Desenvolupament d'una aplicació per a l'estudi de càncer de pulmó a Europa

Gerard Caravaca Ibáñez

Col·laboració amb l'Hospital Clínic de Barcelona
Treball de final de grau

22 de juny del 2022

Director: Josep Solé Pareta (Departament d'Arquitectura de computadors UPC)
Codirectora: Laura Mezquita (Oncologia Hospital Clínic de Barcelona)
Professora de GEP: Carolina Maria Consolación Segura (Organització d'Empreses UPC)
Grau en Enginyeria Informàtica
Menció en Computació
Facultat d'Informàtica de Barcelona - Universitat Politècnica de Catalunya

Taula de continguts

Taula de Figures

Resum

El càncer de pulmó és la primera causa de mort per càncer a escala mundial. A Europa, és un problema sanitari de gran magnitud, ja que cada any aquesta malaltia es diagnostica a 2,6 milions de persones i posa fi a la vida d'altres 1,2 milions de persones, segons informa la Comissió Europea. Encara que el tabac és el factor de risc principal, existeixen altres factors prevalents, tant intrínsecs (predisposició genètica), com extrínsecs o carcinògens (gas radó, agents laborals, etc.), sent molts d'ells sinèrgics amb el tabac. No obstant això, aquesta informació no es registra i estudia de manera detallada en oncologia toràcica i avui dia, es desconeix quin perfil de malaltia s'associa als diferents factors intrínsecs i extrínsecs, més enllà del tabac i el que és més rellevant, si aquests tenen un impacte clínic en l'evolució i supervivència de pacients amb càncer de pulmó. Aquest projecte té com a objectiu desenvolupar una eina per l'estudi d'un d'aquests factors extrínsecs. Concretament, es construirà una aplicació per a la comparació de dades de càncer de pulmó amb dades d'origen ambiental amb el propòsit final de què aquesta pugui ser emprada per l'equip d'oncologia de l'Hospital Clínic per a analitzar possibles relacions entre zones d'altes emissions contaminants i zones amb alts índexs de càncer de pulmó. El resultat ha sigut una aplicació web duta a terme amb Dash, un framework dissenyat per a crear i desplegar aplicacions de dades amb interfícies d'usuari personalitzades, i MongoDB, una base de dades no relacional escalable i d'alt rendiment. L'aplicació final permet visualitzar, actualitzar i descarregar les dades segons les necessitats de l'equip.

Resumen

El cáncer de pulmón es la primera causa de muerte por cáncer a nivel mundial. En Europa, es un problema sanitario de gran magnitud, ya que cada año esta enfermedad se diagnostica a 2,6 millones de personas y acaba con la vida de otros 1,2 millones de personas, según informa la Comisión Europea. Aunque el tabaco es el factor de riesgo principal, existen otros factores prevalentes, tanto intrínsecos (predisposición genética), como extrínsecos o carcinógenos (gas radón, agentes laborales, etc.), siendo muchos de ellos sinérgicos con el tabaco. Sin embargo, esta información no se registra y estudia de manera detallada en oncología torácica y hoy en día, se desconoce qué perfil de enfermedad se asocia a los diferentes factores intrínsecos y extrínsecos, más allá del tabaco y el que es más relevante, si estos tienen un impacto clínico en la evolución y supervivencia de pacientes con cáncer de pulmón. Este proyecto tiene como objetivo desarrollar una herramienta para el estudio de uno de estos factores extrínsecos. Concretamente, se construirá una aplicación para la comparación de datos de cáncer de pulmón con datos de origen ambiental con el propósito final de que esta pueda ser empleada por el equipo de oncología del Hospital Clínico para analizar posibles relaciones entre zonas de altas emisiones contaminantes y zonas con altos índices de cáncer de pulmón. El resultado ha sido una aplicación web desarrollada con Dash, un framework diseñado para crear y desplegar aplicaciones de datos con interfaces de usuario personalizadas, y MongoDB, una base de datos no relacional escalable y de alto rendimiento. La aplicación final permite visualizar, actualizar y descargar los datos según las necesidades del equipo.

Abstract

Lung cancer is the leading cause of cancer deaths worldwide. In Europe, it is a major health problem, with 2.6 million people diagnosed with the disease each year, killing another 1.2 million people, according to the European Commission. Although tobacco is the main risk factor, there are other prevalent factors, both intrinsic (genetic predisposition) and extrinsic or carcinogenic (radon gas, occupational agents, etc.), many of which are synergistic with tobacco. However, this information is not recorded and studied in thoracic oncology and today, it is not known what disease profile is associated with the different intrinsic and extrinsic factors, beyond tobacco and, more importantly, whether they have a clinical impact on the evolution and survival of patients with lung cancer. This project aims to develop a tool for the study of one of these extrinsic factors. Specifically, an application will be built to compare lung cancer data with data of environmental origin with the aim that it can be used by the oncology team of the Hospital Clínic to analyse possible relationships between areas with high pollutant emissions and areas with high rates of lung cancer. The result has been a web application developed with Dash, a framework designed to create and deploy data applications with customised user interfaces, and MongoDB, a scalable, high-performance, non-relational database. The final application allows data to be viewed, updated and downloaded according to the team's needs.

Agraïments

Abans de començar a introduir el projecte volia mostrar la meva gratitud a totes les persones que han col·laborat en la realització d'aquest. Primerament, m'agradaria agrair a la Dra. Laura Mezquita que ha sigut la responsable de què sorgís aquesta col·laboració i, com a codirectora del treball, ha sigut indispensable des del punt de vista de la investigació clínica. Seguidament, també volia donar les gràcies al director d'aquest treball de final de grau, en Josep Solé Pareta, que m'ha guiat en totes les etapes d'aquest projecte. Per una altra banda, caldria reconèixer el treball de la professora de l'assignatura de Gestió de Projectes de la FIB, Carolina Maria Consolación Segura, que em va donar les pautes per a la planificació i la gestió inicial d'un projecte d'aquestes característiques i a en Carles Farré i l'Albert López, membres dels departaments de Software i Arquitectura de computadors de la UPC, que em van ajudar en l'àmbit tècnic. Finalment, volia donar els meus més sincers agraïments tant al Joan Sart Vizcaíno, que ha sigut un gran company de treball, com a tot l'equip de l'Hospital Clínic que m'han permès treballar amb ells com un més de l'equip. Especialment, volia destacar a la Marta Garcia de Herreros que em va facilitar les dades del seu projecte Radon Europe i al Dr. Víctor Albarrán que em va permetre assistir a la seva consulta. *Moltes gràcies a tots. Muchas gracias a todos. Thank you all very much.*

1. Context

1.1. Introducció

El càncer de pulmó és el segon càncer més comú que afecta tant a homes com a dones. Per als homes el més comú és el càncer de pròstata i en dones és el de pit. S'estima que el 13% de tots els càncers nous són càncers de pulmó. A més, aquest tipus de càncer pot arribar a ser mortal per a un 80% de les persones que el pateixen [1]. Tanmateix, l'esperança de supervivència a la malaltia augmenta en els pacients en els quals es detecta el càncer de forma prematura. Per aquest motiu esdevé de gran importància l'estudi per a la prevenció i la detecció primerenca del tumor.

En aquest àmbit, l'Hospital Clínic de Barcelona [2] i l'Institut d'investigacions Biomèdiques August Pi i Sunyer (*IDIBAPS*) [3] treballen per donar resposta, a través de la recerca, als diferents factors de prevenció del càncer, entre altres estudis. Una part important per al grup de treball és la innovació i la implementació tecnològica en la seva recerca per a l'obtenció de millors resultats a gran escala.

Aquest treball de fi de grau de modalitat A (Centre) se situa al marc de la Facultat d'Informàtica de Barcelona (FIB), tot i que estaré actuant en col·laboració amb l'equip de recerca de l'Hospital Clínic. La idea és combinar el coneixement tècnic obtingut per part de la UPC amb un fons teòric biomèdic, que pot aportar l'equip de recerca de l'hospital. Finalment, cal destacar que aquest treball es desenvolupa en paral·lel amb el d'en Joan Sart, un company de la FIB que també duu a terme el seu TFG en col·laboració amb l'Hospital Clínic i que en algunes parts del procés haurem d'actuar en tàndem.

1.2. Descripció del problema

És un fet que les coses que no s'estudien i, per tant, no es poden notificar passen a no ser considerades per a la població general i els òrgans de govern. Això mateix passa amb la prevenció pel càncer de pulmó. De fet, es coneixen centenars de factors, anomenats carcinògens, que poden causar aquesta malaltia, d'entre aquests hi ha nombrosos factors que han de veure amb causes mediambientals. Però la gran majoria d'aquest no han pogut ser estudiats de forma profunda i, per tant, no s'ha pogut demostrar una relació de causalitat. Per una altra banda, el tabac es considera la principal causa del càncer de pulmó. Si més no és la més estudiada i, en conseqüència, la més coneguda. Aquest fet ha causat que la gran majoria de països d'Europa hagin establert pautes de prevenció contra el tabac. D'altra banda, això ha fet que se li hagi donat una menor importància als altres possibles factor, entre els quals podem distingir el gas radó, l'arsènic, el dièsel i altres representats a la Figura 1.1.

Per aquest motiu l'equip de recerca biomèdica va veure interessant iniciar projectes de recerca en aquest àmbit. En aquest moment es van trobar amb el problema de què no disposaven de les eines necessàries, en l'àmbit tecnològic, per a poder utilitzar les dades proporcionades pels països Europeus. Ja que amb les eines que disposaven no aconseguien representar les dades per poder comparar-les de forma adient i no podien predir futurs escenaris. És aquí on encaixa aquest projecte, podem, conseqüentment, desenvolupar una aplicació informàtica que permeti comparar i analitzar les dades de contaminació amb les dades de càncer de pulmó durant les tres últimes dècades als diferents països d'Europa. Amb la finalitat de descobrir i predir punts calents, així com poder veure si els plans de les autoritats sanitàries de cada país estan funcionant.

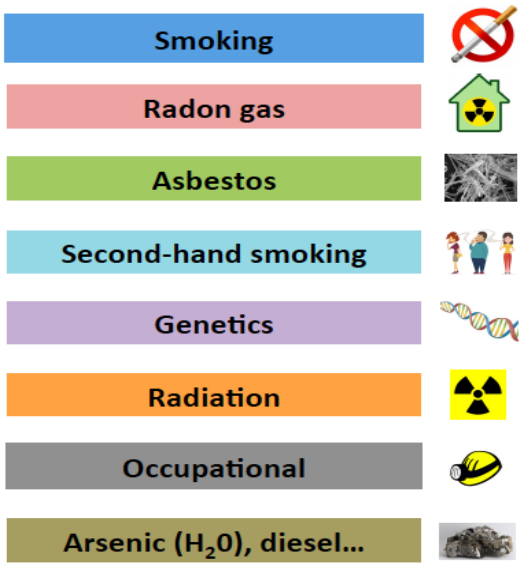

Figura 1.1: Carcinògens de grup 1 pel càncer de pulmó. *Font: base de dades WHO[7]*

1.3. Actors implicats

Els actors implicats en el projecte són els següents:

- **Els proveïdors de les dades:** una part de la riquesa de l'anàlisi dependrà de la qualitat de les dades proporcionades pels diferents països Europeus. En aquest cas hem trobat bases de dades interessants com: Globocan [4], *Eurostat* [5], *Nordcan* [6] i *WHO* [7].

- **L'equip d'investigació i recerca de l'Hospital Clínic:** el personal de recerca vol una aplicació intuïtiva, fàcil d'utilitzar i que els hi ajudi a demostrar una relació entre factors mediambientals i la incidència del càncer de pulmó. Seran els que hauran de beneficiar-se dels resultats obtinguts.

- **El personal del projecte:** en Joan Sart i jo tenim la necessitat de finalitzar el treball a temps i aconseguir els resultats desitjats. A més, suposa un repte per a nosaltres integrar els dos treballs en una sola aplicació.

2. Justificació

2.1. Valoració

Des d'un primer moment l'equip del Clínic ens van proposar que el sistema a desenvolupar sigui una aplicació. Ens van facilitar un projecte semblant, però inacabat que van fer els alumnes de l'assignatura de PAE (Projecte avançat d'enginyeria, assignatura de la FIB) el quadrimestre anterior. Després d'analitzar l'aplicació vam decidir que faríem una de nova amb llenguatges i metodologies que fossin del nostre coneixement. El motiu principal d'aquesta elecció va ser que el primer prototip estava fet en llenguatges que no coneixem i no es disposava d'una documentació detallada del projecte.

Existeixen moltes eines per a dur a terme aplicacions. En el nostre cas utilitzarem *Dash* [8], un *framework* de *Python* [9] pensat per a construir aplicacions web, però que també es pot utilitzar per a crear visualitzacions diverses. *Dash* està basat principalment en *Flask* [10], *Ploty* [11] i *ReactJS* [12], aquestes tres eines fan que tingui les següents característiques interessants per al projecte. Primerament, cal destacar que és una opció Codi obert i que, per tant, és gratuïta. Seguidament, permet renderitzar les aplicacions en el navegador i és multiplataforma. Aquesta característica serà molt important, ja que l'aplicació s'ha de poder fer servir en múltiples sistemes operatius. A més, el fet que el llenguatge emprat sigui *Python* és un avantatge, pel fet que és un llenguatge conegut per nosaltres i que facilita la gestió i anàlisis de dades així com la implementació d'eines de IA com *Aprenentatge automàtic*. Finalment, es tracta d'un *framework* que permet crear aplicacions simples de manera ràpida.

3. Abast

3.1. Objectius

L'objectiu final del projecte és el desenvolupament d'una aplicació manejable per a l'equip de recerca de l'hospital. Que els hi permeti estudiar com evoluciona la incidència i mortalitat del càncer de pulmó en els diferents països d'Europa. I per l'altra banda també poder comparar aquesta evolució amb la dels diferents agents contaminants. Tot plegat ha de permetre demostrar si existeix o no una relació evident entre els diferents agents ambientals i la malaltia. Una idea és mostrar les dades en mapes com els de la Figura 3.1.

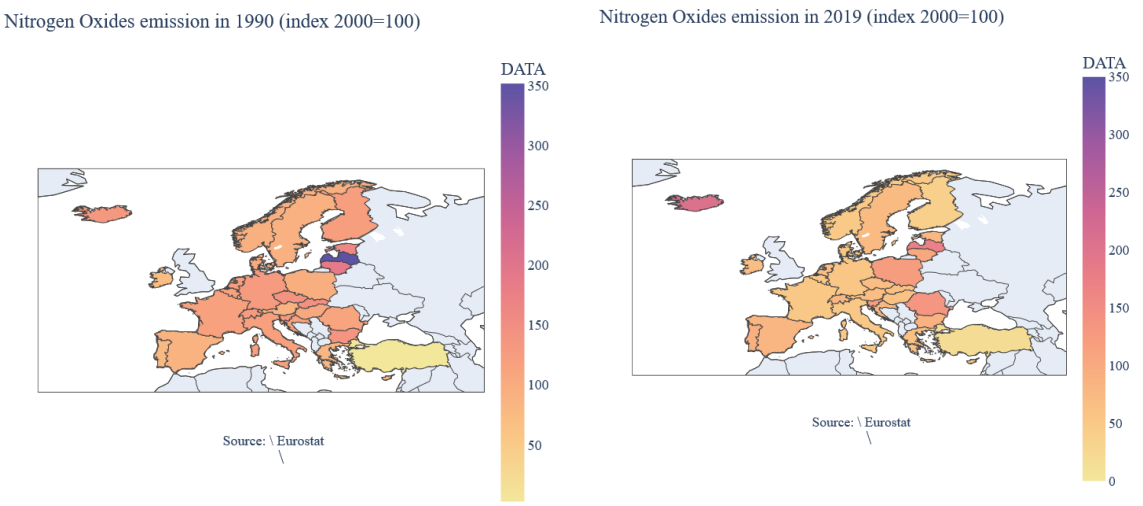

Figura 3.1: mapa d'emissions d'òxids de Nitrogen a Europa en els anys 1990 i 2019. *Font: Elaboració pròpia*

A més d'aquest objectiu principal, juntament amb l'equip de recerca del Clínic hem fixat els següents subobjectius que creiem que són interessants per a enriquir el projecte:

1. Dissenyar i desenvolupar una base de dades que ens faciliti el tractament de dades tant de càncer com d'agents contaminants en Europa.
2. Dissenyar una interfície d'usuari intuïtiva i fàcil d'utilitzar.
3. Implementar un sistema d'ingesta de noves dades des de la mateixa aplicació.
4. Valorar diferents formes de representació de dades i implementar les més útils (mapes, diagrames de barres…).
5. Afegir mecàniques de personalització per facilitar l'exploració de les dades.
6. Obtenir un sistema eficient i escalable que permeti una vida útil de l'aplicació elevada.

3.2. Requeriments

El fet que l'aplicació estigui destinada a un grup de treball en concret extern a la FIB comporta que aquesta hagi de complir uns requisits en concret. Els requisits són els següents:

1. L'aplicació ha de poder ser executada en diferents sistemes operatius tant Linux, com Windows com també IOs.
2. L'aplicació ha de ser senzilla d'instal·lar i de fer servir.

3. Les dades utilitzades han de ser rigoroses.
4. La interfície ha de ser simple, clara i professional, ja que ha de poder ser presentable.
5. Seria convenient que la ingesta de dades es pugui fer dada a dada i també directament amb un fitxer de taules.
6. És indispensable que el sistema sigui robust i no tingui problemes de seguretat.
7. L'aplicació ha d'estar adaptada tant a un ambient professional com també a un ambient acadèmic, ja que l'equip de recerca està format tant per doctors com per estudiants.

3.3. Obstacles

Els principals obstacles que poden dificultar la consecució dels objectius fixats per aquest projecte són:

- **Inexperiència:** els encarregats de desenvolupar l'aplicació no tenim massa experiència en el disseny *FrontEnd* per aplicacions, per tant, això suposarà un repte. Per altra banda, tampoc som experts en oncologia, això suposa que hem de realitzar un estudi previ del tema.

- **No enteniment de les dades:** l'obtenció de les dades s'ha de fer de fonts externes i en conseqüència hem d'entendre la forma en què estan exposades i validar si són o no són adients per l'estudi que volem fer.

- **Dades insuficients:** la majoria de països d'Europa tenen comptabilitzades les dades de càncer així com les de gasos contaminants. Tanmateix, per a segons que país aquestes dades no són públiques o no estan completes.

- **Organització temporal:** juntament amb aquest projecte també estic fent pràctiques en empresa a temps parcial. Això pot comportar que en alguns dies no pugui dedicar tot el temps que voldria al desenvolupament de l'aplicació.

3.4. Riscos

Com en qualsevol projecte professional d'aquesta magnitud poden sorgir imprevistos i problemes. Consegüentment, he realitzat un llistat dels principals contratemps amb els quals ens podem trobar al llarg del treball:

- **Errors de disseny:** aquests tipus d'errors són deguts al fet que en algun moment de la fase de disseny del projecte ens hem equivocat. Per tant, aquests provoquen tornar a modificar el disseny inicial, poden incloure errors d'eficiència, escalabilitat i representació de les dades.

- **Errors d'implementació:** són errors que sorgeixen en la fase de desenvolupament del projecte. Aquests són molt comuns i formen part de tots els projectes de caràcter tècnic. Poden incloure *bugs*, incompatibilitats entre eines, implementacions massa complexes…

- **Gestió del temps:** en aquest cas estem parlant d'una mala previsió del temps que es trigarà a completar una tasca. És degut a previsions poc realistes en la planificació o en successions de riscos consecutius.

- **Qüestions personals:** val la pena també tenir en compte que hi ha molts factors personals que poden afectar en el ritme de treball i, en conseqüència, en la consecució dels diferents objectius.

- **Resultats imprevistos:** el resultat de les dades mostrades pot no ser el desitjat per algun error en el tractament inicial de les dades o en l'obtenció d'aquestes. Això haurà de ser resolt tornant a revisar la fase de tractament de dades.

