BRADEX

Lucía Sánchez Guillén

24 de juny de 2025

Resum– Aquest projecte consisteix en el desenvolupament d'una aplicació web intel·ligent per a la gestió d'immobles i contractes de lloguer enfocada a l'inversió inmobiliaria, integrant tecnologies modernes com Next.js per al frontend i Flask per al backend, desplegament al núvol i emmagatzematge d'arxius mitjançant Google Cloud Storage. L'aplicació incorpora funcionalitats d'intel·ligència artificial per les quals s'han generat dades sintètiques mitjançant la integració amb l'API Gemini de Google i s'ha realitzat un estudi de variabilitat de les dades amb mètriques com la similitut del cosinus i la entropía de Shannon per garantir la qualitat de les dades. Els models d'IA permeten la classificació i extracció de característiques de documents relacionats amb aquest àmbit. A més, el sistema inclou un panell d'anàlisi de la situació actual dels immobles, creació i seguiment d'incidències, així com enviament de notificacions, facilitant la interacció propietari-inquilí. Els resultats mostren una aplicació funcional, desplegada i operativa, amb potencial de millora i ampliació futura.

Paraules clau– Aplicació web, Intel·ligència artificial, Inversió immobiliària, Dades sintètiques, Classificació de documents, Extracció de característiques, Interacció propietari-inquilí

Abstract– This project involves the development of an intelligent web application for real estate and rental contract management, focused on real estate investment. It integrates modern technologies such as Next.js for the frontend and Flask for the backend, with cloud deployment and file storage via Google Cloud Storage. The application incorporates artificial intelligence functionalities for which synthetic data was generated through integration with the Google Gemini API. A data variability study was conducted using metrics like cosine similarity and Shannon entropy to ensure data quality. The AI models enable the classification and feature extraction of relevant documents within this domain. Additionally, the system includes a dashboard for analyzing the current status of properties, creation and tracking of incidents, and sending notifications, facilitating owner-tenant interaction. The results show a functional, deployed, and operational application with potential for future improvement and expansion.

Keywords– Web application, Artificial intelligence, Real estate investment, Synthetic data, Document classification, Feature extraction, Owner-tenant interaction

1 INTRODUCCIÓ

En el sector immobiliari actual, la gestió eficient d'una creixent quantitat de documents i la capacitat d'extreure informació rellevant de manera ràpida i precisa esdevenen reptes crucials per als inversors. Les plataformes de lloguer i gestió de propietats han proliferat, però moltes d'elles encara presenten mancances pel que fa a la integració de tecnologies intel·ligents que permetin optimitzar la classificació documental, la comunicació entre propietaris i llogaters, així com l'automatització de tasques repetitives.

Aquest projecte té com a objectiu abordar aquesta necessitat amb el desenvolupament de BRADEX (Bienes Raíces Adquisición y Explotación): una aplicació web intel·ligent orientada a la gestió immobiliària que facilita la comunicació propietari-inquilí i inclou la gestió de documents, contractes i notificacions d'una manera eficient i assistida per intel·ligència artificial (IA). L'aplicació permet, entre altres funcionalitats, la classificació automàtica de documents mitjançant models de machine learning, l'extracció de característiques rellevants d'aquests documents, i la gestió, anàlisi i manteniment dels immobles.

El projecte combina diferents tecnologies web i serveis externs per tal de construir una solució funcional, escalable i orientada a l'ús real. Així mateix, s'han explorat mètodes per generar dades sintètiques amb l'objectiu d'entrenar els models d'IA, i s'ha dut a terme un estudi sobre la variabilitat dels textos generats, com a part del procés de validació.

Aquest document recull el procés seguit durant el desenvolupament, les decisions preses, els resultats obtinguts, així com les possibles línies de millora i continuació futura.

2 ESTAT DE L'ART

En els darrers anys ha augmentat el desenvolupament d'aplicacions digitals per a la gestió de patrimoni immobiliari, com ara *AppFolio Property Manager*, *Maintenance Connection Software* o *Rentec Direct*. Però la proposta que guanya el primer lloc fins el moment es *Buildium* sense dubte, una plataforma de gestió de propietats immobiliàries pensada principalment per a administradors de finques i propietaris que gestionen múltiples immobles de lloguer. Està considerat com a software comercial (SaaS) que permet gestionar pagaments de lloguers online, sol·licituds de manteniment (incidències), comunicació amb inquilins, gestió de contractes i documents, seguiment de finances i informes, i fins i tot signatura electrònica de documents. Es cert que es tracta d'una aplicació molt completa i prometedora, pero també hem de considerar que està orientada sobretot a mercats dels Estats Units i Canadà, i no està específicament adaptada a sistemes legals o documentals d'altres països com Espanya. A més té una diferència clau respecte a la meva proposta i es que aquesta aplicació no té integrat el sistema de classificació automàtica de documents ni extracció de característiques que BRADEX proposa. Així doncs, quan pujes un document a Buildium has de seleccionar manualment a quina propietat correspon i etiquetar-lo, classificar-lo i organitzar-lo a la carpeta corresponent també de manera manual.

El que presenta BRADEX es automatitzar aquesta gestió i no només això sinó extreure conclusions de la informació dels arxius pujats per saber automàticament per exemple quina es la cuota que ha de pagar mensualment cada inquilí o inclús per fer anàlisis com saber la quantitat de propietats en relació al tipus d'inmoble. A més a més en un format molt visual i fàcil d'entendre.

3 TECNOLOGIES I EINES UTILITZADES

En aquest projecte s'han utilitzat diverses tecnologies i eines per a cobrir les diferents necessitats del sistema: desenvolupament web, entrenament de models d'intel·ligència artificial, generació de dades, emmagatzematge, desplegament i tractament d'informació. A continuació es detallen les principals:

- **Next.js + Typescript** → Frontend + gestió de la base de dades
- **Neon** → Base de dades amb PostgreSQL
- **Flask + Python** → Backend amb IA + pujar documents a la base de dades i al bucket
- **scikit-learn + SpaCy** → Llibreries utilitzades per l'entrenament dels models
- **PDFMiner** → Llibreria utilitzada per l'extracció de text dels documents PDF
- **Google OAuth i Next Auth** → Autenticació de Google
- **Google Buckets** → Per guardar els arxius al núvol
- **Google Gemini API** → Per generació de dades sintètiques
- **Vercel** → Eina per desplegar el frontend de l'aplicació
- **Visual Studio Code** → IDE per programar
- **Git i GitHub** → Sistemes de control de versions

4 ARQUITECTURA DEL SISTEMA

L'aplicació s'ha desenvolupat de mode que el frontend gestiona la base de dades i mostra la informació mentre que el backend gestiona els models i la classificació amb IA, de manera general.

