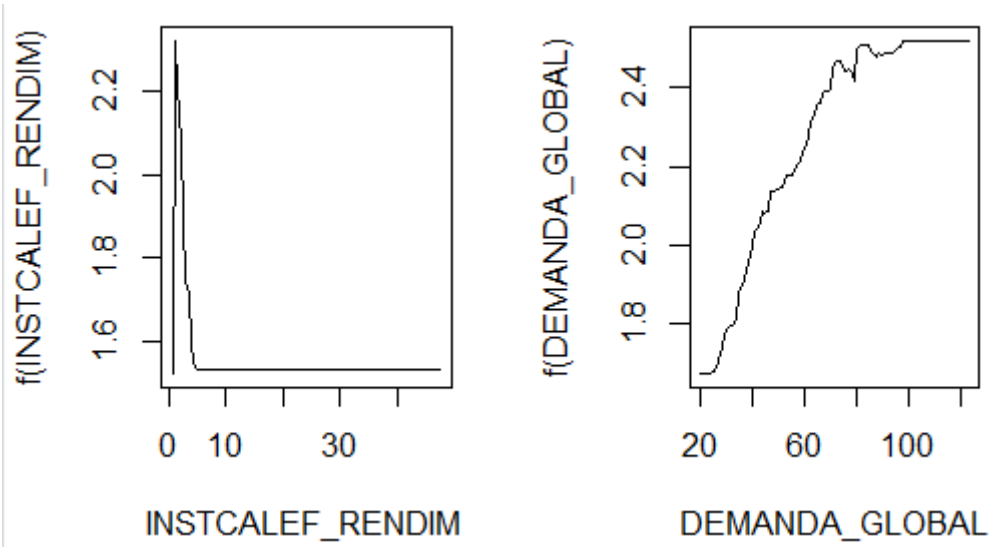
Per últim, podem fer el mateix exercici de l'interval que en el cas de la regressió clàssica. És a dir, agafem el valor real i li construïm un interval de +/- el 10% del valor. Quants encerts tindríem llavors?

Amb un interval del +/- 10% del valor real tindríem una tassa d'èxit d'un 77%. És a dir, el valor predit estaria dins de l'interval del valor real +/- un 10% en un 77% de les vegades.

4.2.2 INTERPRETACIÓ

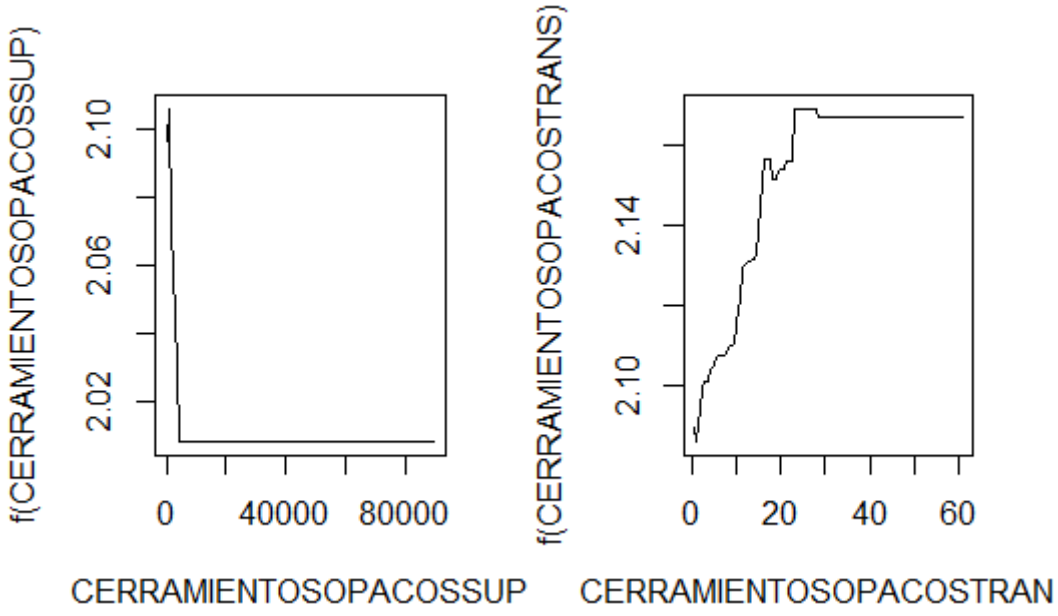
Veient la influència relativa de les variables podem dir que les dues variables que apareixen a quasi totes les iteracions són el rendiment de la instal·lació de calefacció i la demanda d'energia global.

Per la interpretació individual de cada variable, hem de mirar els gràfics de contribució marginal o parcial. Alguns dels exemples de les variables més importants són els següents:



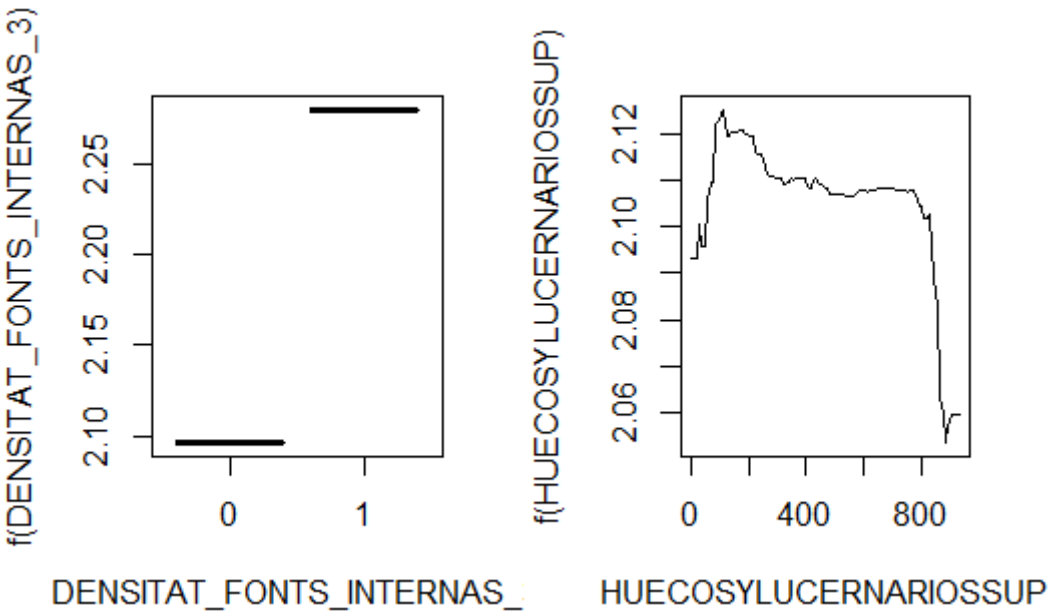
Imatge 12: Contribucions marginals 1

Com podem veure en el gràfic de la dreta, a més demanda global d'energia més emissions de CO2.



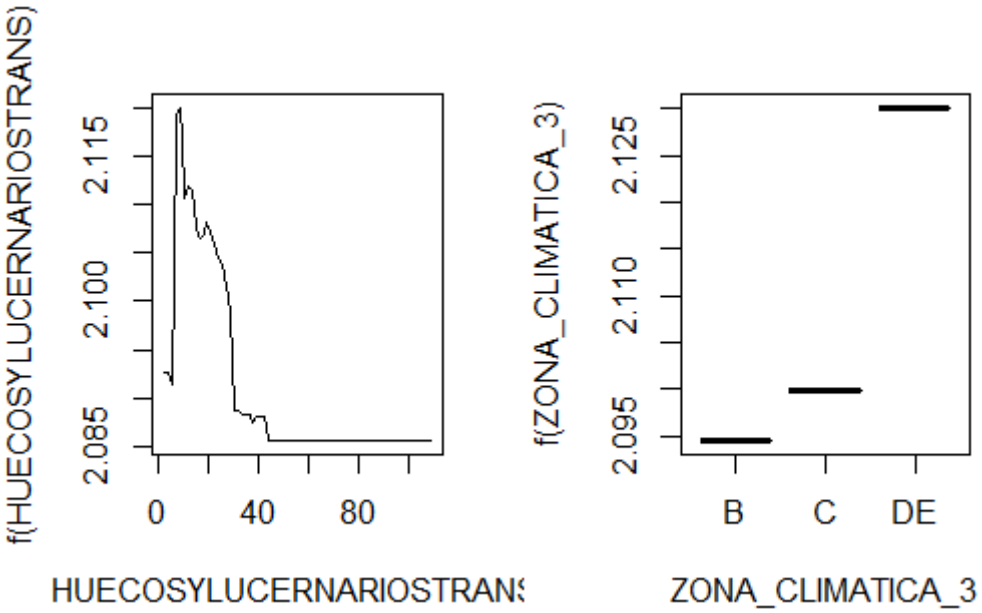
Imatge 13: Contribucions marginals 2

Com podem veure al gràfic de la dreta, la transparència dels tancaments té una relació positiva amb les emissions de C02.