4. Metodologia

Com ja he esmentat diverses vegades aquest és un TFG complex pel fet que estan involucrades moltes persones. Per aquest motiu la metodologia ha de permetre treballar de forma àgil i poder intercalar tasques de treball individual amb tasques de treball en equip. És important destacar que en Joan i jo tenim formes i ambients diferents de treball. Mentre ell elabora el projecte presencialment en els laboratoris del Clínic, jo faig feina de forma remota. Per tant, la metodologia també ha de permetre flexibilitat en aquest aspecte. Addicionalment, s'ha de tenir en compte que el seguiment del treball és dur a terme des de dos marcs totalment diferents, un és la FIB i l'altre és l'Hospital Clínic.

Scrum [14] és una metodologia que pretén arribar a obtenir els millors resultats a partir de la coordinació de l'equip per poder ser prou productius. La idea és que amb *Scrum* es van efectuant entregues o posades en comú regulars i parcials del treball realitzat, de forma prioritària i en funció dels beneficis que aporta cadascuna als receptors del projecte. Per aquest motiu és una metodologia especialment indicada per projectes complexos amb múltiples requisits i més d'un participant, com és el cas d'aquest treball de fi de grau.

La metodologia *Scrum* passa per diferents fases mostrades a la Figura 4.1:

1. **Planificació:** és la fase en la qual s'estableixen les tasques prioritàries i on s'aconsegueix informació breu i detallada sobre el projecte que es desenvoluparà. És necessari per a poder arrencar amb el primer sprint, té permès canviar i créixer tantes vegades com sigui necessari en funció de l'aprenentatge adquirit en el desenvolupament del producte.

2. **Execució (Sprint):** es defineix com un subprojecte on l'equip de treball es focalitza en el desenvolupament de tasques per a assolir l'objectiu que s'ha definit prèviament.

3. **Control:** és la fase en la qual es mesura i es valora el progrés del projecte periòdicament.

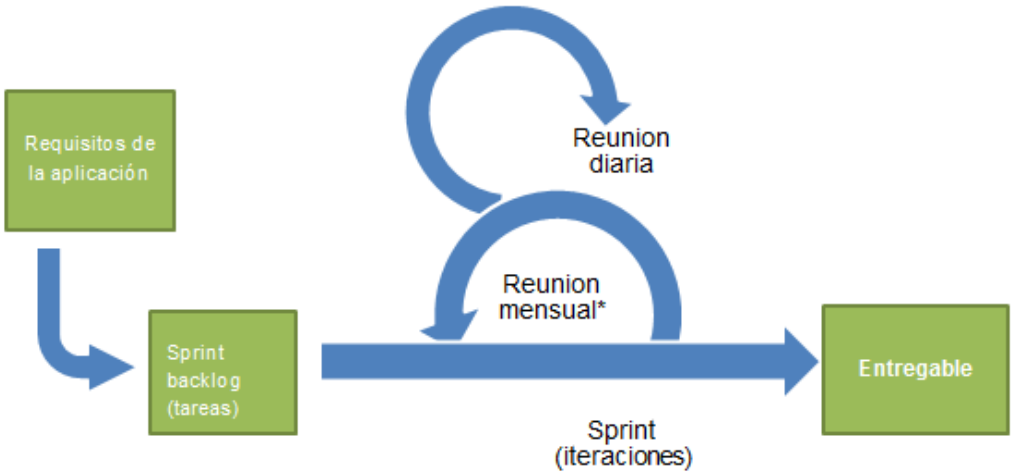

Figura 4.1: gràfic de fases de la metodologia *Scrum Font: Scrum a Wikipedia [14]*

Creiem que *Scrum* és la metodologia més adient pel nostre projecte, ja que ens ha de permetre començar a treballar en l'aplicació al mateix temps que estem estudiant el tema de forma teòrica. A més el fet que sigui un procés cíclic ens proporciona certa flexibilitat a les errades.

4.2. Seguiment

Per posar en comú el treball fet, verificar que el camí assolit és el correcte i resoldre alguns dubtes que sorgeixen, hem establert un seguiment amb la directora del treball al Clínic (Laura Mezquita). Aquest seguiment consta, bàsicament, d'una reunió setmanal de forma remota, en la qual mostrem els avanços fets durant la setmana, els valorem en referència als objectius

establerts i implantem nous objectius i subobjectius per a la setmana següent. A més d'aquesta reunió setmanal també es fan dues trobades al mes de l'equip de recerca de l'hospital on es presenten les actualitzacions dels diferents projectes en desenvolupament. Aquestes ens han d'ajudar a entendre la base teòrica del projecte, quan ens toca veure presentacions, i a validar que el que estem fent és correcte, quan ens toca presentar a nosaltres.

4.3. Eines

Per a poder seguir aquesta metodologia necessitarem fer ús de diferents eines que ens facilitaran el flux del treball i la comunicació. Primer de tot l'eina que utilitzarem per al control de versions de l'aplicació és el *GitHub* [15]. Aquesta elecció ha sigut senzilla, ja que és l'opció més emprada per la comunitat i estem familiaritzats amb ella. La idea en aquest cas és que els dos participants de l'aplicació tinguem una branca en local i puguem anar treballant de forma individual per poder després fer un *merge* de les dues branques per obtenir la versió oficial.

Per una altra banda, per al seguiment de les fites i els objectius farem servir *Trello* [16], una eina de gestió de projectes que permet crear llistes de tasques en una interfície amigable, simple i compartida per a tots els membres de l'equip. Aquesta eina enllaça perfectament amb la metodologia escollida.

Finalment, farem servir *Microsoft Teams* [17] per a fixar les dates de les reunions i realitzar aquelles que siguin de forma remota. Hem escollit aquesta opció perquè va ser la que va proposar la directora del treball i ens vam poder adaptar sense problemes.

5. Extensió temporal i recursos

L'extensió prevista d'aquest projecte és d'unes 450 hores de dedicació. S'espera una dedicació mínima de 4 hores diàries. Com es tracta d'un projecte en modalitat A, tinc bastant llibertat per al repartiment de la càrrega de treball. De totes maneres, la idea és de mantenir un ritme constant d'unes 25 hores setmanals. Per tant, donat que el projecte va ser iniciat el 7 de febrer del 2022, la data de finalització hauria de ser al voltant del 30 de juny del mateix any, com a resultat, l'extensió seria de vint setmanes. La raó d'aquesta extensió és la previsió de què acabin apareixent els obstacles i riscos descrits en els apartats anteriors. Atès que el torn de lectures del TFG de juny comença el dilluns 27 de juny [18], hem decidit que la memòria hauria d'estar acabada un parell de setmanes abans, per poder tenir temps amb marge per preparar la presentació.

Finalment, tot i ser un projecte desenvolupat en col·laboració amb l'Hospital Clínic de Barcelona, no s'han detallat restriccions temporals afegides a les establertes per la normativa de la FIB.

5.1. Recursos necessaris

Per aquest projecte, com per a qualsevol altre projecte informàtic, requerim tant de personal com també de recursos materials. Per la part del personal hem determinat els següents actors fonamentals:

- **Directors (R1):** tant la directora de l'equip de recerca com el ponent de la FIB són peces necessàries per a la guia del projecte, la definició del seu abast i l'assessorament als estudiants.

- **Desenvolupadors de software (R2):** es tracta d'en Joan Sart i jo mateix. Som els encarregats de dur a terme el punt central del projecte, que és el disseny i desenvolupament tècnic de l'aplicació.

Per l'altra banda hem determinat els següents recursos materials necessaris:

- **Hardware (R3):** el principal harware necessari són dos ordinadors portàtils, un per programador. El motiu de la selecció és la seva relació qualitat/preu i la mobilitat que ofereix.

- **Software:** hem decidit que només farem ús de programari de Codi obert, com ara Python, Dash, Visual Studio Code [19] (R4), Github (R5), Trello (R6), Google docs (R7) i Microsoft Teams (R8) [17].

6. Planificació

6.1. Fases del projecte

Hem decidit classificar les tasques en diferents àmbits per poder estructurar-les amb més claredat abans de definir-les de forma individual. Les diferents fases definides són les següents:

- **Gestió del projecte:** és dur a terme dins del marc de l'assignatura de GEP i comprèn totes les tasques de planificació del projecte.

- **Estudi previ del tema:** en aquesta fase estudiarem la base teòrica del projecte. La finalitat és establir una base de coneixement sobre els principals conceptes relacionats amb el càncer i els diferents factors que el poden causar.

- **Extracció de dades:** les dades disponibles en aquest àmbit varia segons la regió. Suposarà un repte de recerca l'obtenció de totes les dades sanitàries. En aquest punt tindré l'ajut dels responsables de l'hospital clínic, que podran demanar dades extra a contactes Europeus.

- **Disseny de l'Aplicació:** l'objectiu final és l'obtenció d'una aplicació informàtica útil per l'anàlisi anteriorment esmentada. La idea és la representació de la informació en un més d'un mapa per poder així comparar els resultats. Addicionalment, l'aplicació hauria de tenir una forma senzilla d'inserir noves dades per poder actualitzar la base de dades. En aquesta fase hauria de quedar clara l'estructura de l'aplicació a partir del desenvolupament de diagrames.

- **Implementació:** aquesta és la fase més tècnica del projecte. Haurem d'implementar en codi totes les estructures dissenyades anteriorment.

- **Proves:** abans de donar per terminada la implementació de l'aplicació cal fer un testatge exhaustiu de totes les funcions.

- **Documentació:** la part tècnica del projecte ha d'anar documentada de la forma indicada. En aquesta fase em dedicaré a redactar la memòria del treball.

- **Presentació:** per finalitzar el projecte cal fer una presentació de tot el treball fet. Aquest apartar constarà de la preparació d'aquesta presentació.

6.2. Descripció de les tasques

A continuació es descriuran les diferents tasques que conformen aquest projecte. Aquestes estan resumides a la figura 6.1.

6.2.1. Tasques de gestió de projecte

- **Abast (G1):** redacció de la primera entrega de gestió de projectes on es descriu una base del context, els requeriments de l'aplicació, els objectius del projecte i la metodologia de treball a seguir per a poder arribar a satisfer les necessitats. (25 hores)

- **Planificació temporal (G2):** redacció de la segona entrega de gestió de projectes. En aquesta entrega es descriuen les tasques i la distribució en el temps d'aquestes. (20 hores)
  Dependències: G1.

- **Pressupost i anàlisis de sostenibilitat (G3):** redacció de la tercera entrega de gestió de projectes. Aquest entrega comporta un estudi de costos del projecte i dels efectes que pot causar en la sostenibilitat. (20 hores)
  Dependències: G2.

- **Reunió setmanal (G4):** cada setmana es convoca una reunió a la qual assistim: la directora del Clínic, el Joan Sart i jo mateix. En aquesta setmana es posa el treball en comú, es pregunten dubtes i es fixen nous propòsits per a la setmana següent. (1 hora)

- **Reunió d'actualització de projectes de recerca (G5):** es tracta de dues reunions setmanals en què els membres de l'equip del Clínic mostren el progrés en els seus projectes. Com a part de l'equip aquest projecte també es presenta en aquestes reunions. (3 hores)

6.2.2. Tasques de documentació

- **Documentació de la fase inicial (D1):** es recullen les tres entregues de gestió de projectes i es corregeixen els errors seguint el feedback de la professora. (5 hores)
  Dependències: G3.

- **Documentació de la fase de seguiment (D2):** es prepara l'entrega i es fa una reunió amb el ponent per a informar del progrés i l'estat del projecte en aquest moment. (40 hores)
  Dependències: D1.

- **Documentació de la fase final (D3):** es redacta la part final del treball i es perfeccionen les anteriors intentant que la memòria quedi de la forma més entenedora possible. (40 hores)
  Dependències: D2.

6.2.3. Tasques de Presentació

- **Preparació de la presentació final (P):** es prepara un document en format presentació i es practica la posada en escena. (30 hores)
  Dependències: D3.

6.2.4. Tasques d'estudi previ

- **Aprenentatge teòric (E1):** estudi dels diferents aspectes teòrics que envolten el projecte. Aquests aspectes engloben: el càncer de pulmó, les seves causes, la representació de les dades, els gasos contaminants més determinants… (15 hores)

- **Estudi d'eines per a l'aplicació (E2):** valoració de les diferents eines, llenguatges de programació i *frameworks* que podem fer servir per a la creació de l'aplicació. (10 hores)

6.2.5. Tasques d'extracció de dades

- **Cerca i extracció de dades (ED1):** cerca de les diferents dades que s'utilitzaran en l'aplicació. Així com la validesa d'aquestes. (20 hores)

- **Organització de les dades (ED2):** segurament en el moment de descarregar les dades ens trobarem amb una distribució bastant caòtica. Per aquest motiu és important organitzar les dades per taules. (10 hores)
  Dependències: ED1.

6.2.6. Tasques de disseny de l'aplicació

- **Disseny de la base de dades (DA1):** disseny de l'estructura de la base de dades tenint en compte les dades obtingudes. (10 hores)
  Dependències: ED2.

- **Disseny de l'estructura de classes (DA2):** disseny de les diferents classes, controladors i més genèricament del *backend* del sistema. (10 hores)
  Dependències: DA1.

- **Disseny de la interfície per usuaris (DA3):** disseny de les diferents vistes de la interfície d'usuari, més genèricament del *frontend* del sistema. (10 hores)

6.2.7. Tasques d'implementació

- **Implementació de les funcions de visualització de dades (I1):** desenvolupament dels algoritmes i del codi necessari per a cobrir les funcions de visualització de les dades en mapes. (20 hores)
  Dependències: DA2.

- **Implementació de les funcions d'ingesta de dades (I2):** implementació de les funcions necessàries per poder introduir dades en la base de dades de forma automatitzada. (20 hores)
  Dependències: DA2.

- **Implementació de la interfície d'ingesta de dades (I3):** implementació de les vistes per a la introducció de noves dades. (20 hores)
  Dependències: DA3.

- **Implementació de la interfície de visualització de dades (I4):** implementació de les vistes per a la representació dels mapes. (20 hores)
  Dependències: DA3.

6.2.8. Tasques de prova

- **Testatge del *backend* (T1):** creació de funcions per a testejar la implementació del *backend.* (10 hores)
  Dependències: I1, I2.

- **Testatge del *frontend* (T2):** creació de funcions per a testejar la implementació del *frontend.* (10 hores)
  Dependències: I3, I4.

TASCA	HORES	DEPENDÈNCIES	RECURSOS
G1: Abast	25	-	R1, R2, R7
G2: Planificació temporal	20	G1	R1, R2, R7
G3: Pressupost i anàlisis sostenible	20	G2	R1, R2, R7
G4: Reunió setmanal	1	-	R1, R2, R8
G5: Reunió d'actualització de projectes de recerca	3	-	R1, R2
D1: Documentació de la fase inicial	5	G3	R1, R2, R7
D2: Documentació de la fase de seguiment	40	D1	R1, R2, R7
D3: Documentació de la fase final	40	D2	R1, R2, R7
P: Preparació de la presentació final	30	D3	R1, R2, R7
E1: Aprenentatge teòric	15	-	R1, R2
E2: Estudi d'eines per a l'aplicació	10	-	R2
ED1: Cerca i extracció de dades	20	-	R1, R2

TASCA	HORES	DEPENDÈNCIES	RECURSOS
ED2: Organització de les dades	10	ED1	R2, R3
DA1: Disseny de la base de dades	10	ED2	R2
DA2: Disseny de l'estructura de classes	10	DA1	R2
DA3: Disseny de la interfície per usuaris	10	-	R2
I1: Implementació de les funcions de visualització de dades	40	DA2	R2, R3, R4
I2: Implementació de les funcions d'ingesta de dades	30	DA2	R2, R3, R4
I3: Implementació de la interfície d'ingesta de dades	30	DA3	R2, R3, R4
I4: Implementació de la interfície de visualització de dades	30	DA3	R2, R3, R4
T1: Testatge del backend	10	-	R2, R3, R4
T2: Testatge del frontend	10	I3, I4	R2, R3, R4

Figura 6.1: Taula de tasques. Font: Elaboració pròpia.

6.4. Representació de la planificació

A la figura 6.2 trobem representada la planificació temporal del projecte, a escala de tasques, utilitzant un diagrama de Grantt.

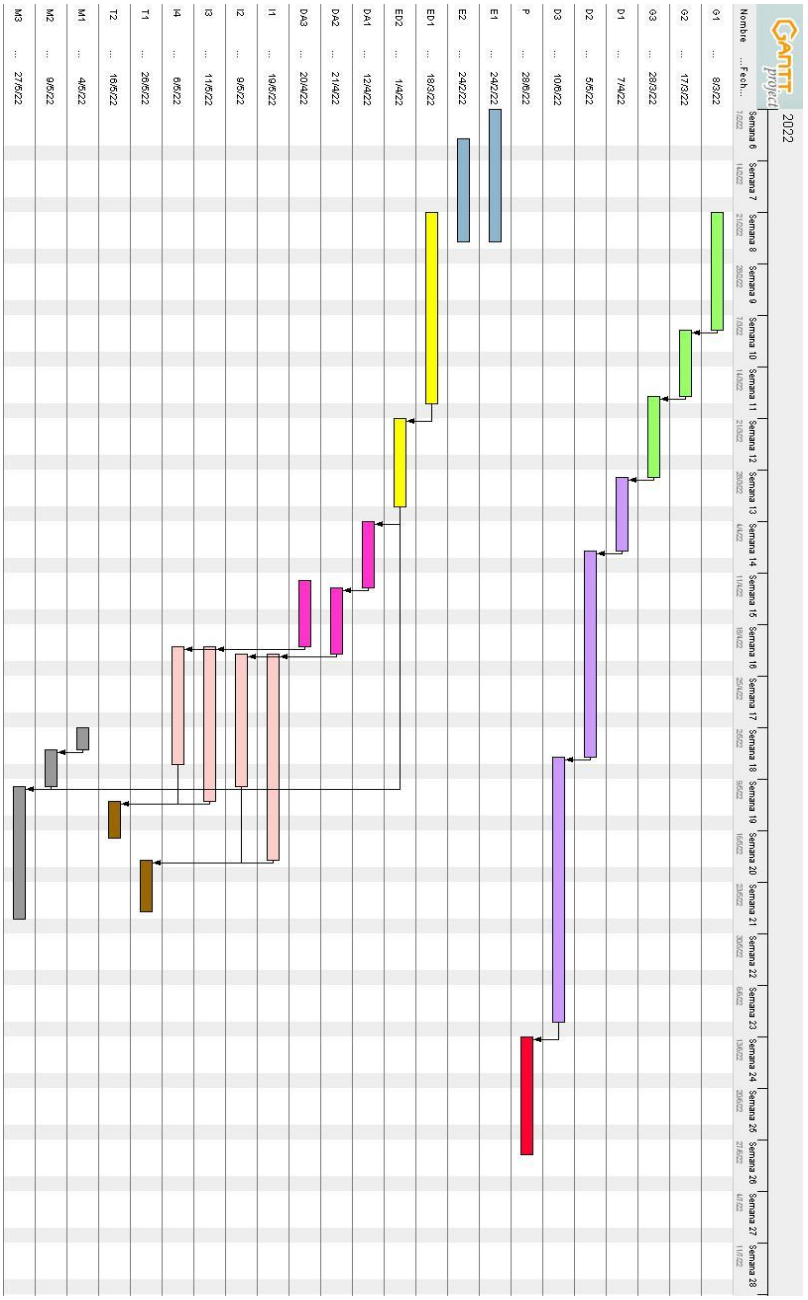

Figura 6.2: Diagrama de Grantt. Font: Elaboració pròpia.

7. Gestió de riscos

Com ja hem vist anteriorment, el projecte compte amb uns riscos que poden endarrerir el compliment de les tasques. A causa d'aquest factor, s'han de proposar possibles solucions a aquest fet.

Per començar, els riscos més habituals en un projecte d'aquestes característiques, són els possibles problemes d'implementació. Aquests ja estan inclosos en la planificació. Donat que estem molt habituats a veure bugs contínuament, hem decidit incloure les tasques de testatge a la planificació temporal del projecte. D'aquesta forma prevenim un retard en les tasques posteriors.

Un altre risc important, com ja he esmentat anteriorment, són els errors de disseny. El fet que la fase d'implementació sigui posterior a la de disseny, està pensat per facilitar la detecció d'aquest tipus d'errors de la forma menys destructiva possible.