• E-mail de contacte: lucia.sanchezg@autonoma.cat
• Menció realitzada: Enginyeria de Computació
• Treball tutoritzat per: Ronny Romero Navarrete (Àrea de Ciències de la Computació i Intel·ligència Artificial)
• Curs 2024/25

4.1 Fluxe de connexions

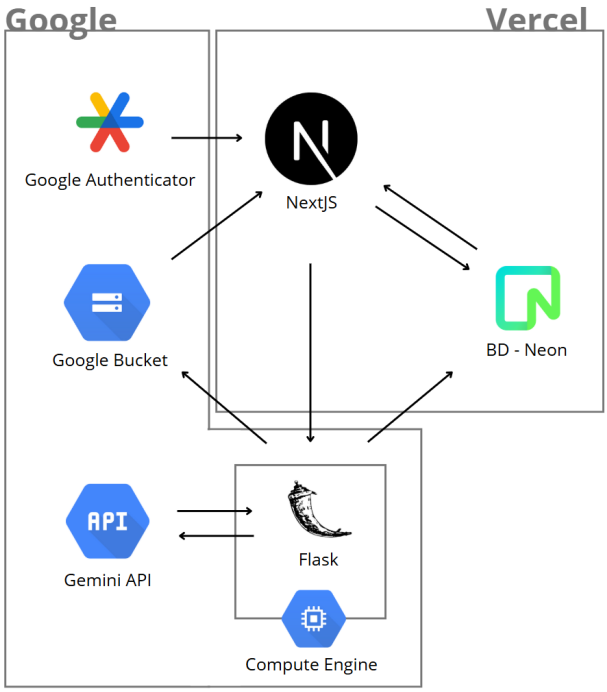


Fig. 1: Diagrama d'arquitectura

Com podem veure el diagrama de l'arquitectura del sistema anterior mostra les connexions entre els diferents components que intervenen en l'aplicació. El fluxe comença amb l'usuari de tipus propietari que inicia sessió a l'aplicació a través de Google Authenticator. Una vegada ha creat el compte o ha iniciat la sessió, pot començar a afegir immobles i inquilins creant-se les instàncies i emmagatzemant la informació a la base de dades. A partir d'aquest punt ja pot començar a carregar els seus arxius com ara contractes, rebuts o documentació tant dels inquilins com dels seus immobles i aquests s'envien al backend a través de Flask on es processaran passant-los pels models prèviament entrenats, es definirà el tipus de document i s'extrauran les seves característiques. Una vegada això, s'enviaran els arxius al bucket de google i es crearan les instàncies corresponents a la base de dades deixant el contingut de l'arxiu al bucket i la informació a la base de dades. Posteriorment l'usuari podrà veure a l'aplicació tots els arxius de manera ordenada automàticament, filtrar per informació rellevant i anàlitzar les dades gràcies als gràfics visuals que s'alimenten d'aquests arxius.

Cal destacar que tot el projecte es troba al núvol, per part del frontend utilitzem Vercel mentre que pel backend s'ha creat una màquina virtual a Google. A més les apis externes que s'utilitzen es troben també a Google. Això permet escalabilitat, disponibilitat d'accés a qualsevol hora i lloc, reducció de costos, seguretat i velocitat en el manteniment, a més a més, també ho considero una bona pràctica.

4.2 Model de dades

Per part de l'aplicació, s'ha dissenyat una arquitectura de base de dades personalitzada i adaptada a les seves necessitats amb un total de vuit taules amb Neon utilitzant PostgresSQL. La informació de la base de dades s'obté des del frontend mitjançant peticions http utilitzant les Route Handlers de Next on he implementat uns formularis com per exemple per afegir immobles i també del backend per part de la taula d'arxius on es generen les instàncies una vegada processats. A continuació tenim un resum de les taules amb la seva funcionalitat.

TAULA 1: TAULES BASE DE DADES

Nom taula	Funcionalitat
Usuario	Emmagatzema informació dels usuaris, tant propietaris com inquilins.
Inmueble	Registra les propietats disponibles a la plataforma.
Archivo	Guarda les metadades associades als documents vinculats a un contracte o a un inmoble concret.
Contrato	Representa els acords establerts entre un propietari i un inquilí.
Incidencia	Guarda les incidències notificades pels usuaris referents a un inmoble.
Incidencia estado historico	Permet fer el seguiment dels canvis d'estat d'una incidència al llarg del temps.
Notificacion	Registra tot tipus de notificacions destinades als usuaris.
Nota	Guarda les notes personals dels inquilins escrites pels propietaris.

A l'annex podreu trobar el diagrama ER on es poden veure tots els camps definits i les relacions entre taules a la *figura 4* de la secció *A.1 Diagrama ER de la base de dades*.

5 METODOLOGÍA

Per a la gestió i organització de les tasques del projecte, s'ha optat per utilitzar una metodologia basada en el sistema **Kanban**. Aquest enfocament és especialment útil per visualitzar el flux de treball, permetent una gestió àgil i eficient de les activitats.

5.1 Què és Kanban?

Kanban és un mètode visual que permet gestionar el treball en progrés de manera eficient. La seva principal característica és l'ús d'un tauler dividit en columnes que representen els diferents estats de les tasques.

5.2 Estructura del tauler Kanban

En aquest projecte, el tauler Kanban s'organitzarà en tres columnes principals:

- **Per fer (To Do)**: En aquesta columna s'inclouran totes les tasques identificades que encara no han començat. Aquestes tasques representen el treball pendent.
- **En progrés (In Progress)**: Un cop una tasca s'inicia, es mourà a aquesta columna. Aquí es reflectiran les activitats que s'estan executant en aquell moment.

