TREBALL DE FI DE GRAU

Grau en Estadística

Títol: Anàlisi de la presència d'elements químics en el fang de les depuradores de rius de Catalunya

Autor: Max Planas i Batllori

Director: Antonio Monleón-Getino. Javier Méndez Viera.

Departament: Secció Estadística. Departament de Genètica, Microbiologia i Estadistica. Fac Biologia. UB

Convocatòria: Juny 2023

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Índex

1. Resum

Aquest treball de final de grau té com a objectiu analitzar el fang en diferents depuradores de diversos rius de Catalunya a través de tècniques estadístiques com sèries temporals i models lineals. L'estudi es centra en la presència de diversos elements químics com el nitrogen, potassi, calci i pH, i busca observar l'evolució de l'estat dels fangs i detectar possibles canvis. Aquesta anàlisi proporcionarà una millor comprensió de la salut dels rius i les implicacions per al medi ambient i la salut pública.

2 .Introducció

La qualitat de l'aigua és un tema de preocupació creixent a escala mundial, ja que la contaminació de les aigües pot tenir efectes negatius en la salut humana, la biodiversitat i la sostenibilitat dels ecosistemes aquàtics. La gestió de les aigües residuals és una de les principals tasques en la prevenció de la contaminació i la protecció de la salut pública. En aquest sentit, les depuradores són un element clau per a la gestió dels residus, ja que permeten reduir la quantitat de contaminants que s'alliberen als rius i altres cursos d'aigua.

El fang que es genera a les depuradores és un subproducte important del procés de depuració, i el seu tractament i gestió són crucials per a la prevenció de la contaminació. Això es deu al fet que el fang absorbeix, en el procés de depuració de les aigües que provenen de les empreses i ciutadans, una gran quantitat de contaminants, com ara nutrients, metalls pesants i compostos orgànics, que poden afectar negativament la salut dels ecosistemes aquàtics.

Per comprendre millor la presència i l'evolució dels elements químics en el fang de les depuradores, aquest treball de final de grau té com a objectiu analitzar el fang de diverses depuradores de rius de tot Catalunya. A través de l'aplicació de diverses tècniques estadístiques, com ara sèries temporals, models lineals i altres mètodes d'anàlisi, es pretén detectar possibles tendències o canvis en la presència dels diferents elements químics en els fangs de les depuradores.

Així doncs, els resultats d'aquest estudi poden ser clau per a la millora de la gestió dels residus i la prevenció de la contaminació en les depuradores. També poden tenir implicacions importants per a la salut dels rius de Catalunya i la gestió sostenible dels seus recursos hídrics. El contingut i els resultats d'aquest treball de final de grau poden ser de gran interès per als professionals que treballen en la gestió de les aigües residuals com per exemple l'Agència Catalana de l'Aigua (ACA) que és l'entitat que ens ha proporcionat les dades que s'utilitzaran per desenvolupar aquest treball, així com per a la comunitat científica que estudia la contaminació de les aigües i les seves conseqüències en la salut pública i l'ecologia aquàtica.

3. Objectius

En aquest estudi es plantejaran diversos objectius per tal d'aprofundir en l'anàlisi del fang de les depuradores a Catalunya. Amb un enfocament multidisciplinari, es buscarà comprendre millor les característiques i les propietats del fang, així com identificar possibles millores en el tractament i la gestió d'aquest recurs valuós.

El primer objectiu fonamental serà estudiar i tractar les dades existents per tal d'aconseguir una base de dades òptima que ens permeti realitzar anàlisis en profunditat i prendre decisions basades en evidències. Això implicarà netejar les dades, solucionar problemes de manca d'informació o valors atípics, i organitzar-les de manera adequada per a l'anàlisi posterior.

En segon lloc es realitzarà una anàlisi tant univariant com bivariant de les dades fent ús d'una eina de Microsoft anomenada Power BI. Un cop s'hagi conclòs aquest punt, l'estudi es centrarà en la part de modelització. En aquesta etapa, la primera qüestió relativa que s'abordarà és la possible tendència pel que fa a les concentracions dels diferents elements que s'estudiaran. L'objectiu serà localitzar aquests punts de canvi i analitzar-los utilitzant Power BI.

Finalment, el treball s'enfocarà a modelar i centrar l'estudi en una variable que pugui ser d'un gran interès, com podria ser el pH o la conductivitat elèctrica. S'aplicaran diverses tècniques de regressió per tal de poder determinar quina és la millor manera de construir un model de predicció entorn algun d'aquests indicadors amb la finalitat de realitzar prediccions de la manera més precisa.

En tots aquests punts, s'aplicaran els coneixements adquirits durant el grau d'estadística de diverses assignatures com mineria de dades, models lineals, models lineals generalitzats, i altres assignatures relacionades amb aquest camp, per tal d'obtenir resultats significatius i fiables.

4. Metodologia

4.1. Selecció de les depuradores i recollida de mostres

L'Agencia Catalana de l'Aigua (ACA) és l'organització encarregada de realitzar la recollida de mostres d'aigua a diferents zones del territori català per controlar i avaluar la qualitat de les masses d'aigua. En aquest apartat us explicarem pas a pas la metodologia de la selecció i recollida de mostres que utilitza l'ACA per recopilar la informació que utilitzarem en aquest estudi:

1-Planificació i selecció de llocs de mostreig:

L'Agència Catalana de l'Aigua (ACA) selecciona depuradores per inspeccions basant-se en risc ambiental, mida, tipus d'establiment i històric de conformitat. També es realitzen anàlisis de risc per avaluar les conseqüències de no complir amb les normatives. Les inspeccions poden ser aleatòries o periòdiques i són essencials per verificar el compliment normatiu i prendre mesures correctores en cas de detectar incompliments. Això contribueix a protegir i millorar la qualitat de les aigües residuals a Catalunya, especialment en zones sensibles o prop de conques hidrogràfiques. L'ACA assegura una supervisió exhaustiva per garantir un tractament adequat dels residus i evitar impactes negatius en el medi ambient. A continuació, es mostra un mapa de les localitzacions de les diferents depuradores:

Imatge 4.1 Depuradores de Catalunya

2- Equip i preparació:

Sobre la metodologia a l'hora de recollir les mostres no s'ha trobat massa informació. Tot i això, es coneix que en aquests casos el més comú és realitzar una recollida de mostres amb instruments de seguretat i equips especialitzats, mesurar part d'aquestes mostres en el mateix moment com per exemple el pH i la conductivitat; i deixar la resta per a una posterior anàlisi al laboratori.

3-Recollida de mostres:

Un cop arribat al punt de mostreig, els tècnics de l'ACA segueixen un protocol estandarditzat. Es recullen mostres representatives de la superfície a diferents profunditats, en funció del tipus de massa de fang i dels paràmetres a analitzar.

4- Manipulació i emmagatzematge de mostres:

Després de recollir les mostres, s'etiqueten adequadament i es transfereixen immediatament a dipòsits d'emmagatzematge refrigerats per mantenir les temperatures adequades i evitar canvis en les seves propietats físiques i químiques. És fonamental garantir que les mostres estiguin lliures de contaminació durant el transport i l'emmagatzematge.

5-Anàlisi de mostres:

Les mostres de fang recollides es porten al laboratori de l'ACA per a l'anàlisi detallat de diversos paràmetres físics, químics i biològics. Aquests paràmetres inclouen oxigen dissolt, concentració de nutrients, presència de contaminants, presència de microorganismes patògens i altres elements. En el següent apartat s'aprofundeix més sobre on i com es realitza aquesta anàlisi químic.

4.2. Anàlisi químic

4.2.1 Laboratori

L'Agència Catalana de l'Aigua té un laboratori d'anàlisi a la planta de potabilització d'Abrera. Aquest laboratori realitza les tasques i determinacions analítiques per controlar la qualitat de les aigües superficials i subterrànies, complint el Pla de Seguiment i Control de la Directiva marc de l'aigua. Analitza mostres d'aigua per al subministrament en alta segons la legislació vigent, controla les depuradores conforme a la Directiva 91/271 , i caracteritza els abocaments industrials per verificar el seu ús i contaminació. A més, realitza anàlisis sol·licitats per altres entitats i estudis relacionats amb la qualitat del medi hídric.

4.2.2 Sistema de qualitat.

El laboratori de l'Agència Catalana de l'Aigua realitza anàlisis per controlar la qualitat de les aigües potables i continentals. S'analitzen paràmetres físics, químics i microbiològics, incloent microcontaminants orgànics i inorgànics. També es realitzen anàlisis de mostres d'abocaments i depuradores. El laboratori segueix els criteris de les normes UNE-EN ISO/IEC 17025:2017 i compta amb acreditacions de l'ENAC (*Entidad Nacional de Acreditación*) per assegurar la qualitat dels resultats. Amb aquestes tasques, contribueix a la preservació del medi aquàtic i proporciona informació per prendre decisions basades en dades científiques. És un actor clau en el control i la caracterització de les aigües a Catalunya.

4.2.3 Funcions del laboratori

El laboratori de l'Agència Catalana de l'Aigua és responsable de realitzar anàlisis per a controlar i avaluar la qualitat de les aigües potables, continentals i residuals. A través d'aparells i instruments d'alta precisió analítica, es determinen paràmetres físics, químics i microbiològics essencials per garantir el compliment dels requisits legislatius i les necessitats dels clients. Els paràmetres analitzats s'inclouen en diverses categories, com ara els físico-químics, els microcontaminants orgànics i inorgànics, i els paràmetres microbiològics. A més del control de les aigües, el laboratori realitza anàlisis de mostres d'abocaments i depuradores, responent a sol·licituds d'altres institucions. Amb una

àmplia gamma de tasques i anàlisis, el laboratori assegura el seguiment sistemàtic de la qualitat de les aigües i la caracterització dels impactes mediambientals causats pels abocaments. També s'encarrega de realitzar estudis específics per avaluar diversos aspectes de la qualitat del medi aquàtic, amb la participació de col·laboradors externs. Així, el laboratori de l'Agència Catalana de l'Aigua exerceix un paper fonamental en la protecció i la gestió sostenible dels recursos hídrics a Catalunya.

4.2.4. Tècniques d'anàlisi dels materials químics utilitzades.

El laboratori està equipat amb tecnologia avançada per realitzar les anàlisis requerides. En química orgànica, es disposa de cromatògrafs líquids i de gasos amb detectors diversos. Per a l'anàlisi d'elements inorgànics, hi ha espectròmetres de masses, fluorescència atòmica, absorció atòmica, cromatògrafs iònics i espectrofotòmetres UV-Vis. També s'utilitzen analitzadors de flux segmentat i robots per a determinacions automatitzades de DQO i DBO. Pel que fa a la biologia, s'utilitzen tècniques com la PCR i la microscòpia, amb microscopis per recomptar quists de protozous, analitzar fitoplàncton i microinvertebrats, i controlar el musclo zebrat. Aquest equipament avançat permet dur a terme les anàlisis de forma precisa i eficient.

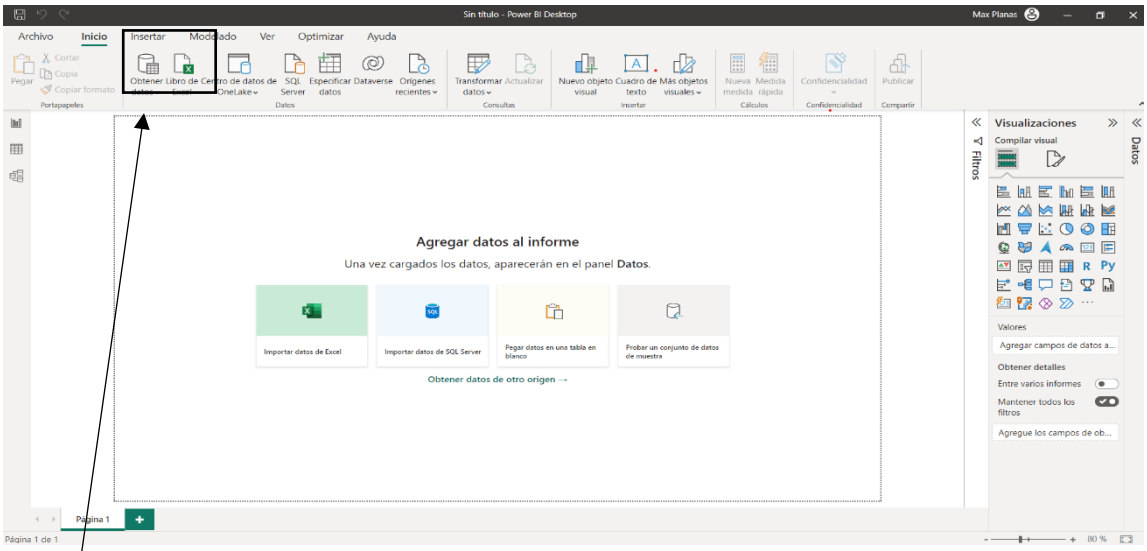
4.3. Eines d’anàlisi estadístiques utilitzades

Les eines d'anàlisi estadístiques són essencials per a l'exploració de dades i la presa de decisions informades. A través d'aquestes eines, es pot tractar, transformar i especificar les variables, així com imputar valors faltants utilitzant tècniques com la imputació mitjana o KNN. … Per a la modelització, s’apliquen models de regressió simple o múltiple, arbres de decisió i altres tècniques per predir i classificar. Pel que fa a la part més analítica s’utilitzarà el software de R-studio. A partir de aquest programa s’analitzaran i es tractaran les dades però pel que fa a la part visual degut a la complexitat de la base de dades la millor opció és fer ús del Power bi que permetrà realitzar informes més dinàmics.

4.4. Power Bi

Power BI és una potent eina desenvolupada per Microsoft de visualització de dades que permet exportar bases de dades complexes de diversos formats, relacionar-les i presentar-les d'una manera clara i entenedora. En aquest estudi es farà ús d'aquesta eina per tal de poder realitzar diferents visualitzacions que poden ser molt útils a l'hora d'interpretar els resultats que s'aniran obtenint a partir de R-studio. Un gran avantatge d'aquest programa, a diferència de l'R-studio, és que els informes que genera seran dinàmics gràcies a la utilització de diversos filtres que permetran modelar i interpretar els diversos models estadístics que s'aniran plantejant durant l'estudi. A continuació es farà una breu explicació sobre les característiques principals d'aquest programa.

4.4.1. Importar les dades



Imatge 4.2 Power BI

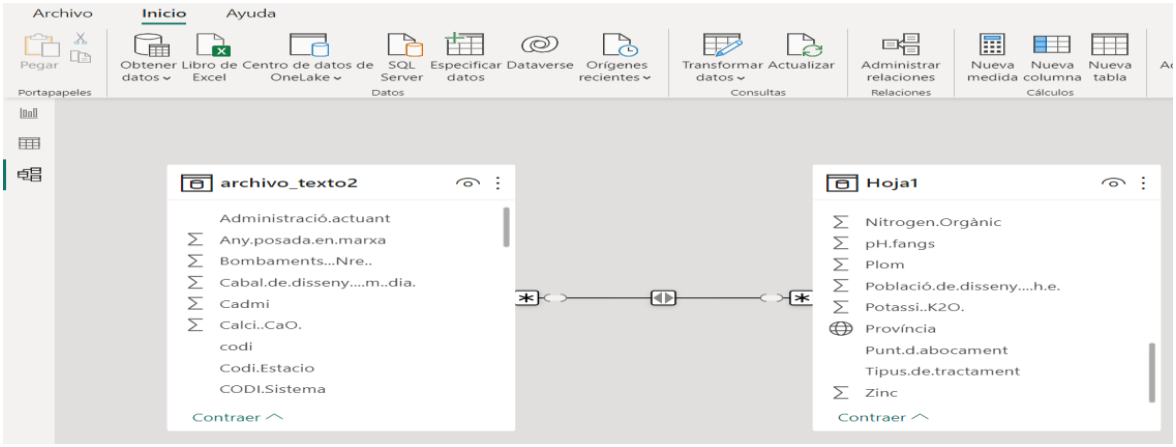
Primer de tot, s'importa la base de dades. En general power bi té un gran número de formes d'importar bases de dades, les cuals es troben des de les més bàsiques com Excel, csv o txt a Mysql, o inclús fer connexions en directe per tal de que les dades s'actualitzin en temps real.

4.4.2 Tractament de variables

Imatge 4.3 Power BI Herramientas

Una vegada importada cal observar si la base de dades està ben definida. A la secció de 'Vista de datos' es mira que les dades estiguin ben categoritzades. És important que les variables que fan referència a una localització estiguin ben especificades així no hi haurà problemes a l'hora de realitzar un mapa.

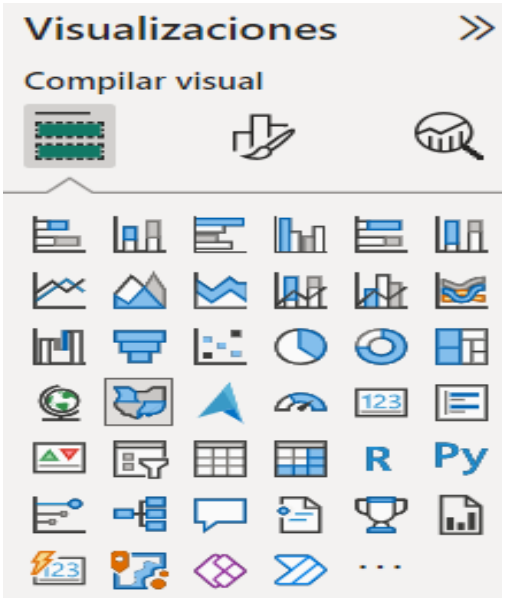
4.4.3 Taules Relacionades



Imatge 4.4 Power BI Relaciones

Si hi ha varies bases de dades s'haurà de mirar com es relacionen entre elles (apartat 'Vista de modelo'). Utilitzant taules relacionals es pot observar com es relacionen les diferents bases de dades. En general, el programa sol detectar les diverses relacions (una taula pot estar relacionada un a un, un a varis, i varis a varis) però en cas contrari s'haurà d'especificar manualment a 'Administrar relaciones'.

4.4.4 Visualitzacions



Imatge 4.5 Power BI Visualizaciones

A la secció de visualització de Power BI, trobem un ampli ventall d'eines per representar i analitzar visualment dades, facilitant la comprensió i detecció de patrons, tendències i relacions. Algunes de les visualitzacions més utilitzades inclouen gràfics de columnes i barres per comparar valors entre categories, gràfics de línies per mostrar l'evolució al llarg del temps, gràfics de dispersió per identificar correlacions, gràfics d'àrea per destacar els canvis en la composició de dades, escales de gràfics circulars i mapes per representar dades geogràfiques. A més, Power BI ofereix opcions avançades com ara mapes d'arbre, gràfics de Gantt i matrius, i us permet personalitzar visualitzacions amb colors, etiquetes i altres elements de disseny per crear informes interactius i atractius.

5. Marc conceptual

5.1. Gestió de les aigües residuals

Les aigües residuals, també conegudes com a aigües residuals municipals o aigües residuals domèstiques, són produïdes per activitats humanes, com ara habitatges, fàbriques, negocis i oficines. Aquest fangs contenen contaminants, com ara nutrients, sòlids en suspensió i patògens, que s'han d'eliminar o reduir abans de l'abocament al medi ambient. La gestió adequada de les aigües residuals és important per protegir la salut pública i el medi ambient. Les aigües residuals no tractades poden ser una font de contaminació de rius, mars i aqüífers, amb efectes negatius sobre la biodiversitat i la qualitat de l'aigua. A més, l'abocament d'aigües residuals no depurades pot danyar la salut de les zones d'aigua i afectar les activitats recreatives i econòmiques que depenen d'aquestes zones. Per gestionar les aigües residuals de manera eficaç, és necessària una depuradora. Una depuradora d'aigües residuals és una instal·lació encarregada de tractar les aigües residuals per eliminar els contaminants i produir aigua que sigui segura per al medi ambient o per a la seva reutilització, com ara l'aigua de reg.

El tractament d'aigües residuals acostuma a implicar diferents processos, que inclouen tractament primari, tractament secundari i, en alguns casos, tractament terciari. Durant aquests processos s'utilitzen diferents tecnologies i mètodes, com ara la sedimentació, la filtració, la desinfecció i la separació biològica, per eliminar o reduir els contaminants i millorar la qualitat de l'aigua calenta. És important destacar que la gestió de les aigües residuals inclou no només un tractament adequat, sinó també la seva correcta recollida i transport. El sistema de clavegueram s'encarrega de recollir les aigües residuals de la zona de producció i transportar-les a la depuradora per a un posterior tractament.