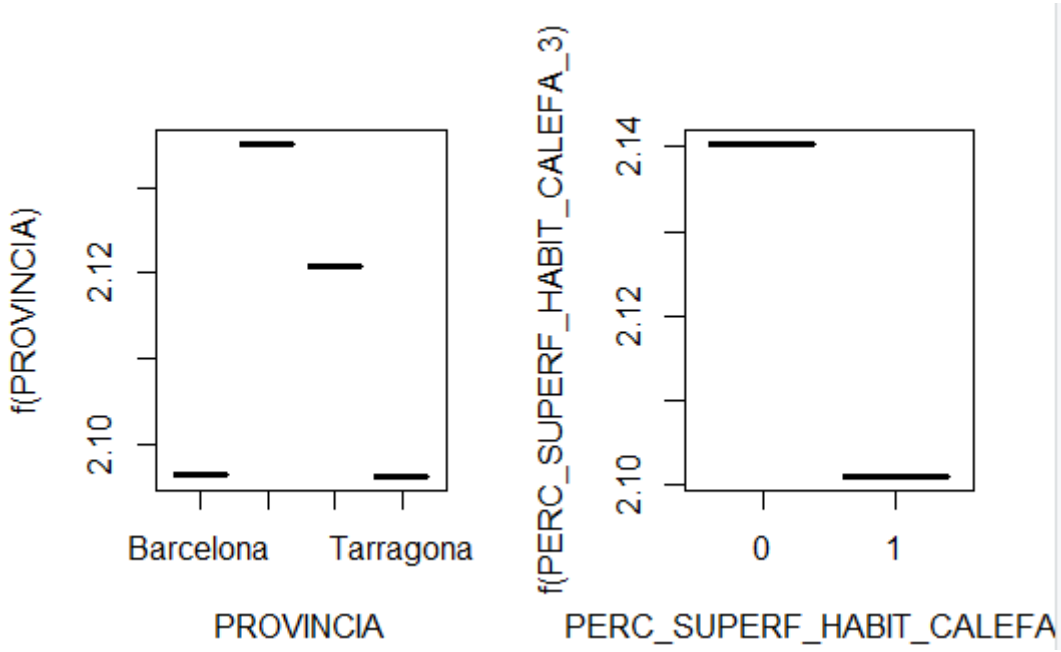
Imatge 14: Contribució marginal 3

Si l'edifici té densitat a les fonts internes llavors les emissions són molt més elevades.



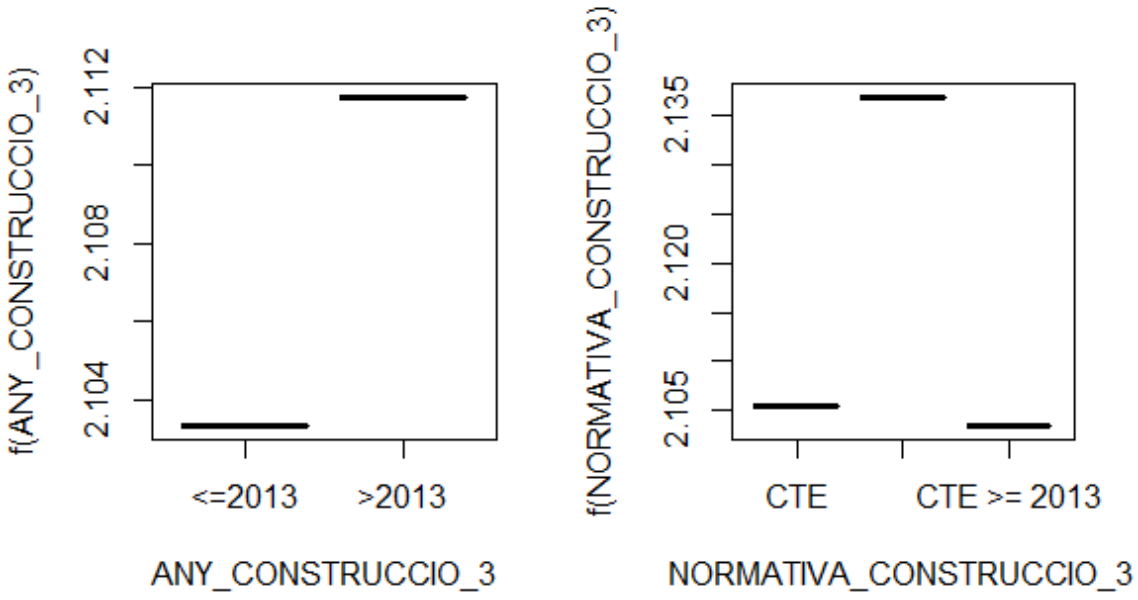
Imatge 15: Contribució marginal 4

Com en el cas de la regressió clàssica, la zona climàtica DE és la de majors emissions de CO2.



Imatge 16: Contribució marginal 5

Com veiem en el gràfic de la dreta, el fet de tenir la superfície de casa 1005 calefactada suposa que estàs en un edifici modern i això fa que les emissions siguin menors.



Imatge 17: Contribució marginal 6

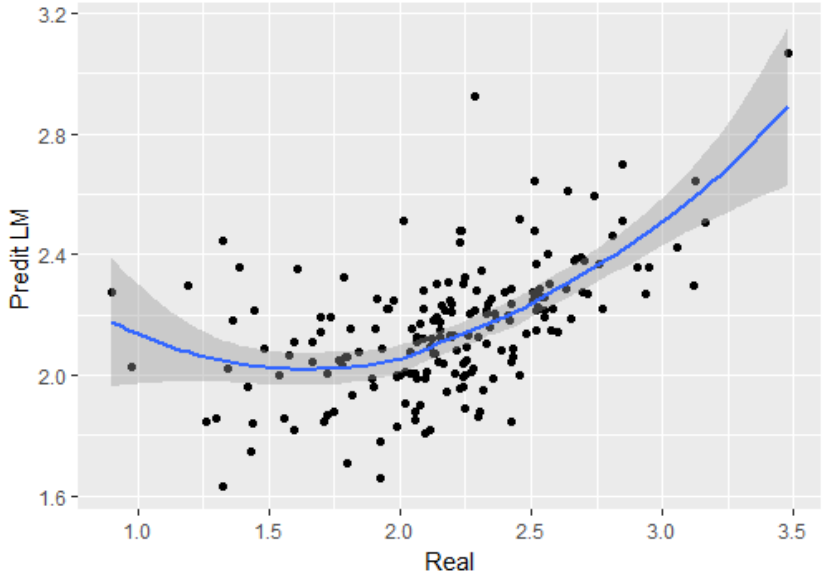
Com veiem en el gràfic de l'esquerra, els edificis construïts més enllà de 2013 emeten més CO2. Aquest podria ser un factor comportamental dels habitants dels edificis moderns i no tan per culpa de les característiques de l'edifici. Igualment, veiem que la diferència no és molta.

Si es volen interpretar la resta de variables, nomes cal consultar la resta de gràfics marginals que estan disposats a l'annex.