- **Completades (Done)**: Quan una tasca finalitza, es mourà a aquesta columna. Això proporciona una visió clara del treball acabat i ajuda a mantenir un registre del progrés general del projecte.

5.3 Implementació en el projecte

El tauler Kanban s'ha implementat en un word de manera manual seguint un sistema de colors (blanc per 'To Do', groc per 'In progress' i verd per 'Done'). Es una manera senzilla pero eficient que permet una actualització ràpida.

6 DESENVOLUPAMENT DE MODELS D'IA

La integració de la IA en aquesta aplicació es clau ja que esdevé la base del projecte i li dona l'objectiu dessitjat que es l'automatització de tasques i manteniment en processos immobiliaris. He decidit dividir els models en dos grans grups, per un costat la classificació de documents i per l'altre l'extracció de característiques. Amb la classificació de documents volem acconseguir una etiquetació determinada on es reconeix el tipus de document mentre que amb l'extracció de característiques volem obtenir la informació important pels anàlisis posteriors.

6.1 Model de classificació

L'objectiu d'aquest model és classificar automàticament els documents pujats per l'usuari segons el seu tipus, a partir d'una llista d'etiquetes predefinides com la següent.

- `cèdula habitabilitat`
- `certificat energètic`
- `contracte lloguer`
- `escriptura immoble`
- `pòlissa segur`
- `rebut`
- `registre catastral`

A partir dels textos generats sintèticament, es realitza una representació vectorial mitjançant la tècnica **TF-IDF** (Term Frequency - Inverse Document Frequency). Aquesta transformació converteix cada paraula en un valor numèric segons la seva freqüència d'aparició, i el conjunt de tots aquests valors forma un vector que representa el document.

Un cop vectoritzats els documents, s'entrena el model de classificació amb les etiquetes definides. S'ha decidit utilitzar el **Naive Bayes multinomial**, un model probabilístic especialment adequat per a dades discretes com les freqüències de paraules. Aquest model estima la probabilitat de cada tipus donat el vector de paraules del document i assigna la categoria amb probabilitat màxima. L'entrenament s'ha realitzat amb una divisió entre train i test molt comú on es deixa un 70% de les dades per a entrenament i un 30% per a prova.

Posteriorment es passen una sèrie de proves de validació començant per una validació creuada K-Fold amb 5 particions tenint com a mètrica l'accuracy. També s'ha considerat inclou-re un **classification report** que es tracta d'un informe automàtic on s'avaluen diferents mètriques. A continuació podem veure el total de mètriques que s'han considerat junt amb la seva explicació.

- **Precisió** (*accuracy*): proporció total d'encerts.
- **Precisió per classe** (*precision*): proporció d'encerts sobre les prediccions per a cada categoria.
- **Recall**: proporció d'encerts sobre els casos reals de cada categoria.
- **F1-score**: mitjana harmònica entre precisió i recall.

D'altra banda, s'ha generat una **matriu de confusió**, que mostra en format visual els errors de classificació entre categories. Aquest gràfic és molt útil per veure si existeix confusió entre les dades i trobar tendències entre classificacions errònies.

Per últim, per tal de guardar els resultats obtinguts, es crea automàticament un informe HTML amb totes les mètriques i gràfics corresponents per tal de poder accedir als resultats de manera més còmode i ràpida.

6.2 Model d'extracció de característiques

Per tal d'extreure les característiques més rellevants dels arxius, com per exemple detectar l'inquilí a un contacte, ordenar els arxius correctament o saber les quotes de pagament, s'ha seguit una sèrie de passos detallats a continuació.

El primer pas va ser definir les característiques dessitjades, pensar què necessitava pel funcionament de l'aplicació i definir-ho. El segon pas va ser la anotació o preparació de les dades per a l'extracció. Aquest pas es fonamental ja que qualsevol errada conduirà a un mal entrenament i per tant a uns mal resultats. Per portar-ho a terme s'ha utilitzat **Gemini API**, una eina que permet estalviar temps i evitar possibles errades. El que s'ha fet ha sigut implementar un script on truquem a la API enviant-li un prompt definit. En aquest prompt indiquem que volem etiquetar el tipus 'X' de document amb les etiquetes 'Y' y que retorni l'etiquetació en format: [nom etiqueta, fracció de text corresponent, fracció ampliada del text corresponent, posició d'inici, posició de fi].

El funcionament aquí s'acconsegueix identificant la posició d'inici i de fi de l'etiqueta al text a més que indica directament el contingut de text que pertany a l'etiqueta en si. Aquest tractament inicial fallava molt degut a que el text que s'enviava a l'API no era el mateix rebut i per tant no s'acconseguia definir bé la posició de les etiquetes al text. Així doncs vaig implementar un procés de validació dels offsets agafant també el context de l'etiqueta, per aquest motiu es demana a l'API també per la fracció ampliada del text que vol dir agafar uns caràcters més al voltant de l'etiqueta per trobar després la posició exacta al text més propera a les posicions d'inici i de fi definides pel model.

També he implementat un control de solapaments pel fet que les etiquetes no es solapin amb altres etiquetes, ja que per l'entrenament posterior donaria errors. Finalment com a part del tractament de l'etiquetament s'ha implementat un control d'etiquetes úniques, deixant només un resultat per etiqueta dins de cada text, ja que això també donaria problemes més endavant.

Una vegada etiquetades les dades podem començar a entrenar el model. En aquest cas, s'ha decidit utilitzar **spaCy**, una llibreria de Processament del Llenguatge Natural (PLN). Més en detall, s'ha decidit implementar un model propi per tal d'ajustar-lo als objectius concrets amb NER (Named Entity Recognition).

El NER és una tècnica fonamental dins del camp del PLN que consisteix a identificar i classificar entitats específiques (com ara ”propietari”, ïnquilí”, ”direcció de l’inmoble”o ”preu”) dins d’un text, assignant-los una etiqueta predefinida. Amb les dades ja etiquetades, spaCy aprendrà els patrons i el context per detectar i classificar aquestes entitats en nous documents, adaptant-se a la terminologia i l’estructura pròpia dels arxius.