Planificar un projecte és una tasca complicada, ja que cal tenir en compte molts factors. En el nostre cas, en utilitzar una metodologia àgil, és més difícil de quadrar les dates amb les tasques necessàries per complir amb els objectius. Per aquest motiu, he decidit fer una sobreestimació en la duració de les tasques, també com a mètode de prevenció.

Els riscos esperats per aquest projecte no haurien de comportar un increment dels recursos necessaris, tenint en compte tant els recursos humans com materials.

8. Gestió econòmica

8.1. Identificació dels costos

En aquest capítol es realitza un estudi econòmic per tal d'identificar i estimar els costos associats a aquest treball amb la finalitat de valorar la seva viabilitat econòmica. Amb l'objectiu de poder identificar correctament els costos que suposarà el projecte, hem de tenir en compte tant els recursos humans, com els de *hardware* i els de *software*. A més, també cal tenir en compte les contingències i els costos carregats a possibles imprevistos.

8.1.1. Recursos humans

Com ja vaig comentar en l'apartat anterior, els perfils de personal que intervindran en aquest projecte són el director de projecte i el de programador. En aquest cas el rol de director serà compartit entre el ponent de la FIB, la directora del Clínic i jo, mentre que el rol de programador serà desenvolupat per mi mateix, l'autor d'aquest TFG. A la figura 7.1

s'indiquen les retribucions en euros per hora que correspondrien a cada perfil. Aquestes dades les he obtingut de la web Indeed [20].

Perfil	**Sou brut**	**Seguretat Social (x1.3)**	**Retribució**
Cap de projecte	16 €/h	4.80 €/h	20.80 €/h
Programador	12 €/h	3.60 €/h	15.6 €/h

Figura 8.1: Resum costos de personal per perfil. Font: Elaboració pròpia.

Finalment, podem calcular els costos per activitat duent a terme una estimació del cost de cada tasca definida en el diagrama de Gantt. A la taula 7.2 podem veure els perfils que participen en cada tasca i el cost total de cadascuna d'ells.

Tasca	**Hores**	**Cap de projecte (h)**	**Programadors (h)**	**Cost Total (€)**
G1: Abast	25	25	0	520
G2: Planificació temporal	20	20	0	416
G3: Pressupost i anàlisis sostenible	20	20	0	416
G4: Reunió setmanal (x20)	1	1	0	416
G5: Reunió d'actualització de projectes de recerca (x5)	3	3	0	312
D1: Documentació de la fase inicial	5	5	0	104
D2: Documentació de la fase de seguiment	40	40	0	832
D3: Documentació de la fase final	40	40	0	832
P: Preparació de la presentació final	30	20	0	416
E1: Aprenentatge teòric	15	0	15	234
E2: Estudi d'eines per a l'aplicació	10	2	8	166,4
ED1: Cerca i extracció de dades	20	5	15	338

Tasca	Hores	Cap de projecte (h)	Programadors (h)	Cost Total (€)
ED2: Organització de les dades	10	0	10	156
DA1: Disseny de la base de dades	10	0	10	156
DA2: Disseny de l'estructura de classes	10	0	10	156
DA3: Disseny de la interfície per usuaris	10	0	10	156
I1: Implementació de les funcions de visualització de dades	40	0	40	624
I2: Implementació de les funcions d'ingesta de dades	30	0	30	468
I3: Implementació de la interfície d'ingesta de dades	30	0	30	468
I4: Implementació de la interfície de visualització de dades	30	0	30	468
T1: Testatge del *backend*	15	0	15	238
T2: Testatge del *frontend*	15	0	15	238
Total CPA	**460**	**182**	**238**	**8590,4**

Figura 8.2: Estimació de costos de personal per tasca. Font: Elaboració pròpia.

8.1.2. Despeses generals

En aquest apartat calcularé els costos genèrics, els quals s'han de calcular de forma general perquè no depenen de les tasques del projecte. Consideraré com a costos generals les amortitzacions, l'espai de treball i el consum d'electricitat i d'internet.

- ***Software*:** tot el programari utilitza't és de codi obert i, per tant, gratuït. L'amortització total per *software* és de 0 €.

- ***Hardware*:** el *hardware* necessari per a aquest projecte consta de dos ordinadors portàtils de gamma mitjana, és a dir, d'un cost de 650 €. Suposant que, en mitjana, el temps de vida d'un ordinador d'aquestes característiques és d'uns cinc anys, l'amortització d'aquests recursos és de 5/60 * 650 = 54,16 €.

- **Espai de treball:** l'espai necessari per a desenvolupar el projecte és una habitació per a treballar i una sala per a reunions. En aquest cas faré ús l'habitació de la meva casa i una sala de reunions del Clínic. El preu d'una habitació similar a la meva a Barcelona és d'uns 400 €/mes i el d'una sala de reunió com la del Clínic és d'uns 16 €/h. En total les despeses en espai de treball són de 400*5 + 16*35 = 2560 €.

- **Factura d'internet:** segons les factures personals la tarifa d'internet és d'uns 43 € mensuals. Amb una jornada laboral de 4 hores diàries la despesa en internet seria d'un total de 5*43*4/24 = 35,8 €.

- **Consum elèctric:** donat que el preu de l'electricitat actualment està en 0.4029 €/kWh [21] i tenint en compte que el consum d'un portàtil en 450 hores de duració és de 13,5 kWh. Podem concloure que la despesa en electricitat seria de 0.4029*13,5*2 = 10,88 €.

Una vegada desglossats els costos genèrics podem calcular el cost total d'aquest apartat. Aquest càlcul es mostra en la figura 7.3.

Concepte	**Cost (€)**
Software	0
Hardware	54,16
Espai de treball	2560
Factura d'internet	35,8
Consum elèctric	10,88
Total CG	**2660,84**

Figura 8.3: Resum de costos genèrics. Font: Elaboració pròpia.

8.2. Estimació dels costos

Finalment per acabar amb l'estimació dels costos hem de tenir en compte els imprevistos i la contingència. La contingència en aquest sector ronda el 15% el que es traduiria en 1687,56 €. Per la banda dels costos per imprevistos es pot observar el càlcul en la figura 7.4. I finalment a la taula 7.5 tenim resumits els costos totals del projecte.

Risc	**Probabilitat**	**Temps estimat (h)**	**Cost estimat (€)**	**Cost esperat (€)**
Error de disseny	40%	20	312	124,8
Error d'implementació	70%	30	468	327,6
Gestió del temps	15%	10	218	32,7
Qüestions personals	10%	10	218	21,8
Resultats imprevistos	10%	10	218	21,8
Total				**532,7**

Figura 8.4: Taula amb els costos de imprevistos estimats. Font: Elaboració pròpia.

Concepte	**Cost (€)**
Cost de personal per activitat (CPA)	8590,4
Cost general (CG)	2660,84
Contingència	1684,56
Imprevistos	532,7
Total	**13468,5**

Figura 8.5: Taula dels costos estimats pel projecte. Font: Elaboració pròpia.

8.3. Control de gestió

Durant el transcurs del treball es portarà un control sobre les possibles desviacions en el pressupost. L'objectiu és assegurar que la diferència entre els costos reals i els estimats sigui la mínima. Per fer aquest control s'utilitzaran les mètriques següents:

DC = (CR - CR) * CHR
DCS = (CHE - CHR) * CE

on:

- **DC** = Desviació del cost
- **DCS** = Desviació del consum
- **CE** = Cost estimat
- **CR** = Cost real
- **CHR** = Consum d'hores real
- **CHE** = Consum d'hores estimat

A l'etapa final del projecte farem aquests càlculs amb la finalitat d'obtenir els desviaments. D'aquesta forma podrem fer una anàlisi del pressupost final i determinar si caldria fer ús de la contingència calculada anteriorment. Tanmateix, si les desviacions fossin degudes als imprevistos, podríem utilitzar la partida dels imprevistos per cobrir-ne el cost.

9. Informe de sostenibilitat

9.1. Autoavaluació

Aquests quatre anys a la Facultat d'Informàtica de Barcelona m'han fet conscient de les implicacions positives i negatives que pot tenir la informàtica sobre la vida de les persones, el medi ambient i l'economia global. La sostenibilitat sempre ha sigut un punt a tenir en compte en les diferents assignatures de la carrera, sobretot en la dimensió ambiental. En aquest punt hem valorat els residus de *hardware* i els cementiris tecnològics existents a països subdesenvolupats on també causa un gran impacte social.

A l'assignatura d'aspectes mediambientals de la informàtica (ASMI) vam treballar tota mena de visions sostenibles del sector, des d'una visió ètica fins a una de més mediambiental, passant també per les conseqüències econòmiques que pot tenir un projecte tecnològic. Vam treballar aquests factors a partir de les valoracions de projectes ficticis.

On vam veure amb més detall l'aspecte econòmic va ser a l'assignatura d'Empresa i entorn econòmic (EEE). En aquesta assignatura vam treballar l'entorn socioeconòmic on es desenvolupa l'activitat de les empreses. Va ser de gran ajuda per fer-me una idea de com funcionava l'economia d'un país.

Finalment, cal destacar també, que nombroses assignatures durant el grau destinaven una part del seu temari a la sostenibilitat, fent ús de les competències transversals. Això va fer que tingués més present aquest tema en els diferents projectes que he hagut de dur a terme.

9.2. Dimensió econòmica

Has estimat el cost de la realització del projecte (recursos humans i materials)?

Abans de dur a terme qualsevol mena de projecte s'ha de veure si aquest és viable econòmicament. En aquest cas hem fet un pressupost de les despeses que comportaria aquest projecte. Com es tracta d'un treball de fi de grau les despeses no són reals, però m'he basat en salaris i costos mitjans al sector tecnològic que he trobat per internet.

Com es resol actualment el problema que vols abordar (estat de l'art)? En què millorarà econòmicament la teva solució a les existents?

Com ja he comentat anteriorment, l'equip de recerca en oncologia de l'Hospital Clínic treballa contínuament amb estudis estadístics. Com no hi ha informàtics com a tal a l'equip, treballen amb eines bàsiques com Excel. Aquestes eines no els hi permeten fer estimacions i representar les dades de forma òptica. Per tant, la nostra aplicació els ajudarà en aquest sentit i els hi facilitarà la feina. Aquesta millora es traduirà conseqüentment en una millora econòmica, ja que augmentarà l'eficiència de l'equip.

9.3. Dimensió ambiental

Has estimat l'impacte ambiental que tindria la realització del projecte? T'has plantejat minimitzar l'impacte, per exemple, reutilitzant recursos?

L'aspecte mediambiental sempre ha sigut un punt central del projecte. Si més no, la contaminació és una de les causes del càncer de pulmó. I com ja he comentat amb anterioritat, aquest projecte estudiarà l'evolució del càncer de pulmó i també la dels efectes mediambientals. Per una altra banda, ens trobem en un tema en el qual no s'ha fet massa recerca i, per tant, era difícil reutilitzar altres eines.

Com es resol actualment el problema que vols abordar (estat de l'art)? En què millorarà ambientalment la teva solució a les existents?

El fet que l'aplicació no només estigui encarada a l'estudi del càncer, sinó a la relació entre el càncer i els diferents agents mediambientals. Deixa clar que aquesta eina també podrà ser utilitzada per estudiar l'evolució de la contaminació en els diferents països d'Europa. Per aquest motiu pot ser una eina emprada per millorar ambientalment els països d'Europa.

9.4. Dimensió social

Què creus que t'aportarà en l'àmbit personal la realització d'aquest projecte?

Aquest projecta m'ha d'enriquir tant professionalment com personalment. No solament m'ensenyarà a com realitzar un producte útil per a una associació externa, sinó també estaré tractant un tema molt delicat, el qual afecta molta gent. Aquest fet m'està fent valorar el treball que duc a terme i m'està ensenyant que cadascú decideix a quina finalitat destina el seu coneixement. I és un fet que destinar el meu treball de fi de grau a la lluita contra el càncer serà un punt d'inflexió en l'àmbit personal i professional.

Com es resol actualment el problema que vols abordar (estat de l'art)? En què millorarà socialment (qualitat de vida) la teva solució a les existents?

En l'actualitat, l'equip es troba en dificultat per donar veu al seu estudi a causa que no es destina massa pressupost a aquestes investigacions. Això porta al fet que l'equip intueixi que hi ha aspectes mediambientals que sembla que tenen efecte en el càncer de pulmó, però que no es pot demostrar. Aquest projecte ajudarà a solucionar aquest fet i, per tant, l'objectiu final és que es pugui reduir la incidència del càncer de pulmó a Europa provocada per aquests factors. Clarament, això millorarà la qualitat de vida de molts ciutadans.

Existeix una necessitat real del projecte?

Tal com he comentat en l'apartat anterior, aquest projecte va sorgir d'una necessitat de l'Hospital Clínic. L'equip d'investigació necessitava una eina efectiva per estudiar relacions entre causes externes i càncer de pulmó. Això justifica la necessitat real de l'aplicació que durem a terme.

10. Estudi previ

Com ja he comentat en apartats previs, aquest projecte és multidisciplinari. Una part important en aquests tipus de treballs, en què s'apliquen els coneixements informàtics a una disciplina totalment diferent, és l'estudi del camp que hem de tractar. Per aquesta raó, en aquest capítol es descriuen els conceptes més rellevants que s'han de conèixer per poder realitzar una aplicació que treballa amb dades de càncer de pulmó i contaminació. Molts d'aquests conceptes els he obtingut de l'assessoria tant de la Dra. Laura Mezquita com del seu equip.

10.1. Carcinògens

El càncer de pulmó és un càncer que es forma en els teixits del pulmó, generalment en les cèl·lules que recobreixen els conductes d'aire. És la principal causa de mort per càncer tant en homes com en dones. En termes generals, els factors ambientals són responsables de la gran part de càncers [22]. Tanmateix, el fet que es tracti d'una malaltia que afecta principalment a l'aparell respiratori fa que aquesta pugui estar produïda i empitjorada per efectes ambientals com la contaminació de l'aire.

Un carcinogen es defineix com un agent físic, químic o biològic potencialment capaç de produir càncer [23]. Existeixen centenars de carcinògens pel càncer de pulmó amb suport científic. Aquesta és una petita llista dels que més afecten en la malaltia:

- **Tabac:** El fum del tabac és una mescla de més de 7000 substàncies químiques, la majoria d'elles tòxiques. Està demostrat que almenys 70 d'aquestes poden causar càncer en persones. A més les persones que fumen habitualment es calcula que són entre 15 i 30 vegades més propens a contreure càncer de pulmó, segons el centre per al control i la prevenció de malalties (CDC) [24]. També s'ha registrat el fum de segona mà com a carcinogen. El risc de càncer de pulmó atribuïble a exposició passiva a tabac a Europa s'ha estimat en un 1.6% per l'Agència Internacional d'Investigació contra el Càncer (IARC).

- **Contaminació atmosfèrica:** Els contaminants sobre els quals se solen basar els estudis són: les partícules en suspensió de diàmetre inferior a 10 μm (PM10) i inferior a 2.5 μm (PM2.5) com a indicador de les partícules de menor grandària, el diòxid de nitrogen, el diòxid de sofre, el monòxid de carboni i l'ozó. També es té informació sobre els nivells de compostos orgànics volàtils com el benzè o de metalls que formen part de les partícules, però amb un grau de cobertura poblacional/geogràfica molt de menor. La majoria d'estudis i revisions suggereixen un risc relatiu entorn d'1.3-1.5 per a càncer de pulmó quan es comparen zones amb nivells alts en contra de zones amb baixos nivells de contaminació. A Europa s'ha proposat que els nivells mitjans de PM2.5, entorn de 15 $\mu g/m^3$, podrien suposar fins al 10% dels càncers de pulmó. En no fumadors o exfumadors de més de deu anys de residència prop de carreteres d'alta densitat de trànsit o en àrees amb nivells de NO2 superiors a 30 $\mu g/m^3$, s'ha estimat un risc atribuïble entorn del 5-7%. Tanmateix, el grau d'evidència respecte a altres tipus de càncers és limitat i no permet extreure conclusions. [25]

- **Contaminació interior:** Els principals contribuïdors a la contaminació en interiors són el gas radó i l'amiant. En primer lloc, el radó és un element químic natural, incolor, inodor, estable químicament i que prové de la degradació radioactiva de l'urani-238. Per la seva elevada densitat tendeix a concentrar-se en les parts baixes dels edificis, com per exemple soterranis mal airejats. Actualment, no existeix evidència científica sobre la relació entre el radó i el risc de càncer. Tot i que Europa s'ha estimat que entre un 4-5% dels càncers de pulmó podria deure's a l'exposició a radó en ambients interiors, segons explica l'Institut nacional de càncer. El terme amiant designa diversos tipus d'asbestos, minerals amb forma filaments flexibles i poc resistents, que químicament es corresponen amb silicats hidratats de magnesi, ferro i altres metalls. L'amiant s'ha utilitzat per a la confecció de vestits resistents a la calor, el recobriment de calderes i equips elèctrics, i la fabricació de materials per a la construcció com el fibrociment (Uralita). L'exposició per via respiratòria a pols d'asbest entre els treballadors de la mineria d'aquest mineral és una causa establerta de càncer de pulmó. [25]

- **Exposició a radiació:** La radiació, d'unes certes longituds d'ona, anomenada radiació ionitzant, té suficient energia per a danyar l'ADN i causar càncer. La radiació ionitzant

> inclou radó, raigs X, raigs gamma i altres formes de radiació d'alta energia. Les formes de radiació d'energia més baixa, no ionitzant, com la llum visible i l'energia dels telèfons cel·lulars, no s'ha trobat que causin càncer en les persones. [25]

Finalment, cal destacar que els carcinògens poden ser sinèrgics. Això vol dir que l'aparició de dos carcinògens que actuen de forma simultània potencia els seus efectes. Per exemple, la contaminació ambiental, el tabac i el radó són carcinògens sinèrgics, per tant, una persona que s'exposa a aquests tres factors és molt més probable que pateixi càncer de pulmó. Això fa encara més important l'estudi de tots els carcinògens i no només del tabac.

10.2. Avanços i descobriments

Fa uns trenta anys s'estudiava el càncer de pulmó com una malaltia única. Això feia que els tractaments fossin generals per a tots els pacients. En conseqüència hi havia pacients que toleraven molt bé el tractament i els hi funcionava, mentre que per altres no s'obtenien bons resultats. Fruit de la investigació i l'estudi ara es concep la malaltia com diverses variacions amb diversos tractaments. Això és rellevant per al projecte, ja que alguns projectes paral·lels del Clínic, que seran explicats al següent punt, treballen per demostrar que algunes variants de la malaltia són més sensibles a segons que carcinògens.

Per un costat es pot dividir el càncer de pulmó en dos tipus, el de cèl·lules petites i el de cèl·lules no petites. Mentre que el primer només representa entre el 10% i el 15% de tots els tumors malignes del pulmó, el de cèl·lules no petites representa la resta. Es pot veure amb més profunditat la divisió a la Figura 10.1. Aquesta divisió s'anomena classificació histològica i es va començar a utilitzar fa quinze anys.

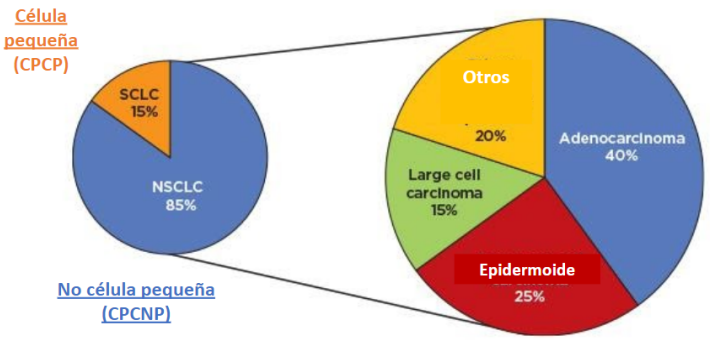

Figura 10.1: Primera divisió en tipus de càncer de pulmó. Font: Dra. Laura Mezquita

Ara per ara, es fa servir una subdivisió molt més acurada. Es tracta d'una classificació molecular, en particular, ara és possible definir subgrups d'adenocarcinoma segons la presència de mutacions específiques en els gens [22]. Aquest fet provoca variacions en els

tractaments segons els perfils moleculars de cada pacient. La Figura 10.2 mostra els diferents perfils moleculars que es treballen actualment.

