

ESCOLA POLITÉCNICA DA UNIVERSIDADE DE SÃO PAULO
DEPARTAMENTO DE ENGENHARIA QUÍMICA
CURSO DE ENGENHARIA QUÍMICA

ANA BEATRIZ MAZZINI
RAPHAEL PELLEGRINI CORREIA
VINÍCIUS CAMPOS CONSTANTINO

Otimização do processamento de documentos impulsionada por Machine Learning: desenvolvimento e implementação de soluções inovadoras e mais eficientes para escritórios

São Paulo, SP
Abril, 2024

ANA BEATRIZ MAZZINI
RAPHAEL PELLEGRINI CORREIA
VINÍCIUS CAMPOS CONSTANTINO

Otimização do processamento de documentos impulsionada por Machine Learning: desenvolvimento e implementação de soluções inovadoras e mais eficientes para escritórios

Versão Original

Trabalho de conclusão de curso de graduação apresentado ao Departamento de Engenharia Química da Escola Politécnica da Universidade de São Paulo como parte dos requisitos para obtenção do grau de Bacharel em Engenharia Química.

Orientador: Prof. Dr. José Luis Pires Camacho
Segunda avaliadora: Profa. Dra. Martina Costa Reis

São Paulo, SP
Abril, 2024

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

Catálogo-na-publicação

Mazzini, Ana Beatriz. Pellegrini Correia, Raphael. Campos Constantino, Vinícius.
Otimização do processamento de documentos impulsionada por Machine Learning: desenvolvimento e implementação de soluções inovadoras e mais eficientes para escritórios / A. B. Mazzini, R. Pellegrini Correia, V. Campos Constantino -- São Paulo, 2024.
50 p.

Orientador: Prof. Dr. José Luis Pires Camacho
Trabalho de Formatura - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia Química.

1.Machine Learning 2.Automação de processos 3.Digitalização de documentos 4.Eficiência operacional I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia Química II.t.

ANA BEATRIZ MAZZINI
RAPHAEL PELLEGRINI CORREIA
VINÍCIUS CAMPOS CONSTANTINO

OTIMIZAÇÃO DO PROCESSAMENTO DE DOCUMENTOS IMPULSIONADA POR
MACHINE LEARNING: DESENVOLVIMENTO E IMPLEMENTAÇÃO DE
SOLUÇÕES INOVADORAS E MAIS EFICIENTES PARA ESCRITÓRIOS

Trabalho de conclusão de curso de graduação
apresentado ao Departamento de Engenharia
Química da Escola Politécnica da Universi-
dade de São Paulo como parte dos requisitos
para obtenção do grau de Bacharel em Enge-
nharia Química.

Aprovado em ____ de _____ de _____

BANCA EXAMINADORA:

Prof. Dr. José Luis Pires Camacho

Profa. Dra. Martina Costa Reis

AGRADECIMENTOS

Ao Professor Dr. José Luis Pires Camacho, orientador deste trabalho, agradecemos pela orientação, paciência e conhecimento disponibilizados ao longo do desenvolvimento deste estudo. Seu apoio foi fundamental para a conclusão deste trabalho, oferecendo uma base sólida de conhecimento e incentivo.

À Professora Dra. Martina Costa Reis, segunda avaliadora, expressa-se gratidão pelo suporte acadêmico, sua dedicação e disponibilidade contribuíram de maneira relevante para o aprofundamento do estudo e para o desenvolvimento pessoal e profissional do autor.

À Escola Politécnica da Universidade de São Paulo, reconhece-se a importância das instalações, recursos e do ambiente acadêmico estimulante que foram cruciais para a formação e realização deste trabalho. A excelência da instituição proporcionou uma base educacional inestimável, permitindo o desenvolvimento deste estudo em condições ideais.

Aos grupos de extensão da Universidade, agradece-se pela oportunidade de aplicar o conhecimento teórico em contextos práticos, contribuindo para uma formação acadêmica mais rica e aplicada. As experiências compartilhadas nos projetos de extensão foram fundamentais para o enriquecimento profissional e pessoal.

Aos amigos e familiares, expressa-se um sincero agradecimento pelo apoio incondicional, compreensão e incentivo constantes. Sua presença e estímulo foram essenciais nos momentos de desafio, constituindo-se como um suporte vital para a superação de obstáculos e a conclusão deste marco acadêmico.

Este trabalho é fruto do esforço conjunto de muitas pessoas que, direta ou indiretamente, contribuíram para sua realização. A todos, o mais sincero agradecimento do grupo.

"The value of an idea lies in the using of it."

Thomas Edison

RESUMO

MAZZINI, Ana Beatriz; PELLEGRINI CORREIA, Raphael; CAMPOS CONSTANTINO, Vinícius. Otimização do processamento de documentos impulsionada por Machine Learning: desenvolvimento e implementação de soluções inovadoras e mais eficientes para escritórios. 2023. Trabalho de Conclusão de Curso (Bacharelado em Engenharia Química) - Escola Politécnica, Universidade de São Paulo, São Paulo, 2023.

No atual cenário empresarial, marcado por avanços tecnológicos como automação, integração na nuvem e adoção de softwares inovadores, a inteligência artificial, notadamente o Machine Learning, desempenha papel essencial na otimização dos processos. Identificou-se a necessidade de uma solução acessível e eficiente para a digitalização de documentos em empresas de diferentes portes, impulsionada pela observação do cotidiano empresarial. A solução desenvolvida, por meio de um código em Python utilizando técnicas de Machine Learning, permite a transcrição de documentos de imagens para formatos de texto. Isso agiliza a extração de informações cruciais para diversas áreas, como financeiro, produção, controladoria, jurídico e compliance. Essa abordagem visa reduzir o tempo dedicado pelos funcionários a tarefas de digitalização, resultando em cortes nos custos operacionais. As informações extraídas são facilmente exportáveis em formatos como Excel, bancos de dados ou JSON, proporcionando flexibilidade conforme as necessidades específicas de cada empresa. Além de otimizar processos, a solução contribui para a agilidade operacional e eficiência financeira, permitindo uma alocação estratégica dos recursos humanos em tarefas mais complexas e de maior valor agregado. A implementação bem-sucedida dessa solução representa um passo significativo em direção à transformação digital e à excelência operacional nas empresas.

Palavras-chave: Machine Learning. Automação de processos. Digitalização de documentos. Eficiência operacional.

ABSTRACT

MAZZINI, Ana Beatriz; PELLEGRINI CORREIA, Raphael; CAMPOS CONSTANTINO, Vinícius. Optimization of document processing driven by Machine Learning: development and implementation of innovative and more efficient solutions for offices. 2023. Final Paper (Bachelor's Degree in Chemical Engineering) - Polytechnic School, University of São Paulo, São Paulo, 2023.

In the current business scenario, marked by technological advances such as automation, cloud integration, and the adoption of innovative software, artificial intelligence, notably Machine Learning, plays an essential role in process optimization. The need for an accessible and efficient solution for document digitization in companies of various sizes was identified, driven by observations in everyday business operations. The developed solution, through Python code employing Machine Learning techniques, allows the transcription of documents from images to text formats. This streamlines the extraction of crucial information for various areas, including finance, production, control, legal, and compliance. This approach aims to reduce the time spent by employees on digitization tasks, resulting in cuts to operational costs. The extracted information can be easily exported in formats such as Excel, databases, or JSON, providing flexibility according to the specific needs of each company. Besides optimizing processes, the solution contributes to operational agility and financial efficiency, enabling a strategic allocation of human resources to more complex and higher-value tasks. The successful implementation of this solution represents a significant step towards digital transformation and operational excellence in companies.

Keywords: Machine Learning. Process automation. Document digitization. Operational efficiency.

LISTA DE ILUSTRAÇÕES

Figura 1 – Distribuição do tipo de documento que os entrevistados lidam no dia-a-dia, relacionado com a porcentagem referente ao total das 78 pessoas da pesquisa	21
Figura 2 – Distribuição das respostas: à esquerda tem-se a divisão das pessoas que realizam ou não a extração de dados dos documentos que lidam no dia-a-dia, e à direita tem-se a segmentação dos motivos pelo qual as 51 pessoas extraem dados de documentos	21
Figura 3 – Distribuição dos tipos de documentos que as pessoas extraem dados para fins de reembolso	22
Figura 4 – Distribuição dos tipos de documentos que as pessoas extraem dados para fins de uso pessoal	22
Figura 5 – Levantamento da quantidade de pessoas que extraem dados de documentos para fins de reembolso ou atividades do trabalho e que possuem ou não acesso a ferramentas para isso	23
Figura 6 – Levantamento das citações de ferramentas para extração de dados, a fim de se concluir o alcance da tecnologia OCR	23
Figura 7 – Nível de satisfação das soluções utilizadas coletado de pessoas que realizam a extração de dados de documentos	24
Figura 8 – Exemplo de utilização do programa para consolidação de valores e datas de viagens realizadas pelo Uber	27
Figura 9 – <i>Software</i> aplicado ao caso de boletos bancários	32
Figura 10 – <i>Graphical User Interface</i> desenvolvida para o <i>software</i>	35
Figura 11 – Faixa etária dos estudantes que responderam a pesquisa de mercado	46
Figura 12 – Gênero dos estudantes que responderam a pesquisa de mercado	46
Figura 13 – Estudantes que exercem ou não atividade remunerada que responderam a pesquisa de mercado	47
Figura 14 – Distribuição dos estudantes que responderam a pesquisa de acordo com a sua área de atuação no mercado	47
Figura 15 – Distribuição das finalidades para qual a digitalização ou extração de informações é utilizada	48
Figura 16 – Distribuição dos tipos de documento que os estudantes presentes na pesquisa costumam lidar no dia-a-dia	48
Figura 17 – Tempo gasto para a extração de dados dos documentos	49
Figura 18 – Quantidade de pessoas presentes na pesquisa que utilizam ou não ferramentas específicas para a extração de dados	49
Figura 19 – Distribuição das ferramentas utilizadas	49

Figura 20 – Quantidade de pessoas que tem a ferramenta oferecida pela empresa em que trabalha para exercer a extração dos dados	50
Figura 21 – Distribuição do nível de satisfação daqueles que utilizam ferramentas com a solução já utilizada	50

LISTA DE ABREVIATURAS E SIGLAS

OCR	<i>Optical Character Recognition</i>
ODS	Objetivos do Desenvolvimento Sustentável
MVP	Produto Mínimo Viável
PAT	Programa de Alimentação do Trabalhador
NPS	<i>Net Promoter Score</i>

