RĪGAS TEHNISKĀ UNIVERSITĀTE
DATORZINĀTNES, INFORMĀCIJAS TEHNOLOĢIJAS UN ENERĢĒTIKAS FAKULTĀTE

PROJEKTĒŠANAS LABORATORIJĀ

Rēķinu apstrāde ar OCR un datu analīze

Darbu izstrādāja:

Vārds, uzvārds	Apliecības nr.	Grupa
Jūlija Zaiceva	181RIC090	4.
Aļina Kosmatinska	221RDB239	3.
Alfs Āboltiņš	221RDB256	2.
Artūrs Apinis	221RDB396	2.
Daniils Aksjonovs	211RDB450	4.

Rīga 2024

Satura rādītājs

Ievads

Darbs paredz izstrādāt risinājumu, kas ļauj uzņēmumiem precīzi ziņot par oglekļa emisijām valsts iestādēm, lai veicinātu ilgtspējīgo attīstību. Uzņēmumiem, augšupielādējot rēķinus PDF formātā par iegādātajām precēm, piemēram, degviela, elektrība, telekomunikācijas abonēšana, aviobiļetes, pārtikas produkti utt., un rezultātā tie vēlas saņemt aprēķināto CO2 emisiju daudzumu katram nopirktam produktam. Rezultātā saņemti dati tiks izmantoti, lai, balstoties uz iepriekš noteiktiem emisijas faktoriem, aprēķinātu kopējās oglekļa emisijas, kuras iegādājas uzņēmums, un noteikt to ‘oglekļa pēdas’ izmēru.

Risinājuma prasības: Automatizēta sistēma, kas izmanto OCR tehnoloģiju, lai izvilktu datus no rēķiniem PDF formātā un aprēķinātu kopējās emisijas, balstoties uz iepriekš definētiem emisijas faktoriem.

Darba jautājums: Kā izveidot lietotājam draudzīgu un efektīvu rīku, kas ļauj uzņēmumiem precīzi un vienkārši aprēķināt un ziņot par savām oglekļa emisijām?

Darba mērķis: Izstrādāt tīmekļa lietotni, kas ļauj uzņēmumiem automatizēt CO2 emisiju aprēķināšanu un ziņošanu, izmantojot OCR tehnoloģiju datu iegūšanai no rēķiniem un iepriekš definētus emisijas faktorus aprēķiniem.

1. Līdzīgo risinājumu pārskats

Pirms risinājuma izstrādāšanas uzsākšanas tika veikts jau eksistējošo OCR risinājumu pētīšana, kurā gaitā tika apkopoti zemāk norādīti tīmekli:

- [Normative.io] (Mājas lapa: https://normative.io/) - tā ir platforma automatizēto oglekļa uzskaitēm, lai palīdzētu uzņēmumiem izmērīt, samazināt oglekļa emisijas un ziņot par tām. Platformā ir iespējams augšupielādēt datus no dažādiem avotiem. Risinājums piedāvā visaptverošus oglekļa atskaites, kas atbilst starptautiskajiem standartiem, piemēram, siltumnīcefekta gāzu (SEG) protokolam. Programmatūras raksturojums: Back-end ir izveidots ar Python vai Node.js, īpaši rēķinu augšupielādei, apstrādei un oglekļa emisiju aprēķiniem. OCR īstenota izmantojot tādas bibliotēkas kā Tesseract vai API, piemēram, Google Cloud Vision, lai izvilktu tekstu no PDF failiem. MySQL vai PostgreSQL ir izmantotas rēķinu datu un emisijas faktoru glabāšanai.
- Carbon Analytics (Mājas lapa: https://www.co2analytics.com/) - platforma piedāvā rīkus uzņēmumiem, lai izmērītu un pārvaldītu to oglekļa pēdas nospiedumu. Tas ļauj lietotājiem augšupielādēt rēķinus vai integrēt finanšu datus, lai izsekotu iegādāto preču un pakalpojumu oglekļa ietekmei. Programmatūras raksturojums: Back-end ir izveidots ar Python vai Node.js. OCR integrācija notiek ar Google Cloud Vision vai Adobe PDF pakalpojumu teksta lasīšanai no PDF failiem. Datu analīze ir veikta ar ar bāzēm. Izstrādāta API integrācija ar finanšu sistēmām, piemēram, QuickBooks, lai tieši iegūtu rēķinu datus. Front-end izstrādāts uz JavaScript balstītas sistēmas, piemēram, Vue.js vai React,
- Amazon Textract (Mājas lapa: https://aws.amazon.com/textract/) - automātiski nolasa tekstu un datus no skenētiem dokumentiem (PDF, attēli), pārsniedzot tradicionālo OCR, atpazīstot formas un tabulas. Programmatūras raksturojums: nodrošina API integrācijai, atbalsta teksta, tabulu, formu un atslēgvārdu pāru ekstrakciju. Integrējas ar AWS pakalpojumiem, automātiski mērogojas, piemērots liela apjoma dokumentu apstrādei.
- Carbon Interface (Mājas lapa: https://www.carboninterface.com) - API, kas ļauj aprēķināt oglekļa emisijas, balstoties uz transporta, enerģijas un citiem datiem. Programmatūras raksturojums: RESTful API, atbalsta pielāgojamus emisiju aprēķinus un dažādus datu avotus. Nodrošina emisiju pārskatus un reāllaika aprēķinus.
- Carbon Cloud Platform (Mājas lapa: https://gbfcalc.azurewebsites.net/gbf/calc/datainput) - CarbonCloud ir platforma, kas palīdz uzņēmumiem efektīvi pārvaldīt oglekļa pēdu lieliem produktu portfeļiem.

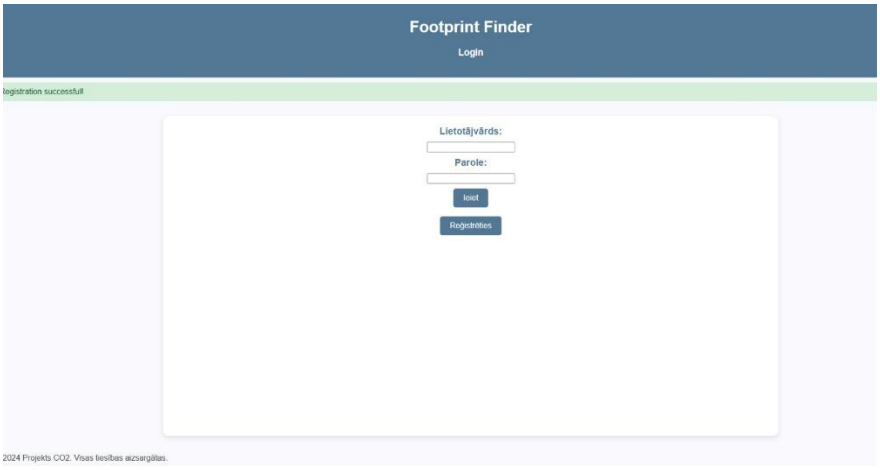
Izmantojot mākslīgo intelektu, tā automātiski savieno produktu materiālu sarakstus ar C02 emisijas datu bāzi. Lietotāji ievada materiālu rēķinu, un platforma nodrošina precīzus 3. līmeņa emisiju aprēķinus, kas ļauj salīdzināt dažādu sastāvdaļu avotus un iegūt pārskatu par visiem produktiem. Tāpat tā piedāvā ļoti interaktīvus un pārskatāmus rezultātus, kas ļauj samazināt CO2 emisiju arī nākotnē.

Secinot pēc visiem augstāk aprakstītiem eksistējošiem risinājumiem, ir konstatēts, ka bieži vien kā Optical Character Recognition jeb OCR risinājums tiek izmantots mākslīgais intelekts un īpaši Google Cloud vision. Izveidojot kontu Google cloud platformā, jauniem lietotājiem ir iespēja izmēģināt izvilkt datus no PDF vai cita formātā un teksta formātā saglabāt savā Google Cloud mākoņkrātuvē. Rīka izmantošana ir maksas pakalpojums, līdz ar to darbā bija nolemts izmantot citu Google pakalpojumu - Google Gemini mākslīgo intelektu.