5.2. Les depuradores i els seus contaminants.

En una depuradora es poden trobar molts contaminants que provenen de les activitats humanes. Aquests contaminants poden ser de diferents tipus, inclosos els orgànics i els inorgànics. Entre els contaminants més comuns es troben els sòlids en suspensió, com els sediments i la matèria orgànica. També es poden trobar contaminants com oli, greixos, detergents, productes químics industrials i agrícoles i compostos orgànics volàtils. A més, les aigües residuals poden contenir nutrients com nitrats i fosfats, que poden provocar problemes d'eutrofització i disponibilitat d'aigua. Altres contaminants comuns inclouen metalls pesants, hormones, fàrmacs i microorganismes patògens. Les depuradores d'aigües residuals estan dissenyades per eliminar o reduir aquests contaminants mitjançant mètodes de tractament adequats com ara la sedimentació, filtració, biodegradació i desinfecció. Això garanteix que l'aigua potable compleix amb bons estàndards i és bona per al medi ambient i per al consum humà a la indústria de l'aigua potable. A continuació es presenta una taula amb la llista d'elements principals que s'han detectat a les depuradores i informació essencial pel que fa a cada un d'aquests elements.

5.3. Els diferents elements químics i altres indicadors

5.3.1 Els diferents elements químics

Els fangs de les plantes de tractament pot contenir una varietat d'elements químics que poden tenir impactes sobre la salut humana i el medi ambient. En aquest estudi s'analitzaran diversos d'aquest elements per intentar estudiar l'evolució pel que fa a les concentracions (mg/Kg) i quins possibles impactes pot tenir sobre el cos humà i sobre el medi ambient. Primerament, s'haurà d'entendre una mica millor quins són aquests elements i quins possibles efectes negatius poden tenir. A continuació es troba un llistat amb els diferents elements i quins són els seus efectes:

- Plom (Pb): El plom és un metall tòxic que pot afectar negativament la salut humana. L'exposició a l'aigua alta en plom pot causar danys al sistema nerviós, problemes de desenvolupament en nens, danys renals i efectes adversos al sistema cardiovascular. A més, el plom pot acumular-se al medi ambient, provocant la contaminació del sòl i de l'aigua.
- Níquel (Ni): El níquel és un metall que pot estar present a l'aigua de les plantes de tractament com a conseqüència d'activitats industrials. L'exposició a alts nivells de níquel pot tenir efectes nocius per a la salut, com ara irritació de la pell, contaminació d'orina, danys pulmonars i efectes cancerígens a llarg termini.

- Zinc (Zn): El zinc és un micronutrient essencial per al cos humà, però les altes concentracions de zinc a l'aigua poden tenir efectes adversos. La ingestió d'aigua alta en zinc pot causar nàusees, vòmits, diarrea i molèsties gastrointestinals. A més, les altes concentracions de zinc alliberades al medi ambient poden afectar negativament els organismes aquàtics.
- Cadmi (Cd): El cadmi és un metall tòxic que es pot alliberar a l'aigua a través de diverses activitats industrials. L'exposició a alts nivells de cadmi pot causar danys renals, pulmonars, problemes ossis i efectes cancerígens.
- Coure (Cu): El coure és un oligoelement essencial per al cos humà, però les altes concentracions de coure a l'aigua poden tenir efectes adversos per a la salut. Consumir aigua alta en coure pot provocar problemes gastrointestinals i danys al fetge. A més, les altes concentracions de coure alliberades al medi ambient poden ser tòxiques per als organismes aquàtics.

- Crom (Cr): El crom pot estar present a l'aigua com a crom hexavalent (Cr(VI)), una forma altament tòxica i cancerígena. L'exposició a alts nivells de crom a l'aigua pot tenir efectes negatius per a la salut, com ara danys pulmonars, problemes renals i danys respiratoris.

- Mercuri (Hg): El mercuri és un metall altament tòxic que pot estar present a l'aigua a causa de les activitats industrials i la contaminació ambiental. L'exposició al mercuri pot tenir greus efectes sobre la salut, especialment en el sistema nerviós. La contaminació de l'aigua pot causar danys cerebrals, problemes de desenvolupament en nens i efectes adversos en el sistema cardiovascular.

- Ferro (Fe): El ferro és un micronutrient essencial que es troba a l'aigua. Si bé els nivells baixos de ferro a l'aigua són necessaris per a la salut, les altes concentracions poden causar problemes en el gust, l'olor i l'aspecte de l'aigua.

- Fòsfor (P): El fòsfor per si sol no té efectes directes sobre la salut ni el medi ambient. No obstant això, l'excés de fòsfor a l'aigua pot provocar l'eutrofització dels ecosistemes aquàtics, la qual cosa pot afectar negativament els organismes aquàtics.
- Potassi (K): El potassi en si no té efectes directes sobre la salut ni el medi ambient. És un nutrient essencial per als éssers vius i es troba de forma natural a l'aigua.

- Nitrogen (N): El nitrogen en forma d'amoni (NH_4^+) actua com un àcid feble que redueix lleugerament el pH de l'aigua quan està present en concentracions elevades. A més, la presència de nitrat (NO_3^-) pot indicar contaminació i degradació de l'aigua, però no afecta directament el pH.

5.3.2 Altres indicadors

Ph : El valor del pH dels fangs de depuradora és un factor important a tenir en compte en el tractament d'aigües residuals. El pH és una mesura que indica l'acidesa o alcalinitat d'una substància (en aquest cas, els fangs produïts durant el procés de purificació). El pH dels fangs de depuradora pot variar degut a diversos factors, com ara la composició de les aigües residuals tractades, el procés de tractament utilitzat i les condicions ambientals. El pH adequat dels fangs de depuradora és important per diverses raons. En primer lloc, el pH afecta l'activitat dels microorganismes responsables de la degradació de la matèria orgànica a les aigües residuals. El pH òptim afavoreix l'activitat bacteriana i la descomposició dels compostos contaminants, ajudant a millorar el rendiment. En general, és important mantenir el pH dels fangs dins d'un rang òptim per garantir el bon funcionament de la depuradora i per obtenir la millor qualitat dels fangs.

Conductivitat elèctrica: La conductivitat elèctrica dels fangs de depuradora és important per mesurar la presència de substàncies dissoltes. L'alta conductivitat indica una major concentració de sals i metalls. El seguiment és fonamental per garantir que els equips de depuració funcionen correctament i per evitar problemes com ara sedimentació o incrustació. La conductivitat òptima s'ha de mantenir mitjançant tractaments complementaris, si cal. Aquest paràmetre ajuda a garantir una qualitat suficient dels fangs i protegir el medi ambient. El control de la conductivitat elèctrica és fonamental per a una gestió eficaç dels fangs de depuradora i el tractament d'aigües residuals.

5.4. Diversos Tractaments.

L'ACA utilitza diferents mètodes de tractament per tractar les aigües residuals. A continuació es pot trobar una breu explicació per a cada un dels mètodes utilitzats pel que fa al tractament de les aigües residuals:

- Biològic: El tractament biològic és un mètode comú de tractament d'aigües residuals. Es basa en l'ús de microorganismes com bacteris i fongs per descompondre i eliminar els contaminants orgànics presents a les aigües residuals. El procés es pot dur a terme mitjançant bioreactors aeròbics o anaeròbics, en funció de les condicions específiques de cada depuradora.
- Eliminació biològica de fòsfor: Aquest tipus de tractament biològic es centra específicament en l'eliminació de fòsfor, un contaminant comú de les aigües residuals. L'ús de processos biològics especials afavoreix el creixement de microorganismes que poden absorbir i emmagatzemar el fòsfor present a l'aigua.
- Eliminació biològica de nitrogen: En aquest cas, el tractament biològic pretén eliminar el nitrogen, un altre contaminant comú. S'utilitzen diferents processos biològics com la nitrificació i la desnitrificació per convertir el nitrogen orgànic en forma gasosa, que després es perd a l'atmosfera.
- Desnitrificació biològica i tractament terciari: A més de la desnitrificació, aquest tractament inclou una etapa de tractament terciari addicional. El tractament terciari normalment implica processos avançats, com ara la filtració de membrana o l'ús de carbó actiu, per eliminar contaminants residuals específics i garantir una qualitat de l'aigua superior
- Eliminació biològica de nitrogen i fòsfor: Aquest tractament pretén eliminar el nitrogen i el fòsfor de les aigües residuals. Per aconseguir-ho s'utilitzen diversos processos biològics combinats, com ara la nitrificació, la desnitrificació i la precipitació química.
- Tractament terciari de desnitrificació biològica i eliminació de fòsfor: A més de l'eliminació de nitrogen i fòsfor, aquest tractament també combina etapes de tractament terciari per millorar encara més la qualitat de l'aigua tractada.
- Tractament biològic i terciari: Aquest tipus de tractament biològic es combina amb etapes de tractament terciari per abordar l'eliminació de contaminants específics o aconseguir una major qualitat de l'aigua mitjançant processos avançats.
- Llacunatge: Una llacuna és un mètode de tractament natural que permet que les aigües residuals flueixin a través d'un sistema de llacuna artificial. En aquestes llacunes es produeixen una sèrie de processos físics, químics i biològics que ajuden a eliminar els contaminants presents a l'aigua.

- Tractament extensiu: Aquest tractament utilitza múltiples etapes i processos per reduir els contaminants a les aigües residuals. Normalment inclou processos biològics, físics i químics per eliminar diversos contaminants abans de l'alliberament final.
- Llacunatge amb eliminació de Fòsfor: Aquesta variació del llacunatge es centra específicament en l'eliminació del fòsfor de les aigües residuals mitjançant processos naturals i biològics que es produeixen a les llacunes.

Imatge 5.5 Tipus de tractaments

6. Tractament de dades /Preprocessing

6.1. Descripció de la base de dades

Per realitzar l'estudi s'ha treballat principalment amb dos bases de dades que s'han extret directament de l'ACA, donat el cas que les dades són obertes (és a dir no estan subjectes a cap tipus de limitació legal per utilitzar-les). Ara s'explicarà el procés d'obtenció d'aquestes. La primera a la que anomenarem 'asar', en la qual trobem els elements principals que es poden detectar en una inspecció, es va extreure directament de l'ACA quan treballava realitzant unes pràctiques curriculars per a la universitat. Es va demanar al director responsable de sistemes i es va concloure que totes aquelles dades que no continguessin informació que pogués implicar a una empresa o a un particular en concret es podien utilitzar per aquest treball sense cap tipus de problema. La segona base de dades es pot obtenir directament a la pagina de l'ACA. Aquesta s'anomenarà 'depuradores' i conté informació detallada de les depuradores (descripció sobre la metodologia que s'utilitza a la hora de realitzar les inspeccions).

La base de dades anomenada 'ASAR' conté informació recopilada durant diversos anys de diferents depuradores. L'objectiu de la base de dades és analitzar una sèrie d'elements químics presents en les mostres d'aigua de les depuradores, com el fòsfor, calci, magnesi i altres compostos rellevants. A més dels elements químics, la base de dades també inclou altres indicadors importants per avaluar la qualitat de l'aigua, com la matèria orgànica, el pH i la conductivitat elèctrica. Aquests indicadors proporcionen informació addicional sobre l'estat i la composició de l'aigua tractada a les depuradores. Per tant, la base de dades consta d'un total de 35 indicadors i un total de 13917 registres en diferents dates.

La base de dades 'depuradores' conté informació més detallada sobre les depuradores. En aquesta base de dades podem trobar informació sobre la comarca on es localitza, l'administració actuant, tipus de tractament que utilitza i altra informació que detalla el funcionament d'aquestes. La base de dades conté informació de 548 depuradores diferents, que es pot relacionar molt fàcilment amb la base de dades 'ASAR' i que gràcies a la informació que ens proporciona cada una de les diferents bases de dades es podrà realitzar un anàlisi més extens.

6.2. Unió de les bases de dades

Com ja s'ha mencionat anteriorment la informació disponible per a l'estudi es trobava en dues bases de dades diferents, el primer pas era unir-les. Per fer-ho s'ha utilitzat la funció 'merge' (Rstudio) i mitjançant una columna d'identificadors s'ha pogut relacionar les bases de dades de la següent manera: ASAR(*)~Depuradores(1).

Aquesta relació estableix que per cada depuradora de la taula 'Depuradores' trobarem diversos registres d'aquest valor a la taula de 'ASAR'.

	Núm. de variables	Núm. d'obs.
Base de dades Inspeccions (ASAR)	39	13916
Base de dades Depuradores	22	547
Base de dades del treball	60	13132

Taula 6.1 Numero de variables de les bases de dades

Les dades que s'observa que s'han perdut no tenen rellevància cap a l'estudi ja que contenien informació que no estava relacionada amb les depuradores sinó amb algun altre sistema de sanejament. Per tant al no poder relacionar-les s'ha decidit descartar-les ja que no es pot obtenir informació més detallada.

6.3. Especificació de les variables

Inicialment fent un diagnòstic ràpid s'observa que totes les variables de la base de dades estan categoritzades com 'characters'. Això pot ser un problema ja que al no estar ben especificades serà impossible poder realitzar cap tipus d'anàlisi. Per això, el primer de tot que es farà és canviar la categoria de les variables. Per a les variables numèriques s'utilitzarà la funció 'as.numeric(variable)', per a les categòriques utilitzarem la funció 'as.factor(variable)' i per a les dates s'haurà d'utilitzar 'as.Date(xxx, format=)'.

Abans:

Tipus de variable	Núm. de variables
Text ('character')	60
Numèriques	3
Factors (categòriques)	0

Taula 6.2 Tipus de variables abans

Després:

Tipus de variable	Núm. de variables
Text ('character')	0
Numèriques	32
Factors (categòriques)	30
Data	1

Taula 6.3 Tipus de variables despres

Aquestes especificacions s'han fet per a totes les variables, tot i que en els següents apartat s'haurà de reduir la dimensionalitat de la base de dades per fer una anàlisi més clar.

6.4 Transformació de les variables

S'ha detectat que varies de les variables de diferents elements químics o elements orgànics estan expressats en % en lloc de concentracions. Tenint en compte que la densitat del fang té un valor de 1.0 a 1.2 g/cm³ (s'assumeix que la densitat és 1.1g/ cm³ durant el treball però aquest podria variar en funció de la depuradora), es fa una transformació de les variables de % de fang a mg/kg de fang per a una quantificació precisa del contingut de diferents components en el fang. Aquesta transformació permet expressar les concentracions en termes de quantitat absoluta per unitat de pes del fang, proporcionant dades més fiables i permetent realitzar anàlisis més rigorosos.

Tipus de variable	Densitat	Descripció camp
Matèria.orgànica	1.2	%
Matèria.orgànica.resistent	1	%
Nitrogen.Amoniacal	1	%
Nitrogen.total	1	%
Nitrogen.Orgànic	1.2	%
Fòsfor..P2O5	1.2	%
Potassi..K2O.	0.8	%
Calci..CaO	1.5	%
Magnesi..MgO.	1.7	%
Ferro	7.9	%

Taula 6.4 Taula amb les densitats dels elements en el fang.

6.5 Selecció de variables rellevants per a l'estudi

Inicialment, la nostra base de dades contenia un total de 60 variables diferents. Per simplificar el nostre estudi i utilitzar un conjunt de variables rellevants, es va realitzar una avaluació preliminar de les variables disponibles. Es van fer dues seleccions amb diferents criteris. La primera va ser identificar variables amb un gran nombre de valors de *missings*. Aquestes variables amb un gran nombre de dades que falten podrien afectar negativament la qualitat de la nostra anàlisi. Per tant, es va decidir eliminar aquestes variables del nostre conjunt de dades per evitar resultats distorsionats o poc fiables. Després, es va procedir a analitzar les variables restants per determinar la seva rellevància per al nostre estudi. S'han considerat diversos factors com ara el propòsit de l'estudi, l'alineació amb la pregunta de recerca i la relació amb altres variables.

6.5.1 Selecció 1

A continuació, es pot observar una llista amb els percentatges de dades que falten per a cada una de les diferents variables. Com que el volum de variables és molt elevat, es va decidir escollir inicialment totes les variables que no tinguessin un % de *missings* massa elevat ja que seria molt interessant per a l'estudi realitzar una imputació de *missings*, i si més del 50% dels valors són mancats, és complicat imputar-los amb precisió i fiabilitat.

Variables	Missing_Percentatge	Variables	Missing_Percentatge
codi	0.00000000	Fòsfor (P2O5)	5.39141030
Codi Estació	0.00000000	Potassi (K2O)	3.29728906
Nom Estació	0.00000000	Calci (CaO)	5.14773073
Data	0.00000000	Magnesi (MgO)	4.75936643
Matèria seca	1.21078282	Ferro	4.98781602
pH fangs	1.67529698	Relació C/N	10.21169662
Conductivitat elèctrica fangs	4.85074627	Mercuri	1.73621687
		Cadmi	2.22357600
Matèria orgànica	4.05117271	Níquel	1.53061224
Matèria orgànica resistent	24.15473652	Plom	1.53061224
		Crom	1.71337192
Grau d'Estabilitat	24.13950655	Coure	1.53061224
Nitrogen Amoniacal	7.12762717	Zinc	1.54584222
Nitrogen total	5.26957051	Salmonella spp	93.96893086
Nitrogen Orgànic	9.68626256	Escherichia Coli	93.83947609

Variables	Missing_Percentatge
Nitrogen no hidrolitzable	63.61559549
Dioxines	99.75632044
Alquilbenzens sulfonats lineals	98.72829729
Compostos orgànics halogenats adsorbits	98.76637222
Hidrocarburs Aromàtics Policíclics	98.97197685
Bifenils Policlorats	98.88821200
Di(2-etilhexil)ftalat	99.10143162
Nonilfenol etoxilats	98.95674688
Nitrogen amoniacal	100.00000000
Nitrogen Orgànic sòlid	100.00000000
Nitrogen orgànic	100.00000000
CODI Sistema	0.00000000
Nom del sistema/EDAR	0.00000000
Municipi EDAR	0.00000000
Comarca	0.00000000
Província	0.00000000
Conca	0.00000000
Administració actuant	0.00000000

Variables	Missing_Percentatge
Empresa explotadora	0.00000000
Any posada en marxa	0.00000000
Any ampliació	64.71215352
Tipus de tractament	0.00000000
Cabal de disseny (m^3/dia)	0.00000000
Població de disseny (h-e)	0.12945477
F. sèptiques (Nre.)	47.12777394
Volum dipòsit emmagatzematge (m^3)	45.73518661
Superfície dipòsit (m^2)	45.73518661
Volumes (Nre.)	45.73518661
Cabals (m^3/dia)	0.00000000
Filtres percoladors (Nre.)	56.24738676
Superfície filtrant (m^2)	56.24738676
Cabals percoladors (m^3/dia)	56.24738676
Cabals assegurats (m^3/dia)	43.43098627
Cabals hiperass.	43.43098627
Nre. mòduls (Nre.)	43.43098627
m^2 mòdul (m^2)	43.43098627
Cabals filtrats (m^3/dia)	43.43098627

Taula 6.5 Variables seleccionades.

6.5.2. Selecció 2

En aquest apartat, es va decidir quines variables serien interessants a l'hora de modelitzar-les i poder seguir amb la investigació. En aquest cas, es va triar separar-les en dues categories: les variables explicatives i les variables resposta. Les primeres fan referència a totes aquelles variables que es creu que són importants a l'hora d'analitzar un possible resultat de les inspeccions i que ens permetin modelitzar les variables resposta, com per exemple la Data, tipus de tractament, població disseny i altres. Mentre pel que fa a les variables resposta s'han seleccionat una sèrie d'indicadors de diferents elements químics .

A continuació, es troben dues llistes de cada una de les variables i quina informació ens proporcionen. Al llistat de les variables explicatives hi trobem la categoria de les dades i una breu descripció que ajuda a entendre que es hi ha dins la variable.

Tipus_de variable	Tipus	Descripció camp
codi	Categòrica	Identifica cada una de les depuradores
Data	Date	Data del dia de la Inspecció
Comarca	Categòrica	Identifica cada una de les comarcas a Catalunya
Província	Categòrica	Tenim les 4 provincies de Catalunya
Conca	Categòrica	Conca del rius on esta localitzada la depuradora
Administració.actuant	Categòrica	Administració actuant
Empresa Explotadora	Categòrica	Empresa Explotadora
Cabal de Disseny	Numèrica	Cabal de Disseny en m3/dia
Població de Disseny	Numèrica	Població de Disseny en h-e
Punt d'Abocament	Categòrica	Punt d'Abocament en el riu
Conductivitat Elèctrica Fangs	Numèrica	Conductivitat Elèctrica Fangs en µS/cm

Taula 6.6 Descripció variables categòriques

En el següent gràfic hi ha una llista de les variables resposta que s'analitzen de manera independent.