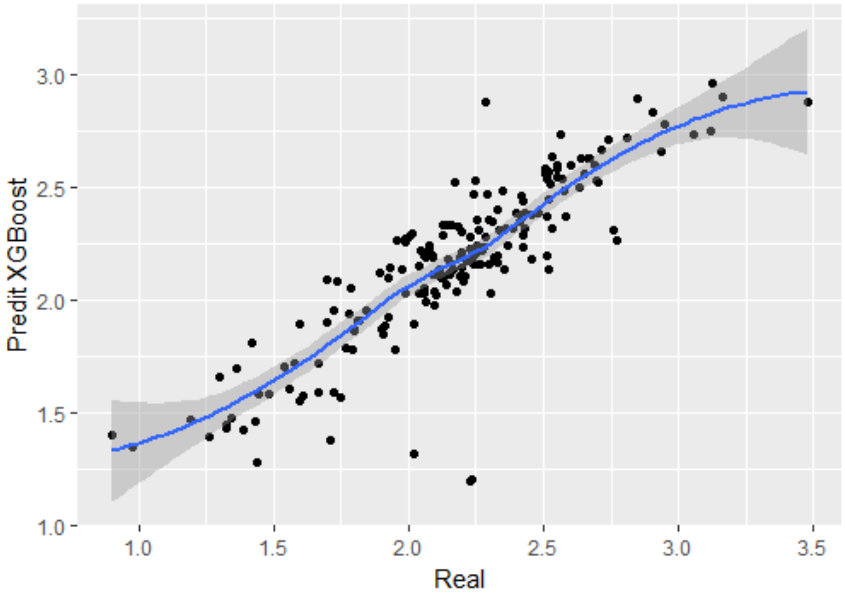
4.3 COMPARACIÓ LM vs GRADIENT BOOSTING

Un cop hem aplicat els dos models, és d'interès fer un resum de les diferències que hi hem trobat.

A nivell visual, el primer que podem fer és ensenyar quina relació hi trobem entre els valors predits i els originals amb les dues metodologies que hem treballat. Gràficament ho representem amb un *scatterplot* i una recta suavitzada ajustada.



Imatge 18: Real vs Predit LM



Imatge 19: Real vs Predit XGboost

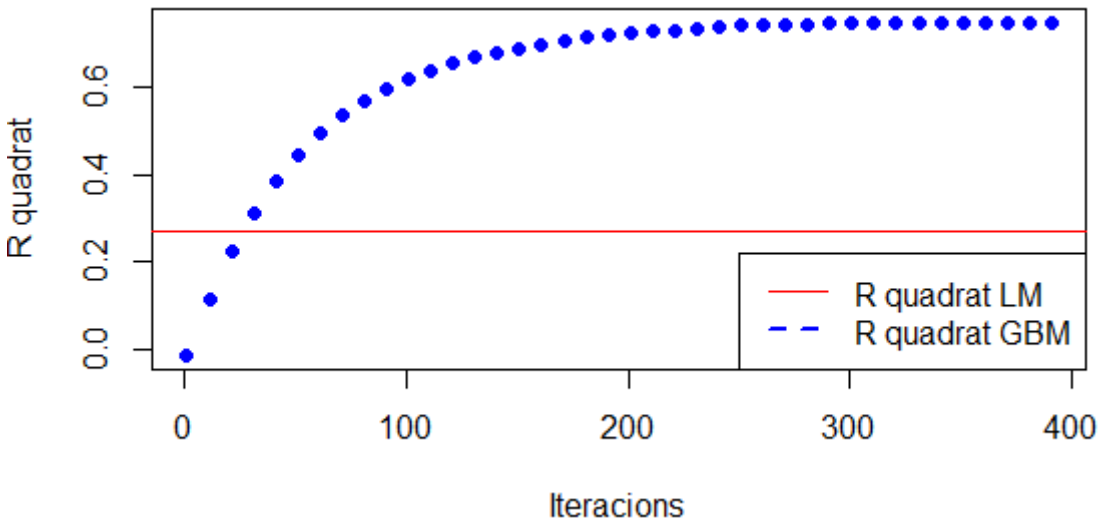
Com podem veure, el model amb XGboost ajusta molt millor els valors i només té un parell *d'outliers* destacats. En canvi, la regressió clàssica veiem que té molts problemes per ajustar tots els valors.

A mes, cal destacar:

1) La regressió lineal clàssica fa només un model complex sense multicol·linealitat. El XGboost en farà 351 de simples i no hi ha problemes amb la presència de col·linealitats entre les covariables.

2) La regressió clàssica ha utilitzat un total de 13 variables mentre que la construïda amb *Machine learning* un total de 24 variables en alguna de les seves iteracions.

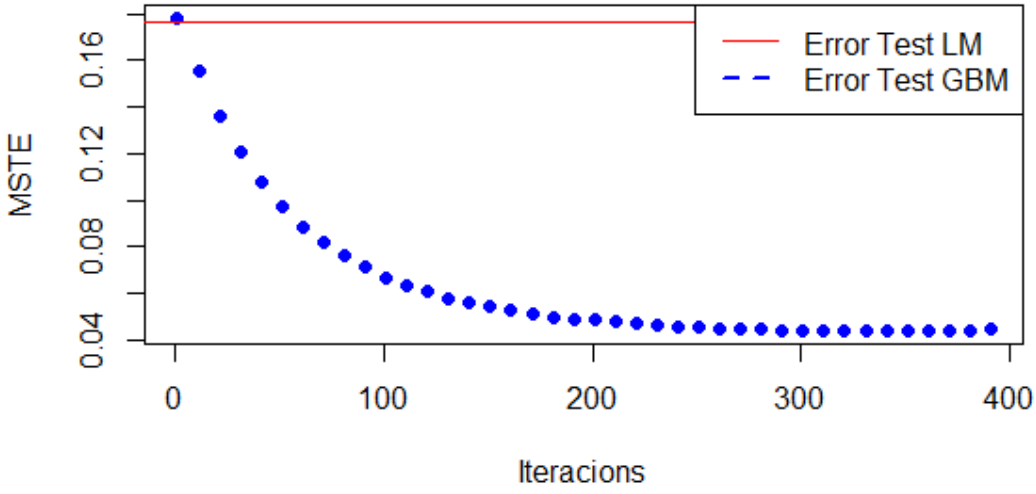
3) Pel que fa al R quadrat, en la regressió clàssica era de 0,27 i en la millor iteració de XGboost és de 0,75. Gràficament podem veure com es va aproximant el R quadrat del XGboost al de la regressió clàssica fins superar-lo.



Imatge 20: Comparació R quadrat LM vs GBM

Com podem veure, el R quadrat del model que es basa en aprenentatge automàtic és molt millor. És més, a partir de la iteració 50 ja teníem un model millor que el de la regressió simple.

4) Pel que fa al MSTE, en la regressió clàssica era de 0,177 i en la millor iteració del XGboost és de 0,0449. Gràficament podem veure com es va reduint el MSTE del XGBoost fins a superar al de la regressió clàssica.



Imatge 21: Comparació de MSTE LM vs GBM

Com podem veure, és molt més petit el del algorisme que utilitza aprenentatge automàtic. És més, podem veure com a partir de la iteració 10 ja teníem un model millor que l'ajustat per MQO.

5) Per últim podem comprovar la capacitat d'encert del valor predit en funció de si encerta en un interval o no. Hem plantejat com a exemple un interval +/- sobre el valor real d'un 5%, un 10% i un 20%.

Els resultats d'encert són els següents:

Model	*5% d'interval*	*10% d'interval*	*20% d'interval*
LM	15%	39%	55%
GBM	51%	77%	94%

Amb aquests resultats, podem afirmar que **el model amb *Machine learning* encerta quasi el doble de vegades que el model clàssic**.

5. CONCLUSIONS

Una vegada finalitzat el treball, podem passar a comentar les conclusions que hem extret de tota la feina realitzada.