2. Tehniskais risinājums

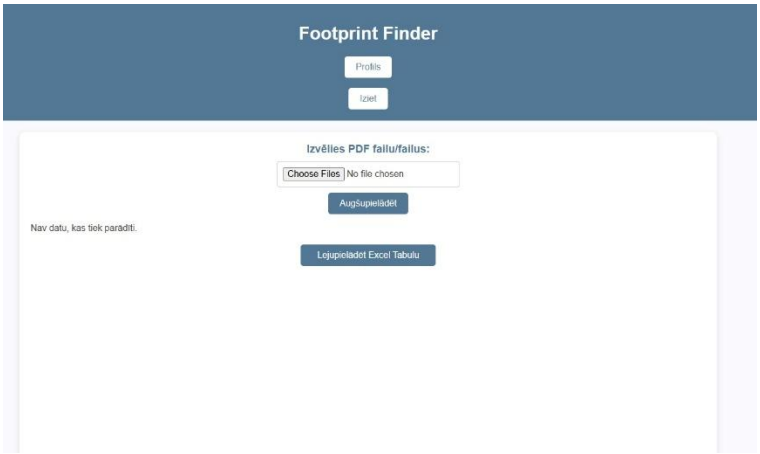
2.1. Prasības

Reģistrācija - lietotājam pirms tīmekļa vietnes lietošanas ir nepieciešams reģistrēt savu profilu reģistrēšanā notiek, nospiežot pogu ‘Reģistrēties’, paša reģistrācija ir vienkāršota - lietotājam ir nepieciešams izveidot savu lietotājvārdu un paroli.



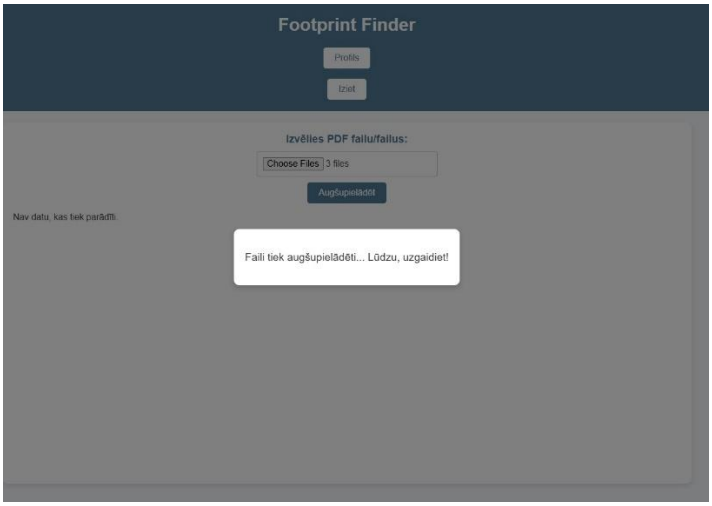
2.1. att. Profila reģistrēšana un pieslēgšana lietotāja profilam

Pieteikšanās - notiek pēc reģistrācija posma, kad lietotājs vēlas ienākt sava izveidotā profilā, izmantojot pogu ‘Ieiet’. Pēc tam lietotājs nonāk uz galveno profila lapaspusi, kur var veikt vairākas tālāk aprakstītas darbības.