En primer lloc per portar a terme l’entrenamet es va dissenyar l’arxiu config.cfg, aquest es tracte d’un detall de tota la configuració que portarà el model incloent els paràmetres d’entrenament. En aquest cas he decidit utilitzar una pipeline de ’tok2vec’ i ’ner’ que serien el tokenitzador i el model de reconeixement d’entitats amb les meves dades previament etiquetades. Per tant el procés començarà amb la tokenització dels textos. El que fa es agafar les característiques de les paraules i transformar-les en vectors numèrics contextualitzats. El que vol dir que aquesta vectorització té en compte el significat semàntic i sintàctic de les paraules així com el seu context. Posteriorment, la part del reconeixement de les entitats pren aquests vectors anteriors com entrada per l’entrenament. Aquest factor es clau pel projecte ja que ens dona la propietat d’identificar el context de les paraules i no només les paraules en si. Un cas d’exemple serien els noms de persona d’un contracte. D’una banda es troba el nom del propietari i d’altre el nom de l’inquilí, tots dos son noms de persona, pero aquest model permet diferenciar concretament entre nom de propietari i nom d’inquilí a partir del seu context.

Parlant dels paràmetres d’entrenament s’ha decidit utilitzar l’optimitzador Adam amb un learning rate de ’0.001’ i un L2 de ’0.01’ a més d’un dropout de ’0.2’. El learning rate es la taxa d’aprenentatge del model, es podria entendre com el tamany dels pasos que donem en l’algoritme en cada iteració per ajustar els pesos del model. El L2 es un altre ajust d’optimització, aquest per regular el sobreajustament, es a dir, que aprengui massa bé les dades. Funciona afegint una penalització al tamany dels pesos del model durant l’entrenament per acconseguir distribuir de manera més equitativa la importància de les característiques. El dropout també ajuda a controlar una mica el sobreajustament mitjançant ’l’apagament’ de neurones aleatories, es a dir, que els càlculs i sortides d’algunes neurones no es porten a terme, en aquest cas el 20% de les neurones.

Finalment, una vegada ajustat l’arxiu de configuració, s’executa la comanda d’entrenament i es produeix l’entrenament dels models d’extracció de característiques. A partir d’aquí ja estan llestos per fer les seves prediccions.

Podreu trobar el conjunt de característiques definides a l’apèndix a la *taula 7* a la secció *A.2 Detall de característiques dels arxius*.

6.3 Generació de dades sintètiques

Després d’una cerca exhaustiva a internet per trobar datasets o informació d’arxius relacionats amb els tràmits immobiliaris vaig determinar que al tractar-se d’informació majoritàriament confidencial no podria obtenir el conjunt de dades desitjat per poder entrenar els models de manera senzilla, així doncs vaig decidir generar dades sintètiques per tal d’acconseguir les dades per a l’entrenament dels models.

Per començar amb la generació d’aquestes dades vaig optar per utilitzar IA Generativa com ChatGPT i Gemini de manera manual. Aquest procés consistia en escriure un prompt manual demanant la generació de textos d’arxius ficticis amb les etiquetes del tipus d’arxiu. Tot anava bé però m’estava portant més temps del esperat ja que havia de generar un nombre suficient de dades per poder entrenar els models de la millor manera. Per aquest motiu vaig decidir avançar una mica més i automatitzar aquesta tasca utilitzant l’API de Gemini de Google. Això va supossar una gran diferència, a més vaig decidir establir un nombre determinat de dades amb una xifra suficient (500 arxius) per a poder entrenar el model.

El problema va venir quan vaig començar a analitzar les dades generades. Em vaig trobar que el contingut de les dades era massa semblant per als diferents casos d’un mateix arxiu on molta informació estava duplicada. Aquest fet es provocava degut a que la IA Generativa sempre dona la resposta més probable o la que creu que es més adecuada portant a deduir que a mateix prompt mateixa resposta o molt semblant.

Vaig descobrir doncs que existeix un paràmetre anomenat temperatura que es el que proporciona aleatorietat en les respostes i que de manera determinada es troba en uns valors baixos o mitjans per garantir coherencia i qualitat el que em provocava dificultats en la variabilitat de les dades. Aquest paràmetre pot prendre els valors de 0 (totalment determinat) a 2 (totalment aleatori). Per tant vaig modificar aquest paràmetre a 1.9, aumentant l’aleatorietat i vaig poder comprovar que efectivament s’havia produit una millora en el contingut de les dades. D’altra banda, per part del nombre de dades, vaig modificar el total de dades per a que em generés els 500 textos amb factor d’aleatorietat de tipus de text, així poder tenir una mica de variació en la quantitat de dades de cada tipus d’arxiu.

Una altre millora que vaig implementar va ser la variabilitat en el prompt. Ho vaig acconseguir afegint variables al prompt, ja que fins el moment utilitzava prompts estàtics variant només el tipus de document a generar. Aquesta millora consisteix en introduir un factor d’aleatorietat en el prompt amb probabilitat equitativa en diferents característiques. El que vaig trobar interessant va ser la zona i la ubicació. Per part de la zona vaig considerar optar per: bona, dolenta, cara, barata i humil. D’altre banda, per part de la ubicació vaig decidir utilitzar totes les províncies d’Espanya extretes d’una api anomenada **GeoAPI**. Així doncs ara el prompt no només tenia aleatorietat en els tipus de documents sinó també en la zona i la província, el que donava més possibilitats de resultats variats.

En conseqüència, cada vegada que he millorat el conjunt de dades he executat el model, per veure la variabilitat en la seva precisió. A continuació podem veure les diferents versions del model i les millores efectuades per cadascú.

6.4 Estudi de variabilitat de les dades

Per analitzar la consistència i homogeneïtat de les dades sintètiques generades s’ha realitzat un estudi de la variabilitat mitjançant la similitud del cosinus i l’entropía de Shannon.