Respecte a l'objectiu inicial que teníem, hem de dir que l'hem complert a mitges. Som capaços de fer prediccions de les emissions de CO2 acceptables i creïbles però això no ens permet solucionar el problema del ICAEN. Hem comprovat que les bases de dades validada i sense validar no tenen el mateix comportament i que la base amb la que hem fet els models no representa una mostra aleatòria de la base sencera de certificats energètics. Això fa que **no puguem utilitzar els models creats per validar la base pendent de validació**. Per tan, s'hauria de pensar un altre mètode per validar de forma massiva tots els certificats energètics (més de 500.000) que té acumulats el ICAEN ara mateix pendents de validació. En el següent apartat, proposem la creació d'una mostra representativa per tal de repetir el procés seguit però que sigui vàlid per a la posada en producció.

Partint d'aquesta base, podem analitzar la resta de resultats aconseguits. El primer que hem de dir és que hem trobat relacions existents entre les emissions de CO2 i les característiques dels edificis de Catalunya. Aquestes relacions, però, han estat molt diverses en tots els sentits en funció de la metodologia aplicada. Tot i tenir una base de dades petita, la recursivitat que ens permet l'algorisme Xgboost fa que els resultats amb aquesta metodologia siguin molt més bons que amb la regressió clàssica. Aquest fet ens demostra que el *Machine learning* no és només útil en bases de dades grans sinó que també pot funcionar amb bases de dades de mida mitjana i petita.

Si passem a comentar els resultats, **l'ajust a les emissions de CO2 de la regressió clàssica és molt pobre** obtenint un R quadrat no superior a 0,3. Per una altra banda**, l'ajust a les emissions de CO2 del gradient boosting utilitzant el XGboost és prou bo** obtenint R quadrats de quasi 0,8 i millorant en totes les proves alternatives a la regressió clàssica.

Tot i que els resultats obtinguts utilitzant la nova metodologia són bons, caldria incloure més variables a l'estudi per tal de que els resultats fossin millors i poder construir així una eina totalment fiable (ara ho és amb un marge d'error de quasi el 20%).

Per últim, cal destacar alguna de les variables més importants en els models presentats com el rendiment de la instal·lació de calefacció o les transparències dels tancaments. Aquest fet ens indica que cal tenir molt en compte com i amb què construïm els edificis per tal d'intentar fer-los el més eficients possibles. És important destacar que tot i treballar amb una base de dades relativament petita, amb una gran quantitat de *missings* i sense l'ajuda d'un expert del domini, hem estat capaços de treure'n uns resultats que poden ser d'interès per a futurs treballs.

Una vegada finalitzat el treball considero, a mode de reflexió personal, que cal conscienciar millor a l'usuari final (llogater o comprador) de què indica un certificat energètic. Avui en dia no li donem importància i entenem que forma part de la burocràcia que hem de passar si volem portar a terme un tràmit d'aquestes característiques.

El certificat energètic, però, inclou molta informació d'interès que hauria de repercutir directament sobre la decisió de llogar/comprar un pis i ara mateix no ho està fent. A més, la promoció d'activitats sostenibles ha de ser cada vegada més important en el món on vivim i a part de regular al constructor a partir de les lleis perquè les seves activitats o els seus edificis

siguin sostenibles també cal conscienciar o educar al consumidor perquè aquest nou concepte entri a formar part de la seva matriu de decisió.

És per això que considero que els organismes públics haurien de fer campanyes de conscienciació per explicar millor què volen dir els certificats energètics i per assegurar-se de que entenem la diferència entre una A o una C i què comporta: més/menys despesa en energia (€) i més/menys emissions de gasos a l'atmosfera del planeta.

6. POSSIBLES FUTURS PASSOS

Una vegada finalitzat tot l'estudi, podem determinar uns possibles futurs passos per tal de millorar en següents treballs la feina realitzada.

El primer que hem de proposar és una reestructuració metodològica de les bases de dades. Ara que hem vist que som capaços de modelitzar correctament les emissions de CO2, ho hauríem de fer amb una base de dades que fos representativa del total de certificats pendents de validació. Proposem, doncs, agafar una mostra aleatòria de la base per validar i validar manualment aquells certificats. Una vegada tinguem validats els certificats, tindrem una base validada i representativa de la població total de certificats. Utilitzarem aquesta nova base per modelitzar les emissions de CO2 (es recomana un model per segments per èpoques per evitar complexitat al model). Una vegada tinguem el model, ara sí que ja podrem executar-lo sobre la base total de certificats pendents de validació i així determinar quins passen el filtre i quins no.

Una altra línia d'actuació seria la del enriquiment de la base de dades per tal de millorar els models. Quan parlem d'enriquiment, ens referim a augmentar el número de variables independents relacionades amb les emissions de CO2. En aquest punt hi podria entrar el coneixement del domini d'experts.

Fins ara, teníem les variables de les característiques del edifici però ens queda una part molt important de l'explicació de les emissions que és la part comportamental de les persones que hi viuen. És a dir, per posar un exemple, no emeten el mateix número d'emissions una família rica de 5 persones que un matrimoni de jubilats de renda baixa. És per això que proposem incorporar variables que informin sobre les famílies que viuen en els edificis a explorar: renta, número de persones, edats, situacions professionals, lloguer/hipoteca, etc. La idea es intentar relacionar la capacitat econòmica/situació demogràfica amb les emissions.

Tenim dues opcions per treure tota aquesta nova informació:

- Pagant a una consultora que tingui informació agregada per secció censal o que tingui informació segons identificador cadastral.

- Buscar-nos nosaltres mateixos la informació pública de la que disposem: aquesta opció, tot i ser molt més complicada té la avantatge de que seria gratuïta i ho podríem fer de la següent manera:

    1) Traiem variables econòmiques/familiars de la EPF (Enquesta de Pressupostos Familiars) des de l'INE (Institut Nacional d'Estadística).
    2) Passem la informació a nivell de secció censal amb l'ajuda del padró i aconseguim variables agregades a nivell de secció censal que ens permetran segmentar més la població.
    3) Amb els portals immobiliaris online extraiem preus de l'habitatge.
    4) Geolocalitzem aquests edificis i fem un model per comparables amb la nostra referència cadastral. Aconseguirem un preu aproximat de l'habitatge per la referència cadastral en concret.

Per últim, plantegem la opció de modificar les metodologies de modelatge aplicades intentant altres algorismes que podrien donar millors resultats com per exemple altres tècniques de *Machine learning* com les xarxes neuronals.

7. BIBLIOGRAFIA

[1] Boe.es. (2018). Directiva 2002/91/CE. [online] Available at: https://www.boe.es/doue/2003/001/L00065-00071.pdf [Accessed 27 Sep. 2018].

[2] Boe.es. (2018). Reial decret 47/2007. [online] Available at: https://www.boe.es/buscar/doc.php?id=BOE-A-2007-2007 [Accessed 27 Sep. 2018].

[3]Boe.es. (2018). Directiva 2010/31/UE. [online] Available at: https://www.boe.es/doue/2010/153/L00013-00035.pdf [Accessed 27 Sep. 2018].

[4]Boe.es. (2018). Reial decret 235/2013. [online] Available at: https://www.boe.es/buscar/act.php?id=BOE-A-2013-3904 [Accessed 27 Sep. 2018].

[5]Institut Català d'Energia. (2018). Inici. [online] Available at: http://icaen.gencat.cat/ca/inici [Accessed 26 Sep. 2018].

[6]Eur-lex.europa.eu. (2018). Directiva 2012/27/EU. [online] Available at: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2012:315:0001:0056:en:PDF [Accessed 27 Sep. 2018].

[7]Statstutor.ac.uk. (2018). Test de Spearman. [online] Available at: http://www.statstutor.ac.uk/resources/uploaded/spearmans.pdf [Accessed 27 Sep. 2018].

[8]Math.mit.edu. (2018). Shapiro test. [online] Available at: https://math.mit.edu/~rmd/465/shapiro.pdf [Accessed 27 Sep. 2018].

[9]Stata.com. (2018). Kolmogorov - Smirnov test. [online] Available at: https://www.stata.com/manuals13/rksmirnov.pdf [Accessed 27 Sep. 2018].