Tipus de variable	Tipus	Descripció camp
Matèria.orgànica	Numèrica	mg/kg
Matèria.orgànica.resistent	Numèrica	mg/kg
Grau.d.Estabilitat	Numèrica	mg/kg
Nitrogen.Amoniacal	Numèrica	mg/kg
Nitrogen.total	Numèrica	mg/kg
Nitrogen.Orgànic	Numèrica	mg/kg
Fòsfor..P2O5	Numèrica	mg/kg
Potassi..K2O.	Numèrica	mg/kg
Calci..CaO	Numèrica	mg/kg
Magnesi..MgO.	Numèrica	mg/kg
Ferro	Numèrica	mg/kg
Relació.C.N	Numèrica	Sin
Mercuri	Numèrica	mg/kg
Cadmi	Numèrica	mg/kg
Níquel	Numèrica	mg/kg
Plom	Numèrica	mg/kg
Crom	Numèrica	mg/kg
Coure	Numèrica	mg/kg
Zinc	Numèrica	mg/kg
Matèria Seca	Numèrica	mg/kg
pH Fangs	Numèrica	pH Fangs

Taula 6.7 Descripció variables Numèriques

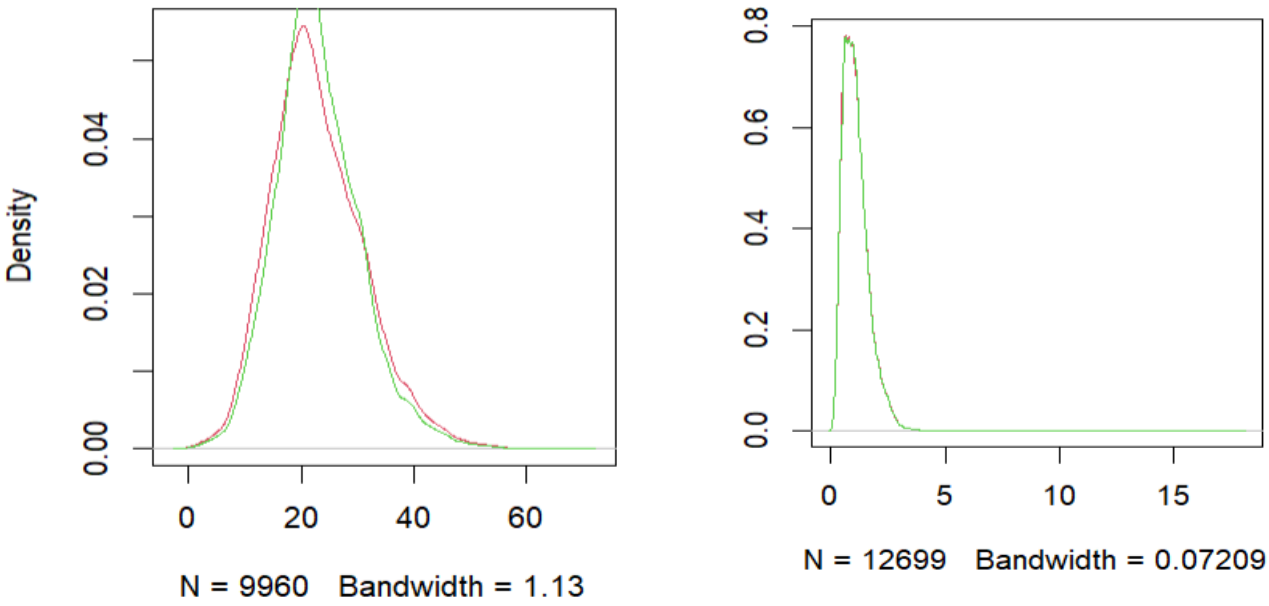
6.6. Imputació de missings

La imputació de dades mancants és un procés important per garantir la integritat i la qualitat del conjunt de dades. Utilitzant el mètode K-Nearest Neighbors (KNN), es busca trobar mostres similars basant-se en altres variables per estimar els valors mancants. Es calcula una estimació mitjançant els valors de la variable objectiu de les mostres similars, com la mitjana o la mediana. S'identifiquen les variables amb dades mancants, es calcula la distància entre mostres i s'aplica el procés per a cada mostra amb dades mancants, assignant valors imputats. La selecció del valor K i la mesura de proximitat són factors importants, i cal tenir en compte les limitacions del mètode, com la sensibilitat a les dades atípiques i la necessitat de tenir suficients mostres per a resultats precisos.

6.5.1 Validació de la imputació.

Una vegada realitzada la imputació de les dades, s'ha pres la decisió d'estudiar aquest procés en detall. Per a això, s'han seleccionat dues variables amb valors mancants. A la imatge de la dreta, es troba la variable de matèria orgànica resistent, la qual presentava un elevat percentatge de dades mancants, concretament un 25%. A l'esquerra, es troba el potassi, amb menys d'un 2% de valors mancants.

Per avaluar l'èxit de la imputació, s'han comparat les distribucions de les dades imputades amb les dades reals. Si les distribucions mostren una similitud significativa, es pot inferir que la imputació ha sigut correcta. A més, s'han utilitzat altres mètodes que es detallen als annexos de l'estudi.



Gràfic 6.1 & 6.2 Densitat Imp vs Densitat originals

En els gràfics de les imputacions per a la variable 'matèria.orgànica' s'observa una amplada de banda (Bandwidth) de 1.13 i 9960 observacions tendeixen a ser més suaus i generalitzades, oferint una visió general de les dades originals. Això significa que les dades s'ajusten generalment bé però no tenen gaire capacitat per detectar i capturar detalls finits, és a dir, es pot veure que les imputacions han capturat l'estructura general de les dades originals però no poden reflectir detalls particulars o variabilitat a nivell més granular.

Les imputacions per a la variable 'Calci' amb un amplada de banda de 0.07 i 12699 observacions tenen més detalls i capturen millor la variabilitat i els canvis en les dades originals. Això significa que les dades imputades poden presentar una estructura i tendències més precises, ja que l'amplada de banda petita permet capturar més detalls a partir de les observacions disponibles. És possible que les imputacions per a aquesta variable mostrin una major sensibilitat als canvis i a les fluctuacions en les dades originals.

Per tant, tot i que en aquest apartat nomes s'han estudiat dos variables, en els annexes es podrà trobar el codi per comprovar que la imputació de dades s'ha realitzat correctament per a la resta de variables.

7. Descriptiva simple

En l'anàlisi simple, s'explora i s'entenen millor les dades. En aquesta secció, es realitzarà una anàlisi simple de les característiques principals de les variables de les bases de dades. Inicialment, es presentarà una petita taula de freqüències per a les variables categòriques, que mostrarà les dades dels cinc primers factors amb més inspeccions (tenint en compte que hi ha variables amb més de 100 nivells). Pel que fa a les dades numèriques, s'indicarà la mitjana, la mediana i altres indicadors rellevants. A més, es mostraran diverses visualitzacions utilitzant Power BI (és important tenir en compte que s'adjunta un informe de Power BI amb totes les dades), proporcionant una anàlisi univariant i bivariant perquè el lector pugui comprendre de manera més dinàmica la base de dades.

7.1. Descriptiva Numèrica simple Categòriques

Com ja s'ha mencionat en l'apartat anterior, es pot visualitzar la informació per a cada una de les variables categòriques. A continuació, es mostren diverses taules de freqüències de les inspeccions en funció de les diferents variables categòriques:

Codi	DBSS	DTEI	DFIG	DVDP	DOLO	Other
Freq	117	111	105	103	101	12595

Taula 7.8 Taula freqüències Codi

Administració Actuant	Agència Catalana de l'Aigua	Consorci Besos Tordera	Consorci d'Aigües Costa Brava Girona	Consell Comarcal d'Osona	Mancomunitat Penedès-Garraf	Other
Freq	3000	1050	1000	897	617	

Taula 7.9 Taula freqüències Administració actuant

Comarca	Osona	Vallès Oriental	Selva	Alt emporda	Vallès Occidental	Other
Freq	897	893	821	792	664	9065

Taula 7.10 Taula freqüències Comarca

Provincia	Barcelona	Girona	Lleida	Tarragona
Freq	5540	3482	1827	2283

Taula 7.11 Taula freqüències Provincia

Conca	El Llobregat	El Ter	El Besòs	El Segre	Rieres de la costa entre Cunit i Vandellòs i l'Hospitalet de l'Infant	Other
Freq	2335	1602	1367	1299	850	5679

Taula 7.12 Taula freqüències Conca

7.2. Descriptiva Numèrica simple variables Numèriques

Variables	Min.	X1st.Qu.	Median	Mean	X3rd.Qu.	Max.
Cab.de.disseny.m..dia	1.00	2.9000	14.4000	132.8000	72.6000	970.0000
Pobl.de.disseny….h.e.	1.00	6.0000	20.000	92.0600	111.630	998.000
pH.fangs	4.10	7.0300	7.5000	7.4800	8.0000	12.7000
Con.elèctrica.fangs	0.06	1.4800	2.6000	243.600	517.250	999.000
Matèria.seca	0.04	10.389	13.776	13.1650	16.8000	81.8160
Matèria.orgànica	2.16	48.800	55.600	53.9300	61.0400	72.9800
Matèria.orgànica.resit	0.00	18.100	22.260	23.0300	27.4600	69.3700
Nitrogen.Amoniacal	0.00	0.4032	0.5936	0.6292	0.7784	3.7128
Nitrogen.Orgànic	0.00	3.2560	4.2240	4.1320	5.0780	10.4370
Fòsfor..P2O5.	0.00	6.0860	7.6800	8.2680	9.8110	68.5900
Potassi..K2O.	0.09	0.6720	0.9984	1.1027	1.3824	17.8944
Calci..CaO.	0.18	7.8408	11.853	13.5751	17.0016	90.5256
Magnesi..MgO.	0.11	1.8144	2.3544	2.5952	3.0528	16.8768
Ferro	0.00	3.4760	6.1940	9.2880	11.6290	81.5280
Mercuri	0.26	0.2300	0.7300	1.1890	1.6800	203.000
Cadmi	0.00	0.6200	1.0000	1.8170	2.0000	133.500
Níquel	0.50	17.200	24.000	42.1400	35.6000	977.000
Plom	0	31.600	47.500	64.6600	75.7000	990.300
Crom	0.39	19.100	30.000	66.5600	54.0000	999.000
Coure	1.00	184.00	267.00	309.450	395.000	999.500
Zinc	0.00	319.00	501.10	469.000	665.000	999.900
Data	22-03-1994	07-11-2002	08-10-2008	10-04-2009	04-10-2016	27-17-2022

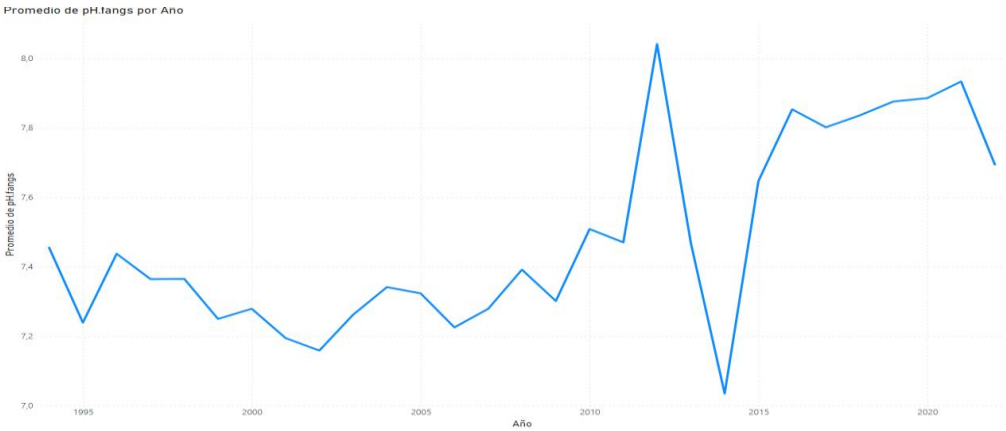
Taula 7.13 Taula descriptiva Numèriques

7.3. Descriptiva simple gràfica

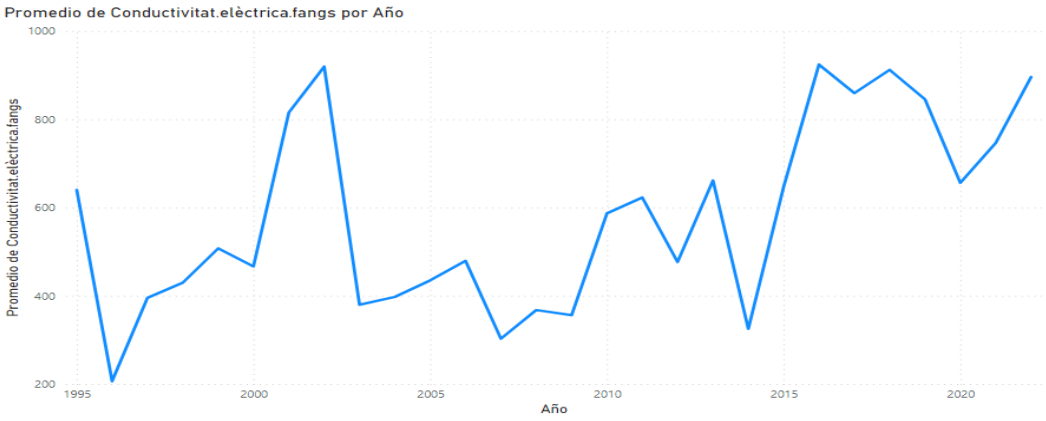
En aquest punt es podran visualitzar diferents indicadors de manera visual. Per tal d'evitar que el treball s'excedeixi de pàgines es posaran un o un parell d'exemples de cada visualització. La resta d'informació i visualitzacions es poden trobar a l'informe de power bi que s'adjuntarà amb el treball.

7.3.1 Gràfics de línies

Els gràfics de línies permeten observar l'evolució temporal de la variables que es volen estudiar. En el 1r gràfic s'observa l'evolució del ph del fang de les depuradores, de manera simple es pot determinar que sí hi ha hagut un augment del ph en els últims anys. Això pot ser degut a l'augment en les concentracions de cert elements o per alguna altra causa que es discutirà més endavant. A la dreta es pot visualitzar la conductivitat elèctrica, en aquest cas sí que s'observa que hi ha un clar augment.



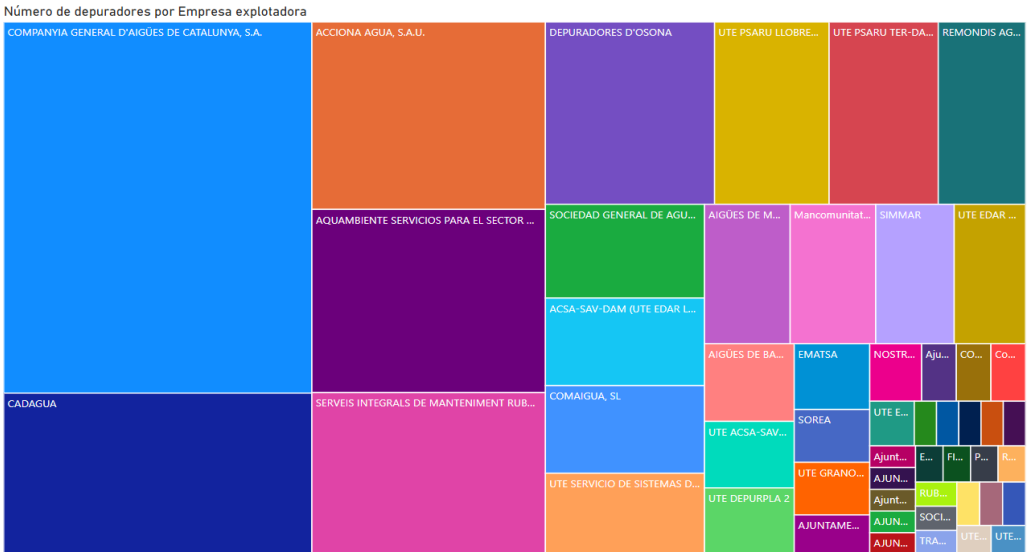
Gràfic 7.3 Gràfics de línies ph



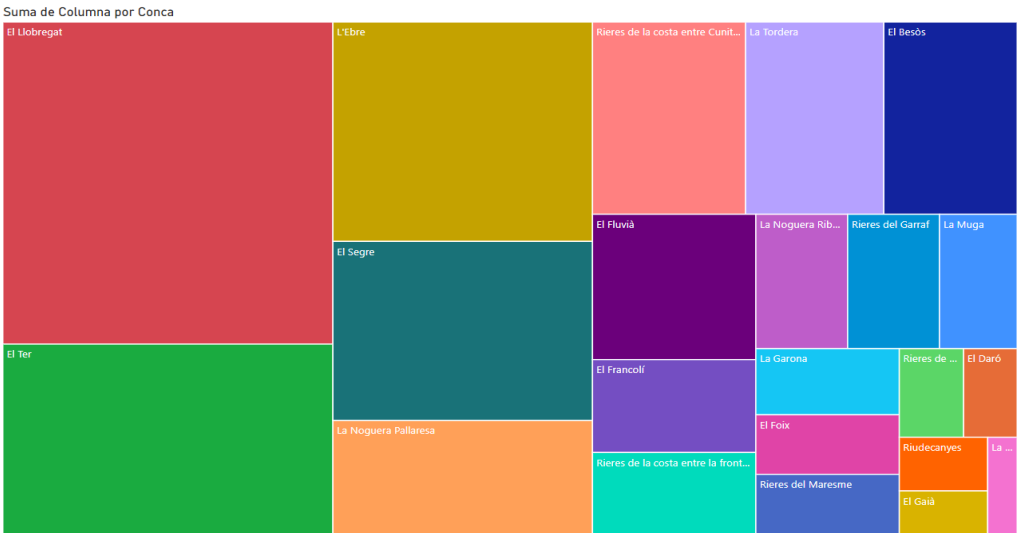
Gràfic 7.3 Gràfics de línies Conductivitat elèctrica

7.3.2 Gràfic de treemap

En un mapa d'arbre, les dades es representen mitjançant rectangles que es subdivideixen en subrectangles més petits. Cada rectangle del treemap representa una categoria o element, i la mida del rectangle reflecteix la mida de la mètrica particular associada a aquesta categoria o element. Això us permet comparar visualment les proporcions relatives de diferents elements en funció d'aquesta mesura. En el primer gràfic s'observa el número de depuradores assignada a cada empresa i en el segon gràfic, el número d'inspeccions que s'han realitzat en cada conca.



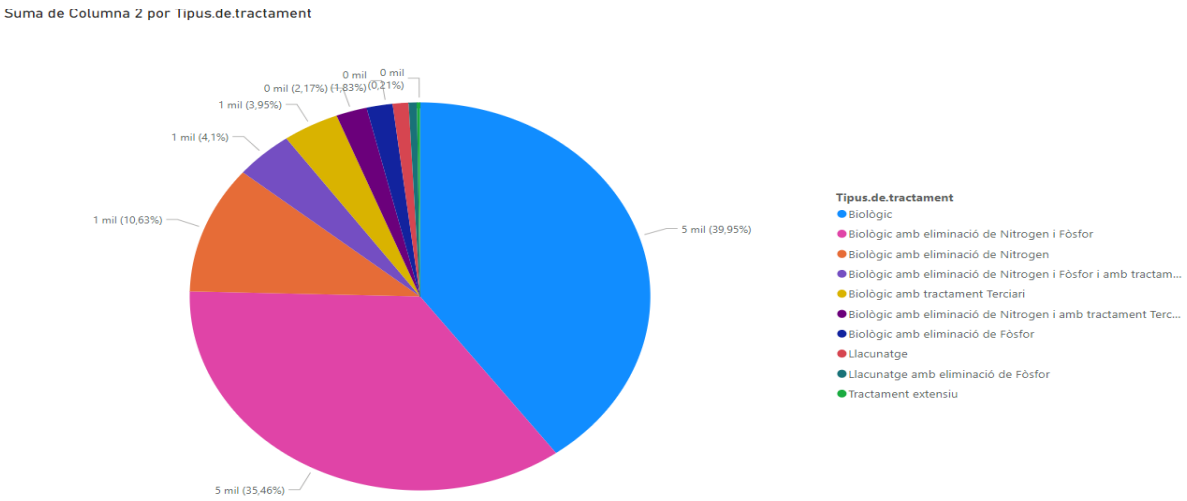
Gràfic 7.4 Gràfic de treemap Empresa explotadora



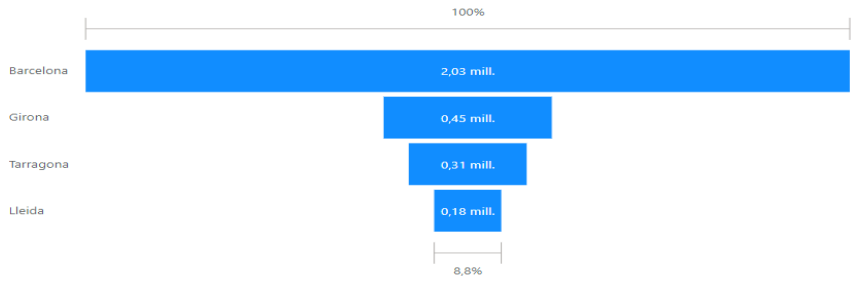
Gràfic 7.5 Gràfic de treemap Conca

7.3.3 Diagrama de pastís i d'embut

Un gràfic circular és una representació gràfica circular que divideix les dades proporcionalment en seccions que mostren la composició relativa de les categories. Aquests desglossaments reflecteixen la mida de cada categoria en relació amb el total. D'altra banda, un gràfic d'embut té una forma trapezoïdal o de piràmide invertida que destaca l'embut de categories. Ambdós efectes visuals ajuden a il·lustrar clarament la distribució i la relació entre les dades. En el primer gràfic podem observar del total de 13.000 inspeccions amb quin tipus de tractament estava associat en cada informe. En el segon gràfic veiem quina és la distribució del disseny del cabal per dia per cada una de les quatre províncies de Catalunya.