La similitud del cosinus és una mètrica utilitzada per comparar documents o cadenes de text representades com

TAULA 2: VERSIONS DEL MODEL I DADES UTILITZADES

Versió model	Explicació dades	Nº de dades
Manual	Dades manuals	91
Gemini v1	Dades amb Gemini API	389
Gemini v2	v1 + modificació del paràmetre de temperatura	500
Gemini v3	v2 + modificació de prompt amb variables	500

vectors. Aquesta mesura calcula el cosinus de l'angle entre dos vectors en un espai vectorial, normalment obtinguts mitjançant una tècnica de representació com TF-IDF o embeddings semàntics.

La fórmula de la similitud del cosinus és:

$$\text{sim}_{\cos}(\vec{A}, \vec{B}) = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\|\|\vec{B}\|}$$

El resultat doncs pot trobar-se entre 0 (completament diferents) i 1 (completament similars). Aquesta mètrica és realment útil per identificar textos amb significat similar encara que estigui expressat amb paraules diferents. Per aquest motiu la vaig escollir per calcular la variabilitat en el meu cas. La seva implementació va ser agafant textos de dos en dos per cada categoria, ja que el que volia era analitzar la variabilitat dins d'un mateix tipus de document.

D'altra banda, l'entropía de Shannon l'he utilitzat per mesurar la incertesa en el contingut d'etiquetes aplicades. Un valor alt vol dir que les etiquetes estan distribuïdes de manera uniforme, es a dir que hi ha més varietat de contingut d'etiquetes. En canvi, un valor baix indica que hi ha poca diveristat en les etiquetes.

La fórmula de la entropía de Shannon és:

$$H(S) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

Per tal de poder analitzar els resultats s'ha portat a terme una normalització posterior on acotem l'entropía a 0 com el valor més baix i a 1 com el valor més alt. Així doncs, un valor proper a 1 indicaria alta diveristat i uniformitat d'etiquetes mentre que un valor proper a 0 indicaria que tota la informació extreta per a les característiques es semblant. Amb aquest estudi podriem identificar doncs si el model ha creat registres catastrals amb la mateixa direcció d'inmoble o contractes amb el mateix nom de propietari, per exemple.

7 FUNCIONALITATS DESENVOLUPADES DE BRADEX

A continuació es presenten totes les funcionalitats que ofereix BRADEX, organitzades per àrees:

- **Gestió d'usuaris**
  - Registre i inici de sessió amb autenticació via Google.
  - Gestió dels rols dels usuaris: propietari o inquilí.
  - Emmagatzematge i visualització de dades del perfil: nom, email, foto i telèfon.
  - Control de la data de registre i el tipus d'usuari per anàlisi intern.
  - Personalització de l'aplicació segons el tipus d'usuari.
- **Gestió d'immobles**
  - Registre de propietats: habitatges, locals, oficines, pàrquings i trasters.
  - Gestió de característiques de l'immoble: adreça, superfície, habitacions, banys, moblatge, terrassa, preu, etc.
  - Associació del propietari a cada immoble.
  - Gestió de l'estat de disponibilitat de l'immoble.
  - Registre de notes i observacions.
- **Gestió de contractes**
  - Creació de contractes entre propietaris i inquilins.
  - Enllaç del contracte amb un immoble concret.
  - Control de l'estat del contracte: pendent, acceptat, rebutjat.
- **Gestió d'arxius**
  - Pujada de documents vinculats a contractes o propietats.
  - Classificació automàtica dels arxius pujats.
  - Accés als arxius per visualització o descàrrega.
- **Gestió d'incidències**
  - Creació d'incidències per part dels usuaris sobre un immoble.
  - Classificació per tipus i motiu (goteres, sorolls, avaries, etc.).
  - Seguiment de l'estat de l'incidència: pendent, en progrés, resolta, cancel·lada.
  - Registre de la data de report i de resolució.
  - Descripció detallada del problema.
- **Històric d'estats d'incidències**
  - Registre dels canvis d'estat de cada incidència.
  - Inclou l'estat anterior i nou, data del canvi i usuari que el va realitzar.
  - Permet un seguiment complet de l'evolució d'una incidència.
- **Gestió de notificacions**
  - Notificar venciments de contractes per evitar incompliments.
  - Enviament de notificacions internes de l'aplicació com per exemple per acceptar contractes de lloguer o notificar incidències.

8 AVALUACIÓ DE RESULTATS

Malgrat les dificultats trobades al llarg del desenvolupament del projecte, puc dir que finalment he arribat a obtenir uns molt bons resultats. A més s'ha conseguit una aplicació completa i totalment funcional que inclou satisfactòriament la funcionalitat dels models entrenats.

8.1 Models d'IA

A nivell dels models d'IA implementats i desenvolupats he acconseguit realment l'objectiu proposat obtenint un classificador que funciona i que a més funciona bé. Entrant una mica més en detall en el model de classificació, podem veure els resultats de la precisió (accuracy) per cada versió de models implementada tant de la validació creuada (k-fold) com del conjunt de proba (test).

TAULA 3: RESULTATS PER VERSIÓ DE MODEL

Versió model	K-Fold	Test
Manual	0.93	1.0
Gemini v1	0.88	0.85
Gemini v2	0.90	0.94
Gemini v3	0.90	0.92

Com podem veure existeixen uns molt bons resultats des del principi amb una mitja del 90% de precisió. Pero s'ha de tenir en compte que aquests resultats avaluen la precisió segons aquestes mateixes dades i no amb dades reals on el canvi pot esdevenir una gran disminució de aquesta precisió. A més, podem observar que per al primer cas de dades manuals teniem moltes menys dades (91 textos) el que ens porta a pensar que degut a l'escassetat s'hagi produit overfiting, es a dir sobreajustament de les dades, donant per tant resultats invàlids. Per aquest motiu també s'ha considerat fer l'estudi sobre la variabilitat de les dades i l'estudi amb noves dades explicats més endavant.

D'altra banda, una vegada amb les dades millorades com les de la versió tres, vaig considerar modificar el **random_state** dels entrenaments. Aquest es tracta d'un paràmetre que determina la aleatorietat mitjançant pseudo-aleatorietat, així doncs podem considerar el factor d'aleatorietat pero sempre que utilitzem el mateix valor per aquest paràmetre obtendrem els mateixos resultats. De manera general s'utilitza el valor 42 que es el valor que he utilitzat per als entrenaments, però aquest valor es pot modificar per tal d'arribar a uns millors resultats. Per tant, vaig decidir comparar els resultats modificant aquest valor per veure si podia obtenir una millor precisió. Com a resultat vaig acconseguir una precisió de fins el 97% amb un valor random state de 32.