[10]Statweb.stanford.edu. (2018). Gradient Boosting. [online] Available at: https://statweb.stanford.edu/~jhf/ftp/trebst.pdf [Accessed 27 Sep. 2018].

[11] Arxiv.org. (2018). Paper XGboost. [online] Available at: https://arxiv.org/pdf/1603.02754.pdf [Accessed 26 Sep. 2018].

Xgboost.readthedocs.io. (2018). XGBoost R Tutorial — xgboost 0.80 documentation. [online] Available at: https://xgboost.readthedocs.io/en/latest/R-package/xgboostPresentation.html [Accessed 26 Sep. 2018].

ANNEX

ANNEX 1: Descripció de variables

Variable	Descripció
ID	Identificador único del edificio
DATA_CREACIO	Fecha de creación del informe
DATA_MODIFICACIO	Fecha de modificación del informe
ZONA_CLIMATICA	Zona climática en la que se situa el edificio
ANY_CONSTRUCCIO	Año de construcción
NORMATIVA_CONSTRUCCIO	Normativa vigente en el momento de la construcción o rehabilitación del edificio o local
NUM_CADASTRE	Referencia o referencias catastrales de la finca o fincas, separadas por comas
US_EDIFICI	Uso/ finalidad del edificio
PROCEDIMENT_UTILITZAT	Procedimiento aplicado para la calificación energética y verificación del cumplimiento del DB-HE
ID_TIPUS_TRAMIT	Tipo de trámite aplicable
DATA_CREACIO_1	Fecha de creación del registro en la tabla
EMISSIONS_GLOBAL	Emisiones de CO2e asociadas al conjunto de servicios del edificio, descontando las aportaciones de energías renovables. Incluye, en uso residencial privado, los servicios de calefacción, refrigeración y ACS; y, en uso terciario, los servicios de calefacción, refrigeración, ACS e iluminación, considerando el impacto derivado de ventiladores, bombas y torres de refrigeración (en kgCO2e/m²·año)
EMISSIONS_CALEFACCIO	Emisiones de CO2e asociadas al servicio de calefacción (en kgCO2e/m²·año)
EMISSIONS_REFRIGERACIO	Emisiones de CO2e asociadas al servicio de refrigeración (en kgCO2e/m²·año)
EMISSIONS_ACS	Emisiones de CO2e asociadas al servicio de ACS (en kgCO2e/m²·año)
EMISSIONS_ILLUMINACIO	Emisiones de CO2e asociadas al servicio de iluminación (en kgCO2e/m²·año)
CODIG_POSTAL	Código postal correspondiente al edificio
MUNICIPI	Municipio correspondiente al edificio
PROVINCIA	Provincia correspondiente al edificio
NOMB_PLANT_BAJORASANT	Número de plantas bajo rasante
NOMB_PLANT_SOBRERASANT	Número de plantas sobre rasante
SUPERFICI_HAB	Superficie de los espacios habitables, tal como se define en el DB-HE (en m2)

COMPACITAT	Cociente del volumen de espacio habitable entre la superficie total de cerramientos que forman parte de la envolvente térmica (en m³/m²)
PERC_SUPERF_HABIT_CALEFA	Porcentaje de la superficie habitable acondicionada con sistema de calefacción en relación a la superficie habitable total (en %)
PERC_SUPERF_HABIT_REFRIG	Porcentaje de la superficie habitable acondicionada con sistema de refrigeración en relación a la superficie habitable total (en %)
PERC_SUPERF_ACRIST_NORD	Porcentaje del área de huecos en relación al área de muro de fachada más área de huecos, computada para las orientaciones de fachada definidas en el DB-HE1: N (en %)
PERC_SUPERF_ACRIST_NORDEST	Porcentaje del área de huecos en relación al área de muro de fachada más área de huecos, computada para las orientaciones de fachada definidas en el DB-HE1: NE (en %)
PERC_SUPERF_ACRIST_EST	Porcentaje del área de huecos en relación al área de muro de fachada más área de huecos, computada para las orientaciones de fachada definidas en el DB-HE1: E (en %)
PERC_SUPERF_ACRIST_SUR	Porcentaje del área de huecos en relación al área de muro de fachada más área de huecos, computada para las orientaciones de fachada definidas en el DB-HE1: S (en %)
PERC_SUPERF_ACRIST_SUREST	Porcentaje del área de huecos en relación al área de muro de fachada más área de huecos, computada para las orientaciones de fachada definidas en el DB-HE1: SE (en %)
PERC_SUPERF_ACRIST_SUROEST	Porcentaje del área de huecos en relación al área de muro de fachada más área de huecos, computada para las orientaciones de fachada definidas en el DB-HE1: SO (en %)
PERC_SUPERF_ACRIST_OEST	Porcentaje del área de huecos en relación al área de muro de fachada más área de huecos, computada para las orientaciones de fachada definidas en el DB-HE1: O (en %)
PERC_SUPERF_ACRIST_NORDOEST	Porcentaje del área de huecos en relación al área de muro de fachada más área de huecos, computada para las orientaciones de fachada definidas en el DB-HE1: NO (en %)
DENSITAT_FONTS_INTERNAS	Promedio de la densidad de fuentes internas del edificio (de uso terciario), tal como se define en el DB-HE1 (en W/m²h)
VENTILACIO_USO_RESIDENCIAL	Tasa de ventilación del edificio o parte del mismo, para uso residencial, excluidas infiltraciones (en ren/h)
VENTILACIO_INFILTRACIONS	Tasa de ventilación total del edificio, incluidas infiltraciones (en ren/h)
DEMANDA_GLOBAL	Demanda energética de los servicios de calefacción, refrigeración y ACS del edificio objeto, obtenida según DB-HE (en kWh/m²año)
DEMANDA_CALEFACCIO	Demanda energética de calefacción del edificio objeto, obtenida según DB-HE (en kWh/m²año)
DEMANDA_REFRIGERACIO	Demanda energética de refrigeración del edificio objeto, obtenida según DB-HE (en kWh/m²año)
DEMANDA_ACS	Demanda energética de ACS del edificio objeto, obtenida según DB-HE (en kWh/m²año)