SUMÁRIO

1	INTRODUÇÃO	11
2	OBJETIVOS	14
3	METODOLOGIA	15
4	ANÁLISE DE MERCADO	16
4.1	CASO I: VR BENEFÍCIOS E A AQUISIÇÃO DA VEXPENSES . . .	17
4.2	CASO II: FLASH E A AQUISIÇÃO DA EXPENSEON	19
4.3	PESQUISA DE MERCADO LEVANTADA	20
5	DESENVOLVIMENTO	25
5.1	IDENTIFICAÇÃO DO PROBLEMA	25
5.2	ESTUDO E SELEÇÃO DO AMBIENTE DE PROGRAMAÇÃO E DAS FERRAMENTAS AUXILIARES	25
5.3	IDENTIFICAÇÃO DOS OUTPUTS E ETAPAS NECESSÁRIAS . . .	25
5.4	IMPLEMENTAÇÃO DO CÓDIGO	26
6	RESULTADOS	31
6.1	TESTE DE HIPÓTESE E FEEDBACK DOS USUÁRIOS	31
6.2	ADAPTAÇÕES NO MVP	31
6.2.1	Incorporação da leitura e processamento de Boletos Bancários	31
6.2.2	Adição de uma <i>Graphical User Interface</i> (GUI)	34
6.2.3	Desenvolvimento de uma <i>Landing Page</i>	39
7	DISCUSSÕES	40
7.1	APLICAÇÃO DO MÉTODO <i>LEAN STARTUP</i> PARA DESENVOL- VIMENTO DE PRODUTOS	40
7.2	BENEFÍCIOS DO <i>SOFTWARE</i> DESENVOLVIDO	41
7.3	PRÓXIMOS PASSOS	42
7.3.1	Criação de um arquivo executável	43
7.3.2	Criação de integrações com plataformas ou serviços externos	43
	REFERÊNCIAS	44
	APÊNDICE A – PESQUISA DE MERCADO LEVANTADA COM ALUNOS DA FACULDADE	46

1 INTRODUÇÃO

A automatização da digitalização e processamento de documentos físicos tornou-se uma imperativa necessidade para as empresas contemporâneas. Sua implementação é impulsionada por diversos motivos, como a busca pela redução de custos, a promoção da preservação ambiental, o reforço da segurança da informação e a otimização dos processos.

A digitalização de documentos pode levar a economias de custos significativas para empresas e organizações, ao reduzir o consumo de papel, mão de obra, manutenção de equipamentos e erros. Uma pesquisa realizada pela AIIM (*Association for Intelligent Information Management*) revelou que um funcionário de escritório médio utiliza anualmente 10.000 folhas de papel, o que custa cerca de \$80, e 45% dos documentos são descartados diariamente. O avanço sobre o tema reflete o compromisso crescente das empresas com responsabilidades sociais e ambientais.

A segurança da informação é outro desafio crítico enfrentado pelas empresas atualmente. Proteger os dados próprios e dos clientes contra vazamentos, fraudes, perdas ou danos é crucial. A automatização da digitalização de documentos apresenta vantagens significativas em comparação ao armazenamento físico, permitindo o uso de criptografia, senhas, backups, sistemas de controle de acesso e auditoria. Além disso, assegura o acesso remoto a esses documentos sempre que necessário.

Por último, a automatização do processamento de documentos agrega agilidade operacional às empresas (AEL Data, 2023). Facilita a busca, compartilhamento, edição e integração dos documentos com outros sistemas e plataformas, proporcionando uma eficiência notável nos fluxos de trabalho.

Tendo em vista esse cenário, decidiu-se pelo desafio de desenvolver uma ferramenta para auxiliar no processo de digitalização de documentos, como notas fiscais, cuja digitalização é fundamental para que seus dados sejam operacionalizáveis e incorporáveis em ambientes digitais de tratamento.

A técnica utilizada para desenvolvimento da ferramenta descrita neste documento chama-se *Optical Character Recognition*, ou simplesmente OCR, que em tradução livre significa Reconhecimento Ótico de Caracteres, a qual serve o propósito de converter diferentes tipos de documentos, como imagens digitalizadas, PDFs ou fotos, em dados editáveis e pesquisáveis por computadores. De modo geral existem três etapas fundamentais no seu funcionamento:

1. **Preparação:** Nesta fase inicial, a imagem destinada ao reconhecimento de caracteres passa por um meticuloso processo de preparação. Esta etapa é crucial para garantir a acurácia do reconhecimento subsequente e envolve técnicas sofisticadas para purificar a imagem de quaisquer elementos distrativos que não sejam texto, tais

como imagens, logotipos e marcas d'água. Ademais, são aplicados ajustes para corrigir possíveis distorções, inclinações ou sombras presentes na imagem, estabelecendo assim uma base sólida para a identificação eficaz de caracteres;

2. **Reconhecimento:** Durante o processo de reconhecimento, o *software* emprega algoritmos avançados para analisar e comparar os caracteres da imagem com uma vasta biblioteca de símbolos pré-definidos. Esta comparação não se limita apenas à identificação direta de formas, mas estende-se à análise detalhada de elementos gráficos, como curvaturas, ângulos e linhas que compõem cada caractere. Essa abordagem permite ao sistema discernir com precisão entre caracteres semelhantes e reconhecer padrões de texto dentro da imagem, mesmo em casos de variação estilística ou de qualidade de imagem comprometida;
3. **Tratamento:** Após o reconhecimento bem-sucedido dos caracteres, o texto é transformado em um formato digital que pode ser facilmente interpretado e manipulado por computadores. Este estágio de tratamento é fundamental, pois converte o texto reconhecido em dados editáveis e pesquisáveis, facilitando a integração com outras aplicações e sistemas. O resultado é um conteúdo textual flexível, acessível para edição, formatação, armazenamento e análise, abrindo um leque de possibilidades para processamento adicional e utilização prática em diversos contextos empresariais e acadêmicos.

Neste documento, será explorado o desenvolvimento de uma ferramenta OCR, utilizando a linguagem de programação *Python* (em sua versão 3.11.6), em conjunto com a *engine open source Tesseract-OCR*¹ e o *Poppler*², destacando os desafios enfrentados, as inovações propostas e os benefícios que essa tecnologia pode proporcionar.

A fundamentação deste trabalho está vinculada aos conhecimentos adquiridos em programação e *machine learning* ao longo da graduação, complementados pelas experiências do grupo no mercado de trabalho. Nesse contexto, observa-se alinhamento com o ODS 9, uma vez que a aplicação de *machine learning* para a otimização do processamento de documentos em ambientes de escritórios promove inovação e produtividade. Essa abordagem contribui para o desenvolvimento de soluções mais eficientes e sustentáveis na gestão de informações.

Além disso, a contribuição se estende para a eficiência operacional das empresas, resultando na redução de custos e na criação de oportunidades de emprego no campo da tecnologia e automação. Isso está em conformidade com a busca por um trabalho decente e o crescimento econômico sustentável, objetivos alinhados com o ODS 8. Desde o setor bancário, que utiliza OCR para processamento de cheques e documentos financeiros, até

¹ <https://github.com/tesseract-ocr/tesseract>

² <https://github.com/oschwartz10612/poppler-windows/releases/>

o varejo, onde facilita a gestão de inventário e a catalogação de produtos, a OCR oferece uma solução prática para a digitalização de operações empresariais. O desenvolvimento de produtos baseados em OCR é um exemplo do empreendedorismo tecnológico voltado para a melhoria contínua de processos, enfatizando a importância da adaptação e inovação constantes. Assim, a OCR não apenas otimiza as operações existentes, mas também abre caminho para a criação de novos modelos de negócios e estratégias de mercado, reforçando o papel do empreendedorismo na liderança da transformação digital nas empresas.

Portanto, ao expandir a análise da solução técnica para uma perspectiva de negócios, antecipa-se a presença de produtos similares no mercado. Esta expectativa deriva do entendimento de que a problemática em questão representa um desafio comum a várias organizações no ambiente corporativo, sugerindo que iniciativas empreendedoras neste domínio têm o potencial de gerar modelos de negócio bem-sucedidos. Diante disso, o presente trabalho se propõe a realizar uma investigação sobre os modelos de negócio aplicáveis ao *software* desenvolvido, explorando como ele poderia se posicionar estrategicamente no mercado. Essa análise visa identificar oportunidades de inovação e diferenciação que possam contribuir para a sua viabilidade e sucesso comercial.

O *software* em discussão pode ser encarado como um Produto Mínimo Viável (MVP), que se refere à versão mais simples de um produto que permite à empresa aprender o máximo possível sobre os clientes com o menor esforço despendido (RIES, 2011). A ideia por trás do MVP é testar hipóteses de mercado rapidamente, adaptando-se com base no *feedback* dos usuários, ao mesmo tempo em que se minimiza o desperdício de recursos. Esse conceito é crucial para startups que buscam validar suas ideias no mercado de forma eficiente.

Neste contexto, a construção do software de OCR pode ser o representativo de um MVP, pois é desenvolvido para satisfazer necessidades específicas de reconhecimento e digitalização de texto. Abordar este software como um MVP permite coletar *feedbacks* valiosos dos primeiros usuários sobre sua utilidade, usabilidade e sobre quais melhorias são necessárias para atender ainda melhor às suas necessidades. Dessa forma, o processo de desenvolvimento neste projeto seguirá a lógica de aprendizado e adaptação rápida, característico da metodologia proposta por Ries. Ao invés de desenvolver um produto final complexo e possivelmente carregado de funcionalidades desnecessárias, nossa equipe buscará focar em validar hipóteses de mercado e direcionar esforços para aquilo que realmente traz valor aos usuários finais.

2 OBJETIVOS

Tendo em vista o cenário abordado sobre a relevância da otimização do processo de digitalização de documentos, o presente trabalho tem como objetivo primário o desenvolvimento de um *software* de OCR no formato de MVP. Este objetivo concentra-se na construção de um núcleo funcional do *software*, que será capaz de realizar o reconhecimento óptico de caracteres com eficiência e precisão. O MVP buscará integrar as características mais valiosas e críticas, com base em um desenvolvimento ágil e iterativo, que permita a adaptação rápida às necessidades emergentes do mercado e *feedbacks* iniciais dos usuários.

Dentro deste processo, a identificação do público-alvo assume um papel complementar, servindo como suporte para o refinamento do MVP. O foco estará na utilização de testes de usabilidade e no acúmulo de aprendizado prático a partir das iterações. Esses testes visam avaliar a eficácia do *software* e identificar oportunidades de aperfeiçoamento.

Além disso, a análise de mercado permanece como uma atividade subjacente, este ponto envolve não apenas o entendimento das tendências e tecnologias emergentes, mas também a identificação dos principais *players* e a avaliação do ambiente competitivo. Compreender as dinâmicas de mercado é fundamental para posicionar estrategicamente o *software* de OCR e identificar as oportunidades de inovação e crescimento.