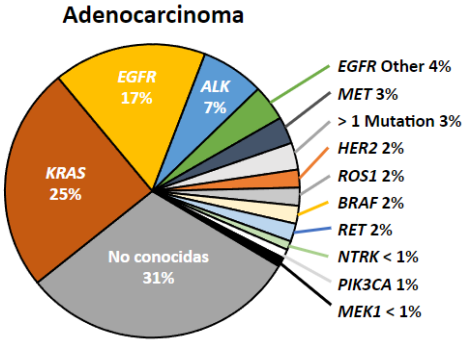

Figura 10.2: Estimació de percentatges per als perfils molecular coneguts. Font: Dra. Laura Mezquita

10.3. Projectes paral·lels

Durant l'estada a IDIBAPS he pogut participar en diversos projectes d'investigació que han pogut fer suport a aquest treball. Alguns d'ells són complementaris i altres m'han servit per ampliar els meus coneixements en el tema. Aquests projectes són els següents:

- **MIRROR:** Estudia si la radiació ionitzant del radó pot provocar mutacions en l'ADN. Paral·lel a aquest estudi ha sorgit un altre anomenat **Radon Europe** el qual estudia directament l'afectació del radó en els pacients de càncer de pulmó. Aquests projectes estan sent elaborat per la Marta Garcia De Herreros.

- **INIVATA:** Estudia la utilitat de biòpsia líquida en temps real per a guiar el tractament de nous pacients de càncer. Aquest projecte està sent elaborat per la Marta Garcia De Herreros.

- **Medicina de precisió:** S'encarreguen de caracteritzar al pacient en molts aspectes, incloent-hi el perfil molecular, la predisposició genètica i l'exposició a carcinògens. En aquest últim punt hem pogut treballar conjuntament proporcionant dades. Aquest projecte està sent elaborat per en Javier Torres Jiménez.

11. Base de dades

Un dels objectius principals del projecte era obtenir una base de dades prou rica per poder fer anàlisis de dades amb rigor. Les dades a assolir es divideixen en dos tipus. En primer lloc, les dades de càncer, en aquest cas ens interessa tenir dades de mortalitat i incidència en el nombre màxim de països d'Europa durant cada any de les 3 últimes dècades. Per l'altra banda tenim la concentració dels principals agents contaminants durant el mateix període. L'obtenció d'aquestes dades no és una tasca fàcil de realitzar, ja que ens hem d'assegurar de

què les dades són perfectament fiables i de què seran aptes per a ser utilitzades en la nostra aplicació. Per aquesta raó he dividit la tasca en dues subtasques: extracció i preprocessament.

11.1. Extracció

El primer pas en aquest apartat és el de cerca i extracció de dades. En aquest cas és important que les fonts siguin fiables i les mínimes possibles. L'objectiu és extreure la informació de càncer d'una mateixa font i la de contaminació d'una altra, si fos possible. D'aquesta forma ens assegurem la concordança entre les dades. L'Organització Mundial de la Salut (OMS) [4] i l'Agència Mediambiental Europea (EEA) [26] seran de gran importància en aquesta tasca.

Inicialment, l'Organització Mundial de la Salut ofereix una base de dades completa amb el sumatori de morts per país dividit entre els diferents motius de la defunció. Aquestes estadístiques procedeixen dels sistemes de registre civil de cada país. Quan es produeix una defunció, aquest es registra en el registre civil local amb informació sobre la causa de la mort. A continuació, l'autoritat nacional recopila la informació i la presenta a l'OMS cada any. L'OMS només publica les morts certificades mèdicament. Aquest procés ens permet verificar les dades obtingudes. A més de les taules a la web de l'OMS també es pot trobar documentació diversa que facilitarà la lectura de les dades i el posterior preprocessament del qual parlaré en apartats següents.

Per la part de les dades d'incidència de càncer de pulmó es poden aconseguir en una base de dades també extreta de l'OMS. Aquesta base de dades es diu CI5plus [27] i conté taxes d'incidència anual actualitzades per a 124 poblacions seleccionades de 108 registres de càncer publicats en CI5, per al període més llarg disponible (fins a 2012), per a tots els càncers. Segons informa la mateixa OMS en la seva web, aquesta base de dades pot utilitzar-se per a analitzar les tendències temporals.

En el cas de la contaminació atmosfèrica, l'Agència Europea de Medi Ambient (AEMA) és una agència de la Unió Europea la comesa de la qual és proporcionar informació sòlida i independent sobre el medi ambient. Aquesta organització exposa una sèrie de conjunts de dades entre els quals es troba la base de dades sobre la qualitat de l'aire. Aquesta consisteix en una sèrie temporal pluriennal de dades de mesurament de la qualitat de l'aire i d'estadístiques calculades per a una sèrie de contaminants atmosfèrics. També conté informació sobre les xarxes de vigilància implicades, les seves estacions i mesuraments, les tècniques de modelització de la qualitat de l'aire, així com les zones de qualitat de l'aire, els règims d'avaluació, els nivells de compliment i els plans i programes de qualitat de l'aire notificats pels estats membres de la UE i els països de l'Espai Econòmic Europeu.

Finalment per a les dades de radó no he aconseguit registres públics. Tenim constància de què hi ha estudis que han recollit aquestes dades. Tanmateix, he demanat aquestes dades , però

demanaven el permís de cadascun dels països implicats. Per aquest motiu he decidit utilitzar una taula de risc de radó proporcionada pel projecte RadonEurope desenvolupat per la Marta Garcia De Herreros, integrant de l'equip d'investigació del Clínic.

11.2. Preprocessament

El fet d'extreure les dades de fonts diverses fa que aquestes requereixin un tractament per poder ser visualitzades. A més, moltes de les taules descarregades de les fonts tenien dades extres que no seran necessàries per al projecte. Per tant, també s'hauria de fer un treball de filtratge. Tots aquests passos seran descrits en aquest capítol.

Les eines fetes servir pel processament de les dades han sigut bàsicament un conjunt de scripts elaborats per mi mateix. Aquests scripts fan servir Pandas [13], una llibreria de python que facilita el tractament de conjunts de dades, i es poden trobar seguint els passos indicats als annexos d'aquesta memòria.

11.2.1. Mortalitat

En el cas de les dades de mortalitat, aquestes venien separades en 8 taules diferents. Cadascuna d'aquestes taules tenia un format diferent i contenia dades de mortalitat de càncer per a tots els càncers registrats i per a tots els països del món. En el nostre cas només ens interessava la mortalitat de càncer de pulmó. Afortunadament, les taules tenien un document amb llistes de codis per a facilitar el tractament de les dades.

El treball de preprocessament a realitzar era el següent:

1. Filtratge de dades per càncer de pulmó i països Europeus.
2. Traducció de codis de país.
3. Inserció de totes les dades en una sola taula per facilitar l'accés.
4. Càlcul de la taxa de morts per cada 100.000 habitants

Pas 1, 2. El treball consistia a llegir les llistes del document i apuntar els codis referents a càncer de pulmó i a països d'Europa. Per als codis de càncer vaig filtrar els codis mostrats a la Figura 11.1.

Taula	Codis
ICD7	A050, 162, 163, B018
ICD8	A051, 162, B019
ICD9	B101, C028, B10
ICD10 (p1-p5)	1034, UE15, C32, C33,C34

Figura 11.1: Codis per a càncer de pulmó per l'OMS. Font: Elaboració pròpia

Un exemple d'una fila d'una d'aquestes 8 taules podria ser el mostrat en la Figura 11.2. Per tant, el que s'havia de fer era seleccionar aquelles files on la columna Cause sigui igual a qualsevol dels codis representats a la Figura 11.1. Les columnes que ens interessaran per als càlculs, són les següents:

- **Country:** el codi del país on s'ha fet la mesura.
- **Year:** any de la mesura.
- **List:** referència a la llista on s'ha de consultar el codi de la causa de les morts.
- **Cause:** codi de la causa de les morts.
- **Sex:** 1 - homes, 2 - dones.
- **Deaths:** nombre de morts.

Country	Admin1	SubDiv	Year	List	Cause	Sex	Frmat	IM_Frmat	Deaths
4303	null	null	2017	101	1000	1	01	08	281784

Figura 11.2: Exemple de format de taula de mortalitat. Font: Elaboració pròpia.

El que vaig fer per filtrar els països Europeus va ser traduir la columna "Country" amb la llista de codis de país i amb una llibreria de python anomenada PyCountry. Aquesta llibreria comptava amb una funció que a partir del nom en anglès de cada país et retornava si era o no Europeu, cosa que va ser molt útil en aquest procés.

Pas 3. Per a fer el merge de totes les taules en una vaig utilitzar una funció *merge* que proporciona la llibreria Pandas.

Pas 4. Per poder comparar les dades de mortalitat entre països calia representar els valors de forma equitativa i en funció de la població de cada país, ja que els països amb més habitants naturalment tindran més morts. Per aquest motiu vaig decidir calcular la taxa de morts cada 100.000 habitants. Això requería aquest càlcul: Taxa = 100000*Morts/Habitants. Les dades poblacionals les vaig extreure de la mateixa web de l'OMS.

El format final de la taula amb tot el procés executat és el mostrat en la Figura 11.3.

STATE	YEAR	TOTAL	MALE	FEMALE
ES	2017	47.54	75.62	20.48

Figura 11.3: Exemple de format final de la taula de mortalitat. Font: Elaboració pròpia.

11.2.2. Incidència

Per a les dades d'incidència de càncer de pulmó, passava una cosa similar que per les taules de mortalitat. En aquest cas, afortunadament, vaig obtenir una sola taula amb totes les dades de països exclusivament europeus. Tanmateix, aquesta taula contenia el recompte de casos

per a tots els tipus de càncer i en un format diferent del desitjat. Per aquest motiu calia una tasca de preprocessament. En aquest cas també vaig poder utilitzar una documentació proporcionada per l'OMS que indicava com treballar amb les dades.

Els passos del preprocés van ser els següents:

1. Filtratge de dades per càncer de pulmó.
2. Traducció de codis de país.
3. Càlcul de la taxa d'incidència per cada 100.000 habitants.

Pas 1, 2. Aquestes dues tasques són anàlogues a les de l'apartat anterior. La diferència és que en aquest cas l'estructura de la taula és diferent i , per tant, s'hauran de tractar les dades de forma diferent. La Figura 11.4 exemplifica el format de la taula. Les columnes representen:

- **Registry**: Codi de país.
- **Year**: Any de la mesura.
- **Sex**: 1 - homes, 2 - dones.
- **Cancer**: Codi del càncer.
- **Total**: nombre de casos totals.

Per a filtrar les files que donaven informació sobre càncer de pulmó calia buscar aquelles amb codi de càncer entre 46 i 54, ambdós inclosos.

Registry	Year	Sex	Cancer	Total
4000000	1998	1	1	18409

Figura 11.4: Exemple de format de taula d'incidència. Font: Elaboració pròpia

Pas 3. Aquest pas és exactament igual al Pas 3 de l'apartat anterior.

El resultat del procés de tractament de les dades és el mostrat a la Figura 11.5.

STATE	YEAR	TOTAL	MALE	FEMALE
DK	2004	72.66	80.22	65.26

Figura 11.5: Exemple de format final de la taula d'incidència. Font: Elaboració pròpia.

11.2.3. Contaminació exterior

La base de dades de contaminació exterior comptava amb 4243349 mostres. Per sort, la web compta amb una eina de filtratge [30]. Fent servir aquesta eina vaig aconseguir descarregar una taula per a cada agent contaminant d'interès per a l'estudi. Els contaminants que vam escollir, amb ajuda i confirmació de l'equip d'investigació del Clínic, van ser els següents:

- **PM2.5:** partícules en suspensió de menys de 2.5 micres. En bona part provenen de les emissions dels vehicles dièsel a la ciutat. D'altra banda, els efectes que tenen sobre la nostra salut són molt greus, per la seva gran capacitat de penetració en les vies respiratòries. [31]

- **PM10:** partícules en suspensió de menys de 10 micres. Les partícules poden ser de pols, cendra, sutge, partícules metàl·liques, ciment o pol·len. Els nivells alts d'aquestes, tenen un efecte advers sobre la salut, provocant malalties greus, com poden ser el càncer de pulmó, altres malalties del sistema respiratori, i malalties del sistema circulatori. [31]

- **O3:** L'excés d'ozó en l'aire pot produir efectes adversos importants sobre la salut humana. Pot causar problemes respiratoris, provocar asma, reduir la funció pulmonar i donar lloc a malalties pulmonars. [31]

- **NO2:** Estudis epidemiològics han revelat que els símptomes de bronquitis en nens asmàtics augmenten en relació amb l'exposició prolongada al NO2. La disminució del desenvolupament de la funció pulmonar també s'associa amb les concentracions de NO2 registrades (o observades) actualment en ciutats europees i nord-americanes. [31]

- **SO2:** El SO2 pot afectar el sistema respiratori i a les funcions pulmonars, i causa irritació ocular. La inflamació del sistema respiratori provoca tos, secreció mucosa i agreujament de l'asma i la bronquitis crònica; així mateix, augmenta la propensió de les persones a contreure infeccions del sistema respiratori. Els ingressos hospitalaris per cardiopaties i la mortalitat augmenten en els dies en què els nivells de SO2 són més elevats. [31]

Aquestes 4 taules tenien el mateix format. Un exemple de fila d'una d'aquestes taules pot ser el mostrat en la Figura 11.6.

Country	Air pollutant	Description	Year	Unit	Air pollutant level
Germany	PM2.5	Particulate matter < 2.5 µm (aerosol)	2004	ug/m3	12.229

Figura 11.6: Exemple de format de taula de contaminació. Font: Elaboració pròpia

Per a procedir a fer el tractament d'aquestes dades calia desenvolupar les següents tasques:

1. Càlcul de la mitjana de totes les mesures d'un mateix any.
2. Unió de totes les taules de cada contaminant en una.
3. Eliminació de files sense dades significants.

4. Càlcul de les coordenades centrals de cada país.

Pas 1. Cada fila de les taules correspon a una mesura donada per una estació amb la mitjana de la concentració del contaminant corresponent durant un any. Per al projecte ens era més interessant tenir una única mostra per any i país. Per tant, vaig fer una mitjana entre les dades de totes les estacions de cada país durant cada any.

Pas 2. La unificació de totes les taules en una es va fer utilitzant la funció *merge* de Pandas explicada en apartats anteriors.

Pas 3. Producte del pas anterior van sorgir files a la taula sense dades de contaminants. El que vaig fer va ser simplement filtrar aquestes files i eliminar-les.

Pas 4. Per a la representació dels contaminants en un mapa és molt útil tenir la latitud i la longitud del centre de cada país. Això ho vaig obtenir fent servir una llibreria de python anomenada Geopy [33] que proporciona dades geogràfiques.

El resultat del procés de tractament de les dades és el mostrat a la Figura 11.7.

STATE	YEAR	CO	NOX	PM10	PM2.5	Lat	Lon
LU	2004	0.41	68.73	16.27	13.27	50.86	22.77

Figura 11.7: Exemple de format final de la taula de contaminació. Font: Elaboració pròpia.

11.2.4. Radó

Finalment per l'apartat de la contaminació en interiors, la taula proporcionada pel projecte Radon Europe tenia informació de la mitjana i de percentatge de nivells perillosos per al desenvolupament de càncer. El format de les dades era el mostrat a la Figura 11.8.

Contry	Rad_AM	Rad_GM	Rad%200	Rad%400
Belgium	29	22	2.20%	0.50%

Figura 11.8: Exemple de format de taula de radó. Font: Elaboració pròpia

Pel cas del radó era més important tenir un valor qualitatiu del risc per radó, per exemple risc alt, baix o mitjà. En el moment del processament de les dades vam pensar que era interessant fer l'anàlisi dels riscos basant-nos en tres criteris: la mitjana aritmètica (Rad_AM), el percentatge de valors per sobre dels 200 Bq/m³ i el percentatge de valors per sobre dels 400 Bq/m³. Per a la classificació per riscos vaig utilitzar l'algoritme de k-Means que explicaré en l'apartat de desenvolupament de l'aplicació.

El resultat del procés de tractament de les dades és el mostrat a la Figura 11.9.

STATE	AM	RIESGO	%200	%400
CZ	140	Alto	Alto	Medio

Figura 11.9: Exemple de format final de la taula de radó. Font: Elaboració pròpia.

12. Disseny de l'aplicació

Una vegada obtingudes les dades, el següent pas en el projecte és el disseny de l'aplicació. En aquest apartat detallaré les diferents decisions de disseny que he pres abans del procés de desenvolupament de l'aplicació web.

12.1. Estructura del codi

Per aconseguir que un projecte sigui escalable i entenedor una bona pràctica sempre és establir una estructura lògica per a la distribució dels fitxers. En aquest cas, com es tracta d'un treball en equip això encara agafa més importància.

L'estructura de carpetes bàsica consta de 6 directoris principals:

- **Assets:** conté els diferents recursos visuals que necessitem per a l'aplicació. Exemples del contingut d'aquest directori són mapes, imatges, logos, estils css…

- **Services:** aquest directori recull una sèrie de fitxers amb funcions utilitzades durant tot el procés d'execució de l'aplicació. Per exemple, hi ha fitxers amb funcions dedicades a la creació de gràfiques. És important tenir separades aquestes peces de codi per a no repetir-les contínuament.

- **Model:** conté els models de dades fets servir a l'aplicació. Aquestes classes s'encarreguen de connectar l'aplicació amb la base de dades, carregar i guardar la informació en les estructures de dades adients.

- **Pages:** en aquest directori es guarden els arxius que generen les diferents pàgines que conté el web.

- **Components:** conté diferents components personalitzats de dash que seran fets servir com a parts d'una pàgina. Els posem en arxius diferents per evitar la repetició de codi.

- **Test:** guarda els arxius de proves que ens serveixen per assegurar-nos de què els algoritmes funcionen de forma adient. A més ens permeten clarificar com hem emprat algunes llibreries i ajuda per la comunicació entre programadors.

Finalment, tenim un fitxer anomenat app.py que s'encarrega de posar a punt el web i d'executar el procés principal.

12.2. Interfície d'usuari

Una part fonamental del disseny d'un web és el disseny de l'estructuració o maquetació de cadascuna de les pàgines. Aquesta estructura serà vital per a obtenir una bona interfície visual i una posterior bona experiència d'usuari. Cal tenir en compte que, tot i que la idea és que el web sigui públic, aquest està pensat per a la recerca professional. Per aquest motiu l'objectiu de la interfície és que sigui funcional, ràpida i molt fàcil d'utilitzar. També cal destacar que aquesta versió podrà ser redissenyada en un futur, ja que la finalitat és que s'ajusti a les necessitats de l'Hospital Clínic i, per tant, es faran diverses proves d'usabilitat.

L'estructura del web està dividida en diverses pàgines de les quals parlaré a continuació.

En primer lloc, tenim l'**estructura general** de la web. Aquesta està composta per les eines principals que hauran de fer de la navegació per la web una cosa trivial. Per una banda, tenim una barra lateral que donarà accés a tots els serveis de l'aplicació. En el nostre cas els serveis són: un resum inicial (pàgina d'inici), l'apartat destinat a l'estudi d'Europa, l'apartat destinat a l'estudi d'Espanya i finalment la pàgina de descàrrega d'informació. Una altra part important per a la navegació són els breadcrumbs o molles de pa. Aquestes "marquen" el camí que s'ha de seguir, a través de l'arquitectura del lloc web, per a arribar des de la pàgina principal del lloc web (això és, la portada) fins a la pàgina que està visualitzant l'usuari. Per exemple, l'estructura d'un rastre de molles de pa (breadcrumbs) tindrà un format semblant al següent: Portada > Pàgina nivell 1 > Pàgina nivell 2 > … > Pàgina actual. L'últim element de l'estructura de la pàgina web és el footer. El footer és la part inferior d'una pàgina web, en la qual s'inclou una sèrie d'elements que poden resultar d'interès per a l'usuari que navega per ella, com a enllaços a les categories principals, informació de contacte, xarxes socials… La maqueta amb aquest disseny és la mostrada a la Figura 12.1.

Figura 12.1: Maqueta de l'estructura de la pàgina. Font: Elaboració pròpia.