2.2.att. Galvenā profila lapaspuse

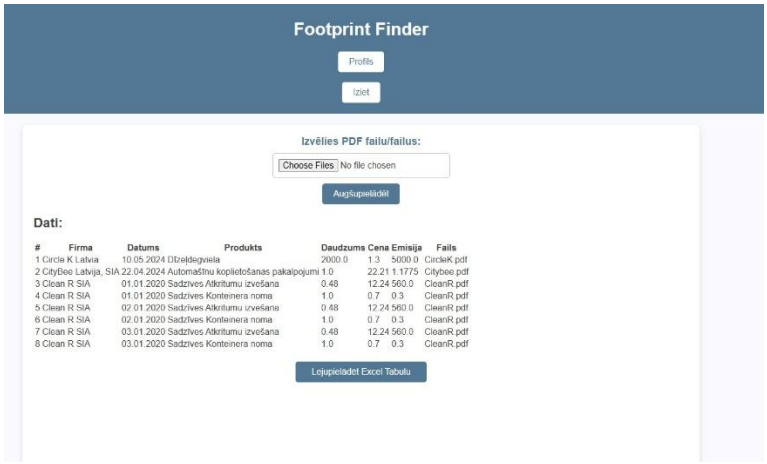
PDF rēķinu augšupielāde - tīmekļa vietne piedāvā lietotājam izvēlēties un augšupielādēt PDF failus jeb rēķinus, no kuriem vēlas saņemt informāciju. ar pogu ‘Choose Files’ lietotājs var pievienot 1 vai vairākus PDF formāta dokumentus (maksimālais pārbaudītais skaits bija 69 dokumenti kopā - maksimālais pieejamais PDF skaits), kad faili ir pievienoti, lietotājam jāuzspiež pogu ‘Augšupielādēt’, lai sāktu PDF lasīšanu.



2.3.att. PDF rēķinu augšupielāde un nolasīšanas gaidīšana

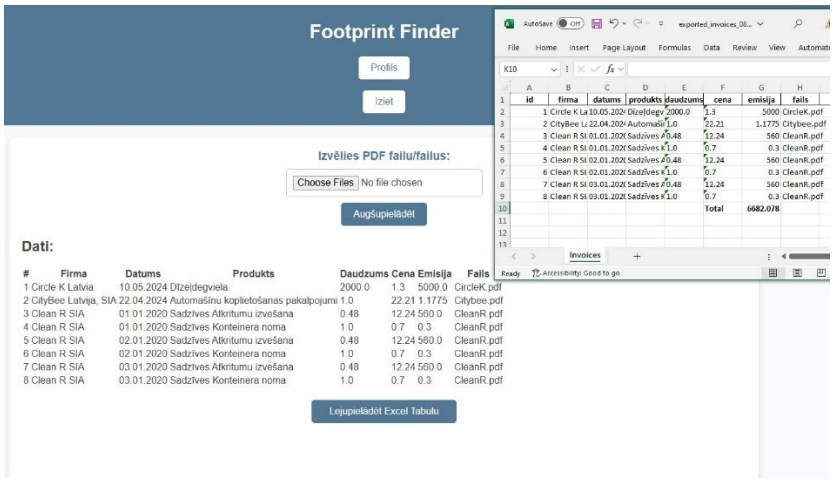
Datu iegūšana no PDF faila - datu iegūšana notiek ar Google Gemini, izmantojot saģenerētu Google API atslēgu. Gemini saņem pieprasījumu, kurā ir aprakstīts, ko ir nepieciešams izlasīt no dokumenta kopā ar piemēru, kāda ir informācija vienā no dokumenta. Kamēr notiek lasīšana, lietotājs redz uz ekrāna paziņojumu ‘Faili tiek augšupielādēti, lūdzu, uzgaidiet’. Dokumentu lasīšanas laiks var variēties no dažām sekundēm līdz maksimāli 5 minūtēm, atkarība no lasāmo dokumentu skaita.