3 METODOLOGIA

A metodologia utilizada como guia para o desenvolvimento do trabalho se baseia na estratégia apresentada no livro *The Lean Startup*, de Eric Ries, a qual foi responsável por revolucionar a maneira como as empresas abordam o desenvolvimento de produtos e a entrada no mercado, focando no aprendizado rápido e adaptação. Primeiramente, é essencial identificar e definir claramente o problema que o negócio pretende resolver, garantindo que o produto ou serviço atenda às necessidades reais do mercado.

Em seguida, desenvolve-se um Produto Mínimo Viável (MVP). Este deve possuir apenas as funcionalidades essenciais que permitam ao produto ser operacional e testável. O objetivo é lançá-lo para um segmento de usuários potenciais para começar a coleta de dados, que informarão se o caminho escolhido está correto ou se ajustes são necessários. Esta etapa é vital para o processo de iteração, onde o produto é refinado e ajustado com base nas interações e respostas dos usuários. Continuar este ciclo de construção, medição e aprendizado forma a base para o crescimento sustentável e a inovação dentro da empresa.

4 ANÁLISE DE MERCADO

No Brasil, o mercado de gestão de despesas corporativas se destaca como o principal aplicador da tecnologia OCR. Isso ocorre porque o OCR automatiza e simplifica a captura e o processamento de informações de recibos e notas fiscais, elementos centrais na gestão de reembolsos e controle de gastos empresariais. A tecnologia proporciona não só um aumento significativo na eficiência operacional, reduzindo a entrada manual de dados, mas também contribui para a diminuição de erros, fraudes e melhoria na precisão dos relatórios financeiros.

As despesas corporativas no Brasil representam um gasto significativo, alcançando cerca de R\$650 bilhões anuais. Uma análise realizada pelo VExpenses identifica que, em média, 23% desses gastos correspondem à alimentação, enquanto combustível e reembolsos por uso de veículos próprios representam, respectivamente, 22% e 21%. Além disso, a hospedagem compreende 11% das despesas, e as categorias restantes somam cerca de 23% do total (VExpenses, 2022).

Para compreender o impacto das estratégias de gestão de despesas corporativas, é crucial analisar as abordagens adotadas pelas empresas. Essas estratégias incluem reembolso, adiantamento, faturamento direto e o uso de cartões corporativos, tanto pós-pagos quanto pré-pagos. Cada método tem suas especificidades: no reembolso, colaboradores antecipam gastos e solicitam reembolsos; no adiantamento, a empresa fornece fundos previamente; despesas faturadas são pagas diretamente pela empresa; e cartões corporativos facilitam a gestão financeira, refletindo a evolução dos serviços financeiros para pessoas jurídicas.

A pesquisa realizada pela VExpenses em 2021 revelou que uma grande maioria das empresas, cerca de 95%, opta pelo método de reembolso para gestão de despesas corporativas, enquanto 51% utilizam cartões corporativos, com várias empresas adotando ambos os métodos. Essas práticas, embora comuns, apresentam desafios significativos, principalmente na necessidade de prestação de contas por parte dos colaboradores e na execução da conciliação financeira pela empresa, afinal, existe a necessidade de adaptação constante das políticas corporativas para atender às necessidades específicas de diferentes departamentos, ampliando a carga de trabalho na aprovação de gastos e na manutenção de registros fiscais correspondentes, bem como na classificação contábil adequada dessas despesas.

Dado o volume de despesas corporativas no Brasil, a tecnologia OCR surge como uma solução promissora para simplificar e automatizar a gestão desses gastos, por esse motivo, o MVP em desenvolvimento afunilará em questões do tema. A fim de expandir o conhecimento das aplicações no setor, serão realizados *benchmarks* com casos de sucesso de empresas que já estão utilizando essa tecnologia para melhorar suas operações financeiras.

4.1 CASO I: VR BENEFÍCIOS E A AQUISIÇÃO DA VEXPENSES

Em 1976, surge no Brasil o PAT (Programa de Alimentação do Trabalhador), um projeto executado pelos Ministérios do Trabalho, Saúde e Fazenda, o qual exerce extrema importância no cenário em análise, pois os benefícios oferecidos pelo PAT não têm natureza salarial e podem ser oferecidos como refeição no próprio local de trabalho, cestas de alimentação, vales ou cartões (restaurante ou alimentação). O valor oferecido pelo empregador deve ser suficiente para atender às exigências nutricionais do PAT.

Conforme publicado pelo Diário da União em 2015, um balanço que mostrava que 19.513.426 trabalhadores foram até então beneficiados pelo Programa de Alimentação do Trabalhador. Segundo as empresas de benefícios flexíveis que oferecem cartões de acordo com as diretrizes do Programa, existem três formas de operacionalização principais no PAT, a primeira é através de refeições prontas no refeitório, a segunda por cestas básicas de alimentos e a última por vales, tíquetes ou métodos eletrônicos de pagamentos.

É nesse cenário em que se enquadra a fundação da VR Benefícios em 1977, empresa que no início elaborou a solução de vales em papel para empregadores, e, depois, evoluíram para cartão com chip, hoje, já exploram o mercado de contas digitais. Atualmente, a empresa conta a prestação de serviços digitais tanto no setor de benefícios, com o fornecimento através de um super aplicativo de vale alimentação, refeição, transporte e cartões multi com funcionalidade de saldo separado, quanto como um apoio no setor de recursos humanos dos clientes, fornecendo ferramentas digitais para controle de despesas corporativas, controle de ponto, escala, férias e folgas.

No fim do ano de 2021, a VR assumiu um posicionamento de ir ao mercado buscar empresas que somassem sinergias na oferta de produtos e serviços à jornada do trabalhador. Pontomais, Audaz Tecnologia e Global Points foram as start-ups adquiridas e permitiram à VR ampliar seu portfólio que agora, além de benefícios corporativos, também conta com gestão de pessoas, mobilidade corporativa e um marketplace com ofertas e cashback para empresas e colaboradores. As aquisições apontadas já se tornaram produtos dentro da VR – a Pontomais virou VR Gente, a Audaz se tornou a VR Mobilidade e Global Points passou a ser VR Shopping. Segundo destacou a VR em nota, os investimentos em portfólio trouxeram resultados, rendendo à companhia um faturamento de R\$ 11 bilhões em 2022, crescimento de 30% sobre o ano anterior.

O movimento mais recente foi no final de 2023 com a aquisição da VExpenses, startup criada com o foco na gestão de despesas corporativas. Segundo a VR, a compra, de valor avaliado em cerca de R\$ 120 milhões, tem como objetivo construir um ecossistema completo para facilitar a vida dos trabalhadores e somar mais facilidades ao ecossistema para gestão de empresas e pessoas. Com o valor pago pela VR, de acordo com reportagem do Valor Econômico, a VExpenses multiplica em mais de 15 vezes o valor investido pela boutique financeira FG/A, que criou a startup a partir de um *spinoff* de uma divisão de

negócios da empresa.

O portfólio de produtos desenvolvido pela VExpenses conta com tecnologia OCR de ponta, voltada para a extração de dados de notas fiscais e preenchimento automático para verificação das políticas de reembolso determinadas pelos clientes. Diante de um leque de clientes da startup, que conta com grandes nomes do mercado como MRV, Alpargatas, Nike e CNN, é possível levantar o estudo de caso da Inpasa Brasil, uma das primeiras indústrias de Etanol de Milho brasileiras, sendo hoje a maior produtora de combustível limpo e renovável à base de milho da América Latina.

Tendo em vista uma demanda crescente de combustível renovável, a Inpasa encontrou espaço para se expandir no país através de três unidades, duas delas em Mato Grosso e outra no Mato Grosso do Sul. Com isso, o número de colaboradores no time aumentou e o controle da prestação de contas se tornou mais desafiador para o time financeiro, que até então usava de ferramentas manuais. Assim, cada prestador de conta deveria preencher uma planilha com o relatório de despesas e detalhes das transações, anexando as notas fiscais dos gastos, e depois enviar para o financeiro. Por sua vez, o financeiro deveria fazer a checagem e conferência de cada item e seu comprovante, a auditoria de gastos de acordo com a política da empresa e, por fim, efetuar o pagamento do reembolso aos funcionários.

Com a adoção da plataforma VExpenses pela Inpasa Brasil, o processo de gestão de despesas corporativas foi completamente digitalizado. Essa integração resultou em um sistema unificado que inclui um cartão corporativo, o Cartão VExpenses. Um dos principais benefícios observados foi a significativa redução no tempo necessário para a aprovação de despesas, passando de quatro dias para apenas um dia. Além disso, a necessidade de enviar comprovantes e notas fiscais em papel foi eliminada, graças à digitalização dos documentos e à mobilidade oferecida pelo sistema, otimizando ainda mais a prestação de contas.

A SMAR APD, empresa voltada para soluções em ERP Público e Cobranças, pode ser utilizada como um outro exemplo de clientes que evoluíram o desempenho das operações diárias após a contratação do OCR na plataforma da VExpenses. Atuando em diversos estados brasileiros, a Smar tem muitos colaboradores que precisam viajar para visitar os clientes em viagens que podem durar de 1 a 2 semanas. Os colaboradores tinham a necessidade de juntar todas as notas fiscais acumuladas durante uma viagem, preencher uma planilha com todos os dados dessas notas e entregar o relatório e as notinhas fisicamente para os responsáveis, com todas as despesas com cartões corporativos, hotéis e passagens.

A empresa levantou a informação com os funcionários de que para montar esses relatórios e apresentar todas as despesas de uma viagem, demorava um tempo mínimo de 4 horas, o qual poderia ser utilizado para outras atividades mais estratégicas. Ainda depois que os relatórios eram entregues para aprovação, existia um tempo de 7 a 10 dias para validação da prestação de contas se não houvesse nenhum erro nos relatórios, caso contrário, demorava ainda mais.

Hoje, com a centralização de todas as informações em um único lugar, as equipes ganharam mais tempo para outras atividades, facilidade para obtenção de qualquer dado e relatórios gerenciais ali na mão. Mas o destaque principal deve ser dado para a facilidade do viajante a partir da implementação do OCR, pois acabaram as preocupações com cálculos errados. E, além de facilitar a produção de relatórios, com o próprio celular, no aplicativo do VExpenses, o colaborador faz a foto da nota, classifica os gastos por tipo de despesas e o sistema realiza todos os cálculos automaticamente pelo reconhecimento da imagem.

De certo, com a solução da VExpenses, a partir da aquisição a VR tem a oportunidades de complementar seus produtos de vale-alimentação e refeição com outros como cartões de despesas de viagens para funcionários, somando mais tecnologia ainda ao seu ecossistema e expandindo soluções para gestão de pagamentos.