A continuació veiem la matriu de confusió final obtinguda amb la versió tres de gemini i aquest canvi de random state a 32.

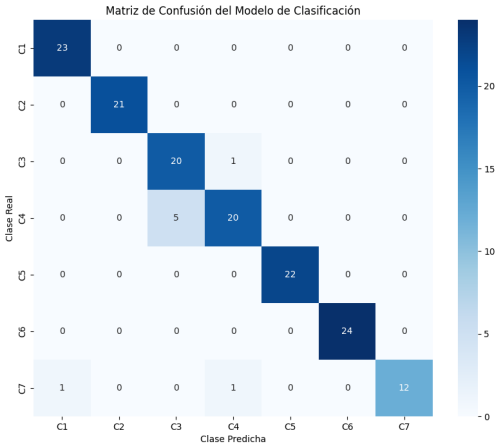


Fig. 2: Matriu de confusió Gemini v3 amb random state 32

La llegenda de clases es:

C1: Cèdula d'habitabilitat
C2: Certficat energètic
C3: Contracte de lloguer
C4: Escriptura de l'inmoble
C5: Pòlissa de segur
C6: Rebut
C7: Registre Catastral

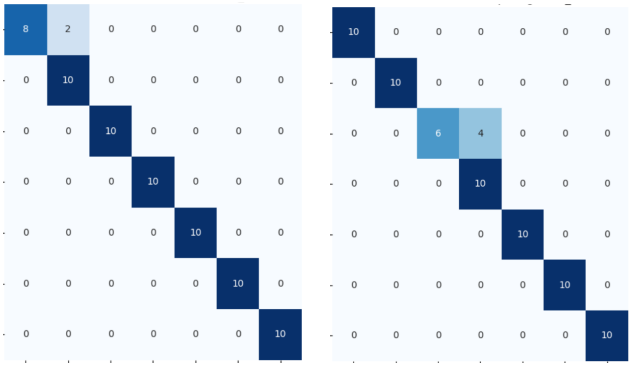
Si analitzem la matriu podem veure que el model tendeix a confondre sobretot els documents de contractes de lloguer amb els d'escritura de l'immoble. Per aquest motiu vaig decidir fer un estudi de totes les versions del model que tenia per veure si potser amb una versió anterior acconseguia millors resultats, no tant en la precisió sinó centrant-me més en el recall. Per aquest estudi el que vaig fer va ser generar noves dades aquest cop en format PDF com si fossin arxius reals, en concret deu de cada tipus d'arxiu i vaig fer la prova d'anàlitzar els resultats de cada model amb aquestes dades. A continuació podem veure el resultat de les mètriques obtingudes, també amb la millor versió de l'últim model amb la variació del random state comentada.

TAULA 4: RESULTATS DE L'ESTUDI AMB NOVES DADES PDF

Versió model	Precisió	Recall	F1-Score
Manual	0.58	0.67	0.60
Gemini v1	0.37	0.46	0.38
Gemini v2	0.98	0.97	0.97
Gemini v3	0.96	0.94	0.94
Gemini v3 rs32	0.95	0.94	0.94

Podem veure que el model que dona millors resultats es el Gemini v2 amb una precisió més fiable de 0.98. A més a més no ens importa només el percentatge de precisió, sino analitzar en quines classes fallava més i quines menys

per donar més pes a segons quins errors. A continuació podem veure la comparativa de les dues matrius de confusió de Gemini v2 i Gemini v3.



(a) Gemini v2 (b) Gemini v3

Fig. 3: Comparativa resultats reals de les dues millors versions

El que veiem es que per a la versió dos obtenim un resultat gairebé perfecte excepte per les clases de cèdula d'habitabilitat i certificat energètic. En canvi, per part de la versió tres obtenim el mateix però en aquest cas confón les clases de contracte de lloguer i escriptura d'inmoble, a més amb una mica més de diferència. Per tant, una vegada vaig veure aquesta comparativa vaig decidir quedar-me amb el model de la versió dos degut a que considero més important la correcta classificació de contractes de lloguer que de cèdules d'habitabilitat.

8.2 Estudi de variabilitat de dades

Un dels majors reptes del projecte ha estat la generació de dades de qualitat per a l'entrenament dels models. Però, gràcies a l'estudi de variabilitat realitzat, s'ha pogut analitzar una mica més en profunditat les dades i poder implementar millores.

A continuació veiem els resultats de l'estudi de variabilitat realitzat on veiem els resultats tant de la similitud del cosinus com de la entropía de Shannon per cada tipus d'arxiu.

TAULA 5: RESULTATS ESTUDI VARIABILITAT

Tipus d'arxiu	Similitud del cosinus	Entropía de Shannon
Contracte de lloguer	0.2089	0.9470
Certificat energètic	0.1981	0.9444
Pòlissa Segur	0.1644	0.9182
Rebut	0.1440	0.9172
Escriptura inmoble	0.1689	0.7458
Registre catastral	0.1570	0.4446
Cèdula d'habitabilitat	0.1547	0.8181

Per part de la similitud del cosinus veiem que mes o menys tots els tipus d'arxius es troben al voltant de 0.15. Tenint en compte que es tracta d'un valor més proper a 0 podem interpretar que els textos de les nostres dades son molt diferents. Això és positiu d'una banda, ja que ens informa que s'estan utilitzant paraules diferents a la majoria del textos i per tant sí que existeix aquesta diversitat que buscàvem. Però, d'altra banda, pot arribar a ser quelcom negatiu pels models ja que si el contingut és massa diferent costarà trobar patrons entre els diferents tipus de textos que siguin útils per a la seva diferenciació.