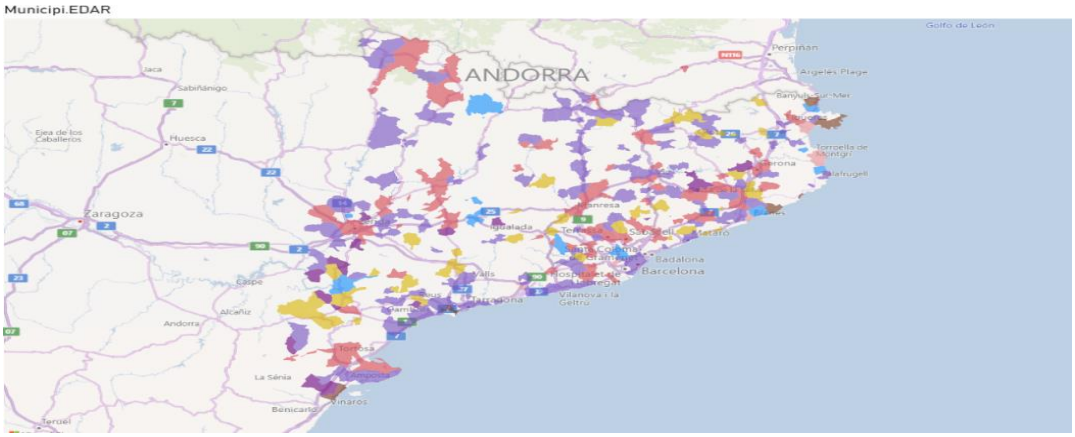
Gràfic 7.6 Diagrama de pastís Tipus de tractament



Gràfic 7.7 Diagrama d'embut Provincia

7.3.4 Mapes

Els mapes són una eina molt útil a l'hora de visualitzar i entendre dades a Power BI. En el context de Power BI, es poden utilitzar mapes per representar informació geogràfica relacionada. En el primer mapa es pot veure quin mètode de tractament utilitza cada depuradora. Mitjançant diferents codis o colors, es pot identificar els diferents tipus de tractament aplicats a una zona concreta de la planta de tractament. Aquesta visualització permet comprendre de manera ràpida i eficient les tècniques de tractament més habituals o prevalents per a cada localització. D'altra banda, a la segona imatge, s'observa el disseny del procés de cada planta de tractament. A través de la representació gràfica al mapa, s'aprecia visualment com canvia el cabal de la depuradora en diferents indrets. Es poden utilitzar diferents mides o colors per indicar la mida del trànsit, cosa que permet identificar de manera visual i intuïtiva zones amb molt i baix trànsit.



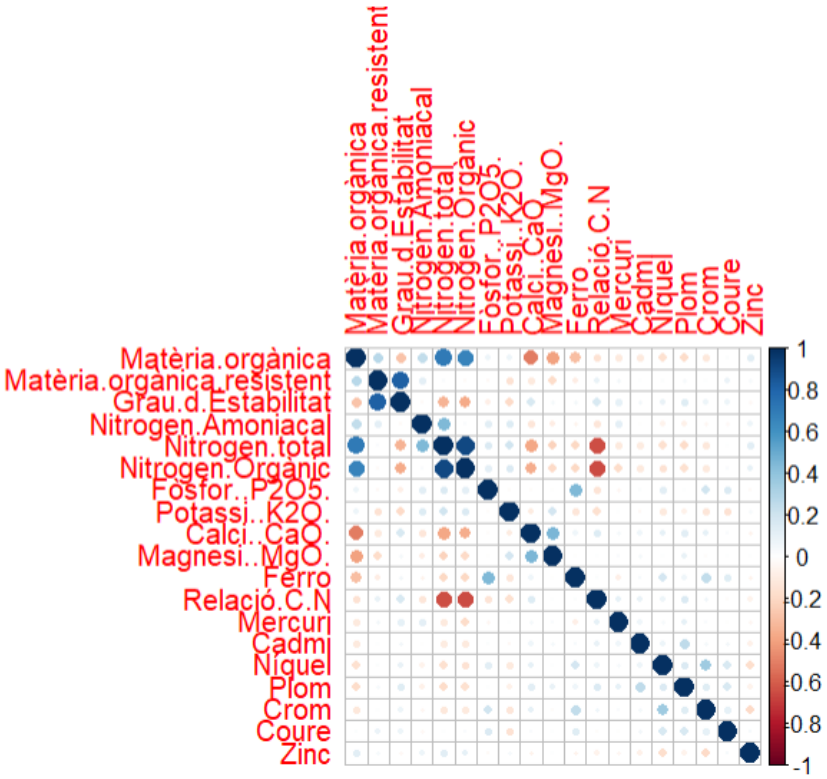
Gràfic 7.8 Mapa tipus de tractament

Gràfic 7.9 Mapa tipus de tractament

8. Tendències de les concentracions dels elements

8.1. Independència entre els diferents indicadors

Abans d'endinsar-nos en el món de la modelització, es comprovarà com es relacionen els diferents indicadors (variables resposta) entre sí. Una possible correlació entre aquestes variables podria suposar un problema a l'hora d'analitzar de manera independent cada un dels diferents elements. Per poder observar possibles correlacions es seleccionarà de la base de dades imputades les variables numèriques i es realitzarà un test de correlacions. Per tal de fer-ho senzill a l'hora d'interpretar els resultats numèrics del test, a continuació es mostra una taula de correlacions que permetrà visualitzar si existeix o no alguna relació de dependència entre les diferents variables.



Gràfic 8.10 Mapa de Correlacions

Es pot observar que, en general, les dades són independents, però es troben un parell de casos en el que es podria pensar que existeix alguna relació. Si existís alguna relació s'hauria de tenir en compte a l' hora de treure conclusions als següents anàlisis. Els punts de la diagonal no s'han de tenir en compte ja que estableix una correlació entre una mateixa variable i aquesta sempre serà 1.

8.2. Tendències pel que fa als diferents elements químics i a les depuradores (regressió simple)

Iniciant l'apartat de modelització, es realitzarà una exploració per identificar els punts de Catalunya on s'han observat augment o disminució significativa en les concentracions dels elements estudiats. Cada element serà analitzat de manera independent, ja que són variables independents i no hi ha correlació entre elles. S'aplicarà una regressió simple per a cada depuradora i cada element, relacionant les seves concentracions amb el temps, sense tenir en compte altres factors. D'aquesta manera, s'obtindrà un conjunt de models de regressió simples.

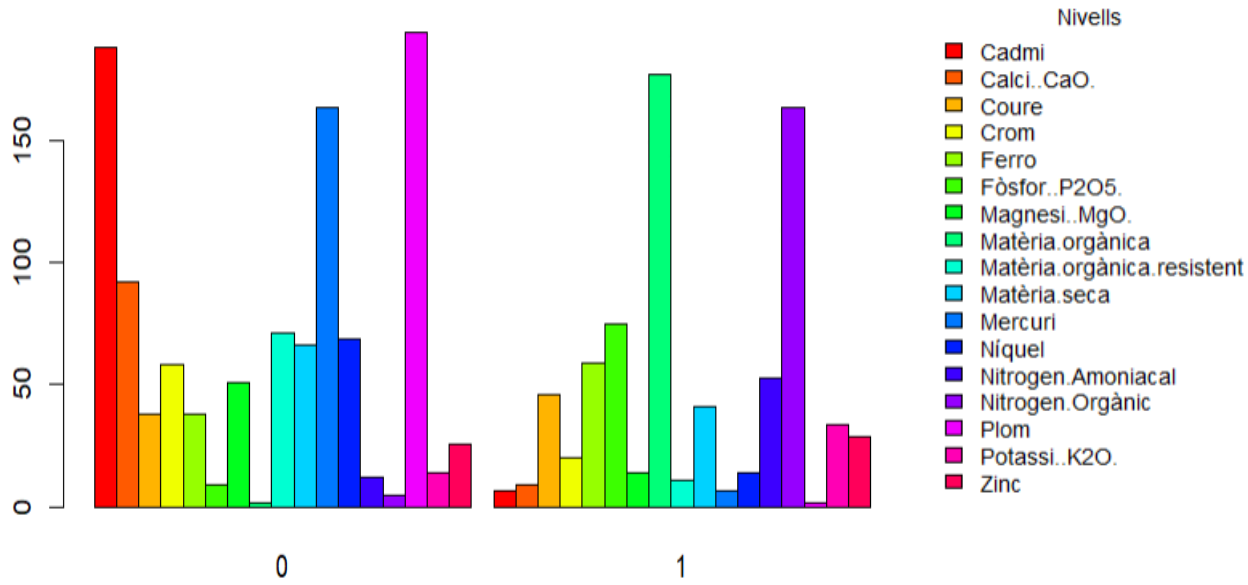
Cal destacar que l'objectiu d'aquesta anàlisi és estudiar les dades de manera independent, centrant-se en cada depuradora i cada element individualment. S'han generat un total de 6426 models lineals diferents, dels quals 1857 han mostrat un valor p inferior al nivell de significació establert. No obstant això, és complex identificar la ubicació exacta d'aquests punts significatius i comprendre els canvis observats en la tendència dels elements. Per facilitar la comprensió i l'organització d'aquesta informació, s'ha creat una nova base de dades que classifica i ordena els elements segons la depuradora, proporcionant informació rellevant sobre el tractament i la ubicació d'aquests elements. A continuació, s'inclou una imatge que il·lustra aquesta nova base de dades perquè el lector pugui obtenir una visió general de les dades analitzades.

V1	V2	V3	V4	V5	V6
DABR	Matèria.seca	Les concentracions d'aquest element han disminuit	Biològic amb eliminació de Nitrogen i Fòsfor	Agència Catalana de l'Aigua	El Llobregat
DABR	Matèria.orgànica	Les Concentracions d'aquest element han augmentat	Biològic amb eliminació de Nitrogen i Fòsfor	Agència Catalana de l'Aigua	El Llobregat
DABR	Nitrogen.Amoniacal	Les Concentracions d'aquest element han augmentat	Biològic amb eliminació de Nitrogen i Fòsfor	Agència Catalana de l'Aigua	El Llobregat
DABR	Nitrogen.Orgànic	Les Concentracions d'aquest element han augmentat	Biològic amb eliminació de Nitrogen i Fòsfor	Agència Catalana de l'Aigua	El Llobregat
DABR	Fòsfor..P2O5.	Les Concentracions d'aquest element han augmentat	Biològic amb eliminació de Nitrogen i Fòsfor	Agència Catalana de l'Aigua	El Llobregat
DABR	Calci..CaO.	Les Concentracions d'aquest element han augmentat	Biològic amb eliminació de Nitrogen i Fòsfor	Agència Catalana de l'Aigua	El Llobregat
DABR	Magnesi..MgO.	Les Concentracions d'aquest element han augmentat	Biològic amb eliminació de Nitrogen i Fòsfor	Agència Catalana de l'Aigua	El Llobregat
DABR	Mercuri	Les concentracions d'aquest element han disminuit	Biològic amb eliminació de Nitrogen i Fòsfor	Agència Catalana de l'Aigua	El Llobregat
DABR	Plom	Les Concentracions d'aquest element han augmentat	Biològic amb eliminació de Nitrogen i Fòsfor	Agència Catalana de l'Aigua	El Llobregat
DAFA	Matèria.orgànica	Les Concentracions d'aquest element han augmentat	Biològic amb eliminació de Nitrogen i Fòsfor	Agència Catalana de l'Aigua	El Llobregat
DAFA	Matèria.orgànica.resistent	Les concentracions d'aquest element han disminuit	Biològic amb eliminació de Nitrogen i Fòsfor	Agència Catalana de l'Aigua	El Llobregat

Imatge 8.6 Models lineals significatius

Tot i que molta informació gràfica es podrà observar en el power bi, en els dos següents subapartats s'analitzarà, en funció dels elements i dels tipus de tractament, si les concentracions en els punts on s'ha detectat un canvi han augmentat o disminuït, per tal de determinar si aquestes variables tenen alguna relació pel que fa al tipus de canvi de tendència.

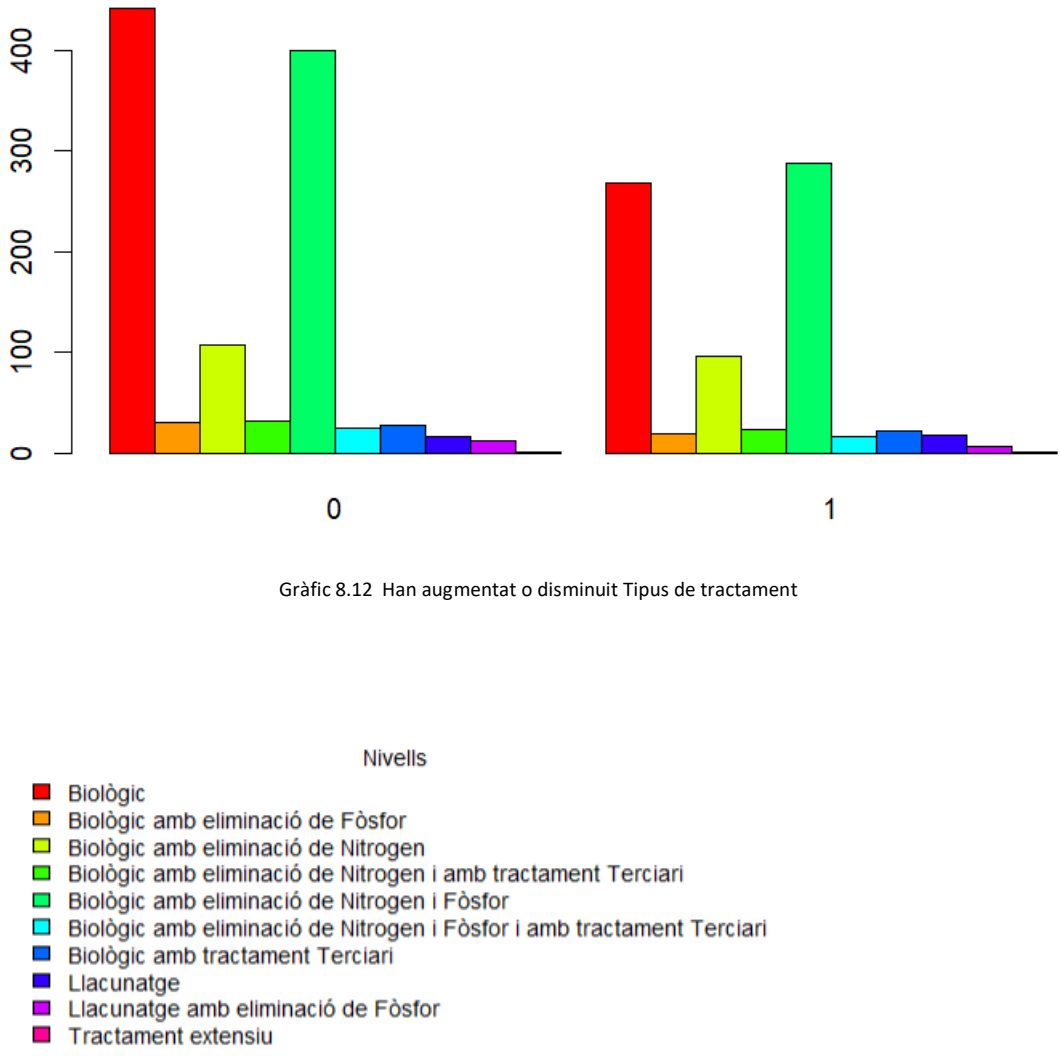
8.2.1 Canvis de tendència pels diferents elements.



Gràfic 8.11 Han augmentat o disminuit Elements

En general, sí es pot observar que hi ha diferències pel que fa a si ha disminuït o augmentat en funció de l'element estudiat. Com a exemple, s'aprecia que el cadmi, en gairebé tots els casos, ha patit una reducció pel que fa a les concentracions, mentre que en el cas de la matèria orgànica les concentracions han augmentat.

8.2.2 Canvis de tendència pels diferents tractaments



Gràfic 8.12 Han augmentat o disminuit Tipus de tractament

En relació als tipus de tractament, no és possible determinar de manera clara si aquests tenen una influència significativa en els canvis de concentració. Tot i que es poden observar algunes diferències en termes totals (hi ha més disminucions que augments), si considerem les proporcions (hi ha més models amb disminució de concentracions en comparació amb els models amb augment), es pot concloure que no hi ha diferències clares entre els diferents tractaments.

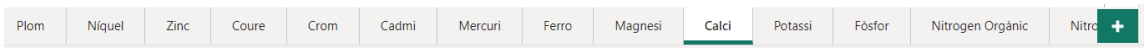
8.3. Anàlisi de la tendència pel que fa als diferents elements químics amb elements gràfics (Calci-DABR)

Una vegada s'ha recopilat la informació sobre els canvis de tendència a la nova base de dades, el següent pas consisteix a realitzar un anàlisi visual més detallat. Cal tenir en compte que hi ha més de 1800 models per analitzar, i seria pràcticament impossible fer-ho mitjançant eines gràfiques a l'R. En aquest sentit, el Power Bi es presenta com una eina molt útil, ja que, gràcies als seus filtres interactius, es pot accedir a la informació rellevant sense generar imatges innecessàries. A continuació, s'exposa un exemple d'interpretació d'un resultat de la taula anterior:

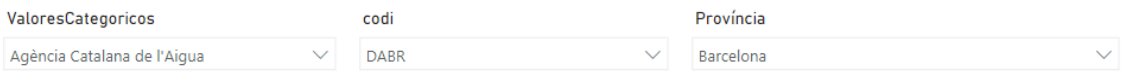
1- Es selecciona la dada que interessaria estudiar. En aquest cas s'ha seleccionat una dada que estudia el Calci. Es veu que el codi de la depuradora és DABR, que està gestionat per l'Agència Catalana de l'Aigua, es pot localitzar a la conca de Llobregat i que les concentracions han augmentat.

DABR	Calci..CaO.	Les Concentracions d'aquest element han augmentat	Biològic amb eliminació de Nitrogen i Fòsfor	Agència Catalana de l'Aigua	El Llobregat

2- Es selecciona la pàgina de l'informe que estudia l'element seleccionat en aquest cas és el Calci.

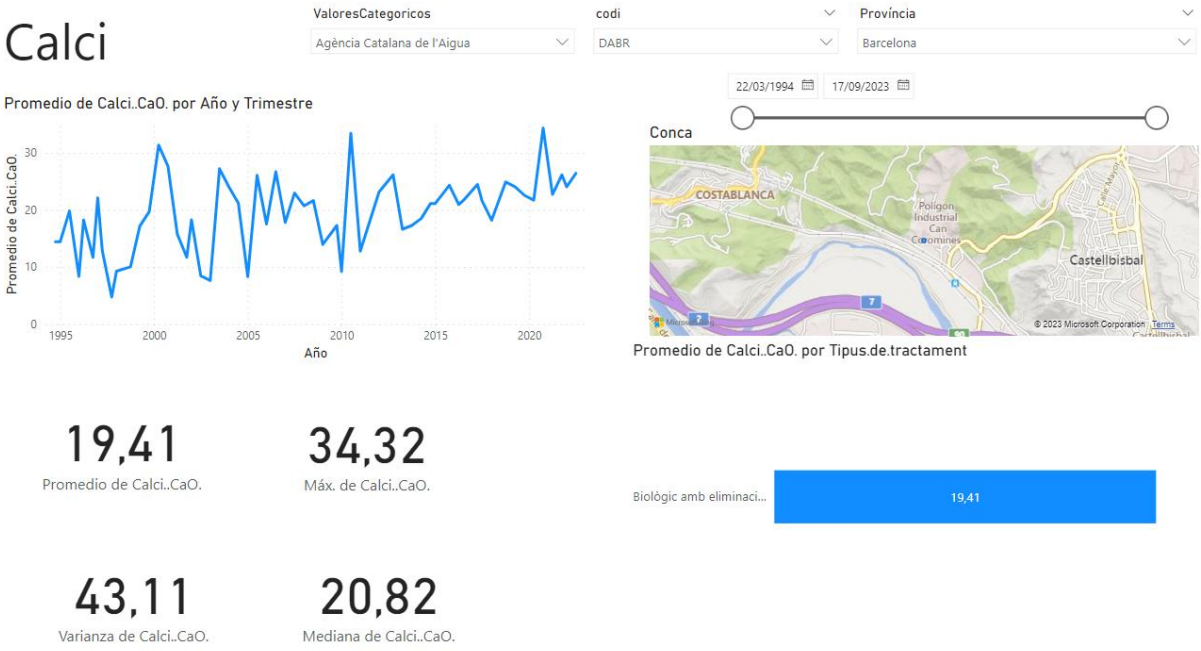


3- Apliquem les característiques del punt seleccionat als filtres de la pàgina de tal manera que l'informe utilitzi només les dades de la consulta:



Una vegada implementada aquesta sèrie de passos es podrà visualitzar el resultat de la nostra consulta. Al final power bi utilitza consultes sql per tal de modificar i representar les dades.