CO2 emisiju aprēķins no iegūtajiem datiem - kopā ar PDF lasīšanu Gemini tiek pieprasīts piešķirt katram nolasītam produktam to tipu - tipam ir piešķirti ID, paši tipi ir salikti Excel failā ‘CO2_koeficienti.xlsx’, kurā ir norādīta tabula ar 29 produktu grupām un tam sameklēto CO2 koeficienti uz 1 vienību. Gemini piešķir katram produktam vistuvāko grupu no saraksta, un līdz ar to piešķir tam CO2 emisiju koeficientu, pēc tam notiek aprēķins - produkta daudzums ir sareizināts ar koeficientu. Lietotājam ir izvadīts rezultāts tabulas veidā: produkta ID, rēķina izsniegšanas firma, datums, kad bija nopirkts produkts, produkta nosaukums, nopirktais daudzums, samaksāta cena par daudzumu, CO2 emisiju daudzums un PDF faila nosaukums, no kura bija nolasīti dati.



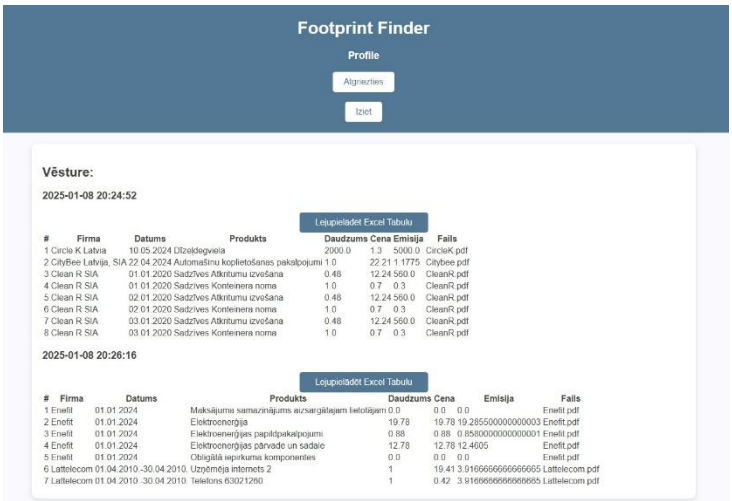
2.4.att. Atskaites izveide

Excel faila atgriešana ar aprēķinātajiem emisiju datiem (Atskaites izveide) - - ar pogu ‘Lejuplādēt Excel Tabulu’ lietotājs var saglabāt izveidoto atskaiti uz sava datorā .xlsx formātā. Saglabātā atskaitē kopā ar jau sagatavotiem datiem būs pieejams vēl kopējais emisiju skaits visiem dotiem produktiem (Pats skaitlis ir redzams pēc uzspiešanas uz pogas ‘Enable Editing’).



2.5.att. Lejulādētais Excel fails ar atskaiti

Kā arī dati tiek automātiski saglabāti lietotāja profilā, kuru viņš var atvērt, izmantojot pogu ‘Profils’. Tur ir norādīta vēsture ar ielādētiem datiem. Katrs vēstures ieraksts ir atšķirīgs ar atskaites izveidošanas datumu un laiku



2.6.att. Lietotāja atskaišu vēsture

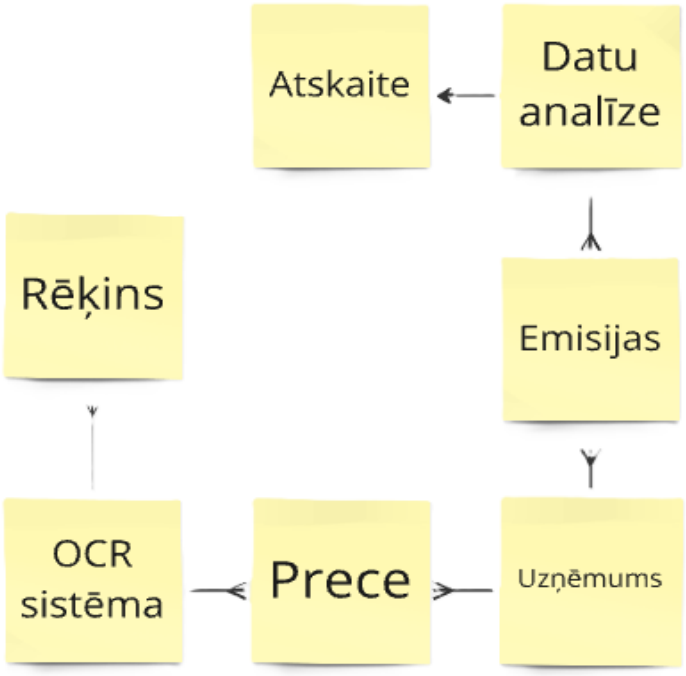
. Pēc vēstures apskates lietotājs var vai atgriezties galvenā lappusē un turpināt ielādēt dokumentus, vai iziet no sava profila.