4.2 CASO II: FLASH E A AQUISIÇÃO DA EXPENSEON

A Flash é uma *startup* que começou suas operações no ano de 2019, como um aplicativo de benefícios flexíveis, a fim de desburocratizar o até então mercado de despesas corporativas. A ideia principal era montar um negócio para resolver a complexidade de inúmeros cartões e vales, destacando que para a solução era preciso um cartão único, um modelo de benefício descomplicado, que unisse todos os segmentos (refeição e alimentação) e pudesse ser usado digitalmente. A depender das categorias escolhidas por cada empresa contratante, os usuários do aplicativo poderiam mover seu saldo entre as vantagens como alimentação, refeição, mobilidade, saúde, educação, cultura e entretenimento. Ademais, é possível destacar duas maneiras de monetização da empresa, por meio da condição de emissora do cartão de crédito, ganha uma comissão a cada transação e pela aplicação de mensalidade por funcionário ativo, dependendo do tamanho da empresa e da complexidade dos produtos contratados.

Atualmente, a Flash tem uma plataforma de pontos para gerenciar essas transferências de valores entre categorias, a qual também permite administrar outros benefícios, oferecidos por parceiros ou pelas próprias empresas. Algumas das mais de 6.000 empresas atendidas pelo serviço são MadeiraMadeira, Loggi e VTEX, para essas companhias, um dos objetivos com a digitalização é tirar a carga operacional de benefícios do departamento de recursos humanos. Já um segundo desafio é transformar os benefícios em uma ferramenta de atração e retenção de talentos. Uma pesquisa encomendada pela Flash com mais de 100 departamentos de recursos humanos de clientes mostrou que, para 62% deles, os benefícios flexíveis melhoraram a retenção de talentos.

A *startup* brasileira destaca sua relevância no mercado através das recentes rodadas de investimento, de acordo com a Pipeline, em 2022 a empresa atingiu uma captação série C de US\$ 100 milhões, atraindo dois fundos de Boston com atuação global. A

Battery Ventures, gestora com quase 40 anos de operação, já investiu em empresas como Glassdoor, Groupon e Coinbase. Já a Whale Rock tem entre suas participações mais conhecidas as gigantes Tesla, Amazon e Facebook. Esse movimento, 10 meses depois da série B, de US\$ 22 milhões, capitaneada pela Tiger Global.

Em paralelo às rodadas de investimento de 2022, ocorreu também a aquisição da ExpenseOn, empresa especializada em gestão de despesas e reembolsos corporativos. Em meios às tecnologias desenvolvidas pela ExpenseOn, destaca-se o mecanismo de OCR voltado para controle de viagens. Além disso, é esperado que exista grande sinergia entre as marcas, assim, potencializando possibilidades de *cross-sell* com a solução de benefícios da Flash, pois a cartela de clientes da ExpenseOn é aproximadamente 200 mil (FLASH... , 2022).

Portanto, esse movimento de mercado não apenas fortalece a posição da Flash frente aos concorrentes, mas também melhora a experiência do próprio usuário, oferecendo um serviço mais completo com tecnologia integrada. Esta aquisição é um claro indicativo de como estratégias alinhadas e complementares podem criar sinergias poderosas, elevando o padrão de serviços disponíveis no setor de benefícios corporativos.

4.3 PESQUISA DE MERCADO LEVANTADA

Através de um estudo dos modelos de negócios já existentes, percebeu-se o domínio de empresas com operações *business to business*, ou seja, soluções desenvolvidas para outras empresas, com pouca presença no mercado direto para o consumidor. Com isso, levantou-se a hipótese de que existe potencial na escalabilidade de um produto voltado diretamente para o cliente, a partir de um *software* aberto. E a partir da aplicação do método para criação de empresas descrito em *The Lean Startup*, (RIES, 2011), o próximo passo se dá pela coleta de informações no mercado por meio de uma pesquisa qualitativa e quantitativa, a fim de encontrar um direcionamento para o MVP e identificar possíveis usuários qualificados para os testes iterativos.

Para o projeto em questão, foi elaborado um *google forms* com o objetivo de encontrar um nicho de pessoas que se identificassem com o problema na conversão de informações do papel para o computador, para isso denominou-se como "*ReadEase*" o nome do projeto em questão para a captação de *leads*. Deve-se notar a importância da pesquisa ter um caráter amplo, não direcional para a solução hipotética, pois só assim é possível validar que a dor existe no público sem viés nas respostas e o produto tem potencial de resolvê-la. A divulgação do formulário foi realizada para alunos do curso de Engenharia Química da Escola Politécnica da Universidade de São Paulo e resultou um total de 78 respostas, as quais terão seu resultado exposto e discutido na sequência, mas a completude das perguntas e respostas estará disponível no Apêndice A - Pesquisa de Mercado levantada com alunos da faculdade.

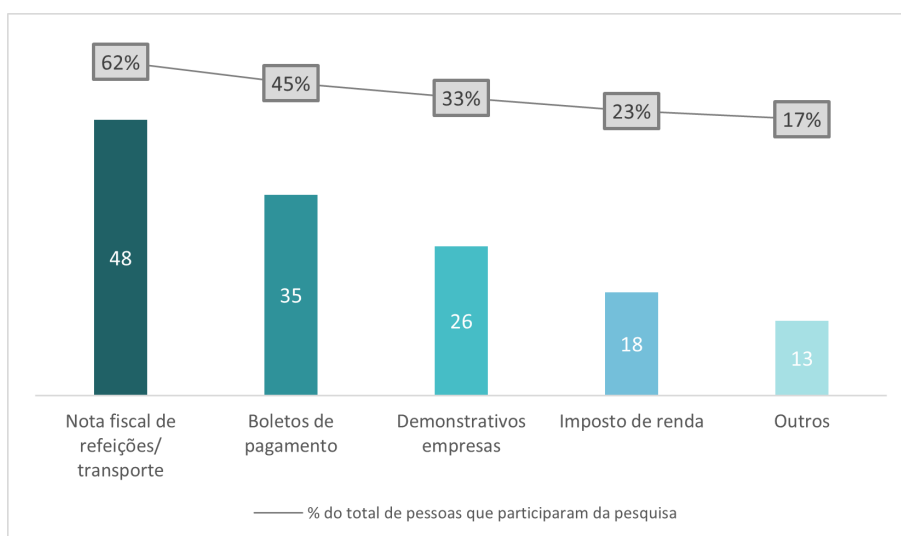


Figura 1 – Distribuição do tipo de documento que os entrevistados lidam no dia-a-dia, relacionado com a porcentagem referente ao total das 78 pessoas da pesquisa

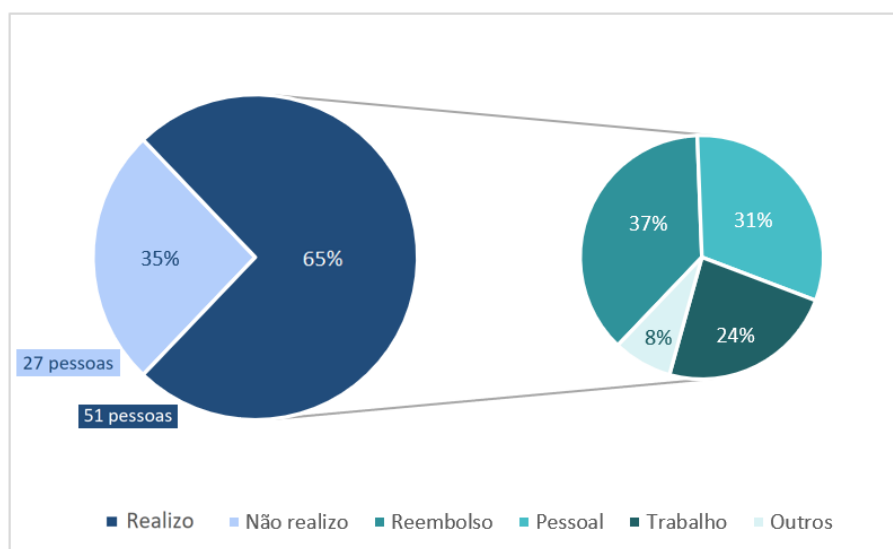


Figura 2 – Distribuição das respostas: à esquerda tem-se a divisão das pessoas que realizam ou não a extração de dados dos documentos que lidam no dia-a-dia, e à direita tem-se a segmentação dos motivos pelo qual as 51 pessoas extraem dados de documentos

A figura 1 apresentada quantifica a presença de cada tipo de documento no dia-a-dia dos respondentes. Notas fiscais de refeições ou transporte lideram, constituindo 62% dos documentos, seguidos por boletos de pagamento com 45%. Demonstrativos empresariais vêm em seguida com 33%, enquanto os comprovantes de Imposto de Renda representam 23%.

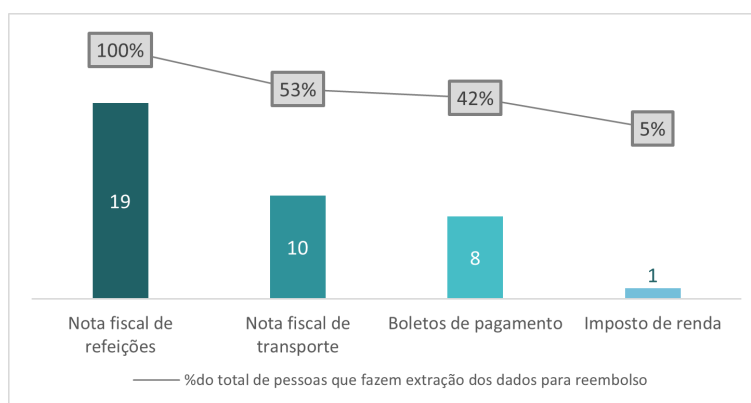


Figura 3 – Distribuição dos tipos de documentos que as pessoas extraem dados para fins de reembolso

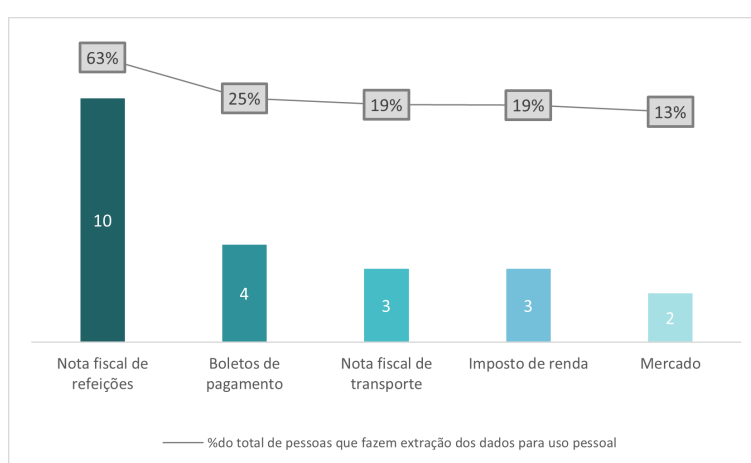


Figura 4 – Distribuição dos tipos de documentos que as pessoas extraem dados para fins de uso pessoal

O gráfico na figura 2 mostra a porcentagem de pessoas respondentes que realizam digitalização e extração de dados de documentos. Observa-se que 65% dos participantes, ou seja, 51 pessoas, realizam a digitalização de documentos, enquanto 35% (27 pessoas) não realizam. No gráfico menor à direita da figura, que detalha as finalidades da digitalização entre aqueles que a realizam, 37% digitalizam para reembolso, 31% para trabalho, 24% para uso pessoal, e 8% para outros fins, esses dados indicam a relevância da digitalização no contexto de reembolsos. Na sequência, pelas figuras 3 e 4, é possível visualizar quais tipos de documentos estão mais presentes dentro das necessidades de reembolso e pessoais, assim, destacam-se notas fiscais de refeições, transporte e boletos de pagamento.