Sobre les visualitzacions que es poden observar en aquest informe tenint en compte que l'objectiu d'aquest apartat era exclusivament analitzar les tendències de les concentracions més avall es troba a l'esquerra un gràfic que representa les concentracions mitjanes al llarg dels anys des del 1995 fins al 2020. En aquest cas, el mapa d'adalt a l'esquerra ens permet localitzar la conca, també podem visualitzar altres indicadors com el valor màxim i mínim, en la visualització de la data podem ajustar la barra per capturar el període que es desitgi estudiar.



Imatge 8.7 Power BI Calci

Com ja s'ha mencionat anteriorment, es pot localitzar un canvi pel que fa a la tendència de les concentracions de Calci al llarg dels anys. Si es volgués estudiar algun altre punt en concret s'haurien de repetir els passos descrits anteriorment.

8.4. Prediccions dels canvis de tendència.

A partir de la funció de l'apartat 8.2, s'ha generat una nova base de dades que conté la informació de les característiques principals de més de 6000 models lineals generats en el primer apartat. S'ha afegit una nova variable que indica si hi ha hagut algun canvi en la tendència dels models, utilitzant el valor 0 per indicar l'absència de canvis i el valor 1 per indicar la presència de canvis. Aquesta variable s'ha seleccionat com a variable resposta i s'ha decidit utilitzar un model de regressió logística, un glm i un model de svm per modelar-la. Un avantatge d'adoptar aquest enfocament és que, utilitzant aquesta base de dades, s'ha reduït la dimensionalitat de les variables resposta de 15 a una única variable que representa l'element estudiat en cada model. Per tant, només cal plantejar un únic model de regressió per a les prediccions, en lloc de generar 15 models diferents.

V1	Tipus de tractement	comarca	Administració.actuant	V5
Matèria.seca	Biològic amb eliminació de Nitrogen i Fòsfor	Baix Llobregat	Agència Catalana de l'Aigua	1
Matèria.orgànica	Biològic amb eliminació de Nitrogen i Fòsfor	Baix Llobregat	Agència Catalana de l'Aigua	1
Matèria.orgànica.resistent	Biològic amb eliminació de Nitrogen i Fòsfor	Baix Llobregat	Agència Catalana de l'Aigua	0
Nitrogen.Amoniacal	Biològic amb eliminació de Nitrogen i Fòsfor	Baix Llobregat	Agència Catalana de l'Aigua	1
Nitrogen.Orgànic	Biològic amb eliminació de Nitrogen i Fòsfor	Baix Llobregat	Agència Catalana de l'Aigua	1
Fòsfor..P2O5.	Biològic amb eliminació de Nitrogen i Fòsfor	Baix Llobregat	Agència Catalana de l'Aigua	1

Imatge 8.8 Totes les regressions

Per començar, les dades s'han separat en conjunts d'entrenament (train) i prova (test) amb l'objectiu de poder entrenar el model i avaluar-ne la qualitat de les prediccions. L'objectiu d'aquest procés és determinar si hi ha hagut algun canvi en la tendència, tenint en compte les característiques proporcionades.

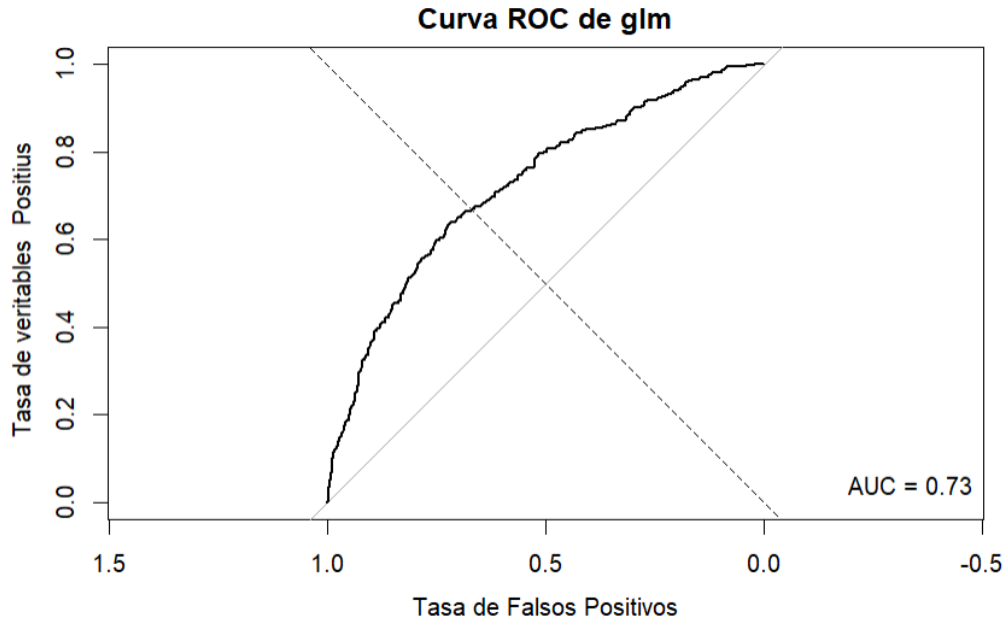
El model de regressió plantejat és el següent:

$$CanvisDeTendència \sim B_0 + B_1 \cdot \text{Element Químic} + B_2 \cdot \text{Tipus de tractament} + B_3 \cdot \text{Comarca} + B_4 \cdot \text{Administració actuant}$$

8.5. Models de regressió logística (glm-svm)

Una vegada s'ha determinat el model, s'aplica la funció del model de regressió logística per obtenir els resultats i avaluar-lo. El model té diversos coeficients, concretament 103, els quals representen les seves característiques. Per avaluar la seva capacitat de predicció o classificació, s'utilitza l'àrea sota la corba de ROC.

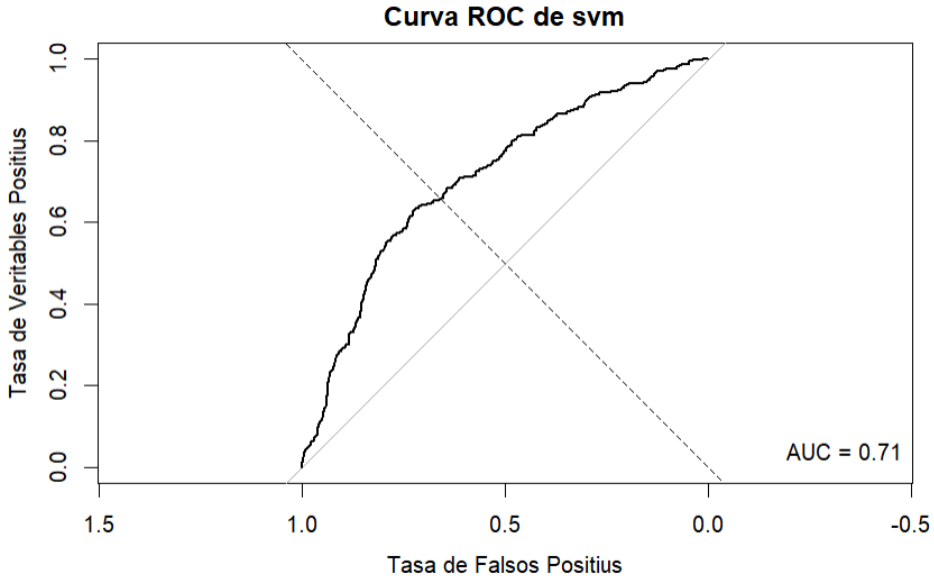
L'àrea sota la corba ROC és una mètrica amplament utilitzada per avaluar models de classificació. Mesura la capacitat discriminativa d'un model per distingir entre dues classes. Per realitzar aquest anàlisi, es realitzen prediccions amb el model plantejat anteriorment. Les prediccions no són valors 0 i 1 com la variable resposta, sinó que proporcionen la probabilitat de que hi hagi hagut un canvi en la tendència. S'estableix un llindar de 0,5, on les probabilitats menors a 0,5 es classifiquen com a 0 i les probabilitats majors es classifiquen com a 1. A partir de les dades predites i classificades correctament, es realitza l'anàlisi de la corba de ROC per als dos models.



Gràfic 8.13 auc glm

Es pot observar que l'AUC és de 0,73, el que indica una capacitat de discriminació moderada del model. L'Àrea sota la corba ROC (AUC) és una mètrica que avalua la capacitat del model per

distingir entre classes, sent un valor de 1 indicatiu d'un rendiment perfecte i un valor de 0,5 indicatiu d'un rendiment aleatori. En aquest cas, el model presenta una capacitat de predicció acceptable, ja que és capaç de classificar correctament el 73% de les dades.



Gràfic 8.14 auc svm

En aquest cas l'AUC està en 0,71 i es pot dir que té menor capacitat discriminant. També s'hauria d'avaluar la sensibilitat i la especificitat , dels models per poder realitzar un anàlisi més complert.

8.6. Especificitat i sensibilitat

La sensibilitat i l'especificitat són dues mètriques importants per avaluar els models de classificació.

La sensibilitat, també coneguda com a taxa de verdaders positius, mesura la capacitat del model per identificar correctament els casos positius. És la proporció de casos positius que el model classifica correctament com a positius. D'altra banda, l'especificitat, la veritable taxa negativa, mesura la capacitat del model per identificar correctament els casos negatius. És la proporció de casos negatius que el model classifica correctament com a negatius.

$$especificitat = \frac{\text{veritables negatius}}{\text{veritables negatius} + \text{falsos positius}}$$

$$sensibilitat = \frac{\text{veritables positius}}{\text{veritables positius} + \text{falsos negatius}}$$

8.6.1. Especificitat i sensibilitat -glm

$$especificitat = \frac{797}{797 + 66} = 0{,}92$$

$$sensibilitat = \frac{113}{113 + 264} = 0{,}3$$

En aquest estudi, només s'analitzarà el model glm segons el criteri de classificació establert en 0,5. S'observa que el model presenta una especificitat molt alta, indicant que té una gran precisió en la detecció dels elements que no han experimentat canvis en la tendència. Això implica que el model ofereix resultats excel·lents en la identificació d'aquests elements.

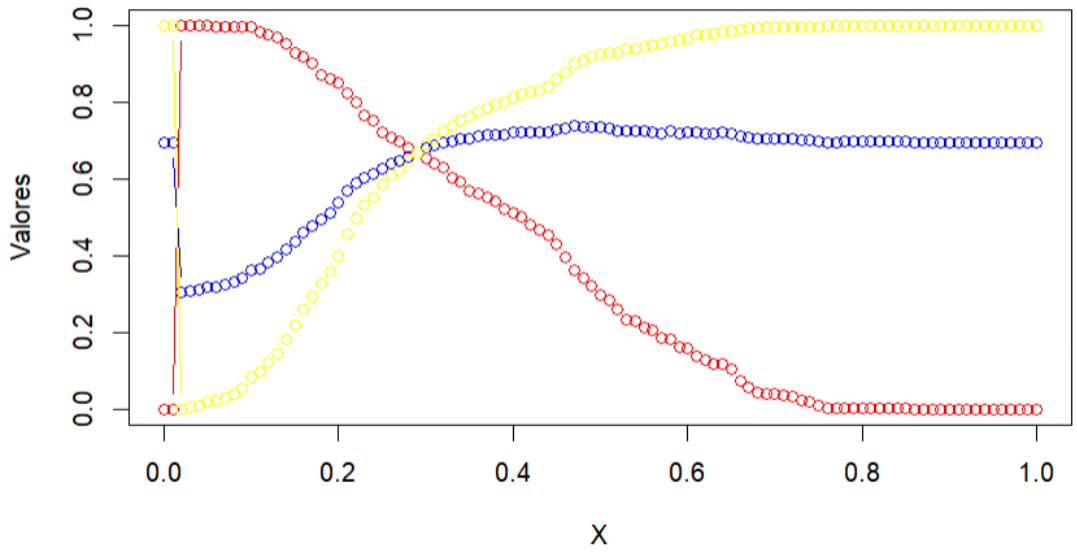
No obstant això, en el cas de la sensibilitat, s'observa un valor molt baix, concretament 0,32. Això indica que el model no té una capacitat adequada per detectar canvis de tendència en els elements. Aquesta situació pot ser problemàtica, ja que hi ha una proporció significativa d'elements amb canvis que el model no és capaç d'identificar.

8.6.2. Es maximitza Auc Especificitat i sensibilitat

En aquest apartat, l'objectiu és generar un vector de valors de 0 a 1 per determinar el llindar de separació òptim. Això permetrà avaluar l'àrea sota la corba ROC (AUC), l'especificitat i la sensibilitat per a cada valor del llindar, amb l'objectiu de determinar el valor òptim a tenir en compte com a criteri de separació.

El criteri de separació és el valor que, si és superat pel valor predit, se li assigna un 1, mentre que si no és superat, se li assigna un 0. Per avaluar aquests paràmetres, es presenta un gràfic que mostra la relació entre els diferents valors del llindar i els tres paràmetres mencionats.

A través d'aquest gràfic, es pot determinar quin valor del llindar ofereix el millor equilibri entre l'àrea sota la corba ROC, l'especificitat i la sensibilitat. Això permetrà seleccionar el valor òptim del llindar com a criteri de separació.



Gràfic 8.15 auc vs sensitivitat vs especificitat

Finalment es pot concloure que si es vol obtenir tenir un nivell de sensibilitat i especificitat iguals s'ha d'establir un llindar entorn al 0,3. Això si, si el lector volgues una sensitivitat mes elevada o per lo contrari una major especificitat haurà d'ajustar un nou llindar per a les dades.

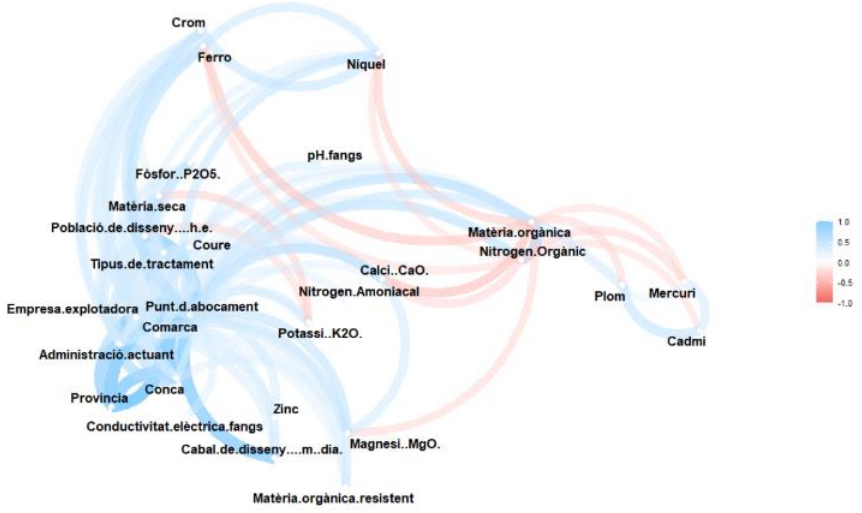
9. Modelització centrada en una variable resposta (ph)

En l'etapa final de l'informe, es centra en l'anàlisi i la investigació d'un indicador clau: el valor del pH. Com s'ha mencionat anteriorment, el pH és una mesura que indica el nivell d'acidesa o alcalinitat d'una substància, i en aquest cas, és la variable de resposta en estudi. L'objectiu principal és construir diversos models estadístics utilitzant les variables explicatives per predir i comprendre com afecten el pH dels fangs residuals. Això permetrà identificar els factors clau que influeixen en el pH. Es procedirà a construir diferents models estadístics utilitzant tècniques com la regressió lineal, els arbres de decisió i la Super Vector Machine Radial. Es faran servir mètodes estadístics per ajustar els models al conjunt de dades i avaluar-ne la capacitat predictiva. Un cop els models estiguin ajustats, es realitzarà la validació mitjançant bootstrap i s'avaluaran diverses mètriques com l'error quadràtic mitjà i el coeficient de determinació R^2 per a les prediccions.

Això ajudarà a comprendre quin model s'ajusta millor a les dades i té una millor capacitat predictiva per al pH dels fangs residuals.

9.1. Correlacions entre les variables

De la mateixa manera que s'ha realitzat en l'apartat 8.1 es vol estudiar les correlacions entre les diferents variables del model i observar si aquestes són independents o pel contrari estan relacionades. En el cas que hi hagi molta correlació entre algunes variables s'haurà de reduir la base de dades excloent les variables innecessàries que no aportin informació respecte al valor del ph, que és l'indicador que estudiarem en aquest apartat. Per realitzar aquest anàlisis s'ha generat un flowmap amb les correlacions que es presenta a continuació:.



Gràfic 9.16 Flow map Correlacions

Es detecta una alta correlació entre diverses variables, la qual cosa indica que generar un model de regressió podria resultar lent i ineficient a causa d'aquesta alta correlació entre les variables predictores. Per tant, el següent pas consistirà en reduir la dimensionalitat de les dades, seleccionant només aquelles variables que aportin informació útil per al model. Això permetrà simplificar el model i millorar-ne l'eficiència i interpretació.

9.2 Reducció de la dimensionalitat de la base de dades

Per reduir la dimensionalitat de la base de dades, s'ha utilitzat el següent mètode. En primer lloc, s'ha exclòs la variable objectiu de la base de dades i s'ha creat una nova base de dades amb les variables restants. A continuació, s'ha definit la variable objectiu (pH dels fangs). S'ha creat un model nul inicial mitjançant regressió lineal. S'ha iniciat un procés iteratiu d'afegir variables al model una per una. Per a cada variable candidata, s'ha ajustat un model nou i s'ha calculat una puntuació per avaluar la seva contribució al model. S'ha seleccionat la variable amb la puntuació més alta i s'ha afegit al conjunt de variables seleccionades.

S'ha repetit aquest procés fins a assolir el nombre màxim desitjat de variables seleccionades. Finalment, s'ha fet un resum dels models seleccionats per avaluar el seu rendiment. Per determinar el nombre màxim de variables a incloure en el model, s'ha creat un vector amb diferents valors (5, 10, 15, 20, 25) que representen el nombre màxim de variables requerides. A continuació, s'ha executat el procés iteratiu descrit anteriorment per a cada valor d'aquest vector. D'aquesta manera, s'han ajustat diversos models amb diferents

nombres de variables i s'ha avaluat el rendiment de cada model. S'ha decidit fixar-se en el coeficient de determinació (R^2):

Numero Variables	5	10	15	20	25.	29
Rsquared	0.4113	0.4215	0.4242	0.4246	0.4246	0.4247

Taula 9.14 rsquared dels models de 5,10,…,29 variables

Per tant, es pot determinar que s'aconsegueix la mateixa capacitat explicativa amb només 5 variables en comparació amb 30. Això és molt positiu, ja que en reduir la dimensionalitat s'exclouen les variables que no aporten informació, es millora el rendiment dels models i s'eviten possibles col·linealitats entre les variables.

El model final plantejat és el següent:

$$pH_{fangs} \sim B_0 + B_1 \cdot \text{Data} + B_2 \cdot \text{Calci} + B_3 \cdot \text{Matèria Orgànica} + B_4 \cdot \text{Coure} + B_5 \cdot \text{Codi}$$

9.3. Test i Prove

Per avaluar el rendiment d'un model predictiu, és fonamental separar les dades disponibles en dos grups: el grup de prova i el grup d'entrenament. Aquesta separació ens permet utilitzar el grup d'entrenament per ajustar els models, és a dir, trobar els millors coeficients i paràmetres. Un cop entrenats, podem fer prediccions amb aquests models utilitzant el grup de prova, que consisteix en dades que no s'han utilitzat durant l'entrenament. Això ens proporciona una mesura objectiva de com els models generalitzen i es comporten amb noves dades, ja que no han estat "vistes" durant l'entrenament. És important mantenir aquests dos grups separats per assegurar-nos que els resultats de les prediccions siguin fiables i reflecteixin el rendiment real dels models.