2.2. Algoritms

- **Lietotāja autentifikācija**
  - **Reģistrācija**
    - Lietotāja ievadītie dati (lietotājvārds, parole)

        - Pārbaudīt vai abas ievadītās paroles, pie reģistrēšanās, sakrīt
        - Pārbaude vai šāds konts jau nepastāv
        - Paroles jaukšana (Hash)
        - Datu saglabāšana datubāzē
    - **Pieteikšanās**
        - Lietotāja ievadītie dati (lietotājvārds, parole)
        - Lietotāja atrašana datubāzē, pēc lietotājvārda
        - Pārbaude vai ievadītā parole atbilst paroles hash no datubāzes
        - Sesijas izveide, ja autentifikācija ir veiksmīga
- **Rēķinu apstrāde**
    - Tiek saņemti lietotāja augšupielādētie PDF faili
    - Katram PDF failam tiek izsaukts Gemini AI, lai no rēķina iegūtu [datums, firma, produkts, daudzums, cena, typeid, correctQuantity], kur dati tiek saglabāti kā JSON un tiek aprēķināta emisija reizinot correctQuantity ar atbilstošo emmisijas koeficientu
    - Datu saglabāšana datubāzē
- **Datu eksports excel failā**
    - Tiek iegūti izvēlētie dati no datubāzes
    - Tiek izveidots excel fails Python atmiņā', kurā tiek ierakstīti:
        - iepriekš izvēlētie dati
        - tiek aprēķināts un ierakstīts kopējais emisiju daudzums
    - Tiek nosūtīts aizpildītais excel fails lietotājam

2.3. Konceptu modelis



2.7.att. Koncepta modeļa shēma

2.4. Tehnoloģiju steks

- **Servera puse**
  - Programmēšanas valoda: Python
  - Satvars: Flask
  - Datubāze: SQLite
  - PDF failu apstrāde (OCR): Google Gemini (google.generativeai)
  - Excel failu apstrāde: pandas un xlswriter
  - Flask veidnes (Templates): Jinja2
  - Lietotāja paroles jaukšanai (Hash): werkzeug
- **Klienta puse**
  - HTML
  - CSS
  - JavaScript

2.5. Programmatūras apraksts

Šajā programmā ir izstrādāta tīmekļa vietne, izmantojot Flask Python bibliotēku, kurā lietotājs var augšupielādēt rēķinu failus, lai aprēķinātu CO2 emisijas katram nopirktajam produktam. PrograŠmmā ir izstrādāta lietotāja autentificēšanās sistēma, kurā lietotājs pats var izveidot sev profilu un pēc tam autentificēties tajā. Datu saglabāšanai tiek izmantota SQLite datubāze, kurā tiek saglabāta informācija par lietotāja profilu (lietotājvārds, parole u.c.), un dati par lietotāja iepriekš analizētajiem PDF rēķiniem. Lai nodrošinātu lietotāja paroļu jaukšanu (Hash), tiek izmantota Python bibliotēka ‘werkzeug’. Lietotājs var augšupielādēt PDF rēķinus, no kuriem, izmantojot Gemini AI API, tiek atgriezti dati par rēķinu, pēc kā, tiek aprēķinātas kopējās CO2 emisijas, kuras tika atrastas vairākos interneta avotos, balstoties uz iepriekš definētiem emisijas faktoriem.