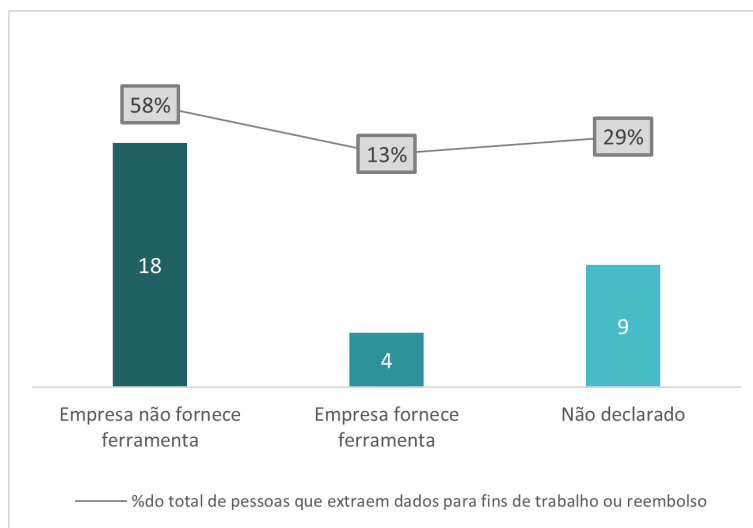


Figura 5 – Levantamento da quantidade de pessoas que extraem dados de documentos para fins de reembolso ou atividades do trabalho e que possuem ou não acesso a ferramentas para isso

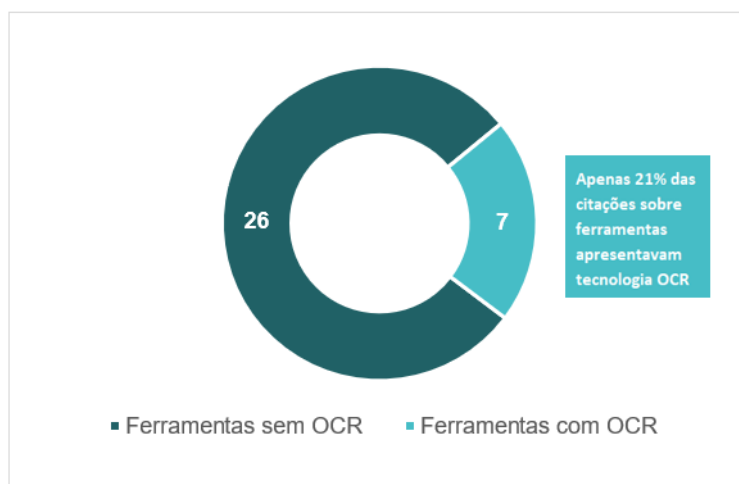


Figura 6 – Levantamento das citações de ferramentas para extração de dados, a fim de se concluir o alcance da tecnologia OCR

A figura 5 traz a disponibilidade de ferramentas para a extração de dados em contexto corporativo, utilizado por funcionários para fins de trabalho ou reembolso. Observa-se que 58% das pessoas que realizam extração de dados para atividades do trabalho ou reembolso, o equivalente a 18 respondentes, indicam que suas empresas não fornecem ferramentas para essa finalidade. Por outro lado, 13%, que correspondem a 4 pessoas, relatam que suas empresas fornecem a ferramenta necessária. Além disso, 29% dos participantes, representando 9 pessoas, não declararam ou não especificaram a situação em suas empresas. Isso indica que a maioria dos trabalhadores realiza a extração de dados sem apoio de ferramentas corporativas especializadas, sustentando a hipótese levantada para uma oportunidade de melhoria no suporte às atividades.

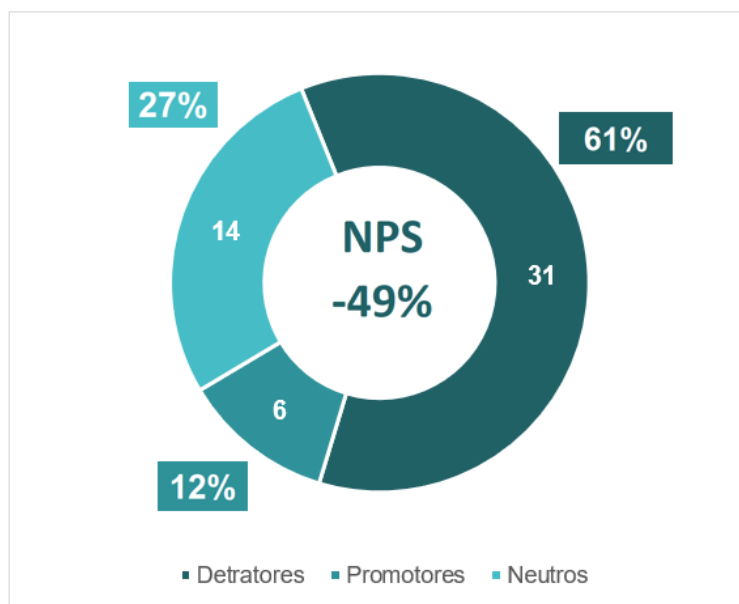


Figura 7 – Nível de satisfação das soluções utilizadas coletado de pessoas que realizam a extração de dados de documentos

O gráfico levantado na figura 6 ilustra a proporção de ferramentas utilizadas para extração de dados que empregam a tecnologia de Reconhecimento Óptico de Caracteres (OCR). Das citações coletadas na pesquisa com as 78 pessoas, apenas 21% das citações de ferramentas, que correspondem a 7 referências, mencionam o uso de soluções com OCR. Em contraste, a grande maioria, ou 79%, equivalente a 26 citações, faz referência a ferramentas que operam sem a tecnologia OCR. Este contraste destaca a subutilização do OCR, apesar de suas vantagens na automação e eficiência na captura de dados.

O último gráfico, apresentado pela figura 7, traz resultados de uma pesquisa *Net Promoter Score* (NPS), que é uma ferramenta para medir a satisfação dos clientes sobre as soluções utilizadas para a digitalização dos dados. A maior parte dos participantes, 61%, são classificados como detratores, o que indica uma tendência negativa na percepção da empresa. Os promotores representam apenas 12%, sugerindo que há poucos clientes altamente satisfeitos que recomendariam ativamente a empresa. Os neutros, que podem ser considerados passíveis, representam 27% dos entrevistados. Com um NPS de -49%, os dados revelam uma área crítica, a qual destaca-se como oportuna para uma nova solução focada nas dores dos clientes.

5 DESENVOLVIMENTO

5.1 IDENTIFICAÇÃO DO PROBLEMA

A partir da análise de mercado realizada, optou-se pela elaboração de uma solução que impactasse o dia-a-dia das pessoas por meio de um *software*. O objetivo principal originou-se por meio da pesquisa levantada com estudantes, pois com isso, a dor a ser resolvida se destacou em: a morosidade na transcrição de notas fiscais de transporte, alimentação e boletos para reportar o reembolso nas empresas em que estagiam. Sendo assim, o próximo passo foi a construção de um MVP, ou seja, um programa simplificado que ficasse pronto de maneira rápida e fosse desenvolvido de forma que seu código fosse iterativo, para que após os testes com usuários em potencial, as alterações com base em *feedbacks* fossem realizadas facilmente.

5.2 ESTUDO E SELEÇÃO DO AMBIENTE DE PROGRAMAÇÃO E DAS FERRAMENTAS AUXILIARES

Um ambiente para produção e desenvolvimento da ideia deve se basear em uma linguagem acessível e de domínio das pessoas que irão desenvolver o projeto. Neste cenário, dado o histórico de projetos da equipe e seus principais desenvolvimentos, no âmbito acadêmico e profissional, o software em destaque seria o *Python*.

A escolha se sustenta no fato de essa ferramenta ser *open source* e gratuita para todos os usuários. Além disso, muito da documentação já está desenvolvida na internet, seja em fóruns ou em artigos acadêmicos. Ademais, um computador com processamento fraco consegue instalar o software e também há ferramentas de adaptação de códigos do *Python* para a nuvem, possibilitando criação de rotinas e *clusters* de um processo automático dentro de uma empresa.

5.3 IDENTIFICAÇÃO DOS OUTPUTS E ETAPAS NECESSÁRIAS

Primeiramente, como as informações iniciais (inputs) do código já são os próprios arquivos PDF, tem-se que analisar quais serão os dados de interesse para obtenção.

No caso exemplificado neste documento, os inputs serão recibos de corridas realizadas no aplicativo Uber que contém as principais informações a respeito da viagem. Assim, desejou-se extrair a data e o valor total gerado no recibo emitido pelo próprio aplicativo da Uber.

Com isso, deve-se observar qual tipo de manipulação de dados irá se utilizar. Isto é, valores ou informações em documentos padronizados que, geralmente, ficam dispostos da mesma forma no documento. Como exemplo, alguns caracteres podem ficar localizados

após determinada frase ou valor numérico, e isso pode ser usado para aplicação do método de extração simplificada.

O método de extração simplificada consiste no armazenamento de parte do texto encontrado durante a varredura completa do documento em apenas uma variável. Por exemplo, ao desejar extrair o valor total de uma corrida do Uber, esse conjunto de números fica localizado, em geral, após o texto "Total", sendo necessário apenas localizar sua posição e armazenar os números dispostos logo em sequência.

Para métodos de extração mais complexos, pode-se usar também *RegEx*, ou seja, expressões regulares. As RegEx são padrões encontrados em conjuntos de informações. Neste caso, a definição de fragmento do texto buscado não será apenas um conjunto de caracteres em sequência, mais sim um padrão de caracteres ordenados. Por exemplo, na busca de um CPF de um usuário, o formato sempre será XXX.XXX.XXX-X, ou seja, sempre será uma sequência numérica separada por pontos, sendo possível esta extração inserindo uma variável de busca como:

```
1 r'\d{3}\.\d{3}\.\d{3}-\d{2}'
```

Sendo que ela localiza a posição e conteúdo de um texto exatamente no formato solicitado.