Grups	Test	Prove
n	10516	2626

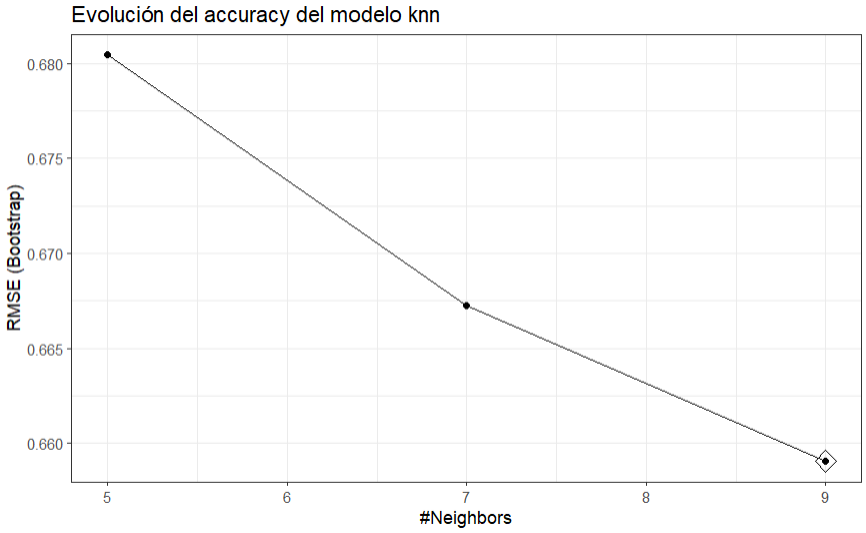
Taula 9.15 Test i prove

9.4. Ajustem els diferents models.

9.4.1 Knn (Model 1)

La regressió k-NN (k-Nearest Neighbors) és un mètode utilitzat en l'anàlisi de regressió per predir valors continus basats en dades dels seus k veïns més propers. L'algoritme intenta trobar els k veïns més propers de punts de dades desconeguts i utilitzar les seves etiquetes de classe per predir el valor de la variable objectiu. En la regressió k-NN, el valor de k és un hiperparàmetre important que determina el nombre de veïns considerats durant la predicció. Un cop trobats els k veïns més propers, el mètode pot utilitzar la mitjana ponderada dels seus valors de resposta per arribar a una predicció final. Aquest mètode és especialment útil quan hi ha una relació no lineal entre les variables independents i objectiu.

9.4.1.1. Característiques del model.



Gràfic 9.17 Accuracy knn

k	RMSE	Rsquared	MAE
5	0.6805	0.1338	0.4742
7	0.6673	0.1404	0.4658
9	0.6590	0.1468	0.4603

Taula 9.16 Característiques del model knn

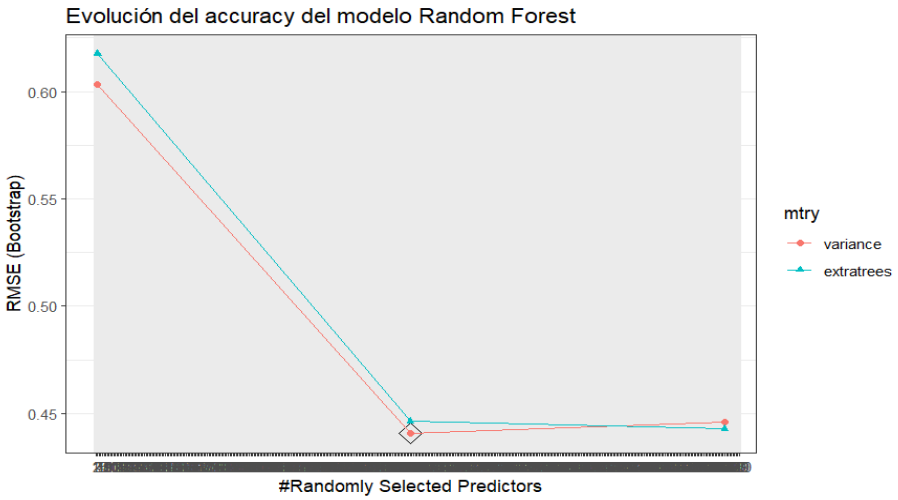
Els resultats dels models KNN que utilitzen 5, 7 i 9 veïns mostren que la precisió i l'ajust del model milloren gradualment a mesura que augmenta el nombre de veïns considerats. Quan k és igual a 5, s'observa una distribució relativament alta d'errors, una capacitat limitada per explicar la variabilitat i una precisió moderada de les prediccions. Quan k era igual a 7, es va observar una lleugera millora en la reducció d'errors i una major potència explicativa, mentre que quan k era igual a 9, es va observar una reducció més de l'error i una major precisió en les prediccions.

9.4.2. Random forest (Model 2)

En aquest estudi, s'utilitzarà un mètode d'aprenentatge automàtic conegut com a Random Forest per avaluar el pH dels fangs residuals. Un conjunt d'arbres de decisió múltiples, anomenat Random Forest, serà utilitzat per prendre decisions basades en múltiples condicions o característiques. Aquest mètode combina les prediccions de diversos arbres de decisió per obtenir una predicció final més precisa.

En aquest cas, es consideren dues estratègies per construir els arbres de decisió: "extratrees" i "variance". L'estratègia "extratrees" selecciona aleatòriament les funcions i realitza divisions en cada arbre per augmentar l'atzar i la diversitat. En canvi, l'estratègia "variance" utilitza totes les funcions disponibles en cada iteració. Aquesta elecció pot afectar el rendiment i la robustesa del model. Els arbres "extratrees" solen ser més robusts a l'excés d'ajustament, mentre que la "variance" pot obtenir millors resultats en problemes amb característiques importants.

9.4.2.1. Característiques del model.



Gràfic 9.18 Accuracy rf

mtry	min.node.size	splitrule	RMSE	Rsquared	MAE
2	5	variance	0.6054	0.3218	0.4759
2	5	extratrees	0.6184	0.2767	0.4872
191	5	variance	0.4667	0.5022	0.3430
191	5	extratrees	0.4640	0.5086	0.3368
381	5	variance	0.4701	0.4982	0.3443
381	5	extratrees	0.4636	0.5112	0.3353

Taula 9.17 Característiques del model rf

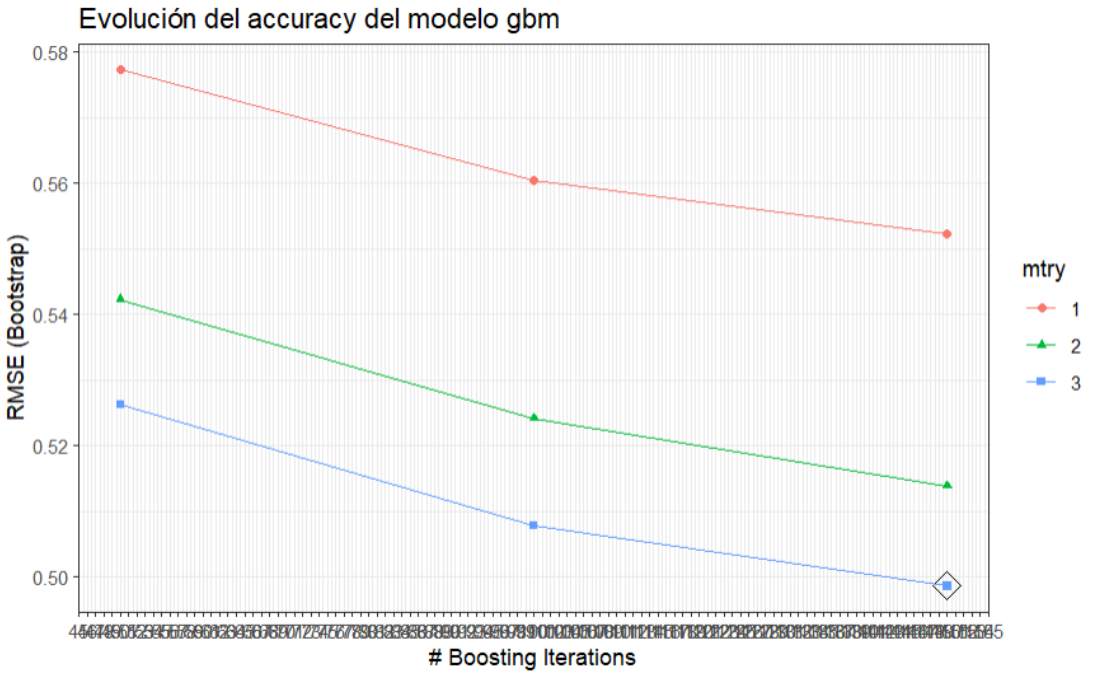
Els resultats de l'algoritme Random Forest mostra diferents combinacions de paràmetres utilitzats en el model. Quan mtry (numero de predictors) és igual a 2 i min.node.size és igual a 5, s'utilitzen dues estratègies per dividir arbres: "variance" i "extratrees". Es pot veure que el model amb "variància" té un RMSE de 0,6054, Rsquared de 0,3218 i MAE de 0,4759, mentre que el model amb "arbres addicionals" té un RMSE de 0,6184, Rsquared de 0,2767 i MAE de 0,4872. Quan mtry és igual a 191 i min.node.size és igual a 5, repetint el mateix patró, el model de "variància" funciona millor en comparació amb "arbre addicional" (RMSE) (RMSE = 0,4667, Rsquared = 0,5022, MAE = 0,3430) = 0,4640, R quadrat = 0,5086, MAE = 0,3368). Finalment, quan mtry és igual a 381 i min.node.size és igual a 5, el model de "variància" (RMSE = 0,4701, Rsquared = 0,4982, MAE = 0,3443) continua superant el model d'"arbre addicional" (RMSE = 0,4636 = 0,5112 Rsquared) , MAE = 0,3353) en termes de rendiment.

9.4.3. Gradient Boosting Machine (Model 3)

Gradient Boosting Machine (GBM) és un algoritme de Machine Learning supervisat per a problemes de regressió i classificació. Es basa en el concepte de combinar models febles per crear models predictius forts.

GBM utilitza una estratègia de construcció iterativa on cada model posterior s'ajusta als residus del model anterior. Això permet que el GBM millori gradualment les seves prediccions amb cada iteració. A cada pas s'afegeix un nou model feble, intentant corregir els errors dels models anteriors, donant més pes als casos mal classificats

9.4.3.1. Característiques del model.



Gràfic 9.19 Accuracy gbm

shrinkage	interaction.depth	n.minobsinnode	n.trees	RMSE	Rsquared	MAE	RMSESD	RsquaredSD
0.1	1	10	50	0.5773377	0.2964522	0.4442066	0.006199403	0.02858854
0.1	2	10	50	0.5422663	0.3737729	0.4205477	0.005663486	0.03425022
0.1	3	10	50	0.5262849	0.4081909	0.4077854	0.005466529	0.03239074
0.1	1	10	100	0.5603438	0.3238934	0.4308041	0.005689163	0.02923894
0.1	2	10	100	0.5241290	0.4081963	0.4051497	0.006952277	0.03513711

shrinkage	interaction.depth	n.minobsinnode	n.trees	RMSE	Rsquared	MAE	RMSESD	RsquaredSD
0.1	3	10	100	0.5077861	0.4423397	0.3918482	0.00609629	0.0303238
0.1	1	10	150	0.5523129	0.3390806	0.4237523	0.00574688	0.0282864
0.1	2	10	150	0.5138650	0.4282222	0.3963413	0.00755 0115	0.0343 0885
0.1	3	10	150	0.4985884	0.4603581	0.3830804	0.00466 4713	0.0261 0130

Taula 9.18 Característiques del model gbm

Els resultats del model Gradient Boost mostrats il·lustren el rendiment del model sota diferents combinacions de paràmetres. Es pot observar que a mesura que el valor de "interaction.depth" augmenta d'1 a 3, els resultats milloren en termes de "RMSE", "Rsquared" i "MAE". Això suggereix que permetre interaccions més complexes entre variables pot millorar l'ajust del model a les dades. D'altra banda, a mesura que el nombre de "n.trees" augmenta de 50 a 150, les mètriques d'avaluació també milloren, la qual cosa indica que més arbres contribueixen a un millor ajust del model. Pel que fa al paràmetre Shrinkage, el valor 0,1 utilitzat en tots els casos es pot considerar baix. Aquest valor indica que els residus disminueixen lentament i poden proporcionar progressivament millors ajustos al model. En resum, els resultats mostren que l'ús de valors "interaction.depth" i "n.trees" més grans i una "contracció" més baixa pot conduir a un millor rendiment del model Gradient Boost en termes de precisió i ajust als dades.

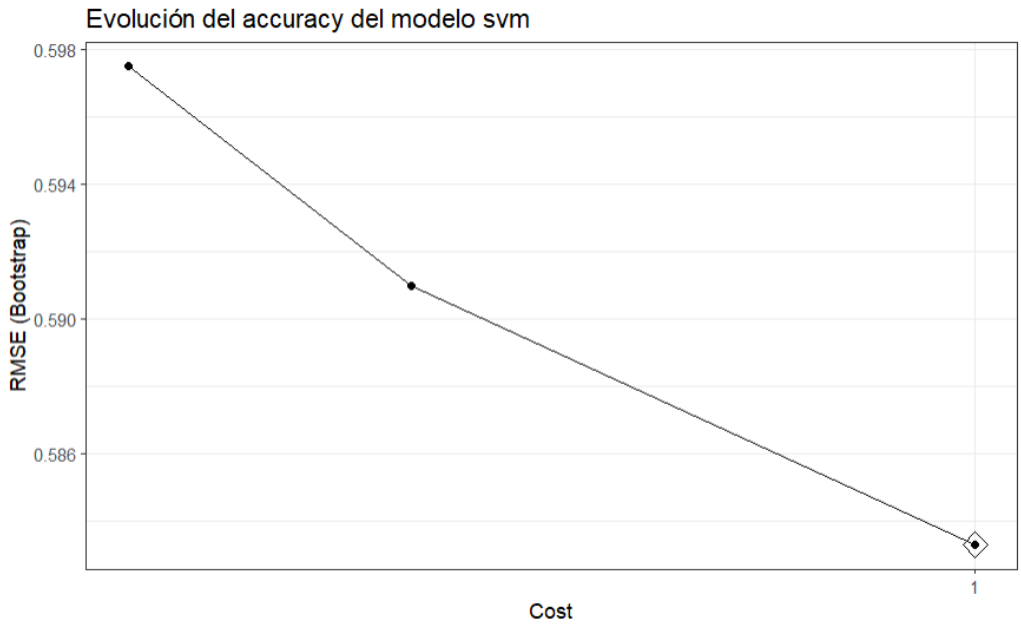
9.4.4. Super Vector Machine (model 4)

El model radial SVM (Support Vector Machine) és un algoritme de classificació i regressió basat en la creació d'hiperplans que separen diferents grups o valors predits en diferents punts. Es diu "radial" perquè utilitza una funció de base radial (RBF) com a funció de nucli per calcular la distància entre les mostres i els vectors de suport.

En aquest exemple, les dades es representen com un punt en l'àrea de visualització, i l'objectiu és trobar un hiperpla que permeti distingir clarament els diferents grups o valors de regressió.

L'algoritme SVM radial pot tenir en compte dades no lineals mentre transforma les dades a un ordre superior utilitzant un nucli radial.

9.4.4.1. Característiques del model.



Gràfic 9.20 Accuracy svm

sigma	C	RMSE	Rsquared	MAE
1.390527e-06	0.25	0.6256036	0.1883511	0.4299957
1.390527e-06	0.50	0.6191010	0.2037089	0.4245266
1.390527e-06	1.00	0.6109024	0.2227415	0.4185305

Taula 9.19 Característiques del model svm

Els resultats del model SVM radial mostren el rendiment del model per a diferents combinacions de paràmetres "sigma" i "C". És important destacar que el valor de "sigma" representa l'amplada radial de la funció de base radial utilitzada pel model SVM. Els resultats

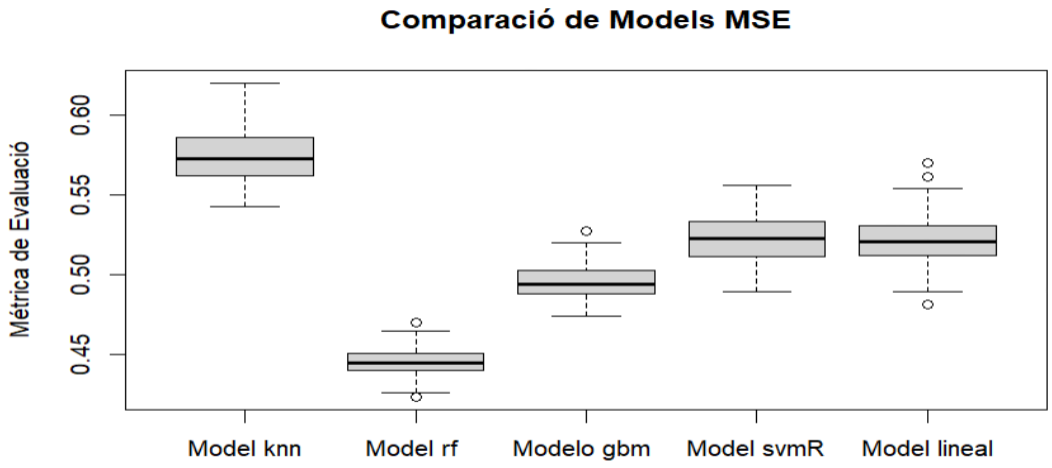
mostren que a mesura que el valor de "C" augmenten de 0.25 a 1 , els indicadors d'avaluació com "RMSE", "Rsquared" i "MAE" milloren gradualment. Això suggereix que valors més grans de "C" poden fer que el model s'ajusti més, mentre que valors més petits de "C" ajuden a evitar aquest fenomen i a obtenir un millor ajust del model a les dades. Pel que fa al paràmetre "sigma", tots els casos mostren resultats similars amb valors molt petits al voltant de 1.390527e-06. Això mostra que l'amplada radial utilitzada a les funcions de base radial no té cap efecte significatiu en les mètriques d'avaluació del model.

9.5. Comparació de models (bootstrap)

Una vegada generat els diferents models, es va realitzar una comparació d'aquests en 100 mostres mitjançant la tècnica bootstrap. Un mètode bootstrap és una eina estadística que permet estimar un paràmetre d'interès mitjançant el remostreig de les dades disponibles. En aquest cas, s'ha aplicat el bootstrap per estimar diferents paràmetres que avaluen la qualitat de les prediccions del model. Els paràmetres utilitzats són l'error quadrat mitjà (MSE), l'error mitjà absolut (MAE) i el coeficient de determinació (R quadrat).

9.5.1. Mse

MSE és una mesura per avaluar la precisió d'un model calculant la mitjana de l'error quadrat mitjà entre els valors predits i reals. Com més baix sigui el MSE, millor s'ajusta el model a les dades observades.

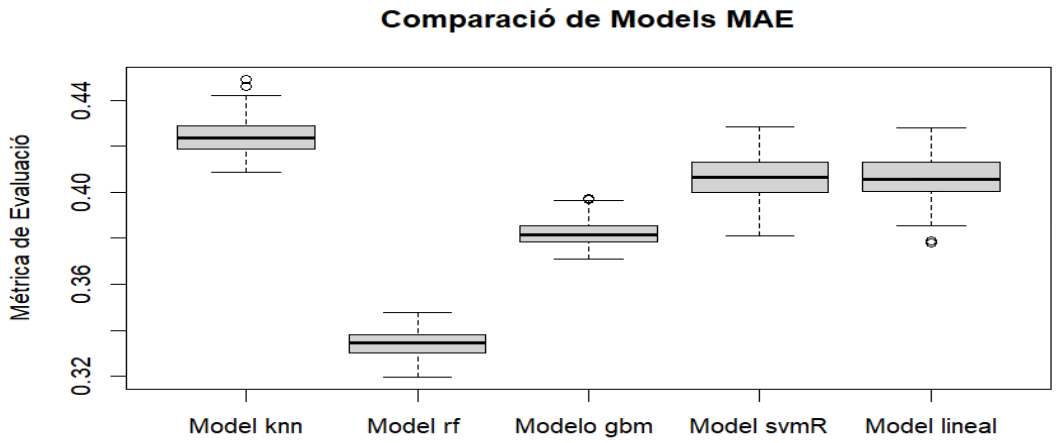


Gràfic 9.21 Comparació Mse

Veiem que pel que fa a la estimació del MSE el Random Forest és el model amb un error quadràtic mitja menor. Pel que fa als models gbm, svmR i lineal no podem determinar que hi hagi masses diferencies mentre que sí que es pot observar que el model knn és el pitjor de tots.

9.5.2. MAE

El MAE, en canvi, és una mesura de l'error absolut mitjà entre els valors predits i els reals. Igual que MSE, MAE també busca minimitzar per obtenir prediccions més precises.