Una altra secció a definir del web és la portada o **pàgina d'inici**. En aquesta pàgina hem decidit, amb la recomanació de l'equip, que podem presentar un resum dels diferents projectes amb imatges per cridar l'atenció dels possibles interessats en el tema. A més, com sol ser habitual, hem afegit una secció de contacte on podem posar informació de l'equip d'oncologia del clínic i dels creadors de l'eina. La maqueta amb aquest disseny és la mostrada a la Figura 12.2.

Seguidament, podem definir la pàgina que apareix a l'hora de seleccionar un estudi. L'estudi Europa té tres parts diferencials: la pàgina d'estudi de càncer de pulmó, la pàgina d'estudi de qualitat de l'aire i la pàgina de comparació entre els dos estudis anteriors. Per aquest motiu es requereix una **pàgina prèvia de selecció**. En aquest cas, he decidit estructurar la pàgina de selecció amb 3 targetes amb informació de cada apartat i amb una imatge descriptiva. Un exemple de l'estructuració d'una d'aquestes pàgines és la Figura 12.3.

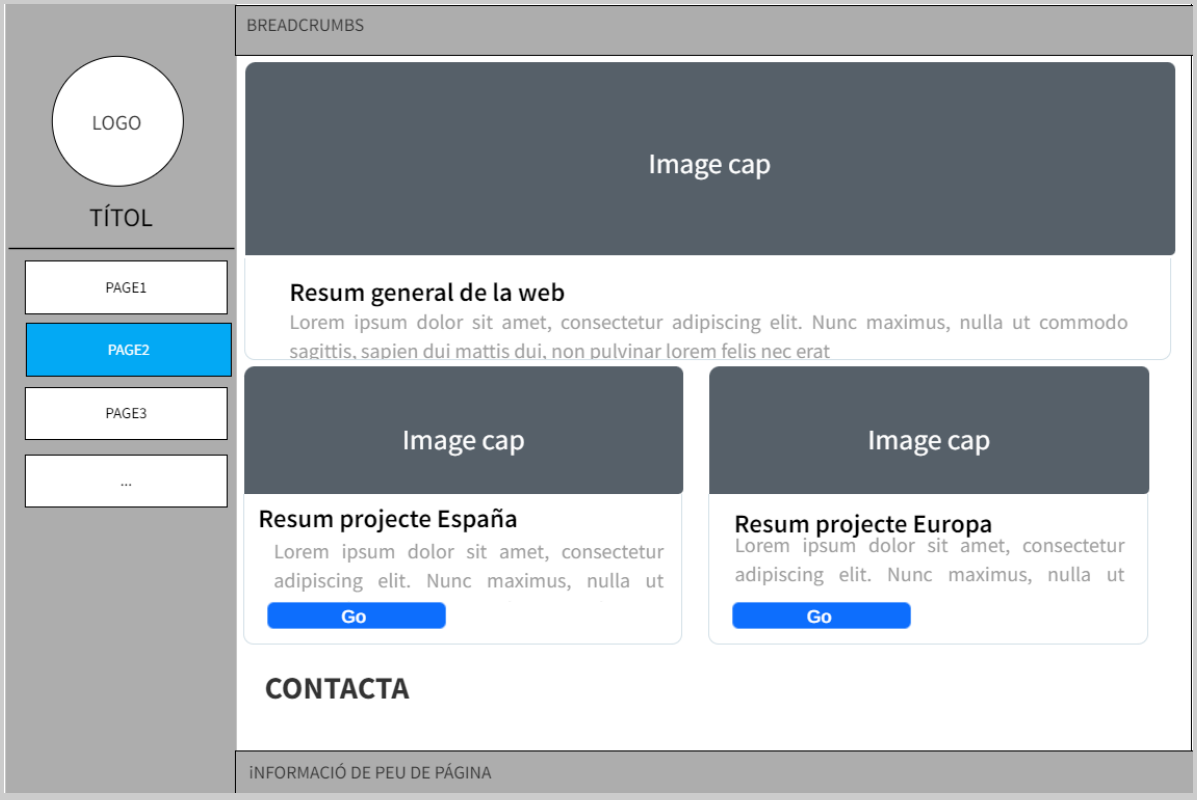

Figura 12.2: Maqueta de la pàgina d'inici. Font: Elaboració pròpia.

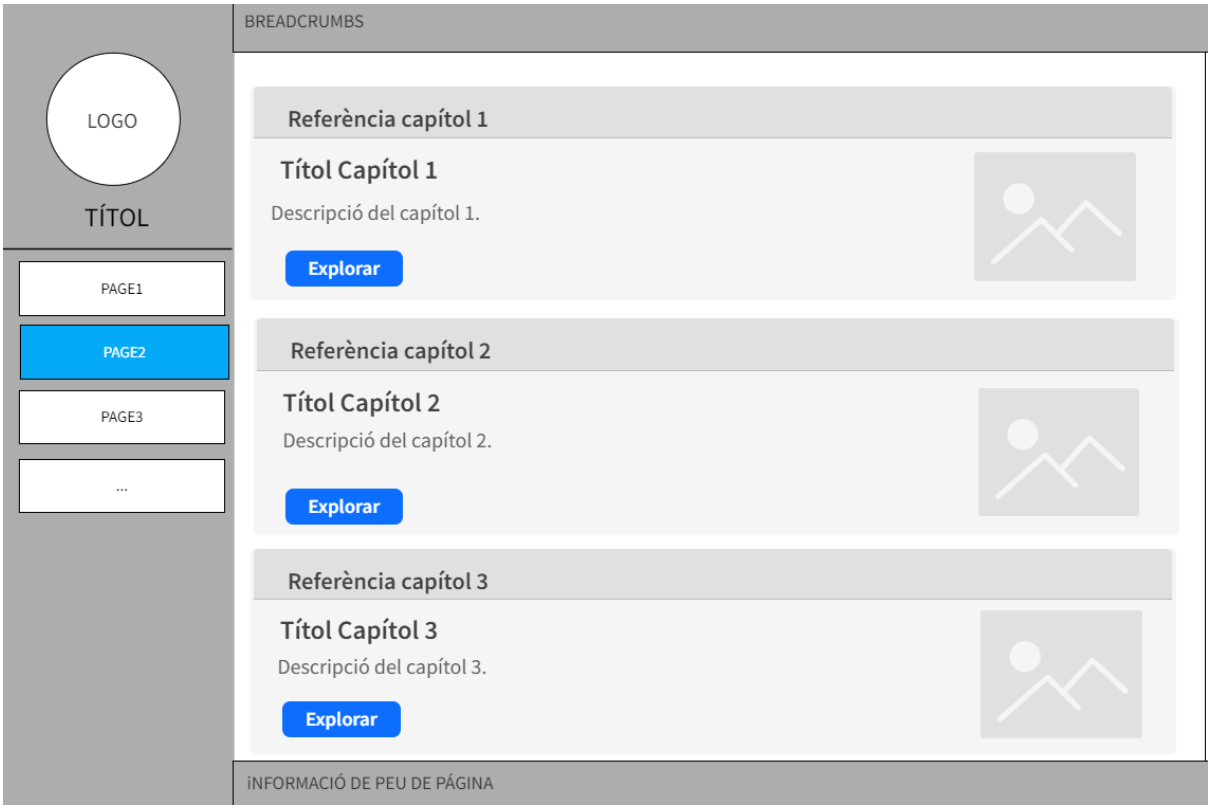

Figura 12.3: Maqueta de la pàgina de selecció. Font: Elaboració pròpia.

Per part de les **pàgines d'anàlisis de dades**, aquestes estaran formades de diverses formes de visualitzar les dades. Per una part tenim els mapes i per l'altra les gràfiques. En els dos casos es despondrà de desplegables per a filtrar per anys, país i tipus de dada. Es pot veure el disseny d'aquests dos conceptes a les Figures 12.4 i 12.5.

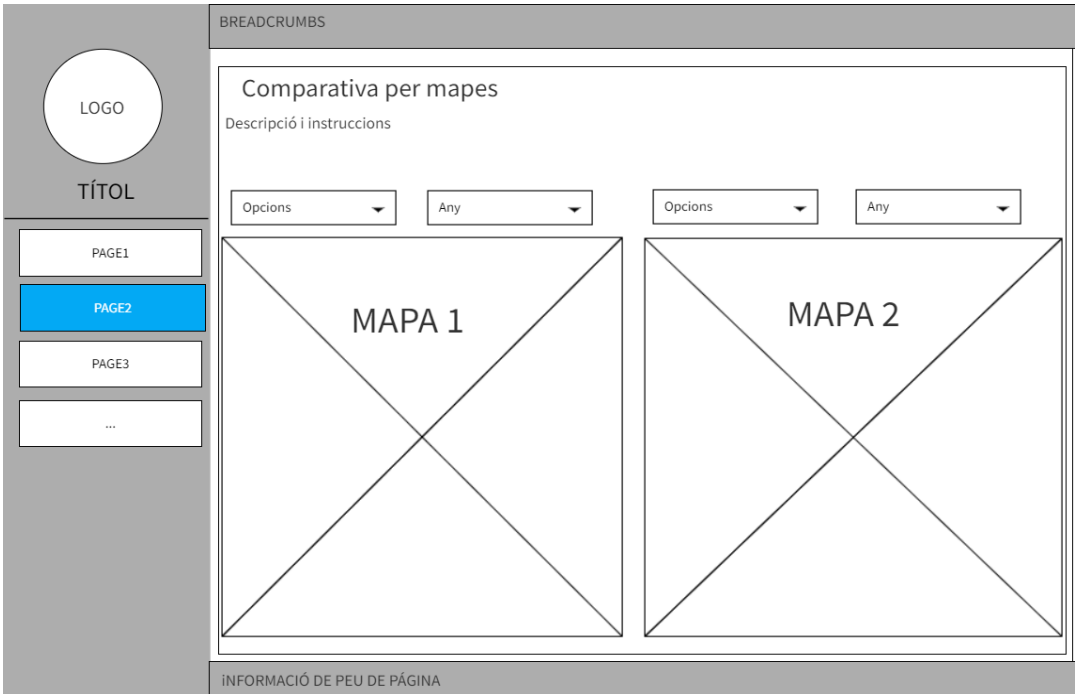

Figura 12.4: Maqueta de la comparativa entre mapes. Font: Elaboració pròpia.

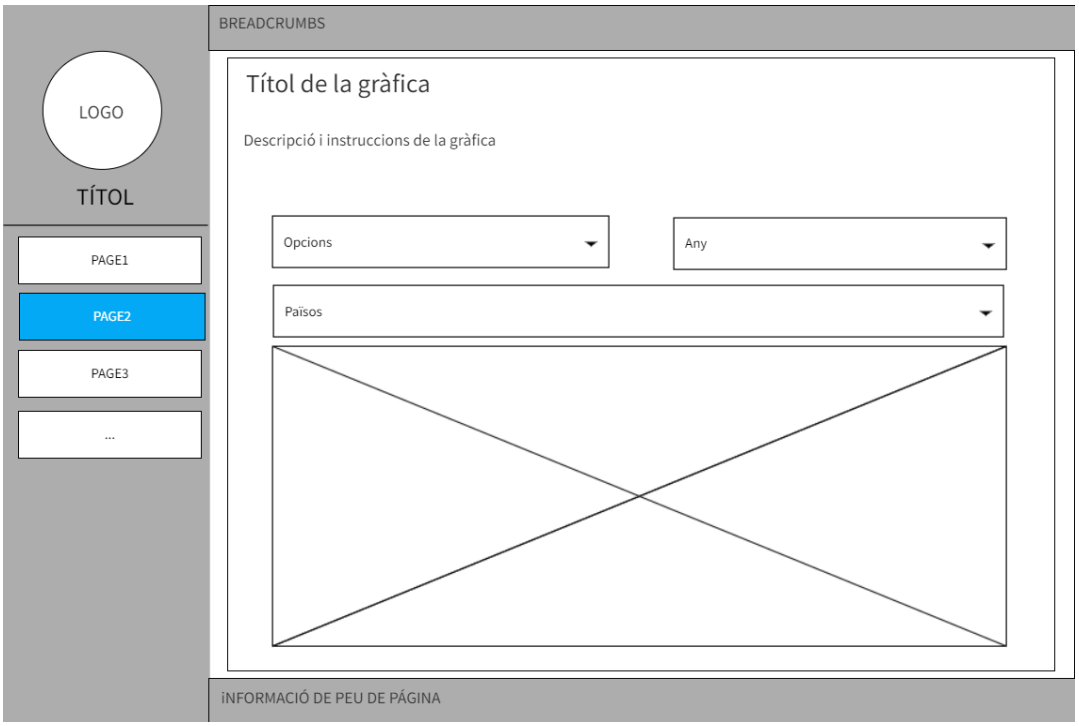

Figura 12.5: Maqueta d'una gràfica. Font: Elaboració pròpia.

Per últim, cal dissenyar la **pàgina de descàrregues**. Aquesta ha de descriure la informació utilitzada amb una petita explicació de les fonts. Per una altra banda, he decidit afegir una taula amb paginació per a mostrar les dades i amb un botó de descàrrega per a obtenir les taules. La Figura 12.6 és un exemple de pàgina de descàrregues.

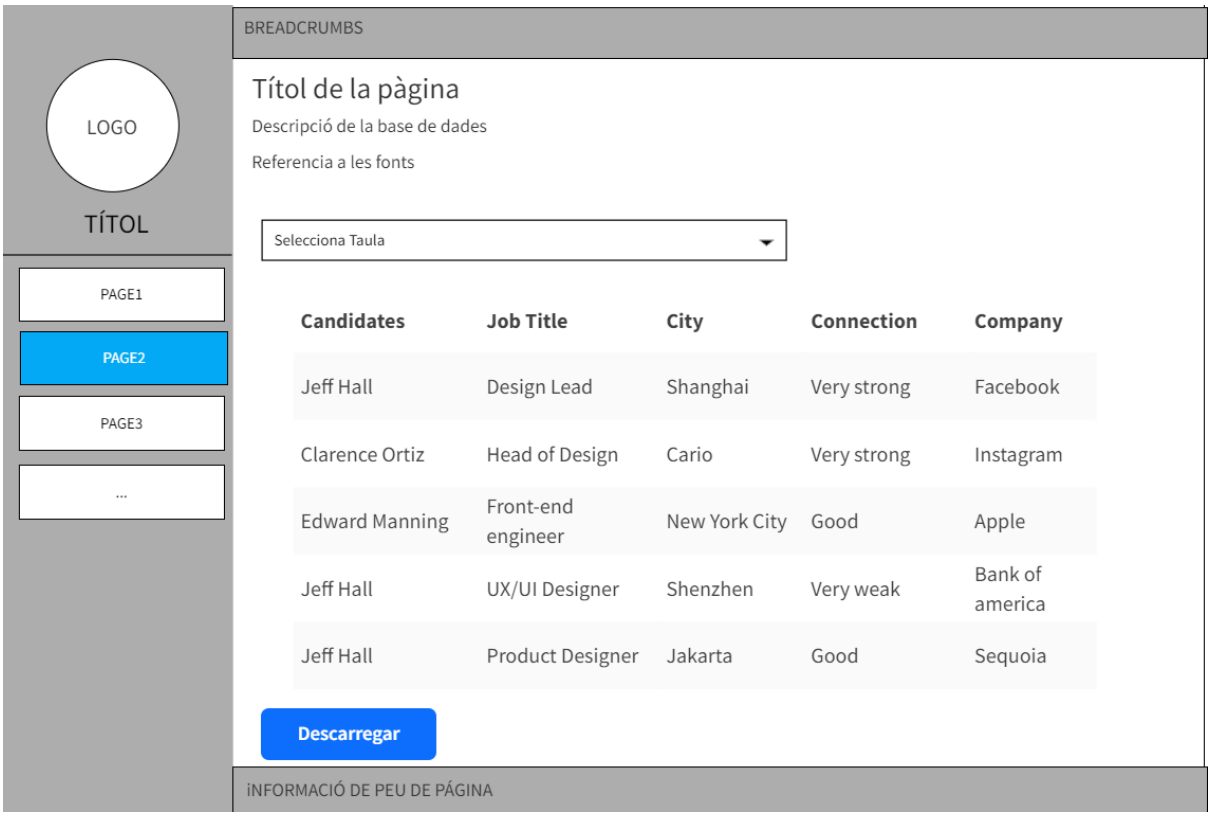

Figura 12.6: Maqueta de la pàgina de descàrregues. Font: Elaboració pròpia.

Aquestes maquetes són orientatives, però es tractaran amb certa flexibilitat a l'hora de desenvolupar el frontend de la pàgina web. Tanmateix, em serveixen per estructura de formes general l'arquitectura de la web i es fan servir com a punt de partida. Això vol dir que pot haver-hi certs canvis visuals entre les maquetes i el resultat final del web.

12.3. Visualització de les dades

En aquest capítol repassaré l'estructura de les gràfiques necessàries per a mostrar la informació de forma que sigui útil per a l'anàlisi. Generalment, l'equip del Clínic demandava 4 formes diferents de visualització de les dades:

- **Mapes:** Els mapes havien de ser una primera forma superficial de visualitzar les dades. Així i tot, això va suposar un repte de disseny, ja que no és fàcil representar més d'un tipus de dada en un mateix mapa, amb el requeriment de què ha de ser

entenedor. Per solucionar aquest apartat vaig decidir representar les dades de dues formes diferents dins d'un mateix mapa. Aquestes dues formes es basen en el color de fons amb una certa transparència per al primer tipus de dada, i per l'altre tipus utilitzaré un cercle sobreposat en què tant el color com el diàmetre determinaran un valor. Aquests mapes es diuen Bubble Maps i es pot veure un exemple en la Figura 12.7.

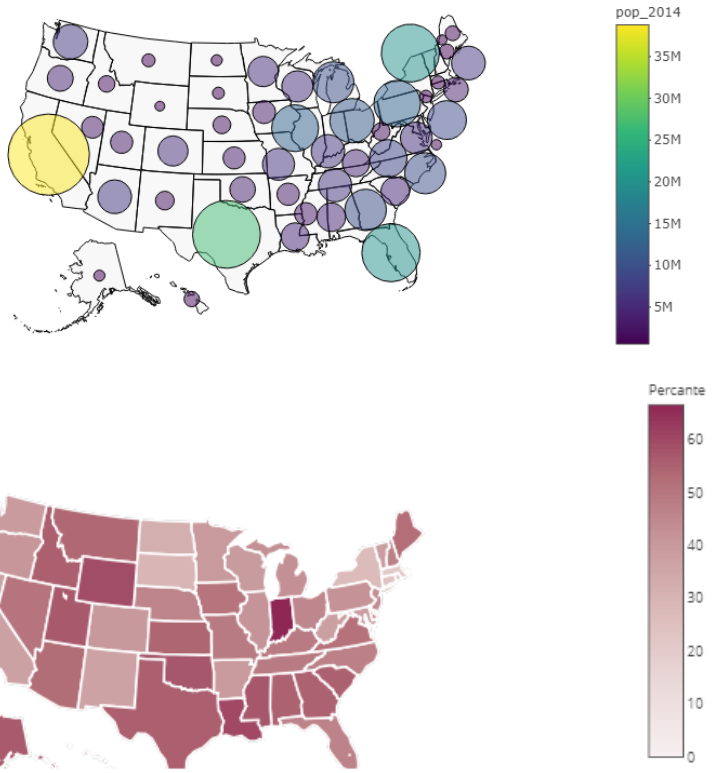

Figura 12.7: Exemples de mapes que hauran de ser combinats. Font: Web Plotly [33]

- **Histogrames:** Els histogrames són una eina que ens permet comparar dades d'un mateix any entre diversos països amb més exactitud que en el mapa. Un exemple del format d'histograma que faré servir és el mostrat a la Figura 12.8.