5.4 IMPLEMENTAÇÃO DO CÓDIGO

Abaixo segue o código comentando utilizado na implementação das aplicações descritas anteriormente, de extração e consolidação de datas e valores de corridas de Uber.

O código pode ser separado em quatro etapas fundamentais:

1. **Importação de Bibliotecas:** Na parte 1, importam-se as bibliotecas utilizadas, um procedimento padrão na construção de softwares a medida que se tornam mais complexos. Nesse caso, como mencionado na seção 4. *Métodos*, as bibliotecas têm função de auxiliar em todas as etapas do código, desde a preparação dos arquivos, até a extração, interpretação e consolidação dos dados.
2. **Definição das funções a serem utilizadas para leitura e adequação dos dados:** A parte 2 do código consiste do uso efetivo da ferramenta de OCR do *Pytesseract* (TESSERACT..., 2018) (I) e na implementação de outras funções auxiliares (II) e de localização do conteúdo de interesse no texto extraído (III). A primeira etapa conta com a conversão de cada uma das páginas do pdf importado pelo programa em imagens no formato .jpeg, usando a biblioteca/*software* Poppler (POPPLER..., 2023) que serão lidas e interpretadas pelo OCR (isso é feito para garantir a uniformidade de formatos de *input* do programa, garantindo maior consistência nos resultados). Em seguida, a método *pytesseract.image_to_string()* é

utilizado para realizar a leitura das imagens, cujos textos são guardados na variável *text* e escritas em um arquivo *.txt*. Já a função de localização usa o princípio de expressões regulares (*RegEx*) mas com uma implementação mais simples, que consiste na localização de *substrings* dentro da leitura extraída da página para identificação da posição da informação relevante - é fundamental, nesse momento, garantir que todas as exceções sejam tratadas de forma adequada, para garantir resultados coerentes independentemente das particularidades da entrada.

3. **Leitura e extração de dados:** A parte 3 do código é a unificação das funções construídas anteriormente, aplicando-as sequencial e oportunamente nos arquivos determinados (e percorridos com o uso da biblioteca *os*) para que os dados corretos sejam extraídos. Nessa etapa, ainda, as informações são guardadas em listas, nas quais as datas e valores advindos do mesmo arquivo possuem índices iguais.
4. **Exportação da análise:** Nesse momento, as listas de valores e datas são transformadas em colunas de um DataFrame (com o uso da biblioteca *pandas*, muito utilizada no tratamento de dados) e em seguida exportadas para um arquivo excel, onde serão úteis ao usuário para seus fins particulares.

Abaixo, na Figura 8 segue um exemplo de utilização do código, com exemplificação de *inputs* e *outputs* do programa.

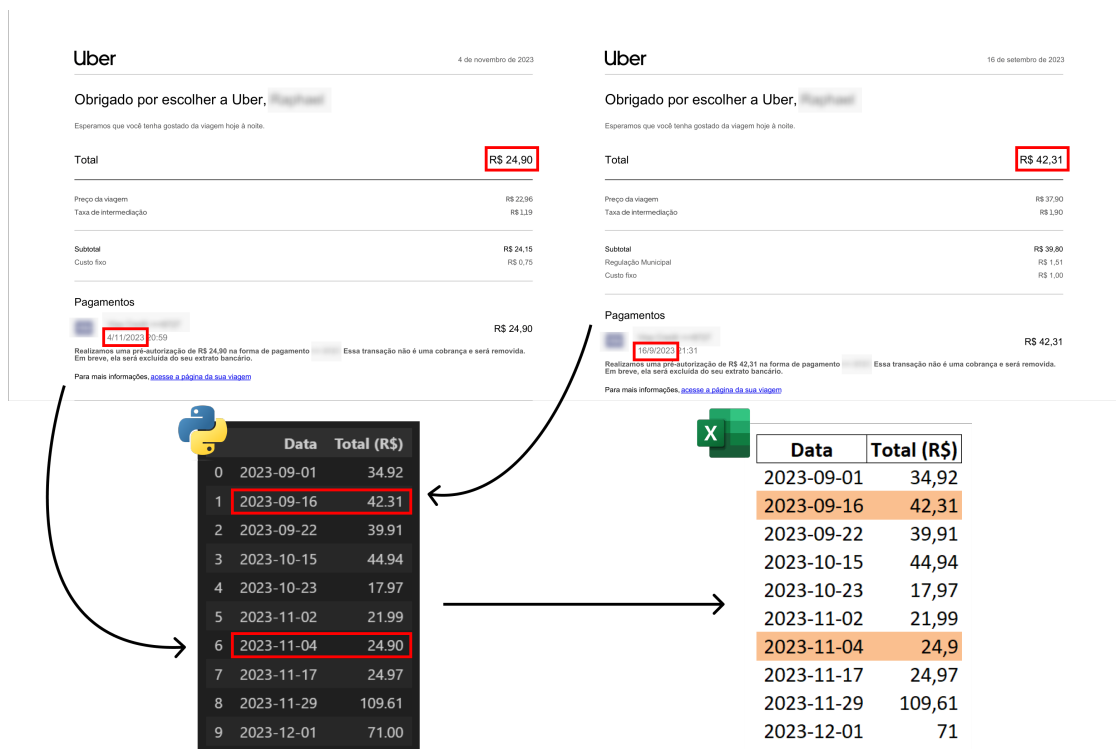


Figura 8 – Exemplo de utilização do programa para consolidação de valores e datas de viagens realizadas pelo Uber

O código referente à aplicação acima é:

```

1  ## 1. Importacao de Bibliotecas
3
5  # I. Importacao das bibliotecas necessarias
7
9  from PIL import Image
11 import pytesseract
13 import sys
15 from pdf2image import convert_from_path
17 import os
19 import re
21 import pandas as pd
23 import time
25
27 ## 2. Definicao das funcoes a serem utilizadas para leitura e adequacao
28     dos dados
30
32 # II. Definicao da funcao que realiza a leitura do arquivo com OCR
34
36 def leitura(path):
37     inicio_machine = time.time()
38     PDF_file = path
39
40     custom_config = r'-l por --psm 6'
41
42     pytesseract.pytesseract.tesseract_cmd = r"tesseract.exe"
43
44     pages = convert_from_path(PDF_file, 500, poppler_path=r'poppler
45 -0.68.0\bin')
46     image_counter = 1
47     for page in pages:
48         filename = "page_"+str(image_counter)+".jpg"
49         page.save(filename, 'JPEG')
50         image_counter = image_counter + 1
51
52     filelimit = image_counter-1
53     outfile = os.path.basename(path)[:4] + "_out_text.txt"
54     f = open(outfile, "a")
55     for i in range(1, filelimit + 1):
56         filename = "page_"+str(i)+".jpg"
57         text = str(((pytesseract.image_to_string(Image.open(filename),
58 config=custom_config))))
59         text = text.replace('-\n', '')
60         f.write(text)
61     f.close()
62
63

```

```

    fim_machine = time.time()
45     print('O tempo de execucao do Machine Learning durou {:.2f}
segundos.'.format(fim_machine - inicio_machine))
    return outfile
47
# III. Definicao de uma funcao auxiliar que transforma numeros guardados
em na forma de string em float
49
def ajeita_num(num):
51     return float(num.replace(',','.'))

53 # IV. Definicao da funcao especifica da aplicacao que identifica os
dados de interesse dentro do arquivo gerado pelo OCR

55 def extrair_info(text_as_string):
    # Pegar total
57
    str_total = "Total"
59     cut_total = text_as_string[text_as_string.find(str_total):]
    start_cut_total_2 = "R$ "
61     end_cut_total_2 = "\n"
    cut_total = cut_total[cut_total.find(start_cut_total_2)+len(
start_cut_total_2):cut_total.find(end_cut_total_2)]
63     cut_total = ajeita_num(cut_total)

65     # Pegar data

67     str_data = ":"
    cut_data = text_as_string[:text_as_string.find(str_data)]
69     if cut_data[0].isdigit() == False:
        cut_data = text_as_string[:cut_data.find(str_data)]
71     start_cut_data_2 = "/"
    cut_data = cut_data[cut_data.find(start_cut_data_2)-2:]
73     end_cut_data_2 = " "
    cut_data = cut_data[:cut_data.find(end_cut_data_2)]
75     if cut_data[:1] == "\n":
        cut_data = cut_data[1:]
77
    return cut_total, cut_data
79
## 3. Leitura e extracao de dados
81
# V. Leitura e extracao de dados de todos os recibos a serem analisados
(guardados em um diretorio especifico)
83
folder_path = r"C:\Users\rpelle02\TCC\receipts "
85

```

```

total_list = []
87 data_list = []

89 for filename in os.listdir(folder_path):
    file_path = os.path.join(folder_path, filename)
91     if os.path.isfile(file_path):
        out_path = leitura(file_path)
93         text_as_string = open(out_path).read()
        total, data = extrair_info(text_as_string)
95         total_list.append(total)
        data_list.append(data)
97
## 4. Exportacao da analise
99
# VI. Consolidacao das informacoes em um Data Frame e exportacao do
    resultado para excel
101
df = pd.DataFrame()
103 df['Data'] = data_list
df['Data'] = pd.to_datetime(df['Data'], format='%d/%m/%Y').dt.date
105 df['Total (R$)'] = total_list
df = df.sort_values(by='Data', ascending=True).reset_index(drop = True)
107
df.to_excel(r'resumo_uber.xlsx', index=False)

```


6 RESULTADOS

6.1 TESTE DE HIPÓTESE E FEEDBACK DOS USUÁRIOS

A partir da disponibilização para teste dentro de uma amostra limitada de usuários, obteve-se os principais pontos de melhoria percebidos por eles.

O primeiro ponto foi a limitação para apenas notas fiscais de transporte, tornando sua utilidade e escopo reduzido. Outro fator limitante foi a dificuldade e familiaridade para conseguir utilizar o *software*, como não havia uma interface clara ao usuário, devia-se alterar propriamente no código a localização dos seus documentos e também a localização de destino do *output*. Por fim, houve dificuldade de se encontrar um suporte às dúvidas e um material explicativo, passo-a-passo, de como utilizar seus funcionalidades.

Com isso, partiu-se para adaptar o *software* para conter as mudanças solicitadas e também criar maior acessibilidade para usar a solução.

6.2 ADAPTAÇÕES NO MVP

6.2.1 Incorporação da leitura e processamento de Boletos Bancários

O primeiro passo em direção ao atendimento da necessidade sinalizada pelos usuários foi de aumento do escopo de atuação do programa. A decisão pelo processamento de Boletos Bancários foi apoiada no *feedback* direto dos usuários, que sugeriram a adição, e nos resultados da pesquisa preliminar realizada pelo grupo. Nessa última, os boletos bancários apareceram como o segundo tipo de documento mais digitalizado, seguidos apenas de notas fiscais de transportes/alimentação¹.