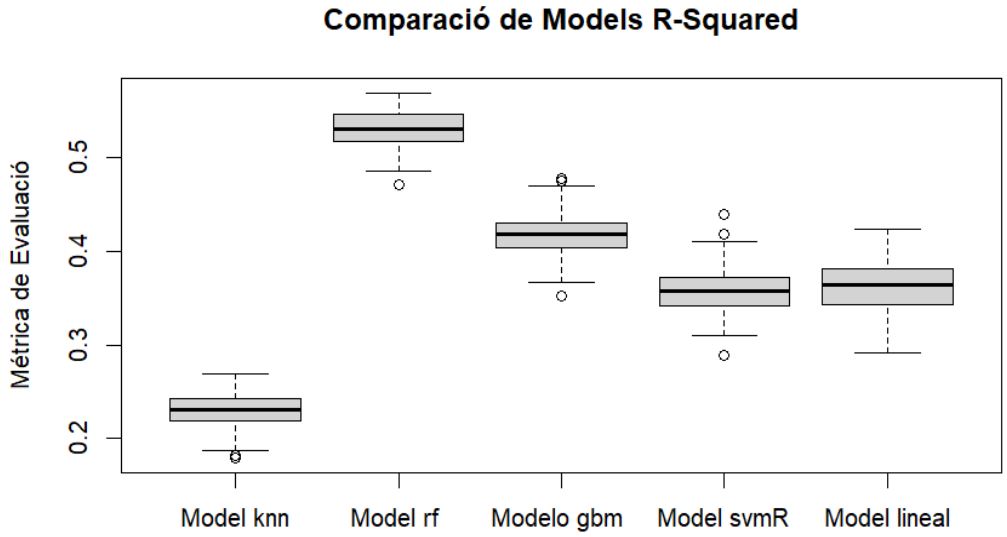


Gràfic 9.22 Comparació MAE

En aquest cas es torna a observar com el model rf té un MAE menor seguit pels tres models mencionats anteriorment (gbm, svmR i lineal). S'observa clarament com en aquest cas el model knn torna a ser el que dona prediccions més imprecises.

9.5.3. Rsquared

El coeficient de determinació, o R-quadrat, proporciona una mesura de la proporció de la variabilitat en la variable de resposta explicada pel model. Un valor R-quadrat més proper a 1 indica un millor ajust del model a les dades observades



Gràfic 9.23 Comparació R-squared

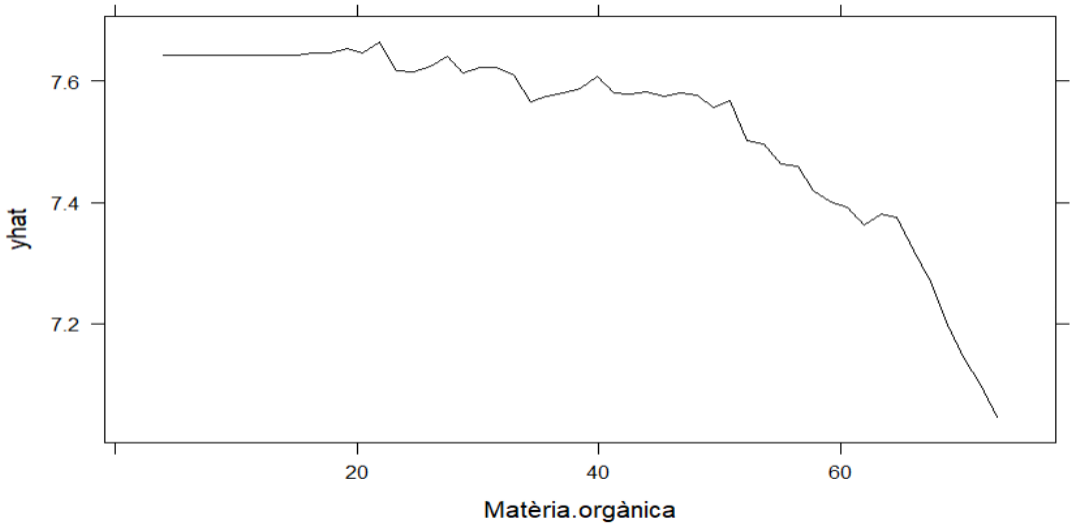
Seguint amb la línia dels dos apartats anteriors novament s'aprecia que el model amb una millor capacitat explicativa es la del rf i gradient boosting. Mentre que a la resta de models tenen un nivell baix pel que fa a la capacitat explicativa de la variabilitat del PH

9.6. Model final

Un cop analitzats els resultats anteriors es pot afirmar que el model que ajusta millor les dades és el model de regressio random forest ja que és el que té millor capacitat explicativa i, un mse i un mae menor. En conclusió, és el model que realitza les millors prediccions comparat amb la resta de models estudiats.

Es vol estudiar l'efecte del pH sobre les variables predictores, concretament sobre els nivells de calci, coure i matèria orgànica. Per realitzar aquesta anàlisi, s'utilitza un gràfic de dependència parcial, una eina que ens permet veure com afecta el pH a aquestes variables predictores i, alhora, com afecten a la variable resposta. Mitjançant aquests gràfics, podrem visualitzar com canvien els nivells de calci, coure i matèria orgànica amb el pH mantenint constants els altres predictors. Això ens ajudarà a entendre millor la relació entre el pH i les variables predictores i identificar tendències o patrons que poden ser importants per a la variable de resposta que s'està estudiant.

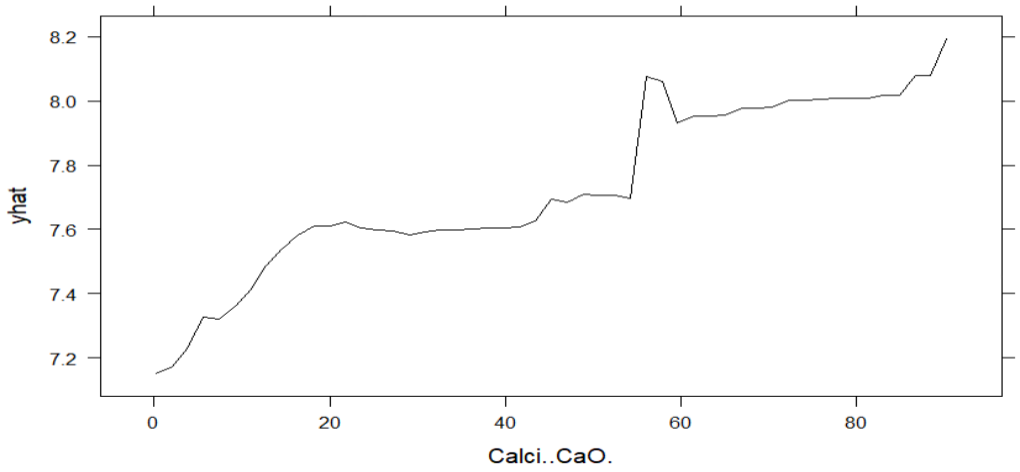
9.6.1. Matèria orgànica



Gràfic 9.24 ph vs Matèria orgànica

Com es pot observar en el gràfic a mesura que el nivell de matèria orgànica augmenta el ph disminueix o dit d'una altre manera si el valor del ph augmenta les concentracions de matèria orgànica disminueixen. Això es deu a que un augment de l'alcalinitat del fang provoca que els microorganismes responsables de la descomposició de la matèria orgànica es puguin tornar menys eficients o fins i tot inactius, la qual cosa porta a una disminució de les concentracions de matèria orgànica.

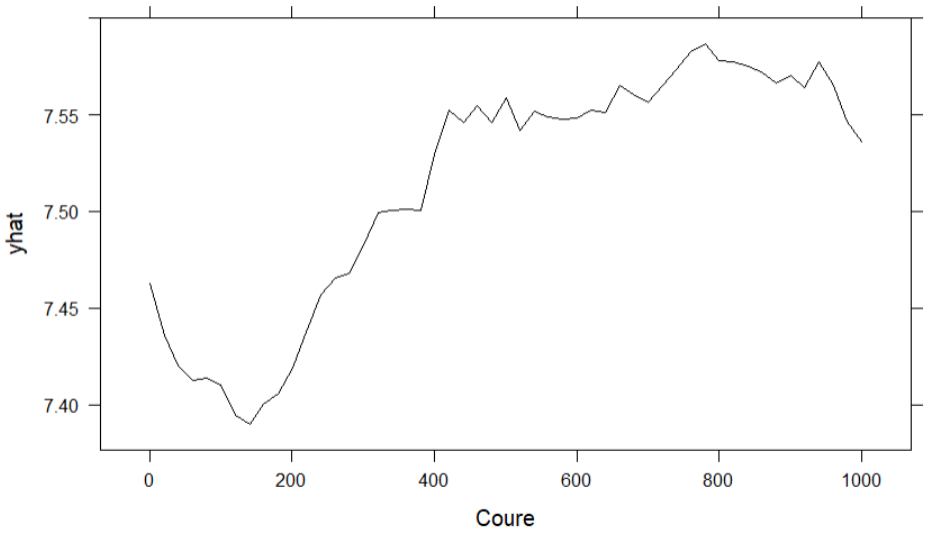
9.6.2. Calci



Gràfic 9.25 ph vs Calci

En aquest cas, s'observa que si augmenta la concentració de calci, també augmenta el pH del fang. Això es deu a la naturalesa alcalina del calci, ja que un augment important d'aquest element fa que augmenti el pH del fang. El calci reacciona amb els àcids presents al medi, neutralitzant-los i augmentant així el pH. La seva interacció amb els components àcids del fang augmenta gradualment el pH a mesura que augmenta la concentració de calci.

9.6.3. Coure



Gràfic 9.25 ph vs Coure

La relació entre el contingut de coure i el pH dels llots de depuradora no és lineal. Això vol dir que no hi ha una relació directa i constant entre la concentració de coure i el pH. Com es pot veure a la gràfica del contingut de coure en funció del pH, en alguns casos el pH pot disminuir primer i després augmentar a mesura que augmenta la concentració de coure. Això es degut a que el coure no té cap influència real sobre el ph o almenys de manera directa ja que el coure pot afectar a les concentracions de matèria orgànica i, per tant, sí que podria provocar un canvi pel que fa al nivell d'aquest.

10. Conclusions

S'han assolit tots els objectius establerts inicialment en l'estudi. En primer lloc, s'ha realitzat una exhaustiva exploració i tractament de les dades. Les dades s'han netejat, s'han abordat les mancances d'informació mitjançant knn i especificat i transformat correctament.

Pel que fa a la segona fase de treball, s'ha dut a terme un anàlisi univariant i bivariant de les dades utilitzant l'eina de Microsoft Power BI. Aquesta anàlisi ha proporcionat una comprensió més profunda del conjunt de dades i ha permès crear un informe molt dinàmic amb possibles aplicacions reals.

A més, s'ha posat l'èmfasi en la modelització, especialment en la investigació de possibles canvis de tendència en les concentracions dels diferents elements. Amb l'ajuda de Power BI, s'han localitzat aquests punts de canvi i s'han analitzat detalladament. Pel que fa a la predicció d'aquets canvis de tendència s'ha pogut observar que els models que s'han plantejat per classificar les dades han assolit un nivell satisfactori pel que fa a la capacitat discriminant.

Pel que fa la tercera fase de l'estudi, s'ha avaluat una variable d'interès rellevant el ph i s'ha pogut reduir la dimensionalitat de les dades de 29 variables explicatives a 5. L'últim fet ha ajudat a millorar el rendiment dels models de regressió que s'han aplicat per tal de construir un model de predicció. S'ha pogut demostrar que el model de regressió que millor ajusta les dades treballades és el Random forest. Per concloure aquest bloc s'ha pogut estudiar la relació de manera individual dels predictors en funció del ph.

Resumidament es pot concloure que tot i que en la base de dades es troba molta informació per poder plantejar altres estudis (com per exemple una caracterització dels fangs mitjançant clusterings, estudiar els microorganismes com la salmonel·la o centrar-nos en altres indicadors), ja que la complexitat de la base de dades permet realitzar una infinitat d'hipòtesis pel que fa aquest camp , s'ha pogut realitzar una anàlisi complerta i diversa enfocada des de diferents camps de la estadística. Això ha permès estudiar i entendre diverses qüestions relacionades amb els fangs de les depuradores de Catalunya.

11. Bibliografia i estudis similars

Estudis similars :

de Medio Ambiente, M., & Marino, Y. M. R. (s/f). *Caracterización de los lodos de depuradoras generados en España*. Cedex.es. Recuperado el 29 de junio de 2023, de https://hispagua.cedex.es/sites/default/files/hispagua_documento/lodos_depuradoras.pdf

Articles:

Gencat.cat. Recuperado el 29 de junio de 2023, de https://aca.gencat.cat/web/.content/10_ACA/J_Publicacions/02-publicacions/01_Fulleto_sanejament_Catalunya.pdf

Ministerio para la Transición Ecológica y el Reto Demográfico. (s.f.). Manual de la Directiva 91/271/CEE relativa al tratamiento de las aguas residuales urbanas. Recuperado de https://www.miteco.gob.es/es/agua/publicaciones/03_Manual_Directiva_91_271_CEE_tcm30-214069.pdf

Pàgines web:

Laboratori. (s/f). Agència Catalana de l'Aigua. Recuperado el 29 de junio de 2023, de https://aca.gencat.cat/ca/laigua/seguiment-i-control/laboratori/index.html

Estacions depuradores d'aigua residual. (s/f). Agència Catalana de l'Aigua. Recuperado el 29 de junio de 2023, de https://aca.gencat.cat/ca/laigua/infraestructures/estacions-depuradores-daigua-residual/index.htm

Estacions depuradores d'aigua residual. (s/f). Agència Catalana de l'Aigua. Recuperado el 29 de junio de 2023, de https://aca.gencat.cat/ca/laigua/infraestructures/estacions-depuradores-daigua-residual/index.html

RPubs - SVM Máquinas de Vector Soporte (Support Vector Machines). (s/f). Rpubs.com. Recuperado el 29 de junio de 2023, de https://rpubs.com/Joaquin_AR/267926

Rodrigo, J. A. (s/f). *Machine Learning con R y caret*. Amazonaws.com. Recuperado el 29 de junio de 2023, de https://rstudio-pubs-static.s3.amazonaws.com/383283_6cc2a3cc3e524e2bbf8b509dd68e8a92.html

12. Annexes

En els annexes trobareu el codi utilitzat durant l'estudi. Si tinguéssiu algun dubte, trobareu anàlisis que s'han anat descartant per la seva falta de rellevància o difícil interpretació, com el pca o clustering.

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

Importem les dades

```{r , warning=FALSE}
library(readxl)
dades<-read_excel("C:\\Users\\maxpl\\OneDrive\\Documentos\\TFG\\asar2.xlsx")
```

```{r }
dades2<-
read_excel("C:\\Users\\maxpl\\OneDrive\\Documentos\\TFG\\LlistatEDAR_Servei.xlsx")
```

2. Creem la base de dades(les unim) i analitzem les dades faltants (seleccionem aquelles variables amb menys de un 25% de missings)

You can also embed plots, for example:

```{r , echo=FALSE}
library(knitr)
dades$codi<-dades$`Codi Estacio`
dades2$codi<-dades2$`CODI Sistema`

df<-merge(dades,dades2, by="codi")

missing_percentages <- colMeans(is.na(df)) * 100
table <- data.frame(Variables = names(df),
                    Missing_Percentage = missing_percentages)
(table)

variables_to_remove <- names(missing_percentages[missing_percentages > 25])
df2 <- df[, !(names(df) %in% variables_to_remove)]

```

#1 especificació de variables (variables numériques)

```{r , echo=FALSE, warning=FALSE}
df3<-df2[,5:26]
library(dplyr)
df3 <- as.data.frame(lapply(df3, function(x) gsub(",", ".", x)))
# Convertir columnas a numéricas
df3 <- as.data.frame(lapply(df3, as.numeric))
df2<-df2[,-(5:26)]
df2<-data.frame(df2,df3)

```

#Transformació de variables

```{r , echo=FALSE, warning=FALSE}
percentatges<-df2[,c(25,28,29,31,33,34,35,36,37,38)]
class(percentatges)
k<-c(25,28,29,31,33,34,35,36,37,38)
densitat<- c(1050, 1000, 1300, 700, 1100, 2400, 2400, 3300, 3600, 7900)
elements <- c("Matèria.seca", "Matèria.orgànica", "Matèria.orgànica.resistent",
              "Grau.d.Estabilitat", "Nitrogen.Amoniacal", "Nitrogen.total",
              "Nitrogen.Orgànic", "Fòsfor..P2O5.", "Potassi..K2O.", "Calci..CaO.",
              "Magnesi..MgO.", "Ferro")
densidad<-1250
for(i in 1:10){

  percentatges[,i]<-percentatges[,i]*densitat[i]/densidad
  p<-k[i]
  df2[,p]<-percentatges[,i]
}
names<-names(percentatges)
names

```

#Selecció de variables interessants

```{r , echo=FALSE, warning=FALSE}
df_definitiva<-df2[,c(1,4,8,9,10,11,12,14,15,16,23,26,27,25,28,29,31,33,34,35,36,37,38,40:46)]
```

#2na Especificació variables

```{r , echo=FALSE, warning=FALSE}
str(df_definitiva)

df_definitiva$codi<-as.factor(df_definitiva$codi)

df_definitiva$Conca<-as.factor(df_definitiva$Conca)
df_definitiva$Administració.actuant<-as.factor(df_definitiva$Administració.actuant)
df_definitiva$Comarca<-as.factor(df_definitiva$Comarca)
df_definitiva$Província<-as.factor(df_definitiva$Província)
df_definitiva$Empresa.explotadora<-as.factor(df_definitiva$Empresa.explotadora)
df_definitiva$Empresa.explotadora<-as.factor(df_definitiva$Empresa.explotadora)
df_definitiva$Punt.d.abocament<-as.factor(df_definitiva$Punt.d.abocament)
df_definitiva$Tipus.de.tractament<-as.factor(df_definitiva$Tipus.de.tractament)

df_definitiva$Data <- as.Date(df_definitiva$Data, format = "%d/%m/%Y")
df_definitiva$Població.de.disseny....h.e.<-as.numeric(df_definitiva$Població.de.disseny....h.e.)
df_definitiva$Cabal.de.disseny....m..dia.<-as.numeric(df_definitiva$Cabal.de.disseny....m..dia.)

```

#3 Implementació de missings

#implementacio de missings

```{r , echo=FALSE, warning=FALSE}

library(VIM)

datos <- df_definitiva

datos_completos <- datos[complete.cases(datos), ]

datos_faltantes <- datos[!complete.cases(datos), ]

k <- 10

iteraciones <- 10

valores_imputados_anterior <- NULL

for (i in 1:iteraciones) {

  valores_imputados <- kNN(datos_faltantes, k = k)

  if (!is.null(valores_imputados_anterior) && all.equal(valores_imputados,
valores_imputados_anterior)) {

    cat("Convergencia alcanzada en la iteración", i, "\n")
```

```
    break
  }

  valores_imputados_anterior <- valores_imputados
  datos[!complete.cases(datos), ] <- valores_imputados
}
```

Validació de la imputació

```{r , echo=FALSE}

plot(df_definitiva$Matèria.orgànica.resistent, datos$Matèria.orgànica.resistent,
     xlab = "Dades originals", ylab = "Dades imputades",
     main = "Comparació entre dades originals i imputades de la Matèria Orgànica Resistent",
     col = "blue", pch = 16)
```

Línies diagonals per a una comparació de referència

```
abline(0, 1, col = "red", lwd = 2)

par(mfrow=c(1,2))

plot(density(df_definitiva$Matèria.orgànica.resistent,na.rm = T),col=2,main="Comparació entre dades originals i imputades de la Matèria Orgànica Resistent")
lines(density(datos$Matèria.orgànica.resistent),col=3)
```

```
par(mfrow=c(1,2))

plot(density(df_definitiva$Potassi..K2O.,na.rm = T),col=2,main="Comparació entre dades originals i imputades del Potassi(K2O)")

lines(density( datos$Potassi..K2O.),col=3)
```

```
# exportem la base de dades (power bi)
```

```{r , echo=FALSE}
# install.packages("openxlsx")

library(openxlsx)

datos$municipi<-df2$`Municipi EDAR`

df<-datos

# Especificar la ruta y el nombre del archivo Excel de destino

ruta_archivo <- "C:/Users/maxpl/OneDrive/Documentos/TFG/dades4.xlsx"

wb <- createWorkbook()

# Agregar el data.frame al libro como una hoja de cálculo

addWorksheet(wb, "Hoja1")

writeData(wb, sheet = "Hoja1", x = df)

# Guardar el libro de Excel en el archivo especificado

saveWorkbook(wb, file = ruta_archivo, overwrite = TRUE)
```

Part de la anàlisis simple es realitza al power bi aquí nomes mirarem si les variables numèriques estan relacionades

```{r , echo=FALSE}

library(knitr)

resumen <- summary(datos)

tabla_resumen <- as.data.frame(resumen)

kable(tabla_resumen)

tabla_elegante <- kable(tabla_resumen , caption = "Resumen de estadísticas descriptivas")

tabla_elegante

resumen <- summary(datos[,c(9,10,12:30)])

resumen

mean_val <- resumen[["Mean"]]

median_val <- resumen[["Median"]]

max_val <- resumen[["Max"]]

min_val <- resumen[["Min"]]

tabla2 <- data.frame(Variable = colnames(data),
                     Mean = mean_val,
                     Median = median_val,
                     Max = max_val,
                     Min = min_val)

```

Correlació entre las variables a estudiar

```{r , echo=FALSE}
#install.packages("corrplot")
library(corrplot)
matriz_cor <- cor(datos[, c(14:30)])corrplot(matriz_cor, method = "circle")
```