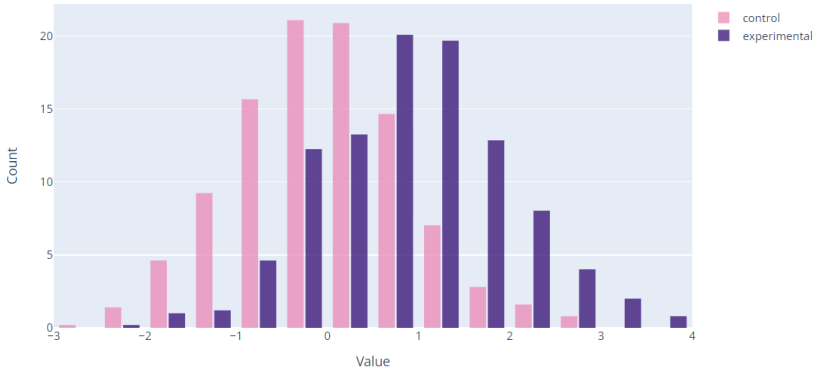

Figura 12.8: Format d'histograma. Font: Web Plotly [33]

- **Gràfics lineals:** A part de la comparativa entre països també ens interessa veure l'evolució de les dades durant les tres dècades que tractem. El gràfic més adient en aquest cas és un gràfic lineal on cada línia pot representar l'evolució d'una dada diferent o d'un país diferent, segons el cas (Figura 12.9).

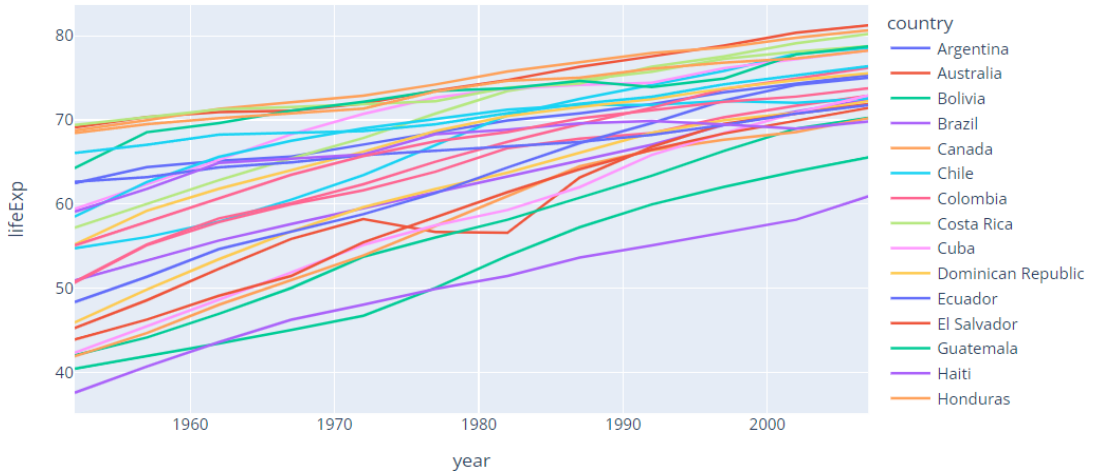

Figura 12.9: Format d'un gràfic de línies. Font: Web Plotly [33]

- **Taules:** finalment seran necessàries les taules per a consultar informació molt concreta. Com per exemple, per a quins anys hi ha dades de càncer disponibles.

13. Desenvolupament de l'aplicació

13.1. Descripció del funcionament de Dash

Dash és una llibreria gratuïta de Python destinada a crear aplicacions web. El codi de l'aplicació Dash és declaratiu i reactiu, la qual cosa facilita la construcció d'aplicacions complexes que contenen molts elements interactius. En aquest apartat explicaré els diversos aspectes a tenir en compte a l'hora de desenvolupar una aplicació amb aquesta eina.

13.1.1. Arquitectura

Les aplicacions Dash són servidors web que executen Flask i comuniquen paquets JSON a través de peticions HTTP. El frontend de Dash renderitza els components utilitzant React.js, la llibreria d'interfície d'usuari de JavaScript escrita i mantinguda per Facebook.

Flask [35] és un Framework escrit en Python i dut a terme per a simplificar i fer més fàcil la creació d'Aplicacions Web sota el patró MVC (model vista controlador). El fet que Dash tingui automatitzat el patró MVC facilita el nostre treball de desenvolupament de l'aplicació, ja que no hem de preocupar-nos pel disseny dels models.

React [36] és la llibreria que fa servir Dash per a la realització de components visuals de l'aplicació web. El fet que React estigui inclòs en l'eina que fem servir facilita el nostre treball de construcció de la interfície d'usuari. Això és molt important per a nosaltres pel fet que no som experts en el desenvolupament frontend.

El concepte crucial d'aquesta arquitectura és que aprofita la potència de dues llibreries molt emprades per al desenvolupament web, posant-les al servei de treballadors que poden no ser experts programadors web, que és el nostre cas.

13.1.2. Concurrència. Aplicacions multiusuari

L'estat d'una aplicació Dash s'emmagatzema en el navegador web. Això permet que les aplicacions es facin servir en un entorn multitasca: diversos usuaris poden tenir sessions independents mentre interactuen amb una aplicació al mateix temps. El codi de l'aplicació Dash és funcional: el codi de l'aplicació pot llegir valors de l'estat global de Python , però no pot modificar-los. Aquest enfocament funcional és fàcil de raonar i fàcil de ser utilitzat. Aquesta característica és fonamental en qualsevol aplicació web, ja que permet que diversos usuaris el facin servir alhora sense cap mena de registre.

13.1.3. Estils CSS

El CSS i els estils per defecte es mantenen fora de la biblioteca central per a la modularitat, el versionat independent, i per a animar als desenvolupadors de Dash a personalitzar l'aspecte i la sensació de les seves aplicacions. Aquest aspecte ens dona una llibertat absoluta per a editar els estils dels components React.

13.1.4. Visualització de conjunts de dades

Dash inclou un component anomenat Graph que representa gràfics amb plotly.js. Plotly.js encaixa perfectament amb Dash: és declaratiu, de codi obert, ràpid i admet una àmplia gamma de gràfics científics, financers i empresarials. Plotly.js està construït sobre D3.js (per a l'exportació d'imatges vectoritzades amb qualitat de publicació) i WebGL (per a la visualització d'alt rendiment). Aquest component comparteix la mateixa sintaxi que la biblioteca de codi obert plotly.py, per la qual cosa es pot canviar fàcilment entre els dos. El component Graph de Dash s'enganxa al sistema d'esdeveniments de plotly.js, permetent als autors d'aplicacions Dash escriure aplicacions que responguin a passar el ratolí, fer clic o seleccionar punts en un gràfic de Plotly. Aquest comportament és molt interessant, ja que ens permet proporcionar gràfics de gran valor analític a la nostra pàgina web. Alguns exemples d'aquests gràfics són els mostrats a la Figura 13.1. [34]

Figura 13.1: Gràfics d'exemple elaborats amb Plotly. Font: Pàgina web de Plotly.js [34].

13.2. Emmagatzematge en MongoDB

Una vegada tenia totes les dades necessàries processades i en un format adient per a poder ser utilitzades, el següent pas és emmagatzemar aquestes taules en algun tipus de servei de base de dades. Aquest pas és necessari pel fet d'haver escollit fer una aplicació web, ja que necessitem tenir les dades disponibles en tot moment.

Per aquesta tasca existeixen bàsicament dos tipus de bases de dades que es podrien fer servir:

- **Base de dades relacional:** emmagatzema i organitza conjunts de dades que estan relacionats entre si. Es basa en el model de dades relacional i presenta conjunts de dades com una col·lecció de taules, proporcionant operadors relacionals per a manipular les dades en forma tabular. MySQL, PostgreSQL i MariaDB són alguns exemples.

- **Base de dades no relacional:** es diferencien de les bases de dades relacionals tradicionals, ja que emmagatzemen les seves dades de forma no tabular. En canvi, les bases de dades no relacionals poden basar-se en estructures de dades com els documents. Un document pot ser molt detallat i, al mateix temps, contenir diferents tipus d'informació en diferents formats. No segueix el model relacional que ofereixen els sistemes tradicionals de gestió de bases de dades relacionals. MongoDB i Cassandre són alguns exemples.

13.2.1. Base de dades MongoDB

Finalment, hem decidit utilitzar una base de dades no relacional anomenada MongoDB [28]. Aquesta tecnologia té els següents avantatges i inconvenients:

Avantatges

- **Organització massiva de dades:** permet emmagatzemar quantitats massives d'informació i consultar-la amb facilitat i velocitat.

- **Flexibilitat per a l'ampliació:** Les dades no són estàtiques. A mesura que es recopila més informació, una base de dades no relacional pot absorbir aquests nous punts de dades, enriquint la base de dades existent amb nous nivells de valor granular, fins i tot si no s'ajusten als tipus de dades de la informació prèviament existent.

- **Múltiples estructures de dades:** Sigui com sigui el format de la informació, les bases de dades no relacionals accepten diferents tipus d'informació en un mateix document.

- **Dissenyada per a ser allotjada en el núvol:** MongoDB ofereix servei d'allotjament de base de dades en el núvol de forma gratuïta.

Inconvenients

- **No permeten relacions entre taules:** per la mateixa definició del model no existeixen relacions entre informació de diferents taules.

- **Problemes de compatibilitat entre instruccions SQL:** Les noves bases de dades utilitzen les seves pròpies característiques en el llenguatge de consulta i no són 100% compatibles amb el SQL de les bases de dades relacionals.

El fet que estem creant una aplicació de zero i no estem reutilitzant cap altre projecte fa que la qüestió de la compatibilitat no sigui un problema. En el cas de les relacions entre taules, per la mateixa natura de les dades no és massa necessari establir relacions entre informació de

diferents taules. Tot plegat fa que els avantatges proporcionats per aquest tipus de base de dades tinguin més importància que els inconvenients.

13.2.2. Importació a un clúster

MongoDB proporciona un servei anomenat MongoDB Atlas Cluster [29]. Aquest permet oferir una base de dades no relacional com a servei en el núvol públic. Aquest servei està disponible en Microsoft Azure, Google Cloud Platform i Amazon Web Services. En el nostre cas hem decidit utilitzar Amazon Web Services [32], ja que era el que, segons les proves de la web de MongoDB Atlas Cluster, oferia millors velocitats d'accés a dades. L'allotjament de les dades en un clúster ens permet escalabilitat, accés immediat a les dades i control de seguretat a través del programari de gestió proporcionat per MongoDB Atlas Cluster. Seguint el punt de la seguretat amb aquest clúster som capaços de limitar l'accés a les dades per adreça IP. Això serà de gran ajuda en el moment en què la web estigui desplegada, ja que podrem restringir l'accés a les dades al servidor de l'aplicació.

El funcionament d'un clúster en MongoDB és molt senzill, MongoDB Atlas és una plataforma autogestionada, la qual cosa vol dir que ells s'encarregaran d'absolutament tots els aspectes relacionats al hosting, instal·lació i actualitzacions, de tal forma que nosaltres només ens encarreguem de les configuracions més bàsiques, com gestionar els accessos, crear les bases de dades, crear alertes, etc.

13.2.3. Estructura de la base de dades

Finalment, cal definir l'estructura de la base de dades. Mongodb estructura les dades per documents en comptes de taules relacionals. Per aquest motiu he decidit crear 4 documents un per a cada tipus de dades, com es mostra en la Figura 13.2. Una altra opció hauria sigut incloure les dades de Mortalitat i Incidència de càncer de pulmó en un mateix document. Després de realitzar diverses proves de rendiment he decidit la primera opció, ja que el rendiment era idèntic i d'aquesta forma era més entenedor.

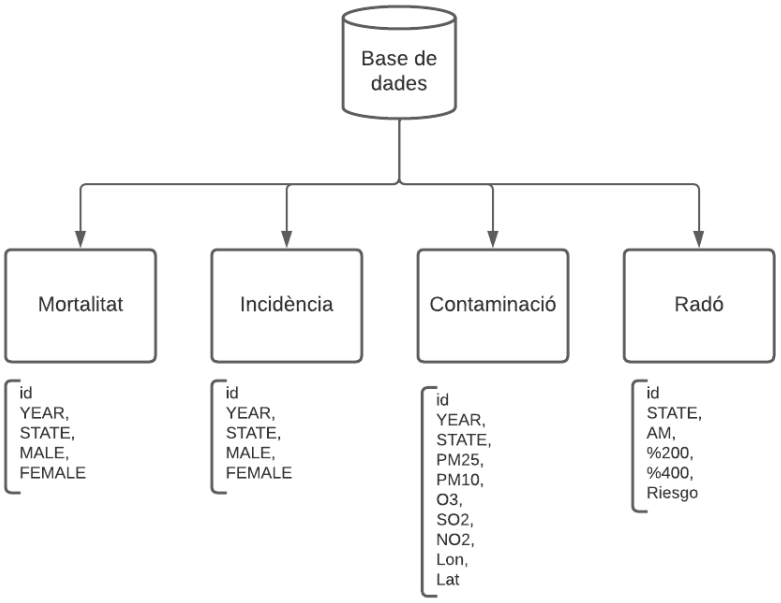

Figura 13.2: Estructura de la base de dades. Font: Elaboració pròpia.

13.3. Algorismes de classificació

Una vegada la fase de desenvolupament del sistema estava certament avançada va sorgir una nova necessitat en l'equip d'investigació. Per a ells era molt interessant, a més de tenir els valors purs de cada contaminant en cada any, també tenir algun tipus de classificació de resultats. És a dir, d'alguna forma veure si el nivell de NO2 a França durant el 2015, per exemple, va ser un nivell alt, baix o mitjà en comparació als resultats històrics.

Una primera idea va ser comparar els resultats amb les recomanacions que fa l'OMS, que són les mostrades en la Figura 13.3, informació extreta directament de la web de l'OMS [31]. Tanmateix, la gran majoria de resultats estaven per sobre dels valors recomanats per l'OMS i , per tant, aquesta classificació no donava massa valor de comparació. Tot i això, vam decidir mantenir aquest criteri mostrant les recomanacions en els gràfics. Com veurem en el següent capítol.

	PM10	PM25	O3	NO2	SO2
Recomanació de l'OMS	15 µg/m³	5 µg/m³	60 µg/m³	10 µg/m³	40 µg/m³

Figura 13.3: Recomanacions de nivells anuals dels principals contaminants per l'OMS. Font: Pàgina web de l'OMS[31].

La solució que vaig trobar va ser establir una classificació per nivells comparant els resultats. Per a fer això cal un algoritme de classificació o de clustering. El clustering és una tècnica d'Aprenentatge automàtic que implica l'agrupació de punts de dades. Donat un conjunt de punts de dades, podem utilitzar un algorisme d'agrupació per a classificar cada punt de dades en un clúster específic. En teoria, els punts de dades que estan en el mateix clúster han de tenir propietats o característiques similars, mentre que els punts en diferents clústers han de tenir propietats o característiques molt diferents. L'agrupació és un mètode d'Aprenentatge no Supervisat i és una tècnica comuna per a l'anàlisi de dades estadístiques feta servir en molts camps. Existeixen nombrosos algoritmes de clustering, dels quals hem analitzat els més populars per veure quin s'ajustava més a les nostres necessitats.

13.3.1. Algoritme de K-Means

K-Means [37] és un algoritme de classificació no supervisada que agrupa objectes, representats com punts en dues dimensions, en k grups. El criteri d'aquest algoritme és la suma de les distàncies entre cada objecte i el centre del grup o clúster. Aquest algoritme consta de tres passos:

1. **Inicialització:** una vegada determinat el nombre k de clústers, que ha de ser determinat per l'usuari, es determinen de forma arbitrària k punts que seran els centres dels primers k clústers.

2. **Agrupació:** cada punt és assignat al seu clúster més proper. Els punts es representen com a vectors de 2 dimensions (x1, x2) i el càlcul de la distància és el mostrat a la Figura 13.4, quan S és el conjunt de dades, k és el nombre de clústers i μ és el centre del clúster.

$$\min_{\mathbf{S}} E(\boldsymbol{\mu_i}) = \min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

Figura 13.4: Representació del problema de càlcul de distància en K-means. Font: Pàgina de Wikipedia de l'algorisme [37]

3. **Actualització del centre dels clústers:** es recalcula el centre de cada grup agafant com a nou centre la posició de la mitjana de punts que pertanyen a aquest grup.

- Es repeteixen els passos 2 i 3 fins que els clústers no canvien o si es mouen per sota d'una distància llindar determinada.

Els principals avantatges del mètode K-Means són que és un mètode senzill i ràpid. Però és necessari decidir el valor de k i el resultat final depèn de la inicialització dels centroides.

13.3.2. Algoritme d'agrupament jeràrquic d'aglomeració

En aquest algoritme cada mostra es tracta com un sol clúster i després es fusionen, o aglomeren, successivament parelles de clústers fins que tots els clústers s'hagin fusionat en un. En aquesta tècnica, inicialment cada punt de dades es considera com un clúster individual. A cada iteració, els grups similars es fusionen amb altres grups fins que es forma un grup o k grups. L'algorisme bàsic d'aglomeració és el següent [38]:

1. Calcular la matriu de proximitat
2. Deixar que cada punt de dades sigui un clúster
3. Repetir: fusionar els dos clústers més pròxims i actualitzar la matriu de proximitat Fins que només quedi un únic clúster

A la Figura 13.5 es pot veure un exemple d'una possible execució d'aquest algoritme.

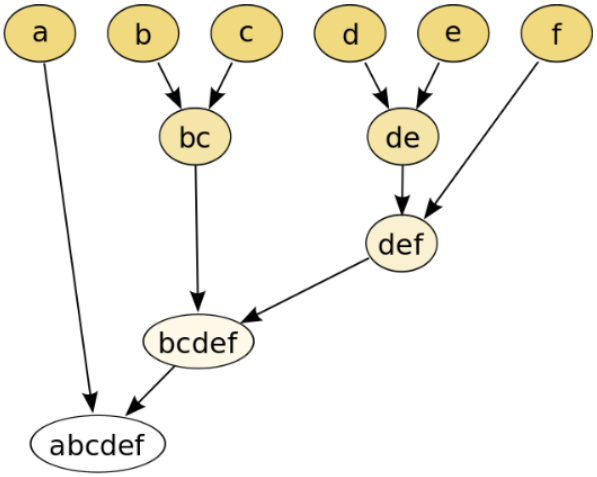

Figura 13.5: Estructura d'exemple d'una execució de l'algoritme d'agrupament jeràrquic d'aglomeració. Font: Pàgina de Wikipedia de l'algoritme [38].

L'agrupació jeràrquica no requereix que especifiquem el nombre de clústers i fins i tot podem seleccionar quin nombre de clústers es veu millor, ja que estem construint un arbre. A més, l'algorisme no és sensible a l'elecció de la mètrica de distància, tots ells tendeixen a funcionar igualment bé, mentre que, amb altres algorismes d'agrupament, l'elecció de la mètrica de distància és crítica. Un cas particularment útil dels mètodes d'agrupament de jeràrquica és quan les dades subjacents tenen una estructura jeràrquica i es desitja recuperar la jerarquia, altres algorismes d'agrupació no poden fer això. Aquests avantatges de l'agrupació jeràrquica venen a costa d'una menor eficiència, a diferència de la complexitat lineal de K-Means i el model de mescla gaussiana. Per una altra banda, el resultat d'aquest algoritme és sempre el mateix i no depèn de cap factor, això li dona consistència als resultats.

13.3.3. Algoritme de mescla gaussiana (GMM)

Els models de mescla gaussiana (GMM) [39] ens donen més flexibilitat que els K Means. Amb els GMM suposem que els punts de dades estan distribuïts per Gauss, això és una suposició menys restrictiva que dir que són circulars usant la mitjana. D'aquesta manera, tenim dos paràmetres per a descriure la forma dels clústers, la mitjana i la desviació estàndard. Prenent un exemple en dues dimensions, això significa que els clústers poden prendre qualsevol mena de forma el·líptica, ja que tenim una desviació estàndard tant en la direcció X com en la Y. Per tant, cada distribució gaussiana s'assigna a un sol grup.

Per a trobar els paràmetres del gaussià per a cada clúster, per exemple, la mitjana i la desviació estàndard, utilitzarem un algorisme anomenat maximització d'expectatives (EM, per les seves sigles en anglès). Llavors podem procedir amb el procés d'agrupació de maximització d'expectatives usant GMM:

1. Comencem seleccionant el número de clúster, com es fa en K-Means i inicialitzant aleatòriament els paràmetres de distribució gaussiana per a cada clúster.

2. Donades aquestes distribucions gaussianes per a cada clúster, calcular la probabilitat que cada punt de dades pertanyi a un clúster en particular. Com més a prop estigui un del centre dels gaussians, més probable serà que pertanyi a aquest grup.