Per part de l'entropía de Shannon veiem que gairebé per tots els tipus d'arxius tenim un valor molt proper a 1. Amb això podem dir que sí que existeix una varietat en el contingut de les etiquetes i tenim dades diverses amb les que treballar. Pel contrari, trobem que el tipus d'arxiu 'Registre catastral' ha obtingut una entropía molt baixa en comparació a la resta. Després d'analitzar el motiu d'aquest desajust trobem que es deu al fet que el model ha inventat etiquetes no definides a l'hora de l'etiquetació i per tant porta a errors de càlcul. Es a dir, per exemple, crea l'etiqueta 'Ascensor' amb un sol registre, el que porta a una entropía de 0, ja que només un arxiu conté aquesta etiqueta i per tant, aquest contingut. Així ha passat també amb més etiquetes el que com a resultat afecta a la mitja final. Per evitar això he implementat un ajust agafant només les etiquetes vàlides que serien les definides inicialment. A continuació veiem els resultats finals de la entropía de Shannon.

TAULA 6: RESULTATS ENTROPÍA CORREGIDA

Tipus d'arxiu	Entropía de Shannon
Contracte de lloguer	0.9572
Certificat energètic	0.9444
Pòlissa Segur	0.9182
Rebut	0.9172
Escriptura inmoble	0.8453
Registre catastral	0.8387
Cèdula d'habitabilitat	0.8671

Ara amb els nous resultats podem veure que efectivament s'ha produït un increment general de la entropía, sobretot al tipus 'Registre catastral' on hem passat d'un 0.44 a un 0.83. Per tant, ara podem dir que tenim una alta entropía de manera general amb les etiquetes definides i això indica que tenim una bona diversitat en el contingut de característiques en els nostres arxius de manera satisfactòria.

8.3 Generals - web

Pel que fa a l'aplicació web en el seu conjunt, s'ha pogut desplegar una plataforma funcional, usable i robusta per a la gestió automatitzada del manteniment d'immobles. Aquesta integració efectiva entre la part de backend i el frontend ha estat clau per portar a terme el projecte. Finalment s'ha aconseguit una solució que funciona correctament i integra els models d'IA entrenats satisfactòriament.

Per l'aplicació s'ha desenvolupat una petita marca on he decidit dissenyar una mica la identitat visual incloent els colors i el logo per donarli estil a la marca creada. Així doncs

tota l'aplicació segueix aquest estil aportant uniformitat visual a la interfície.

A continuació deixo l'enllaç directe a la pàgina web.

`https://bradexnextjs.vercel.app`

9 MILLORES I TREBALL A FUTUR

Tot i que el projecte desenvolupat ha acconseguit ofererir una funcionalitat completa, els models es poden millorar i BRADEX es pot ampliar molt més. A continuació presento les millores i treball a futur que penso que acabarien per perfeccionar l'aplicació:

- **Entrenament del model amb dades reals**: Degut a que el tipus de dades amb les que havia de tractar eren dades amb contingut confidencial i informació restringida no s'ha pogut entrenar els models amb dades reals, per tant una millora molt important es poder entrenar els models amb dades reals per així poder aproparnos més a un resultat de la vida real.

- **Anàlisi d'inversió**: Un anàlisi detallat de propietats per poder trobar bones oportunitats, calcular les inversions i veure prediccions sobre com es desenvoluparia cada possible inversió. Així com poder veure l'estat actual del nostre diner i el nostre patrimoni a més de l'evolució històrica. Veure la diversificació i qué ens està donant més o menys beneficis. També es podría fer un anàlisi predictiu utilitzant IA per predir tendències del mercat immobiliari i oferir recomanacions d'inversió o predicció d'impagaments. Per últim es podría implementar funcionalitats investigació del mercat actual per veure informació detallada del mercat actual inmobiliari.

- **Processament intel·ligent de documents amb visió per computador**: Extracció de text en documents de diferents formats com imatges.

- **Integració amb APIs de tercers per pagament**: Connectar l'app amb plataformes de pagament (PayPal, Bizum) per facilitar el cobrament de lloguers o altres pagaments.

- **Chat amb IA**: Implementar un chatbot per fer preguntes tant com de questions específiques d'inversió i rentabilitat com d'ajuda i informació de l'aplicació com on es troba cert arxiu o quin es el pròxim venciment.

- **Plantilles de documents**: Creació de plantilles per contractes, notificacions de venciment, renovacions, etc. Inclús generació automàtica d'aquests a partir de dades d'entrada.

- **Millora d l'infraestructura al núvol**: Configuració de bucket amb IP privada i accés segur mitjançant Google Cloud SQL Proxy, desplegament a Google Cloud Run.

- **Millorar l'escalabilitat**: Utilitzar eines per poder atendre una quantitat més gran d'usuaris i de peticions. A més de millorar el temps de resposta i eficiència.

- **Gestió de múltiples propietaris o inquilins**: Permetre que un immoble tingui més d'un propietari o inquilí amb gestió de percentatges o responsabilitats compartides.

- **Distinció entre tipus de propietari**: Afegir una categoria per identificar si un propietari és una empresa o un particular i adaptar el contingut com impostos.

- **Ampliació de dades**: Afegir més informació a les taules com per exemple a la taula 'Immoble' el camp 'orientació' (nord, sud, est, oest) o un motiu específic pel qual un immoble no està disponible (en reforma, okupat, intervenció policial, etc.) per ampliar els anàlisis.

- **Sistema de comentaris o valoracions**: Permetre als inquilins valorar immobles o propietaris per afavorir la transparència.

- **Sistema de recordatoris**: notificacions automatitzades per venciments, tasques pendents, renovacions, etc.

- **Internacionalització**: Traducció de la interfície a diversos idiomes per adaptar-se a usuaris internacionals.

- **APP Mòvil**: Desenvolupar una app mòbil per facilitar l'accés en qualsevol moment i lloc.

10 CONCLUSIONS

La realització d'aquest projecte ha suposat un repte tècnic i personal que m'ha permès posar a prova i consolidar coneixements adquirits al llarg del grau. A més, ha estat una oportunitat per enfrontar-me a situacions de disseny i desenvolupament d'un sistema complex, complet i real.

Un dels aprenentatges més significatius ha estat la importància de la planificació i estructuració prèvia del projecte. Definir bé l'arquitectura, anticipar les necessitats de la base de dades i tenir clar el flux d'informació entre els diferents components ha estat clau per evitar problemes en fases posteriors.