Os passos seguintes, envolveram um novo *sprint* de desenvolvimento, com retomada dos mesmos conceitos abordados anteriormente. Nesse momento, foi fundamental identificar quais as informações-chave a serem extraídas (como a data de vencimento, o valor total do boleto e o seu código de barras), que tornariam a aplicação relevante. Com a escolha feita, o desenvolvimento se sucedeu pela adaptação do código e do *Machine Learning* aos modelos de boleto bancário, para que, com isso, fosse possível manipular as funções do código e extrair as informações de interesse, como segue na Figura 9

¹ Os resultados completos da pesquisa podem ser acessados no Apêndice A

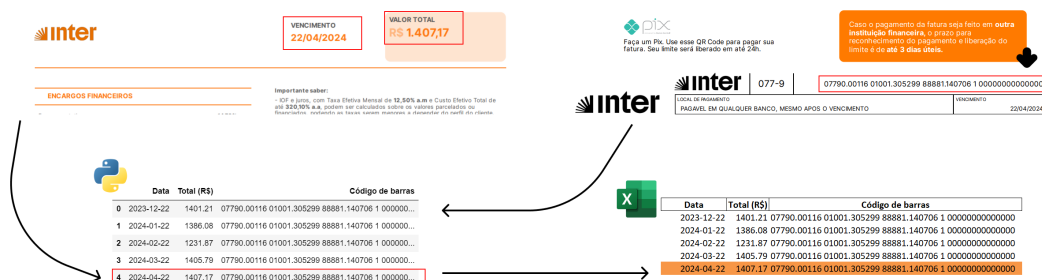


Figura 9 – *Software* aplicado ao caso de boletos bancários

Dessa forma, mostra-se que o código é simples e maleável, sendo possível ser inserido em diversos contextos, apenas aplicando as mudanças específica que cada documento necessita.

Abaixo, está a implementação em python da funcionalidade de leitura e processamento de boletos:

```

1 from PIL import Image
  import pytesseract
3 import sys
  from pdf2image import convert_from_path
5 import os
  import re
7 import pandas as pd
  import time
9
10 def leitura(path):
11     inicio_machine = time.time()
12     PDF_file = path
13
14     custom_config = r'-l por --psm 6'
15
16     pytesseract.pytesseract.tesseract_cmd = ~caminho~
17
18     pages = convert_from_path(PDF_file, 500, poppler_path=~caminho~)
19     image_counter = 1
20     for page in pages:
21         filename = "page_"+str(image_counter)+".jpg"
22         page.save(filename, 'JPEG')
23         image_counter = image_counter + 1
24
25     filelimit = image_counter-1
26     outfile = os.path.basename(path)[: -4] + "_out_text.txt"
27     f = open(outfile, "a")
28     for i in range(1, filelimit + 1):
29         filename = "page_"+str(i)+".jpg"
30         text = str(((pytesseract.image_to_string(Image.open(filename)),

```

```

config=custom_config))))
31     text = text.replace('-\n', '')
        f.write(text)
33     f.close()

35     fim_machine = time.time()
    print('O tempo de execu o do Machine Learning durou {:.2f}
segundos.'.format(fim_machine - inicio_machine))
37     return outfile

39 def ajeita_num(num):
    return float(num.replace(',','.'))

41

43 def extrair_info(text_as_string):
    # Pegar total

45     str_total = "VALOR DO DOCUMENTO"
    cut_total = text_as_string[(text_as_string.find(str_total)) + len(
str_total):]
47     cut_total_2 = cut_total[(cut_total.find(str_total)):]
    str_rs = 'R$ '
49     cut_total_3 = cut_total_2[(cut_total_2.find(str_rs)) + len(str_rs):]
    cut_total_4 = cut_total_3[:cut_total_3.find("\n")]
51     # Inserir um ponto antes dos dois ltimos n meros
    if "," in cut_total_4 and "." in cut_total_4:
53         cut_total_5 = float(cut_total_4.replace(".", "").replace(",", "."
    ))
    elif "," in cut_total_4 and "." not in cut_total_4:
55         cut_total_5 = float(cut_total_4.replace(",", "."))
    elif "." in cut_total_4 and "," not in cut_total_4:
57         cut_total_5 = float(cut_total_4.replace(".", ""))
    else:
59         cut_total_5 = float(cut_total_4[:-2] + "." + cut_total_4[-2:])

61     # Pegar data

63     # Express o regular para encontrar a data no formato "dd/mm/aaaa"
    regex_data = r"\b\d{2}/\d{2}/\d{4}\b"
65     # Encontrar correspond ncias
    matches_data = re.findall(regex_data, text_as_string)
67     cut_data = matches_data[0]

69     # Pegar codigo de barra
    regex_codigo = r"Inter\s\d+\-\d\s(\d{5}\.\d{5}\s\d{5}\.\d{6}\s\d
{5}\.\d{6}\s\d\s\d{14})"
71

    # Encontrar correspond ncias

```

```

73     matches_codigo = re.findall(regex_codigo, text_as_string)

75     # Se houver correspondências, armazenar o primeiro resultado
76     if matches_codigo:
77         cut_codigo = matches_codigo[0]
78         cut_codigo

79     return cut_total_5, cut_data, cut_codigo

81 # Replace 'your_directory_path' with the path to your folder
82 folder_path = ~caminho~

83
84
85 codigo_list = []
86 total_list = []
87 data_list = []

88
89 # Iterate through all files in the specified directory
90 for filename in os.listdir(folder_path):
91     file_path = os.path.join(folder_path, filename)
92     # Check if it's a file and not a directory
93     if os.path.isfile(file_path):
94         out_path = leitura(file_path)
95         text_as_string = open(out_path).read()
96         total, data, codigo_valor = extrair_info(text_as_string)
97         total_list.append(total)
98         data_list.append(data)
99         codigo_list.append(codigo_valor)

101     else:
102         print(f'Directory found: {filename}')

103
104 df = pd.DataFrame()
105 df['Data'] = data_list
106 df['Data'] = pd.to_datetime(df['Data'], format='%d/%m/%Y').dt.date
107 df['Total (R$)'] = total_list
108 df['Código de barras'] = codigo_list
109 df = df.sort_values(by='Data', ascending=True).reset_index(drop = True)

111
112 df.to_excel(~caminho~, index=False)

```

6.2.2 Adição de uma *Graphical User Interface* (GUI)

A decisão pela adição de uma GUI partiu da insatisfação dos usuários testados com a usabilidade do programa, principalmente em termos de *User Experience* (UX) e *User Interface* (UI). Isso, pois na primeira formulação do MVP, o usuário deveria, no código

fonte, incluir os caminhos referentes às pastas dos arquivos corretos, o que causou confusão durante o uso. Nesse sentido, a adição de uma interface gráfica atende à necessidade de incorporação de um modo de uso intuitivo e amigável a usuários que não estejam familiarizados com programação. Abaixo, na Figura 10, está a interface desenvolvida após a primeira sessão de *feedbacks*.

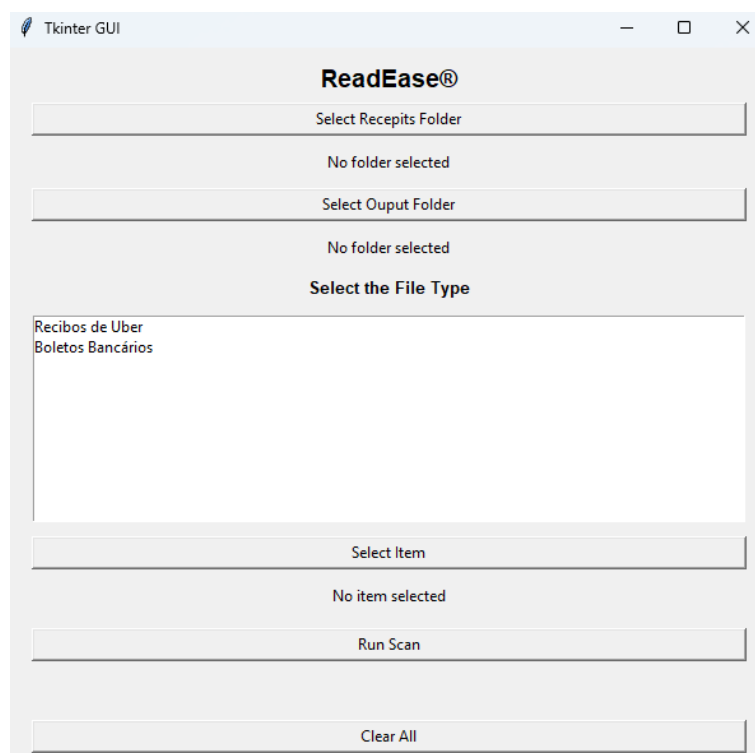


Figura 10 – *Graphical User Interface* desenvolvida para o *software*

As opções funcionam da seguinte forma:

- **Select Receipts Folder:** Permite selecionar o diretório em que os recibos de transporte ou boletos estão
- **Select Output Folder:** Permite selecionar o diretório em que o arquivo de excel resultante deverá ser salvo
- **Select the File Type:** Permite selecionar qual tipo de documento o programa irá processar (atualmente com as opções de Recibos de Uber e Boletos Bancários)
- **Run Scan** Aciona o programa, que processa os arquivos selecionados e retorna o arquivo resultante de excel no diretório correta
- **Clear All:** Limpa as seleções dos diretórios dos recibos/boletos, do diretório de salvamento do resultado e da seleção do tipo de documento

Para tanto, o seguinte código foi desenvolvido em Python:

```

import tkinter as tk
2 from tkinter import filedialog, messagebox, font, Scrollbar, Listbox

4 # Create the main window
root = tk.Tk()
6 root.title("Tkinter GUI")

8 # Configure the grid for the main window
root.grid_columnconfigure(0, weight=1) # Make column 0 stretchable
10
11 for row_index in range(13): # Adjust for additional rows
12     root.grid_rowconfigure(row_index, weight=1) # Make all rows
        stretchable

14 # Define custom fonts
large_bold_font = font.Font(size=14, weight="bold")
16 italic_font = font.Font(size=12, slant="italic")

18 # Global variables to store folder paths and selected item from the
    listbox
folder_path1 = None
20 folder_path2 = None
selected_item = None
22

# List of options to populate the Listbox
24 options_list = ["Recibos de Uber", "Boletos Banc rios"]

26 # Function to be called when Button 1 is clicked
def on_button1_click():
28     global folder_path1
    folder_path1 = filedialog.askdirectory()
30     if folder_path1:
        label_folder1.config(text=f"Folder 1: {folder_path1}")
32

# Function to be called when Button 2 is clicked
34 def on_button2_click():
    global folder_path2
36     folder_path2 = filedialog.askdirectory()
    if folder_path2:
38         label_folder2.config(text=f"Folder 2: {folder_path2}")

40 # Function to get the selected item from the listbox
def get_selected_item():
42     global selected_item
    try:
44         index = listbox.curselection()[0] # Get index of current