3. Basant-nos en aquestes probabilitats, calculem un nou conjunt de paràmetres per a les distribucions gaussianes de manera que maximitzem les probabilitats dels punts de dades dins dels clústers. Calculem aquests nous paràmetres emprant una suma ponderada de les posicions dels punts de dades, on les ponderacions són les probabilitats del punt de dades que pertany a aquest clúster en particular.

- Els passos 2 i 3 es repeteixen iterativament fins a la convergència, on les distribucions no canvien molt d'iteració en iteració.

L'ús dels models de mescla gaussiana presenta dos avantatges clau. En primer lloc, són molt més flexibles en termes de covariància de clústers que els K-Means, a causa del paràmetre de desviació estàndard, els clústers poden adoptar qualsevol forma d'el·lipse, en lloc de limitar-se a cercles. En segon lloc, atès que els models de mescla gaussiana fan servir probabilitats, poden tenir conglomerats múltiples per punt de dades. Per tant, si un punt de dades està enmig de dos grups superposats, podem simplement definir la seva classe dient que pertany al X per cent de la 1 i a l'Y per cent de la 2. És a dir, els models de mescla gaussiana donen suport a una composició mixta.

13.3.4. Proves

Per a decidir quin algoritme utilitzar vaig realitzar diverses proves amb les dades de contaminació. A la figura 13.6 es pot veure una comparativa de resultats per les mostres de PM2.5 on cada color representa un clúster i tant l'eix x com l'eix i representa la mitjana de concentració de PM2.5 mesurada durant un any en cada país. Com es pot veure a primera vista els resultats són molt similars.

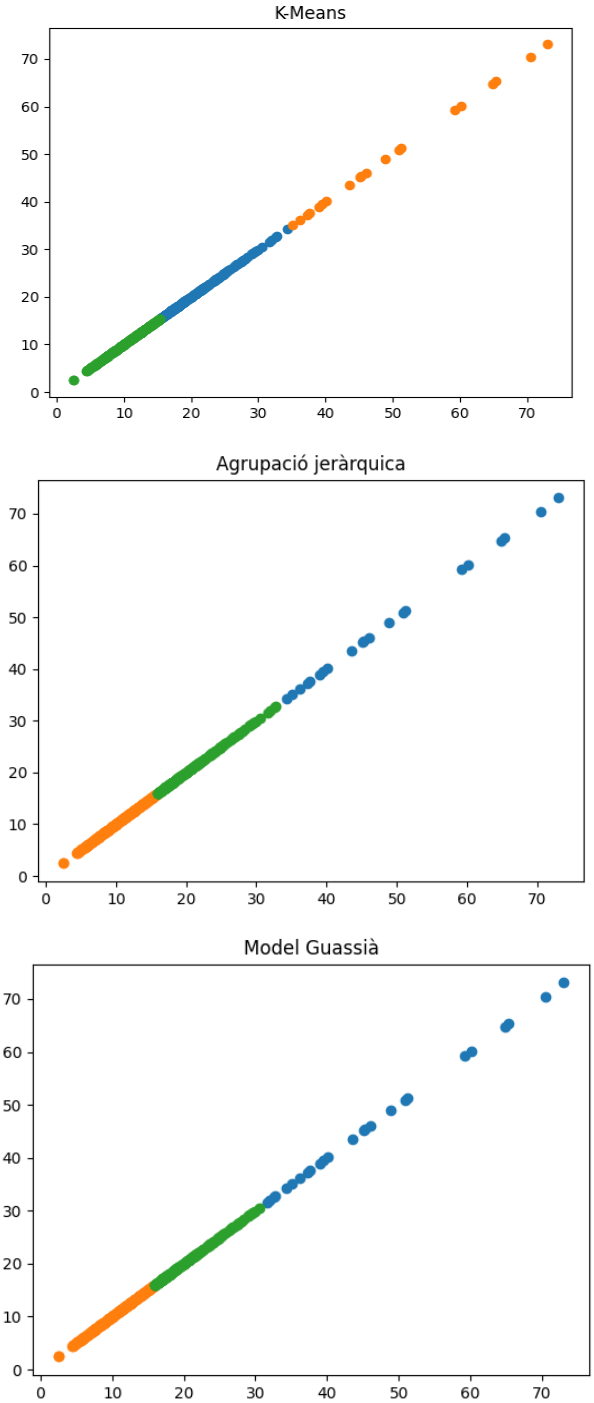

Figura 13.6: Comparativa d'agrupacions per a les dades de PM2.5. Font: Elaboració pròpia.

Una vegada vista la comparativa i valorats els punts forts i febles de cada algoritme he decidit aplicar l'algoritme d'agrupació jeràrquica per a la classificació de nivells de les mostres. Aquesta decisió es deu bàsicament a què el principal problema d'aquest algoritme és que no és tan eficient com els altres dos. Tanmateix, aquest treball es farà en preprocessament, per tant, l'eficiència de la tasca no afectarà el rendiment de l'aplicació. Per una altra banda, per a l'equip d'investigació és important la consistència de les dades i el fet que aquest mateix estudi pugui ser realitzat per una altra persona obtenint els mateixos resultats. Aquesta consistència només te l'assegura l'algoritme d'agrupació jeràrquica, ja que els altres dos tenen una certa aleatorietat en la inicialització.

13.4. Algorismes d'encriptat

Un altre algorisme que vaig necessitar va ser un algorisme per encriptar contrasenyes. Això va sorgir del fet que es va implementar la funció d'editar les dades. Aquesta opció només havia d'estar disponible per a usuaris de l'equip d'investigació, com és obvi. Per implementar aquesta gestió d'usuaris vaig afegir una nova taula a la base de dades que guarda el nom d'usuari i la contrasenya.

Les contrasenyes havien de ser guardades de forma segura. Per aquest motiu no valia amb guardar-les en un format pla. L'opció més segura i ràpida en aquest cas és utilitzar un algorisme de hashing. El terme "hashing" fa referència a la conversió d'una cadena en una altra (que també es denomina "hash") mitjançant una funció hash. Independentment de la grandària de la cadena d'entrada, el hash tindrà una grandària fixa que està predefinit en el mateix algorisme de hashing. L'objectiu és que el hash no se sembli en res a la cadena d'entrada i que qualsevol canvi en la cadena d'entrada produeixi un canvi en el hash. D'aquesta forma les contrasenyes no poden ser violades, ja que l'atac de força bruta prendria tant de temps per a calcular tots els possibles hashes, que ni tan sols val la pena intentar realitzar-ho.

L'algorisme de hashing que he decidit fer servir és el bcrypt. Bcrypt [40] és un algorisme dissenyat específicament per a hash de contrasenyes. Altres algoritmes com MD5 i SHA1 són algorismes de hash de propòsit general. Per disseny, a més, bcrypt permet agregar un salt al procés de generació del hash, la qual cosa, d'entrada, ho fa immune a atacs de diccionari. El poder de còmput requerit per a generar els hashes amb bcrypt és altíssim comparat amb MD5 i SHa1. A més, la quantitat de rondes que corre l'algorisme és adaptable com un paràmetre de la funció. El fet que l'algorisme faci ús de hashing i salt el fa encara més difícil de violar.

13.5. Generació de gràfiques

En aquest apartat mostraré els resultats després d'aplicar els dissenys de gràfics explicats anteriorment amb Plotly. A més donaré algunes indicacions per entendre com mostren la informació aquests gràfics.

En primer lloc, tenim els mapes. Tenim dos tipus de mapes, el primer és un mapa simple en el qual cada país és pinta d'un color en gradient segons el valor que se li doni. A la Figura 13.7 es mostra un exemple amb un mapa de mortalitat de càncer. L'altre tipus de mapa és aquell que combina la forma de representar les dades de l'anterior amb uns cercles que mostren una informació diferent, tant amb la mida com amb el color, ja que a més gran sigui el cercle més alt serà el valor. A la Figura 13.8 tenim un exemple d'aquest segon mapa amb les dades de Radó representades al color de fons i les de PM2.5 amb els cercles.

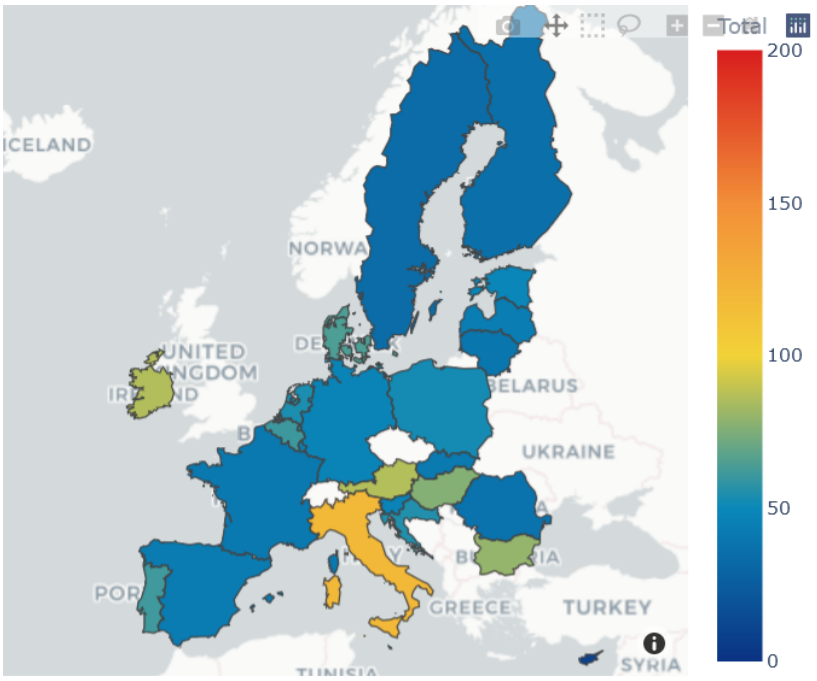

Figura 13.7: Mapa de càncer de pulmó l'any 2000. Font: Elaboració pròpia.

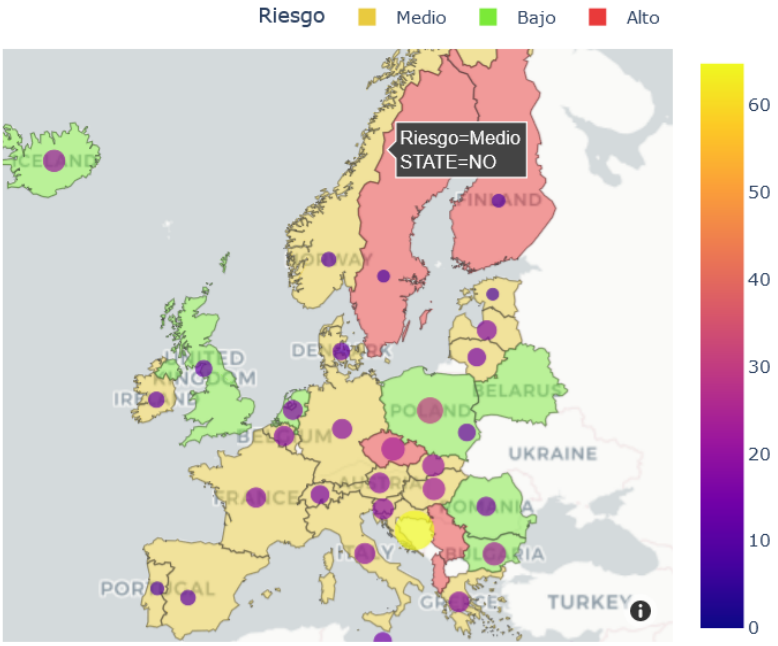

Figura 13.8: Mapa de radó i PM2.5 l'any 2010. Font: Elaboració pròpia.

Seguidament, cal presentar els resultats pels histogrames. He decidit afegir l'opció de visualitzar histogrames amb diverses dades alhora per poder comparar i a més he afegit les recomanacions de l'OMS per als histogrames de contaminació exterior. Un exemple d'aquest gràfic és l'exemplificat a la Figura 13.9 amb les dades d'O3 i SO2.

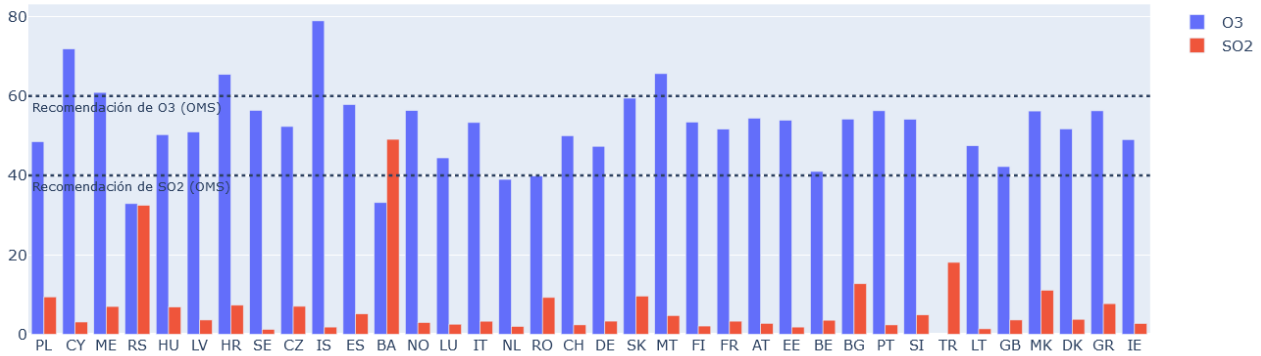

Figura 13.9: Histograma comparatiu de SO2 i O3 l'any 2010. Font: Elaboració pròpia.

L'altre gràfic implementat ha sigut el gràfic lineal. Aquest mostra l'evolució de les dades a un país. Addicionalment, també he posat la informació de les recomanacions establertes per l'OMS. Un exemple és la Figura 13.10 que representa l'evolució de les dades de mortalitat de càncer total (en blau), NO2, SO2, i O3 a Àustria.

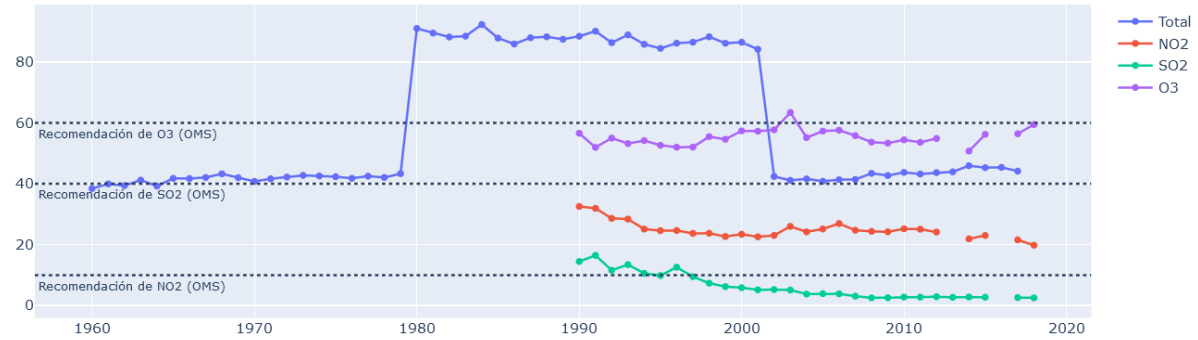

Figura 13.10: Gràfic de línies de mortalitat de càncer, NO2, SO2, O3 a Àustria. Font: Elaboració pròpia.

Finalment, l'última forma de representar les dades és la visualització per taules. Aquestes estan implementades amb paginació i amb la possibilitat de filtrar per paraula i ordenació. A més poden ser descarregades i modificades, tot i que aquesta última opció requerirà permisos. La Figura 13.11 mostra una taula de contaminació exterior.

STATE	YEAR	O3	NO2	SO2	PM10	PM25
filter data...						
PL	1997	47.08	30.93	26.77	56.08	
PL	1998	54	26.84	21.69	51.91	
PL	1999	53.51	23.23	21.76	41.76	
PL	2000	52.88	21.03	16.95	39.07	
PL	2001	51.39	20.91	14.95	34.72	
PL	2002	53.74	22.3	15.91	41.71	50.92
PL	2003	56.19	23.72	16.21	42.55	73.06
PL	2004	51.4	18.74	8.95	31.13	29.05
PL	2005	53	20.66	8.37	34.21	30.55
PL	2006	54.37	21.63	9.57	38.29	28.99
PL	2007	50.94	18.86	7.19	31.36	25.3
PL	2008	50.27	17.88	5.55	29.56	23.82
PL	2009	47.66	18.66	6.4	32.71	27.57
PL	2010	48.51	19.26	9.42	37.1	29.39
PL	2011	49.25	19.46	7.58	36.12	27.59

« ‹ 1 / 56 › »

Figura 13.11: Taula de contaminants. Font: Elaboració pròpia.

14. Proves

Abans de fer el sistema públic aquest haurà de passar dues proves que hem acordat amb la codirectora del projecte. Aquests processos de prova ens permetran rebre un feedback per perfeccionar la web abans de ser publicada.

Primerament, l'aplicació serà restringida a personal de l'equip d'investigació del clínic. En aquest procés es tractaran d'extreure alguns resultats preliminars amb l'aplicació. Així com ajustar l'estètica (colors, imatges, logos…) a la desitjada per l'equip.

Una vegada l'aplicació hagi passat aquest filtre la intenció és poder presentar el projecte amb els primers resultats al congrés mundial de càncer de pulmó [41] (IASLC WCLC World Conference on lung cancer). L'objectiu d'aquest congrés és millorar la comprensió del càncer de pulmó entre els científics, els membres de la comunitat mèdica i el públic. IASLC publica els seus articles en el Journal of Thoracic Oncology, un valuós recurs per a metges especialistes i científics que se centren en la detecció, prevenció, diagnòstic i tractament del càncer de pulmó. Aquest reconeixement pot permetre continuar amb el projecte d'una forma més ambiciosa gràcies a les beques atorgades i també perfeccionar l'aplicació a partir dels consells dels membres. Un exemple d'aquest congrés és el que es va celebrar a Barcelona l'any 2019, on també van ser presentats projectes de l'Hospital Clínic. La figura 14.1 mostra el logo de presentació d'aquell congrés.

Figura 14.1: Imatge del congrés de càncer de pulmó de Barcelona. Font: Web del IASLC [42]

15. Conclusions

Una vegada finalitzat el projecte em proposo a valorar-lo. L'objectiu més important del projecte era dissenyar i implementar un sistema que permeti analitzar dades de càncer i contaminació a Europa durant les últimes dècades. A grans trets, el resultat obtingut ha sigut validat per l'equip d'investigació i és funcional. Tanmateix, en aquest apartat parlaré més a fons dels propòsits satisfets i dels que no han pogut ser satisfets, així com quedaran descrites les limitacions amb les quals m'he trobat i el treball a fer en un futur.

15.1. Contribucions

La realització d'aquest projecte ha sigut, és i serà de gran ajuda al camp de l'oncologia. Com ja vaig explicar, el sistema desenvolupat en aquest projecte permet realitzar un estudi que mai s'ha fet abans. Les afectacions dels carcinògens ambientals en els pacients de càncer de pulmó és un fet que l'equip d'investigació intueix des de fa temps, però que per falta d'eines mai ha pogut ser estudiat. Abans de l'aparició d'aquest projecte els investigadors duien a terme estudis d'aquest tipus utilitzant eines com Excel. La qual cosa pot ser bastant tediós per a la representació visual de les dades. A més en la majoria dels casos feien servir dades d'unes 300 mostres obtingudes del mateix hospital. Això comporta una mostra bastant reduïda i d'una zona molt seleccionada.

El que estem oferint amb l'aplicació web és un sistema personalitzat per aquest projecte en concret amb una base de dades aconseguida d'organismes de tota Europa i, per tant, molt més àmplia. A més també s'ofereix la possibilitat de ser editada per l'usuari. Aquest primer punt resoldrà el problema de les mostres. Per l'altra banda la web permet la comparació de les dades amb els gràfics demanats per l'equip de forma eficient i molt més còmode. Finalment, el fet que sigui una aplicació web i que despongui de permisos d'usuari per editar la base de dades permet que la web sigui pública i, per tant, pot incrementar l'atenció del sector en aquest tipus d'estudis.