Pel que fa al desenvolupament, he guanyat fluïdesa i criteri a l'hora d'utilitzar frameworks moderns i he descobert nous llenguatges i tecnologies molt interessants.

Finalment, puc concloure que aquest projecte no només ha cobert els objectius inicials, sinó que m'ha ajudat a créixer com a desenvolupadora, a entendre millor la complexitat d'un projecte complet i real i a valorar la importància del detall, tant a nivell tècnic com d'experiència d'usuari.

AGRAÏMENTS

En primer lloc vull donar-li les gràcies a tota la meva família per sempre donar-me suport en tots els meus propòsits, per escoltar-me i per donar-me sempre la seva millor opinió. En especial al meu pare qui s'ha implicat molt i m'ha ajudat a definir bé els objectius i la proposta del projecte. D'altra banda vull agraïr també al meu tutor per guiar-me durant tot aquest procès i ajudar-me a millorar per acconseguir un bon resultat final.

REFERÈNCIES

[1] Llibre de Carlos Galán Rubio. *Libertad Inmobiliaria: Consigue la libertad financiera con la inversión inmobiliaria*.
Ús: S'ha llegit aquest llibre per aprendre sobre inversió inmobiliaria de forma general.

[2] Tutorial de Youtube. *Domina Vercel POSTGRES & Next.js*. Recuperat de `https://www.youtube.com/watch?v=ipbuxL2zUNo`
Ús: S'ha visualitzat aquest tutorial per aprendre NextJS i Postgres i com aplicar-lo en el projecte.

[3] Tutorial de Youtube. *Primeros pasos en Google Cloud — Aprendemos a usar Buckets*. Recuperat de `https://www.youtube.com/watch?v=cncZQ5rbZBA`
Ús: S'ha visualitzat aquest tutorial per aprendre què es i com utilitzar Buckets.

[4] Tutorial de Youtube. *Aprende Docker ahora! curso completo gratis desde cero!*. Recuperat de `https://www.youtube.com/watch?v=4Dko5W96WHg`
Ús: S'ha visualitzat aquest tutorial per aprendre Docker.

[5] Tutorial de Youtube. *Set up Google OAuth with Next.js using Next-Auth!*. Recuperat de `https://www.youtube.com/watch?v=ot9yuKg15iA`
Ús: S'ha visualitzat aquest tutorial per aprendre a configurar Google OAuth per l'autentifiació d'usuaris.

[6] IA Generativa. **Ús**: S'han utilitzat diferents eines d'IA generativa com OpenAI - ChatGPT, DeepSeek o Gemini per consultes generals i concretes, resolució de dubtes, generació d'idees durant el desenvolupament del projecte i també per a generar el text de les dades sintétiques.

[7] Tutorial de Youtube. *NLP avanzado con spaCy · Un curso en línea gratis*. Recuperat de `https://www.youtube.com/watch?v=RNiLVCE5d4k`
Ús: S'ha visualitzat aquest tutorial per aprendre NLP amb spaCy des de l'inici.

[8] Tutorials de Youtube. *NER for DH*. Recuperat de `https://www.youtube.com/watch?v=E9h8qVm2uNY&list=PL2VXyKi-KpYs1bSnT8bfMFyGS-wMcjesM`
Ús: S'han visualitzats alguns videos d'aquesta playlist per aprendre NER amb spaCy i veure exemples.

A Apèndix

A.1 Diagrama ER de la base de dades

Fig. 4: Diagrama ER de la base de datos.

A.2 Detall de característiques dels arxius

Tipus d'arxiu	Característiques (Etiquetes)
Contrato	P_NOMBRE, P_TELEFONO, P_CORREO, P_DIRECCION, P_CIUDAD, P_CP, I_NOMBRE, I_TELEFONO, I_CORREO, I_DIRECCION, I_CIUDAD, I_CP, FECHA_FIRMA, FECHA_INICIO, FECHA_FIN, CUOTA, MONEDA, DEPOSITO, DIRECCION_INMUEBLE, CIUDAD_INMUEBLE, CP_INMUEBLE
Recibo	RECEPTOR, EMISOR, FECHA_PAGO, CONCEPTO, MONTO, MONEDA, BANCO, DIRECCION_INMUEBLE, CIUDAD_INMUEBLE, CP_INMUEBLE, FORMA_PAGO
Escritura	NUM_ESCRITURA, FECHA_ESCRITURA, V_NOMBRE, V_TELEFONO, V_CORREO, C_NOMBRE, C_TELEFONO, C_CORREO, TIPO_INMUEBLE, DIRECCION_INMUEBLE, CIUDAD_INMUEBLE, CP_INMUEBLE, REF_CATASTRAL, PRECIO_CV, FORMA_PAGO, ANO_CONSTRUCCION
Seguro	TS_NOMBRE, TS_TELEFONO, DIRECCION_INMUEBLE, CIUDAD_INMUEBLE, CP_INMUEBLE, COMPANIA, FECHA_INICIO, FECHA_FIN, CUOTA, MONEDA
Cédula de Habitabilidad	T_NOMBRE, T_TELEFONO, T_CORREO, T_NUMERO_COLEGIADO, FECHA_FIRMA, DIRECCION_INMUEBLE, CIUDAD_INMUEBLE, CP_INMUEBLE, SUPERFICIE_UTIL, SUPERFICIE_CONSTRUIDA, N_BANOS, N_HABITACIONES, TIPO_INMUEBLE
Certificado Energético	CLASIFICACION, CONSUMO_ENERGIA, CO2_EMISIONES, DIRECCION_INMUEBLE, CIUDAD_INMUEBLE, CP_INMUEBLE, REF_CATASTRAL, ANO_CONSTRUCCION
Registro Catastral	REF_CATASTRAL, DIRECCION_INMUEBLE, CIUDAD_INMUEBLE, CP_INMUEBLE, TIPO_INMUEBLE, SUPERFICIE_CONSTRUIDA, ANO_CONSTRUCCION, VALOR_CATASTRAL, VALOR_REFERENCIA_MERCADO, COORD_X, COORD_Y, CRU_REGISTRO

TAULA 7: DETALL DE CARACTERÍSTIQUES