3. Novērtējums

Novērtēšanas mēri:

- Rēķinu skaits, kas tiek apstrādātas vienlaicīgi;
- Rēķinu skaits, kur precīzi nolasīts produkta nosaukums un produkta daudzums;
- Pareiza vai iespējami pareizas produktu veidu grupā piešķiršana;
- Precīzi sagatavotu pārskatu ar aprēķiniem skaits.

Novērtēšanas plāns:

1. Lietotājs izveido savu profilu, izmantojot lietotājvārdu un paroli;
2. Lietotājs ienāk savā profilā. Ja lietotājs ir tikko reģistrējis, viņam nebūs iepriekšējo ielādēto rēķinu, savukārt jau esošajam lietotājam būs iespēja apskatīt iepriekšējo rēķinu datus;
3. Lietotājs ielādē (pievieno) rēķinu vai vairākus rēķinus, lai saņemtu aprēķināto CO2;
4. Notiek laika skaitīšana, kamēr algoritms lasa datus;
5. Lietotājam ir parādīti dati: rēķina izveidotājs (firma), rēķina datums, produkts, daudzums, cena un emisiju daudzums uz produkta daudzumu
6. Lietotājs var eksportēt .xslx formāta failu ar aprēķinātiem datiem. Failā būs norādīts saraksts ar produktiem un emisiju daudzums par katru, kā arī būs norādīta kopēja emisiju summa.

Novērtēšanas rezultāti:

2.1. Tabula

Tīmekļa vietnes galveno funkciju novērtēšanas rezultāti

Novērtēšanas aspekts	Novērtēšanas rezultāts
Rēķinu skaits, kas tiek apstrādātas vienlaicīgi	Neierobežots
Rēķinu skaits, kur precīzi nolasīts produkta nosaukums un produkta daudzums	81% no rēķiniem tika nolasīti pareizi
Rēķinu skaits, kur precīzi nolasīts produkta nosaukums un produkta daudzums	81% no rēķiniem tika nolasīti pareizi
Pareiza vai iespējami pareizas produktu tipa piešķiršana CO2 emisiju aprēķināšanai	78% tika pareizi piešķirta grupa
Skaits ar precīzi sagatavotiem pārskatiem ar aprēķiniem	72% no pārskatiem bija pareizi

Kā arī zemāk ir redzami lietotāja prasības realizēšanas rezultāti.

2.2. Tabula

Lietotāju prasību realizēšanas rezultāti

Lietotāju prasības	Novērtēšanas rezultāts
Lietotājs izveido savu profilu, izmantojot lietotājvārdu un paroli	Jā
Lietotājs ienāk savā profilā. Ja lietotājs ir tikko reģistrējis, viņam nebūs iepriekšējo ielādēto rēķinu, savukārt jau esošajam lietotājam būs iespēja apskatīt iepriekšējo rēķinu datus	Jā
Lietotājs ielādē (pievieno) rēķinu vai vairākus rēķinus, lai saņemtu aprēķināto CO2	Jā
Notiek laika skaitīšana, kamēr algoritms lasa datus	Nē
Lietotājam ir parādīti dati: rēķina izveidotājs (firma), rēķina datums, produkts, daudzums, cena un emisiju daudzums uz produkta daudzumu	Jā
Lietotājs var eksportēt .xslx formāta failu ar aprēķinātiem datiem. Failā būs norādīts saraksts ar produktiem un emisiju daudzums par katru, kā arī būs norādīta kopēja emisiju summa.	Jā

Secinot pēc novērtēšanas rezultātiem, var teikt ka tīmekļa vietne strādā labi, tomēr to vēl varētu uzlabot. PDF failu lasīšanas precizitāte ir virs 80%, pārējiem rēķiniem, kuriem bija nepareizi nolasīti dati, bija sarežģītākā pret MI struktūrā, līdz ar to tam bija grūtības nolasīt. Jo mazāk rēķinā ir informācijas, jo precīzāk to nolasa MI. Virs 70% nolasītiem produktiem tikai piešķirts pareizs produkta tips un līdz ar to pareizs CO2 emisiju koeficients. Produkta tipi ir vispārināti un izvēlēti subjektīvi, ir iespējams, ka MI varēja izvēlēties nepareizo produkta tipa, jo tas piešķirto to produktu citam tipam. Apkopojot visu iepriekš konstatētu, ir saprotams, kāpēc tikai 72% no atskaitēm ir izveidoti pareizi. Lai risinātu programmas problēmu, nepieciešams papildināt un konkretizēt pieprasījumus, kuri tika veikti MI, lai tam būtu vairāk informācijas par prasīto rezultātu.