Finalment, els usuaris podran obtenir diversos profits d'aquesta eina a banda de la facilitat per realitzar l'estudi de correlacions entre el càncer de pulmó i l'exposició ambiental. Des del punt de vista mèdic, la web permet visualitzar de manera senzilla l'exposició dels pacients als diferents factors ambientals durant tota la seva vida. Des del punt de vista governamental, els resultats assolits de l'anàlisi de les dades poden ajudar a proposar mesures de prevenció diferents depenent de la informació de cada país.

15.2. Objectius satisfets

Quan es va planificar el projecte es van estipular una sèrie d'objectius i subobjectius. Donat per satisfet el propòsit principal, que era aconseguir crear una eina que serveixi per a la investigació del càncer de pulmó i l'exposoma ambiental, caldria valorar els altres objectius i subobjectius que em vaig marcar a la planificació del projecte. A continuació es llisten aquests propòsits:

- *Dissenya i desenvolupar una base de dades que ens faciliti el tractament de dades tant de càncer com d'agents contaminants en Europa.* Com s'explica al punt 11 d'aquest document, es va fer una cerca i un anàlisis de les diferents fonts públiques de dades sobre el tema i es va generar una base de dades completa amb dades de càncer i contaminació.

- *Dissenyar una interfície d'usuari intuïtiva i fàcil d'utilitzar.* La interfície d'usuari dissenyada té un aspecte simple però professional, intuïtiu i fàcil de fer servir. A més, aquest disseny va ser aprovat i va rebre un bon feedback de l'equip al qual va destinat l'eina. El procés de disseny de l'aplicació està definit a l'apartat 12.

- *Implementar un sistema d'ingesta de noves dades des de la mateixa aplicació.* La web compta amb un apartat anomenat Dades (o pàgina de dades) que permet descarregar i editar cadascun dels documents de la base de dades. Tanmateix, una ingesta de dades massiva podria ser un treball llarg de fer directament des de la mateixa aplicació, per aquest motiu descric aquest treball com a treball futur.

- *Valorar diferents formes de representació de dades i implementar les més útils (mapes, diagrames de barres...).* Durant l'etapa de desenvolupament de l'aplicació vaig presentar diverses formes de representar les dades i l'equip mèdic va decidir utilitzar els mapes, els histogrames i els gràfics lineals presentats a l'apartat 13.5 d'aquesta memòria.

- *Afegir mecàniques de personalització per facilitar l'exploració de les dades.* Cadascun dels gràfics generats pel sistema compta amb una sèrie de paràmetres personalitzables per l'usuari (selecció de tipus de dades, filtratge per anys i país...). A més els mateixos gràfics proporcionen la possibilitat de guardar una imatge, fer zoom, activar i desactivar un traça i altres opcions pròpies de Plotly.

- *Obtenir un sistema eficient i escalable que permeti una vida útil de l'aplicació elevada.* El fet d'haver fet servir les eines de Dash i MongoDB m'ha permès elaborar un sistema ràpid i escalable. Un punt important en aquest aspecte és que aquestes dues eines m'han permès generar una web que precarrega les dades a memòria la primera vegada que es desplega la web i no les actualitza fins que no són modificades. Aquest fet permet una navegació molt més ràpida i eficient. Afegint això a una estructura de codi senzilla basada en l'arquitectura MVC ha resultat en una aplicació eficient i escalable.

15.3. Limitacions

Tot i que els objectius han sigut generalment satisfets, aquest projecte s'ha trobat amb una sèrie de limitacions que han dificultat l'optimització d'alguns d'aquests objectius.

El primer punt que cal esmentar com a una limitació rellevant del projecte és el fet que les dades són incompletes. Una primera idea de l'equip d'investigació va ser obtenir dades de contaminació i càncer de les últimes tres dècades. Si bé les dades de càncer s'han fet públiques des del 1960, en el cas dels contaminants les mesures no comencen a ser públiques fins als anys 1990-2000. En el cas de les dades de Radó, es van demanar les mesures de radó a Europa al Comitè De Seguretat Nuclear, ja que actualment no són públiques, però aquest demana un permís signat per a tots els països de la Unió Europea, cosa que no era viable en el

temps d'aquest projecte. Això representa un petit inconvenient per a l'estudi evolutiu de les dades històriques.

Una altra limitació que té aquesta eina a l'hora d'investigar en aquest camp és que no es tenen en compte les dades personals dels pacients de càncer de pulmó. Això provoca que es puguin comparar les dades de contaminants amb les de càncer, però sempre buscant una relació relativa a escala global, tot tenint present que a més dels d'origen ambiental existeixen centenars de carcinògens més que no s'estan tenint en compte, com per exemple l'índex de massa corporal, els hàbits esportius, el tabac...

Finalment, una última limitació que s'ha de tenir en compte a l'hora d'utilitzar aquesta eina és que segons l'any i el país d'on provenen les dades aquestes poden ser poc fiables. Com he explicat a l'apartat d'extracció de les dades (11.1), aquestes provenen directament dels recomptes oficials de cada país. Per aquest motiu és d'esperar que les dades de fa trenta anys no siguin tan fiables com les actuals. Concretament a la web de la base de dades de l'OMS indica que aquestes dades han de ser tractades per ser analitzades de forma evolutiva i no pel seu valor en concret. Totes aquestes limitacions seran proposades en el següent apartat com a possible continuació per aquest projecte.

15.4. Treball futur

Donat aquest treball com a finalitzat, cal destacar alguns aspectes que han quedat pendents de finalitzar o perfeccionar.

En referència a les limitacions descrites a l'apartat anterior, caldria actualitzar les dades en el moment en què es poguessin aconseguir unes de millor qualitat. Per exemple quan s'aconsegueixi obtenir el permís per a les dades de radó. Aquest procés d'ingesta de noves dades es podria fer des de la mateixa web o directament a la base de dades seguint els passos descrits a l'apartat 11.

Un altre aspecte que es podria perfeccionar és generar una eina per automatitzar el procés d'extracció, preprocessament i ingesta de dades de les principals fonts públiques. Això permetria tenir sempre les dades actualitzades automàticament.

Finalment, com ja vaig comentar en l'apartat de disseny de l'aplicació, l'aspecte visual d'aquesta és una primera aproximació que haurà de ser perfeccionada seguint un futur feedback dels usuaris.

15.5. Valoració personal

Abans de donar la memòria com finalitzada m'agradaria fer una valoració personal d'aquests cinc mesos que he dedicat a aquest projecte. Des del punt de vista professional i acadèmic aquest treball m'ha instruït com a enginyer informàtic. Els primers passos de planificació del projecte em van fer veure la importància de destinar el temps necessari a la gestió i planificació de qualsevol projecte tecnològic abans de posar-me a picar codi. Seguidament, la part de desenvolupament de l'aplicació m'ha obligat a valorar diferents opcions desconegudes per afrontar el desenvolupament d'una aplicació web. Aquest punt m'ajudarà en un futur a afrontar altres projectes de software. Finalment, el treball en equip en aquest TFG ha sigut primordial, ja que a més de treballar en tàndem amb un company de la Universitat he hagut d'aprendre a col·laborar amb un equip de professionals d'una àrea diferent de la informàtica, aquesta és una habilitat totalment necessària per a qualsevol enginyer. Per acabar, des del punt de vista emocional, sempre he desitjat destinar els meus coneixements informàtics a una causa tan necessària en la nostra societat com és la investigació sobre el càncer. Ha sigut un absolut plaer fer feina en aquest entorn i tant de bo pugui col·laborar en reptes com aquest en un futur.

Referències

[1] American Cancer Society. Descripció amb dades del Càncer de pulmó. American Cancer Society. Consultat el 25 de febrer del 2022, a https://www.cancer.org/es/cancer/cancer-de-pulmon.html [en línia].

[2] Hospital Clínic de Barcelona. Web informativa de l'Hospital Clínic. Hospital Clínic Barcelona. Consultat el 25 de febrer del 2022,a https://www.clinicbarcelona.org/ca [en línia].

[3] Instituto de Investigaciones Biomédicas August Pi i Sunyer (IDIBAPS). Web de presentació d'IDIBAPS. Universitat de Barcelona. Consultat el 25 de febrer del 2022, a https://www.ub.edu/web/ub/es/recerca_innovacio/recerca_a_la_UB/instituts/institutsparticipats/idibaps.html [en línia].

[4] World Health Organization. Base de dades de càncer a Europa. Global Cancer Observatory. Consultat el 30 de gener del 2022, a https://gco.iarc.fr/ [en línia].

[5] European Statistical. Base de dades d'Eurostat. European Commission. Consultat el 30 de gener del 2022, a https://ec.europa.eu/eurostat/data/database [en línia].

[6] The Association of the Nordic Cancer Registries. (2019, March 26). Base de dades dels països nòrdics. NORDCAN. Consultat el 7 de febrer del 2022, a https://www-dep.iarc.fr/NORDCAN/english/frame.asp [en línia].

[7] World Health Organization. Base de dades de mortalitat mundial. WHO | World Health Organization. Consultat el 30 de gener del 2022, a https://www.who.int/data/data-collection-tools/who-mortality-database [en línia].

[8] Dash Overview. Web de contingut de Dash. Consultat el 21 de febrer del 2022, a https://plotly.com/dash/ [en línia].

[9] Python Software Foundation. Web de presentació del llenguatge Python. Welcome Python.org. Consultat el 21 de febrer del 2022, a https://www.python.org/ [en línia].

[10] Python Software Development. Web de presentació i manual de Flask. Welcome to Flask — Flask Documentation (2.0.x). Consultat el 21 de febrer del 2022, a https://flask.palletsprojects.com/en/2.0.x/ [en línia].

[11] Plotly. Web de presentació de Plotly. Plotly: The front end for ML and data science models. Consultat el 21 de febrer del 2022, a https://plotly.com/ [en línia].

[12] Meta Platforms. Web de presentació de React. React – A JavaScript library for building user interfaces. Consultat el 21 de febrer del 2022, a https://reactjs.org/ [en línia].

[13] Pandas Team. Web d'informació de la llibreria Pandas. pandas - Python Data Analysis Library. Consultat el 21 de febrer del 2022, a https://pandas.pydata.org/ [en línia].

[14] Wikipedia. Scrum (desenvolupament de software). Wikipedia. Consultat el 21 de febrer del 2022, https://es.wikipedia.org/wiki/Scrum_(desarrollo_de_software) [en línia].

[15] 2022 GitHub. Documentació de GitHub. GitHub Docs. Consultat el 24 de febrer del 2022, a https://docs.github.com/es [en línia].

[16] Trello.org. Web de presentació de Trello. Trello. Consultat el 24 de febrer del 2022, a https://trello.com/ [en línia].

[17] Microsoft 365. Videoconferències, reunions, trucades. Microsoft. Consultat el 24 de febrer del 2022, a https://www.microsoft.com/es-es/microsoft-teams/group-chat-software [en línia].

[18] Universitat Politècnica de Catalunya & Facultat d'Informàtica de Barcelona. Treball de Fi de Grau | Facultat d'Informàtica de Barcelona. FIB-UPC. Consultat el 2 de febrer del 2022, a https://www.fib.upc.edu/ca/estudis/graus/grau-en-enginyeria-informatica/treball-de-fi-de-grau [en línia].

[19] Microsoft. (2021, November 8). Blog sobre l'editor de Codi VSC. Visual Studio Code. Consultat el 2 de febrer del 2022, a https://code.visualstudio.com/blogs/2021/11/08/custom-notebooks [en línia].

[20] Indeed. Salaris al sector tecnològic a Catalunya. Indeed. Consultat el 9 de març del 2022, a https://es.indeed.com/career/programador-junior/salaries/Catalunya [en línia].

[21] TarifaLuzHora. Preu mitjà de l'electricitat a Espanya. Precio de la tarifa de luz por horas HOY | Consulta ahora. Consultat el 9 de març del 2022, a https://tarifaluzhora.es/ [en línia].

[22] Hospital Clínic de Barcelona. (2018, November 20). Causes del Càncer. Hospital Clínic Barcelona. Consultat el 2 de març del 2022, a https://www.clinicbarcelona.org/ca/asistencia/malalties/cancer/causes-i-factors-de-risc [en línia].

[23] National Human Genome Research. Carcinogen. National Human Genome Research Institute. Consultat el 2 de març del 2022, a https://www.genome.gov/genetics-glossary/Carcinogen [en línia].

[24] Centres per al control i prevenció de malalties. Càncer de pulmó. CDC. Consultat el 2 de març del 2022, a https://www.cdc.gov/spanish/cancer/lung/ [en línia]

[25] Mezquita, L. Cáncer y medio ambiente [Article sobre les afectacions del medi ambient en diversos càncers]. Consultat el 5 de març del 2022.

[26] Agència Europea del Medi ambient. European Environment Agency's home page — European Environment Agency. Consultat el 5 de maig del 2022, a https://www.eea.europa.eu/ [en línia].

[27] Organització Mundial de la Salut. Pàgina de descàrregues de CI5plus. Download. Consultat el 25 de febrer del 2022, a https://ci5.iarc.fr/CI5plus/Pages/download.aspx [en línia].

[28] MongoDB. Web de MongoDB. MongoDB: The Application Data Platform | MongoDB. Consultat el 15 d'abril del 2022, a https://www.mongodb.com/ [en línia].

[29] MongoDB. Lloc web del servei Atlas. MongoDB Atlas | Multi-cloud Application Data Platform. Consultat el 15 d'abril del 2022, a https://www.mongodb.com/atlas [en línia].

[30] European Environment Agency. Eina de filtratge de dades. Discomap EEA. Consultat durant tot el projecte, a https://discomap.eea.europa.eu/App/AirQualityStatistics/index.html [en línia].

[31] Organització Mundial de la Salut. (2021, September 22). Descripció de la qualitat de l'aire i les afectacions en salut. WHO | World Health Organization. Consultat el 2 d'abril del 2022, a https://www.who.int/es/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health [en línia].

[32] Amazon. Cloud Computing - Serveis informàtics. AWS. Consultat el 18 d'abril del 2022, a https://aws.amazon.com/es/ [en línia].

[33] Geopy team. Documentació de Geopy. Welcome to GeoPy's documentation! — GeoPy 2.2.0 documentation. Consultat el 25 d'abril del 2022, a https://geopy.readthedocs.io/en/stable/ [en línia].

[34] Plotly. Web de mostra de gràfics en python. Plotly. Consultat el 25 d'abril del 2022, a https://plotly.com/python/ [en línia].

[35] The Pallets Projects. Web oficial de Flask. The Pallets Projects. Consultat el 3 d'abril del 2022, a https://palletsprojects.com/p/flask/ [en línia].

[36] Facebook Open Source. Web oficial de React. React – A JavaScript library for building user interfaces. Consultat el 3 d'abril del 2022, a https://reactjs.org/ [en línia].

[37] Wikipedia. Pàgina de Wikipedia sobre K-Means. Wikipedia. Consultat el 2 de maig del 2022, a https://en.wikipedia.org/wiki/K-means_clustering [en línia].

[38] Wikipedia. Pàgina de Wikipedia de l'algorisme jeràrquic. Wikipedia. Consultat el 2 de maig del 2022, a https://en.wikipedia.org/wiki/Hierarchical_clustering [en línia].

[39] Wikipedia. Pàgina de Wikipedia del Model Guassià. Wikipedia. Consultat el 2 de maig del 2022, a https://en.wikipedia.org/wiki/Gaussian_network_model [en línia].

[40] Boterhoven, D. (8 de Desembre del 2016). Blog sobre algorismes de hashing. Why you should use BCrypt to hash passwords. Consultat el 18 de maig del 2022, a https://danboterhoven.medium.com/why-you-should-use-bcrypt-to-hash-passwords-af330100b861 [en línia].

[41] THE INTERNATIONAL ASSOCIATION FOR THE STUDY OF LUNG CANCER. Web informativa sobre la IASLC. IASLC | International Association for the Study of Lung Cancer. Consultat el 28 de maig del 2022, a https://www.iaslc.org/ [en línia].

[42] THE INTERNATIONAL ASSOCIATION FOR THE STUDY OF LUNG CANCER. Web del congrés mundial del càncer de pulmó a Barcelona. IASLC 2019 World Conference on Lung Cancer (WCLC 2019). Consultat el 28 de maig del 2022, a https://wclc2019.iaslc.org/ [en línia].

Apèndix A

A.1. Aclariments sobre el codi

En aquest apèndix descriuré els continguts addicionals que acompanyen aquesta memòria. Aquest TFG compta amb aquesta memòria i un directori amb fitxers de dades, scripts i tots els arxius utilitzats per a desenvolupar la pròpia aplicació.

El contingut del directori extra compta amb la següent estructura:

- app: inclou el codi de la mateixa aplicació. L'estructura d'aquesta carpeta i el contingut dels seus arxius està descrit a l'apartat 12 de la memòria. Cal especificar que els arxius de càncer i contaminació a Espanya no formen part d'aquest treball sinó que són part del TFG d'en Joan Sart, el company amb el qual he estat treballant de forma paral·lela, per tant, pot ser que alguna part dels apartats de l'estudi Espanya estigui per acabar en el moment en què envio la memòria.

- **preprocessament:** conté les dades inicials i finals, així com els scripts fets servir per transformar aquestes dades. Aquest procés està detallat en l'apartat 11 de la memòria. Addicionalment, en aquest directori també s'inclouen els scripts emprats per a testejar els algorismes de classificació, explicats al capítol 13.3 d'aquest document.

A.2. Indicacions per execució

Existeixen dues formes generals d'executar aquesta aplicació per a visualitzat el resultat final del projecte. Primerament, es pot executar l'aplicació de forma local, però en aquest cas caldrà la instal·lació del programari necessari. Per una altra banda, es pot accedir directament a la web desplegada sense la necessitat de cap mena d'instal·lació prèvia. En aquest apartat es troben descrits els dos processos.

Local

Per a facilitar l'execució de l'aplicació en un entorn local he descrit dues maneres de fer-ho, la primera utilitzant un contenidor Docker i la segona utilitzant l'intèrpret de Python. Seguidament explicaré els dos processos:

- Docker:
    1. S'ha de tenir instal·lat el software Docker. Aquest es pot instal·lar des de el següent enllaç: Instal·lació Docker.
    2. Una vegada instal·lat el software cal obrir un terminal dins del directori app.

3. En aquest directori cal executar les següents comandes en ordre:
    a. docker build -t tfg-test .
    b. docker run -p 8000:8000 tfg-test
4. Això pot trigar uns minuts. El que fan aquestes comandes és generar un contenidor Docker amb la instal·lació de el que és necessari, en el contenidor, no en el teu ordinador, per a poder executar l'aplicació.
5. Es pot accedir a la web cercant en el navegador http://localhost:8000/ i inserint les següents credencials:
    Usuari: UPC
    Contrasenya: 12345

- Python:
    1. S'ha de tenir instal·lat el software Python. Aquest es pot instal·lar des del següent enllaç: Instal·lació Python.
    2. Una vegada instal·lat el software cal obrir un terminal dins del directori app.
    3. En aquest directori cal executar les següents comandes en ordre:
        a. pip install -r requirements.txt
        b. python app.py
    4. Això pot trigar uns minuts. El que fan aquestes comandes és instal·lar tots els paquets necessaris en el teu ordinador i executar l'aplicació.
    5. Una vegada executada l'aplicació sortirà un missatge en el terminal amb l'adreça http en la que pots visualitzar-la.
    6. Es pot accedir a la web cercant en el navegador l'adreça http obtinguda al pas anterior i inserint les següents credencials:
        Usuari: UPC
        Contrasenya: 12345

Cal destacar que la primera opció no descarrega cap mena de llibreria en l'ordinador local, com si és el cas de l'execució via Python.

Accés a la web

La web ha sigut desplegada en Google Cloud Platform, que és un servei gratuït de Google per desplegar aplicacions durant el període del seu testatge. Aquesta és la forma més senzilla per explorar la aplicació i és la que utilitzarà l'equip d'investigació, ja que no cal cap mena d'instal·lació. Es pot accedir fent ús d'aquest enllaç Accés a la web.

L'usuari i contrasenya que he creat per a la correcció del treball és:
Usuari: UPC
Contrasenya: 12345

Aquestes credencials són demanades en el moment de l'accés a la web amb l'objectiu que aquesta no sigui pública fins després de la presentació al Mundial d'oncologia justificat a l'apartat 14 de la memòria.