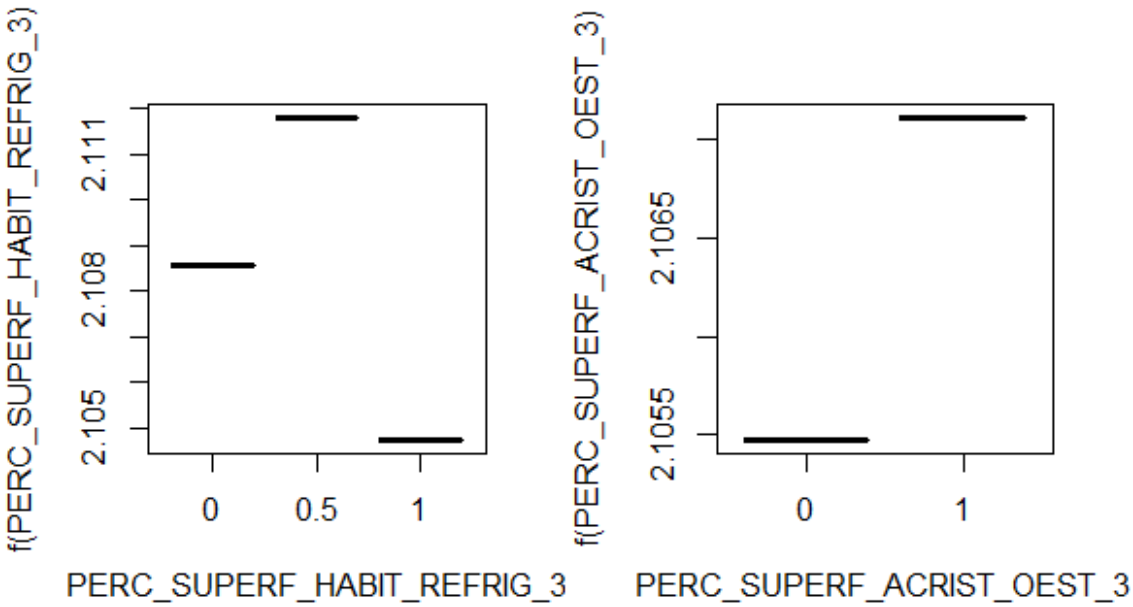
DEMANDA_CONJUNTA	Demanda energética conjunta del edificio objeto, obtenida según DB-HE (en kWh/m²año)
POTENCIA_TOTAL_INSTAL	Potencia total instalada de iluminación en el conjunto de espacios (en W/m²)
INSTAL_TERM_CONSUM	Consumo energético del equipo (en kWh/año)
CONSUMGASNATURAL_GLOBAL	Consumo de energía final asociado al gas natural, para todos los servicios (en kWh/m²·año)
CONSUMGASNATURAL_CALEF	Consumo de energía final asociado al gas natural, para el servicio de calefacción (en kWh/m²·año)
CONSUMGASNATURAL_REFRIG	Consumo de energía final asociado al gas natural, para el servicio de refrigeración (en kWh/m²·año)
CONSUMGASNATURAL_ACS	Consumo de energía final asociado al gas natural, para el servicio de agua caliente sanitaria (en kWh/m²·año)
CONSUMGASNATURAL_ILU	Consumo de energía final asociado al gas natural, para el servicio de iluminación (en kWh/m²·año)
CONSUMELECT_GLOBAL	Consumo de energía final asociado a la electricidad, para todos los servicios (en kWh/m²·año)
CONSUMELECT_CALEF	Consumo de energía final asociado a la electricidad, para el servicio de calefacción (en kWh/m²·año)
CONSUMELECT_REFRIG	Consumo de energía final asociado a la electricidad, para el servicio de refrigeración (en kWh/m²·año)
CONSUMELECT_ACS	Consumo de energía final asociado a la electricidad, para el servicio de agua caliente sanitaria (en kWh/m²·año)
CONSUMELECT_ILU	Consumo de energía final asociado a la electricidad, para el servicio de iluminación (en kWh/m²·año)
CONSUMCARBON_GLOBAL	Consumo de energía final asociado al carbón, para todos los servicios (en kWh/m²·año)
CONSUMCARBON_CALEF	Consumo de energía final asociado al carbón, para el servicio de calefacción (en kWh/m²·año)
CONSUMCARBON_REFRIG	Consumo de energía final asociado al carbón, para el servicio de refrigeración (en kWh/m²·año)
CONSUMCARBON_ACS	Consumo de energía final asociado al carbón, para el servicio de agua caliente sanitaria (en kWh/m²·año)
CONSUMCARBON_ILU	Consumo de energía final asociado al carbón, para el servicio de iluminación (en kWh/m²·año)
CONSUMGASOLEO_GLOBAL	Consumo de energía final asociado al gasoleo, para todos los servicios (en kWh/m²·año)
CONSUMGASOLEO_CALEF	Consumo de energía final asociado al gasoleo, para el servicio de calefacción (en kWh/m²·año)
CONSUMGASOLEO_REFRIG	Consumo de energía final asociado al gasoleo, para el servicio de refrigeración (en kWh/m²·año)
CONSUMGASOLEO_ACS	Consumo de energía final asociado al gasoleo, para el servicio de agua caliente sanitaria (en kWh/m²·año)
CONSUMGASOLEO_ILU	Consumo de energía final asociado al gasoleo, para el servicio de iluminación (en kWh/m²·año)
CONSUMRENOVABLE_GLOBAL	Consumo de energía final asociado a las renovables, para todos los servicios (en kWh/m²·año)
CONSUMRENOVABLE_CALEF	Consumo de energía final asociado a las renovables, para el servicio de calefacción (en kWh/m²·año)
CONSUMRENOVABLE_REFRIG	Consumo de energía final asociado a las renovables, para el servicio de refrigeración (en kWh/m²·año)

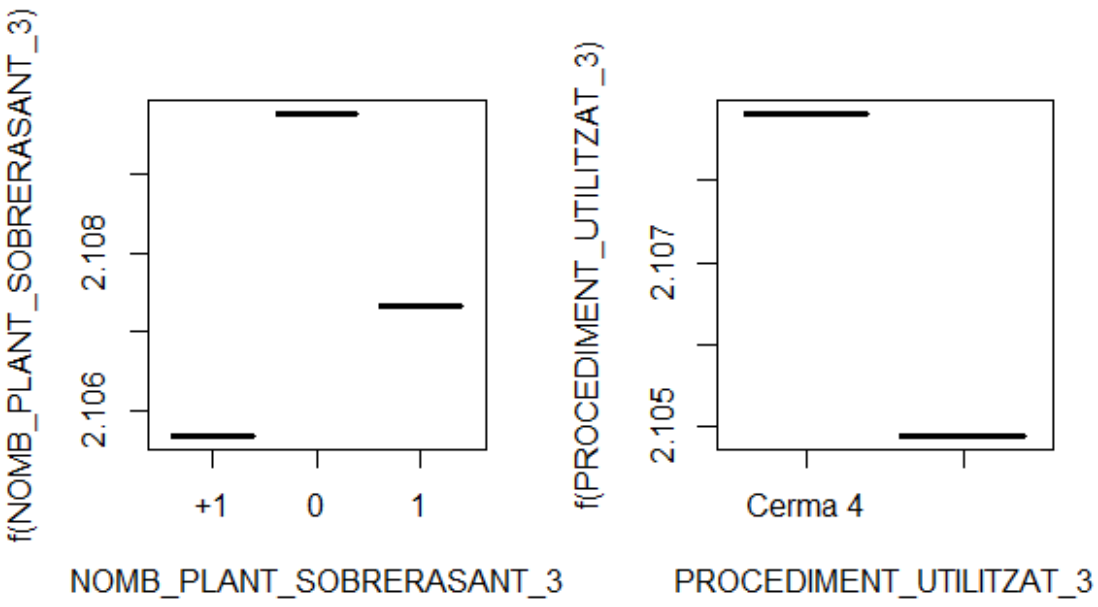
CONSUMRENOVABLE_ACS	Consumo de energía final asociado a las renovables, para el servicio de agua caliente sanitaria (en kWh/m²·año)
CONSUMRENOVABLE_ILU	Consumo de energía final asociado a las renovables, para el servicio de iluminación (en kWh/m²·año)
CONSUMGLP_GLOBAL	Consumo de energía final asociado al GLP, para todos los servicios (en kWh/m²·año)
CONSUMGLP_CALEF	Consumo de energía final asociado al GLP, para el servicio de calefacción (en kWh/m²·año)
CONSUMGLP_REFRIG	Consumo de energía final asociado al GLP, para el servicio de refrigeración (en kWh/m²·año)
CONSUMGLP_ACS	Consumo de energía final asociado al GLP, para el servicio de agua caliente sanitaria (en kWh/m²·año)
CONSUMGLP_ILU	Consumo de energía final asociado al GLP, para el servicio de iluminación (en kWh/m²·año)
CONSUMBIOMASSA_GLOBAL	Consumo de energía final asociado a la biomasa, para todos los servicios (en kWh/m²·año)
CONSUMBIOMASSA_CALEF	Consumo de energía final asociado a la biomasa, para el servicio de calefacción (en kWh/m²·año)
CONSUMBIOMASSA_REFRIG	Consumo de energía final asociado a la biomasa, para el servicio de refrigeración (en kWh/m²·año)
CONSUMBIOMASSA_ACS	Consumo de energía final asociado a la biomasa, para el servicio de agua caliente sanitaria (en kWh/m²·año)
CONSUMBIOMASSA_ILU	Consumo de energía final asociado a la biomasa, para el servicio de iluminación (en kWh/m²·año)
CONSUMBIOCARB_GLOBAL	Consumo de energía final asociado a biocarburantes, para todos los servicios (en kWh/m²·año)
CONSUMBIOCARB_CALEF	Consumo de energía final asociado a biocarburantes, para el servicio de calefacción (en kWh/m²·año)
CONSUMBIOCARB_REFRIG	Consumo de energía final asociado a biocarburantes, para el servicio de refrigeración (en kWh/m²·año)
CONSUMBIOCARB_ACS	Consumo de energía final asociado a biocarburantes, para el servicio de agua caliente sanitaria (en kWh/m²·año)
CONSUMBIOCARB_ILU	Consumo de energía final asociado a biocarburantes, para el servicio de iluminación (en kWh/m²·año)
CONSUMPELLET_GLOBAL	Consumo de energía final asociado al pellet, para todos los servicios (en kWh/m²·año)
CONSUMPELLET_CALEF	Consumo de energía final asociado al pellet, para el servicio de calefacción (en kWh/m²·año)
CONSUMPELLET_REFRIG	Consumo de energía final asociado al pellet, para el servicio de refrigeración (en kWh/m²·año)
CONSUMPELLET_ACS	Consumo de energía final asociado al pellet, para el servicio de agua caliente sanitaria (en kWh/m²·año)
CONSUMPELLET_ILU	Consumo de energía final asociado al pellet, para el servicio de iluminación (en kWh/m²·año)
CERRAMIENTOSOPACOSSUP	Superficie de elemento de la envolvente que delimita todo el espacio habitable del exterior (en m²).
HUECOSYLUCERNARIOSSUP	Superficie del hueco o lucernario (en m²). Incluye la superficie total de marco y vidrio.
CERRAMIENTOSOPACOSTRANS	Valor de la transmitancia térmica del elemento, según defnición del DB-HE1 (en W/m²K).