nepieciešams sniegt vairāk datus par rēķinos iespējamas informācijas klātbūtni un precizēt plašāku produktu sarakstu katram tipam.

Secinājumi

Darba gaitā studentu grupa ir izstrādāja tīmekļa vietni, kuras galvenā funkcija ir PDF rēķinu apstrāde ar OCR un CO2 datu analīze. Tīmekļa vietne izpilda vairākas funkcijas, kurā ietvaros ir lietotāja profila reģistrācija, vairāku vienlaikus PDF augšupielāde un lasīšana, produktu sakārtošana pa grupām, kurai piešķirts vidēji izrēķināts CO2 emisiju koeficients uz 1 vienību, ir izveidota atskaites tabula ar katra produkta CO2 emisiju daudzumu, kura gan saglabājas lietotāja profilā, gan ir iespējams lejuplādēt uz sava datora. PDF failu lasīšana un produkta tipa identificēšana tika veikta ar mākslīga intelekta piesaisti, izmantojot Google Gemini rīku. Tīmekļa vietne ir izveidota ar lietotājam draudzīgo interfeisu, lietošana ir vienkāršota un minimizēta, lai lietotājs ātri varētu sākt izmantot galveno programmatūru.

Pēc novērtējumu analīzes tika secināts, ka tomēr ne visus rēķinus AI var nolasīt pareizi, jo rēķinu struktūra ir atšķirīga, un ir rēķini, kuros ir daudz vairāk pievienotas informācija nekā ir nepieciešams, līdz ar to Gemini nevar to pareizi nolasīt. Kā arī produkta grupas piešķiršana arī nav vienmēr precīza, līdz ar to CO2 emisiju koeficients varētu būt piešķirts nepareizi. Produktu grupu sadalījums bija subjektīvi izveidots grupas ietvaros un ne visi produkti, kuri ir iekļauti rēķinos, varētu piederēt kādai no grupām.

Darba gaitā ir paveikts liels grupu darbs, visi grupas dalībnieki veica gan kopīgus, gan individuālus darbus, lai realizētu projektu. Zemāk var apskatīt katra dalībnieka paveiktais darbs.

Dalībnieki	Paveiktais darbs
Jūlija Zaiceva	Uzdevumu un datu ielādē GitHubā, līdzīgo risinājumu izpēte, koncepta modeļa izstrāde, CO2 emisiju koeficientu noteikšana, datu bāzes arhitektūras izveide, lietotāja reģistrešana, profila izveide ar vēstures saglabāšanu un atspoguļošanu, Excel faila lasīšana ar produktu tipiem un to emisiju koeficientiem, izmantojot Gemini, JSON formāta ieviešana, atskaites izveide
Aļina Kosmatinska	Koncepta modeļa izstrāde, līdzīgo risinājumu izpēte, CO2 emisiju koeficientu noteikšana, atskaites plakāta izveide
Alfs Āboltiņš	Koncepta modeļa izstrāde, PDF datu izvilkšanas kods ar gemini api, OCR metožu meklēšana un testēšana, tīmekļa aplikācijas atkļūdošana, novērtējuma rezultātu iegūšana
Artūrs Apinis	Koncepta modeļa izstrāde, PDF datu izvilkšanas kods ar Gemini api, OCR metožu

	meklēšana un testēšana, tīmekļa vietnes veidošana
Daniils Aksjonovs	Koncepta modeļa izstrāde, OCR metožu meklēšana, lietotāja interfeisa izveide, atskaites plakāta izveide