#PCA no el farem servir

```{r , echo=FALSE}

library(FactoMineR)
dades<-datos[,c(9,10,12:30)]
pca <- PCA(dades)
resultats_pca <- prcomp(dades)

var_explained <- pca$sdev^2 / sum(pca$sdev^2)
num_componentes <- 5
variances <- pca$eig
loadings <- pca$rotation
scores <- pca$x
print(loadings)
print(scores)
```

```{r , echo=FALSE}
datos_pca <- predict(pca, newdata = dades, ncp = num_componentes)
datos_reducidos<-data.frame()
```

datos_reducidos <- cbind(datos_pca$coord)

plot(datos_reducidos, col = iris$Species, pch = 16, xlab = "Componente Principal 1", ylab = "Componente Principal 2")

install.packages("FactoInvestigate")

library(FactoMineR)

library(FactoInvestigate)

data(iris)

datos_reducidos <- as.data.frame(pcaindcoord)

#Clustering

```{r , echo=FALSE}

library(cluster)

library(factoextra)

fviz_nbclust(x = dades, FUNcluster = pam, method = "wss", k.max = 15,

diss = dist(dades, method = "manhattan"))

```

```{r , echo=FALSE}

help(package="mclust")

install.packages("mclust")

library(mclust)

resultado <- Mclust(datos)
```

clusters <- resultado$classification

centroides <- resultado$parameters$mean

print(clusters)

print(centroides)

```

```{r , echo=FALSE}

clusters <- resultado$classification

plot(datos$pH.fangs,datos$Conductivitat.elèctrica.fangs, col = clusters, pch = clusters, main = "Resultados de Mclust")

```

Correlació i pca entre las variables a estudiar

```{r , echo=FALSE}

#install.packages("corrplot")

library(corrplot)

matriz_cor <- cor(datos[, c(14:30)])

corrplot(matriz_cor, method = "circle")

```

```{r , echo=FALSE}

library(FactoMineR)

dades<-datos[,c(14:30)]

pca <- PCA(dades)

resultats_pca <- prcomp(dades)

var_explained <- pca$sdev^2 / sum(pca$sdev^2)

```
num_componentes <- 5

variances <- pca$eig

loadings <- pca$rotation

scores <- pca$x

print(loadings)

print(scores)


```

```{r , echo=FALSE}


contingency_table <- table(datos[,c(3)],datos[,c(10)])

str(datos[,c(3,4,5)])

chi_squared_test <- chisq.test(contingency_table)

print(chi_squared_test)
```


```{r , echo=FALSE}


datos_pca <- predict(pca, newdata = dades, ncp = num_componentes)

datos_reducidos<-data.frame()

datos_reducidos <- cbind(datos_pca$coord)


plot(datos_reducidos, col = iris$Species, pch = 16, xlab = "Component Principal 1", ylab =
"Component Principal 2")
```

`

Modelització

Anàlisi tendència elements per a cada depuradora

```{r , echo=FALSE}
datos$Data <- as.Date(datos$Data, format = "%d/%m/%Y")
t<-data.frame()
base<-data.frame()
dif<-c()
p<-0
nam<-names(datos[,c(14:30)])
count<-0
o<-length(levels(datos$codi))
names<-levels(datos$codi)
length(levels(datos$codi))
for(j in 1:378){
for (i in 14:30){
  p<-p+1

df_sin_missings <- datos
df_ni<-df_sin_missings[df_sin_missings$codi==names[j],]

if(length(df_ni$codi)>1){
modelo <- lm(df_ni[,i] ~ Data, data = df_ni)
```

```
resultado <- summary(modelo)


base[p,1]<-nam[i-13]
base[p,"Tipus de tractement"]<-df_sin_missings[p,"Tipus.de.tractament"]
base[p,"comarca"]<-df_sin_missings[p,"Comarca"]
base[p,"Administració.actuant"]<-df_sin_missings[p,"Administració.actuant"]


# Obtener el valor p
p_valor <- resultado$coefficients[2, "Pr(>|t|)"]


if(is.na(p_valor) ) p_valor<-1



pendiente <- coef(modelo)[2]
#base[p, "pendent"]<-pendiente
library(ggplot2)
if(p_valor<0.01 ){
  count<-count+1
 #print(summary(modelo))
 t[count,1]<-names[j]
 t[count,2]<-nam[i-13]
 i[]
if(pendiente>0){
 t[count,3]<-"Les Concentracions d'aquest element han augmentat"
 t[count,4]<-1
}else{
  t[count,3]<-"Les concentracions d'aquest element han disminuit"
  t[count,4]<-0
```

```
}
 base[p,5]<-1
  t[count,5]<-df_sin_missings[p,"Administració.actuant"]
   t[count,6]<-df_sin_missings[p,"Conca"]
    t[count,7]<-df_sin_missings[p,"Empresa.explotadora"]
    t[count,8]<-df_sin_missings[p,"Comarca"]
     t[count,9]<-df_sin_missings[p,"Tipus.de.tractament"]
}else{
  base[p,5]<-0
}
}
}
}
```

```
v1=codi depuradora
v2=element
v3= han augmentat o disminuit
v4=1: augment 0 : disminucio
```

 # Analisis del Tipus de tractament per a les dades que s'ha detectat que existeix tendencia

```{r , echo=FALSE}
t$V2<-as.factor(t$V2)
nam<-levels(t$V9)

t2<-t
target <- t2$V4
```

```
var1 <- t2$V2

var2 <- t2$V5

var3 <- t2$V9


barplot(table(var1, target), beside = TRUE,main = nam[i], col = rainbow(17))

legend("topleft", legend = nam , fill = rainbow(10), title = "Nivells",cex = 0.7)

#barplot(table(var2, target), beside = TRUE, legend = TRUE, main = "Variable 2")

barplot(table(var3, target), beside = TRUE, main = nam[i], col = rainbow(10))



```

```{r , echo=FALSE}

x<-table(predicciones_clases, valores_reales)

auc<-(x[1,1]+x[2,2])/(x[1,2]+x[2,1]+x[2,2]+x[1,1])

sens<-x[2,2]/(x[2,2]+x[1,2])

esp<-x[1,1]/(x[1,1]+x[2,1])

auc

x

esp

sens

x<-0

j<-seq(-1, 1, by = 0.01)

sens<-c()

auc<-c()

esp<-c()

for(i in 1:length(j)){

predicciones_clases <- ifelse(predicciones >= j[i], 1, 0)
```

```
predicciones_clases<-as.factor(predicciones_clases)

levels(predicciones_clases)<-c(0,1)

valores_reales<-as.factor(valores_reales)

x<-table(predicciones_clases, valores_reales)


auc[i]<-(x[1,1]+x[2,2])/(x[1,2]+x[2,1]+x[2,2]+x[1,1])

sens[i]<-x[2,2]/(x[2,2]+x[1,2])

esp[i]<-x[1,1]/(x[1,1]+x[2,1])

}
```

#auc

```
plot(j, auc, type = "b", col = "blue", ylim = range(c(auc, sens)), ylab = "Valores", xlab = "X")

lines(j, sens, type = "b",  col = "red")

lines(j, esp, type = "b",  col = "yellow")
```

```{r , echo=FALSE}
summary(predicciones)
```

```{r , echo=FALSE}
set.seed(1)

svm_cv <- tune("svm", V5 ~ ., data = datos_train, probability = TRUE,

            kernel = 'linear',

            ranges = list(cost = c(0.001, 0.01, 0.1, 1, 5, 10, 20, 50, 100,

                                   150, 200)))

summary(svm_cv)
```

#Anàlisi del Ph

#escollim les variables importants (no executar aquesta part)

```{r , echo=FALSE}

datos_entrenamiento <- datos[, -12]

target <- datos$pH.fangs

max_variables<-5

modelo <- lm(target ~ 1, data = datos_entrenamiento)

variables_seleccionadas <- c()

while (length(variables_seleccionadas) < max_variables) {

 mejores_puntuaciones <- c()

 for (variable in setdiff(colnames(datos_entrenamiento), variables_seleccionadas)) {

  formula <- as.formula(paste("target ~", paste(c(variables_seleccionadas, variable), collapse =
"+")))

  modelo_temporal <- lm(formula, data = datos_entrenamiento)

  mejores_puntuaciones[variable] <- summary(modelo_temporal)$r.squared

 }

 mejor_variable <- names(mejores_puntuaciones)[which.max(mejores_puntuaciones)]

 variables_seleccionadas <- c(variables_seleccionadas, mejor_variable)
```

```
  formula <- as.formula(paste("target ~", paste(variables_seleccionadas, collapse = "+")))
  modelo_r <- lm(formula, data = datos_entrenamiento)
}



resumen_modelo <- summary(modelo)
print(resumen_modelo)
```````


#Reduccio de la dimensionalitat

```{r , echo=FALSE}

datos<-datos[,c(1,2,12,15,21,29)]

y<-lm(pH.fangs~. , data=datos)
summary(y)


``

## # separem test i train

```{r , echo=FALSE}

library(ranger)


```
```

```
library(dplyr)

library(caret)



set.seed(123)

indices_train <- sample(1:nrow(datos), round(0.8 * nrow(datos)), replace = FALSE)

datos_train <- datos[indices_train, ]

datos_test <- datos[-indices_train, ]

datos_train <- na.omit(datos_train)

datos_test <- na.omit(datos_test)
```

Model d'arbre-Random forest

```
```{r , echo=FALSE}



control <- trainControl(method = "boot", number = 1)



modelo_rf <- ranger(pH.fangs ~ ., data = datos_train, importance = "impurity", num.trees=10)

#modelo_rf <- train(pH.fangs ~ ., data = datos_train, method = "ranger", trControl = control)

(modelo_rf$variable.importance)

#install.packages("randomForest")

#install.packages("pdp")

library(randomForest)

library(pdp)
```

Ajustar el model Random Forest
```

```
# Generar el gráfico de dependencia parcial

plot <- partial( pred.var = "Matèria.orgànica" ,modelo_rf , plot = TRUE)


plot


plot(datos$Data, datos$pH.fangs)
```
```

```{r,warning=FALSE}

predicciones <- predict(modelo_rf, newdata = datos_test)


valores_reales <- datos_test$pH.fangs

mse <- mean((predicciones - valores_reales)^2)

r_squared <- 1 - sum((valores_reales - predicciones)^2) / sum((valores_reales - mean(valores_reales))^2)



print("Resultados bootstrap:")

print(modelo_rf$results)


print("MSE:")

print(mse)


print("R-squared:")

print(r_squared)
```
```

```

## # Svm_radial

```{r }

library(e1071)

library(dplyr)

library(caret)

# Paso 1: Crear los datos

control <- trainControl(method = "boot", number = 1)

modelo_svm <- train(pH.fangs ~ ., data = datos_train, method = "svmRadial", trControl = control)

```

```{r,warning=FALSE}

predicciones <- predict(modelo_svm, newdata = datos_test)

valores_reales <- datos_test$pH.fangs

mse <- mean((predicciones - valores_reales)^2)

r_squared <- 1 - sum((valores_reales - predicciones)^2) / sum((valores_reales - mean(valores_reales))^2)


```
```

```
print("Resultados bootstrap:")

print(modelo_svm$results)


print("MSE:")

print(mse)


print("R-squared:")

print(r_squared)


```

Model multivariant

```{r,warning=FALSE}

library(dplyr)

library(caret)


# Paso 1: Crear los datos




# Paso 2: Crear el control de entrenamiento con bootstrapping

control <- trainControl(method = "boot", number = 100)


# Paso 3: Entrenar el modelo de regresión lineal con bootstrapping

modelo_lineal <- train(pH.fangs ~ ., data = datos_train, method = "lm", trControl = control)

```

```{r,warning=FALSE}

# Paso 4: Realizar predicciones con el modelo entrenado

predicciones <- predict(modelo_lineal, newdata = datos_test)


# Paso 5: Evaluar el rendimiento del modelo

valores_reales <- datos_test$pH.fangs

mse <- mean((predicciones - valores_reales)^2)

r_squared <- 1 - sum((valores_reales - predicciones)^2) / sum((valores_reales - mean(valores_reales))^2)


# Paso 6: Visualizar resultados y métricas

print("Resultados bootstrap:")

print(modelo_lineal$results)


print("MSE:")

print(mse)


print("R-squared:")

print(r_squared)

```

#knn

```{r,warning=FALSE}

library(dplyr)

library(caret)


set.seed(123)

control <- trainControl(method = "boot", number = 1)

modelo_knn<- train(pH.fangs ~ ., data = datos_train, method = "knn", trControl = control)


```{r,warning=FALSE}

predicciones <- predict(modelo_knn, newdata = datos_test)



valores_reales <- datos_test$pH.fangs

mse <- mean((predicciones - valores_reales)^2)

r_squared <- 1 - sum((valores_reales - predicciones)^2) / sum((valores_reales - mean(valores_reales))^2)



print("Resultados bootstrap:")

print(modelo_knn$results)


print("MSE:")

print(mse)


print("R-squared:")

print(r_squared)

```
```

#Gradient Boosting

```{r,warning=FALSE}

library(dplyr)

library(caret)

control <- trainControl(method = "boot", number = 10)

modelo_gbm <- train(pH.fangs ~ ., data = datos_train, method = "gbm", trControl = control)

```

```{r,warning=FALSE}

predicciones <- predict(modelo_gbm, newdata = datos_test)

valores_reales <- datos_test$pH.fangs

mse <- mean((predicciones - valores_reales)^2)

r_squared <- 1 - sum((valores_reales - predicciones)^2) / sum((valores_reales - mean(valores_reales))^2)

print("Resultados bootstrap:")

print(modelo_gbm$results)

print("MSE:")

print(mse)

print("R-squared:")
```

```
print(r_squared)
```

```{r }
xlim <- c(5, 10)
ylim <- c(5, 10)

library(ggplot2)

resultas <- data.frame(Valors_Reals = valores_reales, Prediccions = predicciones)

ggplot(resultats, aes(x = Valors_Reals, y = Prediccions)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, linetype = "dashed", color = "red") +
  labs(x = "Valors Reals", y = "Prediccions") +
  ggtitle("Gràfic de Prediccions del SVM Radial") +
  coord_fixed()
ggplot(modelo_knn, highlight = TRUE) +
  scale_x_continuous(breaks = 1:400) +
  labs(title = "Evolución del accuracy del modelo knn") +
  guides(color = guide_legend(title = "mtry"),
         shape = guide_legend(title = "mtry")) +
  theme_bw()

ggplot(modelo_svm, highlight = TRUE) +
  scale_x_continuous(breaks = 1:400) +
```

```
  labs(title = "Evolución del accuracy del modelo svm") +
  guides(color = guide_legend(title = "mtry"),
         shape = guide_legend(title = "mtry")) +
  theme_bw()

ggplot(modelo_gbm, highlight = TRUE) +
  scale_x_continuous(breaks = 1:400) +
  labs(title = "Evolución del accuracy del modelo gbm") +
  guides(color = guide_legend(title = "mtry"),
         shape = guide_legend(title = "mtry")) +
  theme_bw()

ggplot(modelo_rf, highlight = TRUE) +
  scale_x_continuous(breaks = 1:400) +
  labs(title = "Evolución del accuracy del modelo Random Forest") +
  guides(color = guide_legend(title = "mtry"),
         shape = guide_legend(title = "mtry")) +
  theme_bw()
```

```{r }
modelos <- list(KNN = modelo_knn, rf = modelo_rf,
                boosting = modelo_gbm, SVMradial = modelo_svm,
                lineal = modelo_lineal)
```

summary(modelo_knn)

```

## # Comparació de models

```{r,warning=FALSE}

num_repeticiones <- 100

rmse_boot <- c()

rsquared_boot <- c()

mae_boot <- c()

#for (i in 1:length(modelos)) {

  for (j in 1:1) {

    muestra <- sample(nrow(datos_test), replace = TRUE)

   datos_bootstrap <- datos_test[muestra, ]

   predicciones <- predict(modelo_svm, newdata = datos_bootstrap)

   valores_reales <- datos_bootstrap$pH.fang

   rmse_boot[j] <- sqrt(mean((valores_reales - predicciones)^2))

   rsquared_boot[j] <- cor(valores_reales, predicciones)^2

   mae_boot[j] <- mean(abs(valores_reales - predicciones))

  }

#}

```
```

```{r,warning=FALSE}
r2<-data.frame()
for(i in 1:100){
muestra <- sample(nrow(datos_test)*0.5, replace = TRUE)
   datos_bootstrap <- datos_test[muestra, ]
   predicciones <- predict(y, newdata = datos_bootstrap)
   valores_reales <- datos_bootstrap$pH.fangs
   r2[i,1] <- sqrt(mean((valores_reales - predicciones)^2))
   r2[i,2]<- cor(valores_reales, predicciones)^2
   r2[i,3] <- mean(abs(valores_reales - predicciones))
}
```{r,warning=FALSE}

rmse_estimado <- apply(rmse_boot, 2, mean)
rsquared_estimado <- apply(rsquared_boot, 2, mean)
mae_estimado <- apply(mae_boot, 2, mean)

```

```{r,warning=FALSE}
for (i in 1:5) {
  cat("Modelo", i, "\n")
  cat("RMSE estimado:", rmse_estimado[i], "\n")

```
  cat("R-squared estimado:", rsquared_estimado[i], "\n")
  cat("MAE estimado:", mae_estimado[i], "\n")
  cat("\n")
}
```

```{r,warning=FALSE}
library(cluster)
library(factoextra)
fviz_nbclust(x = datos2 , FUNcluster = pam, method = "wss", k.max = 15,
             diss = dist(datos, method = "manhattan"))
```

```{r , warning=FALSE}
library(readxl)
 rmse_boot<-read_excel("C:\\Users\\maxpl\\OneDrive\\Documentos\\TFG\\rmse_boot.xlsx")
 rsquared_boot<-
read_excel("C:\\Users\\maxpl\\OneDrive\\Documentos\\TFG\\rsquared_boot.xlsx")
 mae_boot<-read_excel("C:\\Users\\maxpl\\OneDrive\\Documentos\\TFG\\mae_boot.xlsx")

 rmse_boot$'1'<-as.numeric(rmse_boot$'1')
 rmse_boot$'2'<-as.numeric(rmse_boot$'2')
 rmse_boot$'3'<-as.numeric(rmse_boot$'3')
 rmse_boot$'4'<-as.numeric(rmse_boot$'4')
 rmse_boot$'5'<-as.numeric(rmse_boot$'5')
  rmse_boot$'5'<-r2$V1
 rmse_boot$'4'<-r$V1
  rsquared_boot$'1'<-as.numeric(rsquared_boot$'1')
 rsquared_boot$'2'<-as.numeric(rsquared_boot$'2')
 rsquared_boot$'3'<-as.numeric(rsquared_boot$'3')
```

```
 rsquared_boot$'4'<-as.numeric(rsquared_boot$'4')
 rsquared_boot$'5'<-as.numeric(rsquared_boot$'5')
  rsquared_boot$'5'<-r2$V2
 rsquared_boot$'4'<-r$V2
 mae_boot$'1'<-as.numeric(mae_boot$'1')
 mae_boot$'2'<-as.numeric(mae_boot$'2')
 mae_boot$'3'<-as.numeric(mae_boot$'3')
mae_boot$'4'<-as.numeric(mae_boot$'4')
 mae_boot$'5'<-as.numeric(mae_boot$'5')
 mae_boot$'5'<-r2$V3
mae_boot$'4'<-r$V3
 str(rmse_boot)



 rmse_estimado <- apply(rmse_boot, 2, mean)
rsquared_estimado <- apply(rsquared_boot, 2, mean)
mae_estimado <- apply(mae_boot, 2, mean)
```

#Boxplots

```{r,warning=FALSE}

#modelos <- list(KNN = modelo_knn, rf = modelo_rf,
                # boosting = modelo_gbm, SVMradial = modelo_svm,
                 #lineal = modelo_lineal)
```

```
metricas <- data.frame(RMSE = rmse_boot, R_squared = rsquared_boot, MAE = mae_boot)

nombres_modelos <- c("Model knn", "Model rf ", "Modelo gbm ","Model svmR","Model lineal")  # Agrega los nombres de tus modelos aquí

boxplot(rmse_boot, main = "Comparació de Models MSE", ylab = "Métrica de Evaluació",
        names = nombres_modelos)

boxplot(rsquared_boot, main = "Comparació de Models R-Squared", ylab = "Métrica de Evaluació",
        names = nombres_modelos)
boxplot(mae_boot, main = "Comparació de Models MAE", ylab = "Métrica de Evaluació",
        names = nombres_modelos)

```

```{r,warning=FALSE}
library(ggplot2)

# Crear un data frame con las predicciones y los valores reales
df <- data.frame(Predicciones = predicciones, Valores_Reales = nuevos_datos$variable_objetivo)

# Crear el gráfico de dispersión
ggplot(data = df, aes(x = Valores_Reales, y = Predicciones)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, linetype = "dashed", color = "red") +
  labs(x = "Valores Reales", y = "Predicciones") +
  ggtitle("Predicciones vs. Valores Reales")
```