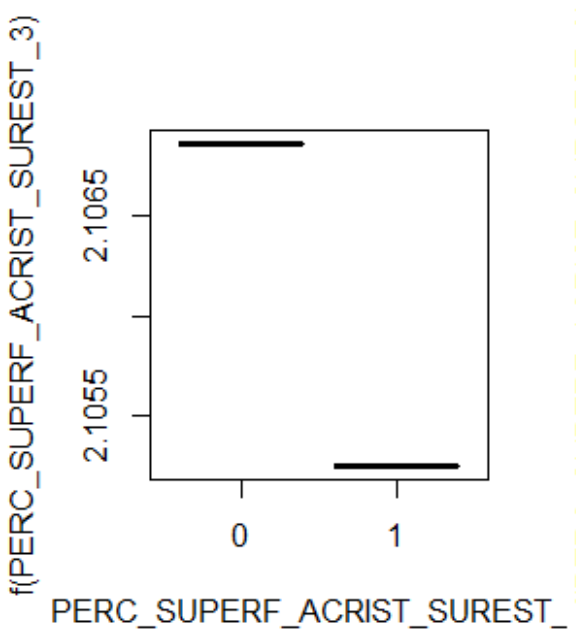
PUENTESTERMICOSTRANS	Valor de la transmitancia térmica (lineal) del elemento, según defnición del DB-HE1 (en W/mK).
HUECOSYLUCERNARIOSTRANS	Valor de la transmitancia térmica del elemento, según defnición del DB-HE1 (en W/m²K).
PUENTESTERMICOSLONG	Dimensión (longitud) del puente térmico (en m).
CERRAMIENTOSOPACOSESPESSOR	Espesor de la capa (en m) del cerramiento
PUENTESTERMICOSESPESSOR	Espesor de la capa (en m) de los puentes térmicos
HUECOSYLUCERNARIOSESPESSOR	Espesor de la capa (en m) de los huecos y lucernarios
CERRAMIENTOSOPACOSCOND	Conductividad térmica del material (W/m·K) del cerramiento opaco
PUENTESTERMICOSCOND	Conductividad térmica del material (W/m·K) de los puentes térmicos
HUECOSYLUCERNARIOSCOND	Conductividad térmica del material (W/m·K) de los huecos y lucernarios
CERRAMIENTOSOPACOSRESIS	Resistencia térmica de la capa (en m2·K/W) (para materiales sin inercia térmica significativa: p.e. cámaras de aire) en cerramientos opacos
PUENTESTERMICOSRESIS	Resistencia térmica de la capa (en m2·K/W) (para materiales sin inercia térmica significativa: p.e. cámaras de aire) en puentes térmicos
HUECOSYLUCERNARIOSRESIS	Resistencia térmica de la capa (en m2·K/W) (para materiales sin inercia térmica significativa: p.e. cámaras de aire) en huecos y lucernarios
CERRAMIENTOSOPACOSDENS	Densidad del material de la capa (kg/m3) de cerramientos opacos
PUENTESTERMICOSDENS	Densidad del material de la capa (kg/m3) de puentes térmicos
HUECOSYLUCERNARIOSDENS	Densidad del material de la capa (kg/m3) de huecos y lucernarios
CERRAMIENTOSOPACOSVAPO	Factor de resistencia a la difusión del vapor de agua de la capa(adimensional) en cerramientos opacos
PUENTESTERMICOSVAPO	Factor de resistencia a la difusión del vapor de agua de la capa(adimensional) en puentes térmicos
HUECOSYLUCERNARIOSVAPO	Factor de resistencia a la difusión del vapor de agua de la capa(adimensional) en huecos y lucernarios
CERRAMIENTOSOPACOSCALOR	Calor específico (en J/kg·K) en cerramientos opacos
PUENTESTERMICOSCALOR	Calor específico (en J/kg·K) en puentes termicos
HUECOSYLUCERNARIOSCALOR	Calor específico (en J/kg·K) en huecos y lucernarios
CO2_GLOBAL	PUNTUACION CO2 GLOBAL
CO2_CALEF	PUNTUACION CO2 CALEF
CO2_REFRIG	PUNTUACION CO2 REFRIG
CO2_ACS	PUNTUACION CO2 ACS
ENERPRIMARIANOR_GLOBAL	Consumo de energía primaria no renovable para todos los servicios, descontando las aportaciones de energías renovables. Incluye, en uso residencial privado, los servicios de calefacción, refrigeración y ACS; y, en uso

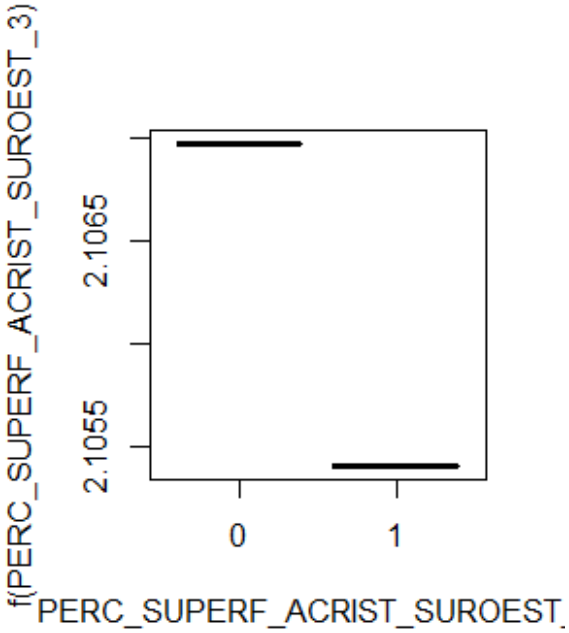
	terciario, los servicios de calefacción, refrigeración, ACS e iluminación, considerando el impacto derivado de ventiladores, bombas y torres de refrigeración (en kWh/m²·año)
ENERPRIMARIANOR_CALEF	Consumo de energía primaria no renovable para el servicio de calefacción (en kWh/m²·año)
ENERPRIMARIANOR_REFRIG	Consumo de energía primaria no renovable para el servicio de refrigeración (en kWh/m²·año)
ENERPRIMARIANOR_ACS	Consumo de energía primaria no renovable para el servicio de ACS (en kWh/m²·año)
INSTCALEF_POTENCIA	Potencia nominal del equipo (en kW). En el caso de equipos ideales, de referencia o sustitución se recomienda señalarlo introduendo el valor 9999999999.99
INSTCALEF_RENDIM	Rendimiento o COP nominal del equipo. En el caso de equipos ideales, de referencia o sustitución con rendimiento constante este valor equivale al rendimiento medio y se puede considerar igual al rendimiento estacional.
INSTACS_POTENCIA	Potencia nominal del equipo (en kW)). En el caso de equipos ideales, de referencia o sustitución se recomienda señalarlo introduendo el valor 9999999999.99
INSTACS_RENDIM	Rendimiento o COP nominal del equipo. En el caso de equipos ideales, de referencia o sustitución con rendimiento constante este valor se puede considerar igual al rendimiento estacional y medio.
INSTREFRIG_POTENCIA	Potencia nominal del equipo (en kW)). En el caso de equipos ideales, de referencia o sustitución se recomienda señalarlo introduendo el valor 9999999999.99
INSTREFRIG_RENDIM	EER nominal del equipo (aplicado a la potencia sensible). En el caso de equipos ideales, de referencia o sustitución con rendimiento constante este valor se puede considerar igual al rendimiento estacional y medio.
EMISSIONS_GLOBAL_LOG	Logaritmo neperiano de las emisiones globales en CO2

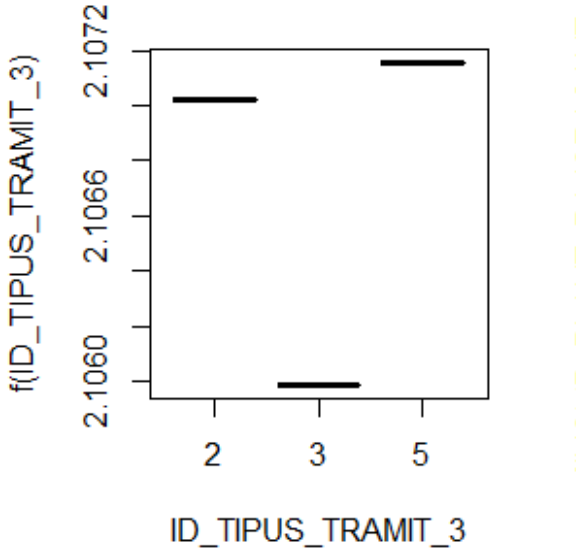
ANNEX 2: Contribucions marginals restants del XGboost

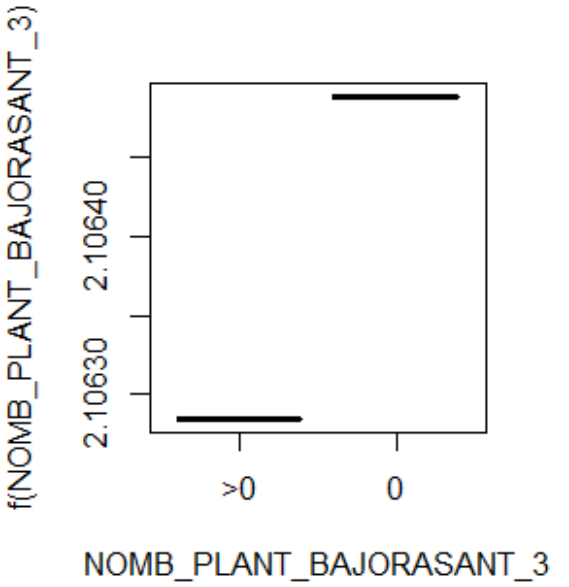


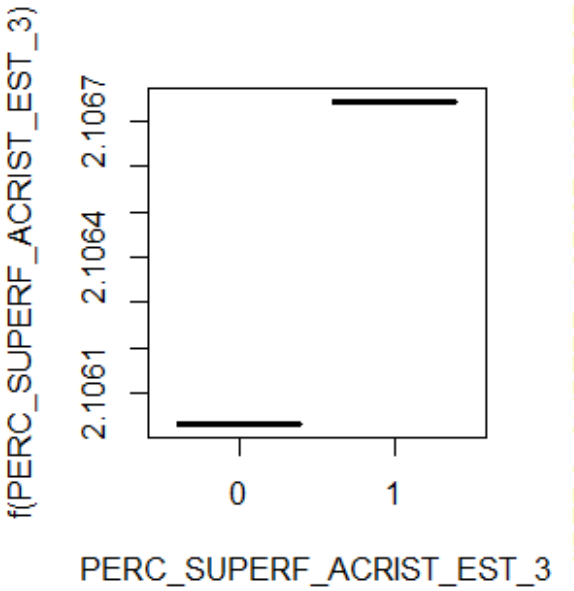


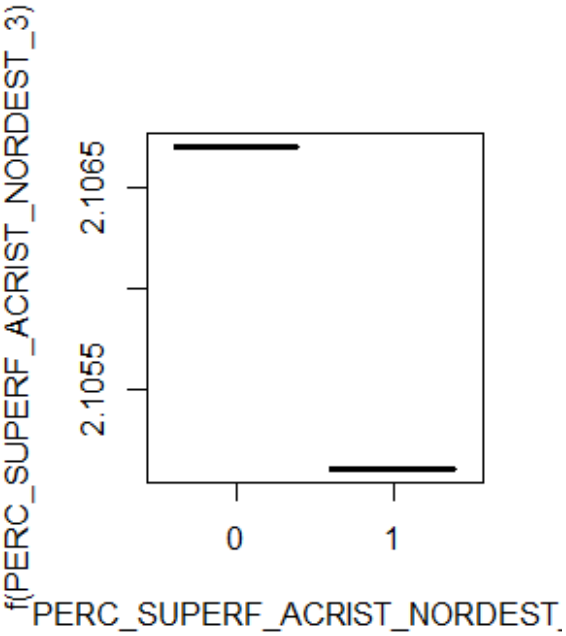

f(PERC_SUPERF_ACRIST_SUREST_3)
2.1055
2.1065
0
1
PERC_SUPERF_ACRIST_SUREST_

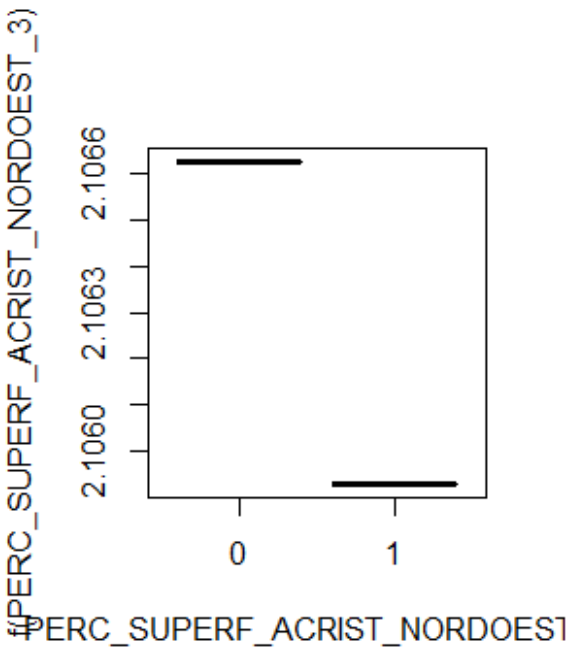

f(PERC_SUPERF_ACRIST_SUROEST_3)
2.1055
2.1065
0
1
PERC_SUPERF_ACRIST_SUROEST_


f(ID_TIPUS_TRAMIT_3)
2.1060
2.1066
2.1072
2
3
5
ID_TIPUS_TRAMIT_3


f(NOMB_PLANT_BAJORASANT_3)
2.10630
2.10640
>0
0
NOMB_PLANT_BAJORASANT_3


f(PERC_SUPERF_ACRIST_EST_3)
2.1061
2.1064
2.1067
0
1
PERC_SUPERF_ACRIST_EST_3


f(PERC_SUPERF_ACRIST_NORDEST_3)
2.1055
2.1065
0
1
PERC_SUPERF_ACRIST_NORDEST_


f(PERC_SUPERF_ACRIST_NORDOEST_3)
2.1060
2.1063
2.1066
0
1
PERC_SUPERF_ACRIST_NORDOEST