```

```

selection
    selected_item = listbox.get(index) # Get the list item using
the index
46     label_selected_item.config(text=f"Selected: {selected_item}") #
Display selected item
except IndexError:
48     messagebox.showinfo("Selection Error", "No item selected. Please
select an item from the list.")

50 # Dummy function for Button 3 with validation
def on_button3_click():
52     # Check if folders have been selected for both buttons
    if "No folder selected" not in label_folder1.cget("text") and "No
folder selected" not in label_folder2.cget("text") and "No item
selected" not in label_selected_item.cget("text"):
54         label_success.config(text=f"Success! The excel file is available
at the selected Output Folder", fg="green") # Update the success
label with the result
    else:
56         # Show a message box if folders haven't been selected
        messagebox.showinfo("Folder and File Type Selection Required", "
Please select both folders and the file type before proceeding")

58
# Function to clear all selections and labels
60 def on_clear_click():
    global folder_path1, folder_path2
62     folder_path1 = None
    folder_path2 = None
64     label_folder1.config(text="No folder selected")
    label_folder2.config(text="No folder selected")
66     listbox.selection_clear(0, tk.END)
    label_selected_item.config(text="No item selected")
68

# Create a label widget with the larger and bold font text
70 label_large_bold = tk.Label(root, text="ReadEase ", font=
    large_bold_font)
label_large_bold.grid(row=0, sticky="nsew", padx=20, pady=(10, 0))
72

74 # Create Button 1 and place it in the window
button1 = tk.Button(root, text="Select Receipts Folder", command=
    on_button1_click)
76 button1.grid(row=2, sticky="nsew", padx=20, pady=5)

78 # Create a label to display the selected folder path for Button 1
label_folder1 = tk.Label(root, text="No folder selected", justify="
    center")

```

```

80 label_folder1.grid(row=3, sticky="nsew", padx=20, pady=5)

82 # Create Button 2 and place it in the window
button2 = tk.Button(root, text="Select Ouput Folder", command=
    on_button2_click)
84 button2.grid(row=4, sticky="nsew", padx=20, pady=5)

86 # Create a label to display the selected folder path for Button 2
label_folder2 = tk.Label(root, text="No folder selected", justify="
    center")
88 label_folder2.grid(row=5, sticky="nsew", padx=20, pady=5)

90 # Create a label to display the selected folder path for Button 2
label_folder2 = tk.Label(root, text="Select the File Type", justify="
    left", font=font.Font(size=10, weight="bold"))
92 label_folder2.grid(row=6, sticky="nsew", padx=20, pady=5)

94 # Create a Listbox with the options list
listbox = Listbox(root)
96 listbox.grid(row=7, column=0, sticky="nsew", padx=20, pady=5)
for option in options_list:
98     listbox.insert(tk.END, option)

100 # Button to select an item from the listbox
button_select_item = tk.Button(root, text="Select Item", command=
    get_selected_item)
102 button_select_item.grid(row=8, sticky="nsew", padx=20, pady=5)

104 # Label to display the selected item
label_selected_item = tk.Label(root, text="No item selected", justify="
    center")
106 label_selected_item.grid(row=9, sticky="nsew", padx=20, pady=(5, 10))

108 # Create Button 3 and place it in the window
button3 = tk.Button(root, text="Run Scan", command=on_button3_click)
110 button3.grid(row=10, sticky="nsew", padx=20, pady=5)

112 # Create a label to display the result of the dummy operation by Button
    3
label_success = tk.Label(root, text="", justify="center")
114 label_success.grid(row=11, sticky="nsew", padx=20, pady=(5, 10))

116 # Create a new button to clear all selections and labels
button_clear = tk.Button(root, text="Clear All", command=on_clear_click)
118 button_clear.grid(row=12, sticky="nsew", padx=20, pady=(5, 10))

120 # Start the GUI event loop

```



```
root.mainloop()
```

6.2.3 Desenvolvimento de uma *Landing Page*

Esta etapa ficou associada com a criação de um *google site* simples, onde seria possível entrar em contato em caso de dúvidas e também fazer o *download* dos arquivos para executar o programa. Também foi disponibilizado na página o manual do usuário, um documento contendo o passo-a-passo de como utilizar e manipular o programa. Nele será encontrado como configurar e inserir os documentos de interesse, manipular o código e direcionar os resultados gerados pela IA. Para melhor experiência do usuário, foi construída uma identidade visual de acordo com a empresa fictícia criada, a "*ReadEase*".

Ao introduzir um MVP por meio de uma *Landing Page*, as startups e empresas podem rapidamente construir uma audiência e colher percepções valiosas sem ter que desenvolver o produto por completo. Isso é particularmente útil para startups que ainda não têm uma base estável de usuários e precisam validar sua ideia de produto no mercado. Uma *Landing Page* de MVP permite que você apresente seu conceito de produto e comece a construir uma conversa com potenciais clientes, ao mesmo tempo que capta feedback detalhado e preciso sobre suas escolhas em situações quase reais, sem perguntar explicitamente o que eles desejam.

Para conferência desta *Landing Page* segue o link a seguir: ReadEase, para acessar basta realiza o login por meio do email usp.

7 DISCUSSÕES

7.1 APLICAÇÃO DO MÉTODO *LEAN STARTUP* PARA DESENVOLVIMENTO DE PRODUTOS

A metodologia Lean Startup revolucionou o planejamento e desenvolvimento de novos produtos, principalmente em ambientes de startup, onde os recursos são limitados e a necessidade de inovação rápida é alta. Originada por Eric Ries, a metodologia enfatiza a importância de aprender com os clientes o mais rápido possível para reduzir os riscos e evitar o desenvolvimento de produtos que os consumidores não querem. A ideia central é construir um Produto Mínimo Viável (MVP) para testar hipóteses de mercado com o mínimo esforço necessário, permitindo que a empresa valide ideias e entenda as demandas do cliente sem investimentos excessivos. Este enfoque inicial no aprendizado em vez de na execução perfeita incentiva uma cultura de agilidade e adaptabilidade, características essenciais em mercados voláteis e altamente competitivos.

O ciclo de "Construir-Medir-Aprender" da metodologia Lean Startup proporciona um roteiro iterativo e eficiente para o desenvolvimento de produtos. Ao construir uma versão inicial do produto, medir como os clientes reagem a ele e aprender com esses dados, as empresas podem fazer ajustes rápidos e informados. Este ciclo não apenas economiza tempo e recursos, mas também alinha as equipes de desenvolvimento e produto com os objetivos do negócio, mantendo o foco na satisfação do cliente e na solução de problemas reais. A capacidade de iterar rapidamente baseando-se no feedback do usuário garante que a empresa permaneça relevante e proativa diante das mudanças nas preferências e necessidades dos consumidores.

Além disso, a Lean Startup promove uma abordagem baseada em dados para a tomada de decisões, que é crucial na minimização do risco associado ao lançamento de novos produtos. Com uma estratégia que valoriza o feedback dos clientes e o aprendizado contínuo, as startups podem evitar o ciclo comum de longos e custosos períodos de desenvolvimento seguidos por lançamentos de produtos falhos. A metodologia encoraja as empresas a falharem rapidamente e a aprenderem com esses fracassos, conduzindo a um refinamento constante do produto. Essa adaptabilidade não só é essencial para a sobrevivência e crescimento em ambientes dinâmicos, mas também oferece uma vantagem competitiva significativa, permitindo que as empresas se antecipem às tendências de mercado e se ajustem de forma proativa às novas oportunidades.

No caso desse trabalho, a metodologia foi especialmente fundamental para priorização de tarefas e direcionamento de esforços para o desenvolvimento de funcionalidades que fizessem sentido para o usuário. A fundamentação teórica para o processo de inovação garante um respaldo importante para validação dos esforços, para garantir que haja maior

aproveitamento das oportunidades identificadas pelo grupo.

7.2 BENEFÍCIOS DO *SOFTWARE* DESENVOLVIDO

Os benefícios advindos do uso do *software* desenvolvido nesse trabalho relacionam-se, principalmente, à economia de tempo gasto em atividades automatizáveis de leitura e processamento de dados encontrados em documentos. Nesse sentido, a proposta de valor elaborada refere-se à desoneração dos usuários, com maior disponibilização de tempo para dedicação a outras tarefas que requeiram sua atenção.

Nesse caso, é possível destacar dois casos principais de benefícios dos usuários:

- **Economia de tempo de trabalho:** Esses são os casos de aplicação do *software* desenvolvido no ambiente corporativo, de trabalho. No caso da solução para processamento de Boletos Bancários, pode-se desonerar um analista de contas a pagar/-contas a receber, que deve compilar todos os boletos de fornecedores e/ou clientes para garantir um controle contábil mais robusto. Nesse sentido, a quantificação do retorno financeiro do uso do *software* é mais direta, podendo ser calculada segundo a seguinte fórmula:

$$Retorno (R\$) = Tempo economizado (h) \cdot compensacao (R\$/h)$$

Sendo:

$$\begin{aligned} Tempo economizado (h) = & Tempo gasto na atividade sem software \\ & - Tempo gasto na atividade com software \end{aligned}$$

Isso, considerando a premissa de que o tempo economizado pelo trabalhador pode ser dedicado a outras atividades dentro da empresa que sejam de igual relevância na atividade econômica da mesma.

Sendo assim, é possível quantificar, monetariamente, os benefícios de utilização da solução fornecida, o que facilita a tomada de decisão e percepção de valor da ferramenta.

- **Economia de tempo pessoal:** Esses são os casos em que o usuário consegue economizar tempo utilizando a solução em seu tempo pessoal, performando tarefas fora do período de sua jornada de trabalho (e não relacionadas a ela). No caso do processamento de Recibos de Transporte, pode-se desonerar o usuário em seu tempo livre pessoal, que antes deveria ser dedicado a horas de trabalho manual de leitura e processamento dos dados referentes ao reembolso de viagens, por exemplo. A pesquisa realizada revelou que um terço dos respondentes dedicavam seu tempo

pessoal realizando solicitações de reembolso para suas empresas¹. Nesses casos, a quantificação dos benefícios associados ao uso do *software* é menos direta, e não necessariamente pode ser feita em termos monetários. A proposta de valor, nessas situações, é de desoneração do usuário em seu tempo livre, para que esse possa dedicar-se a atividades de lazer e descanso, o que pode trazer ganhos de qualidade de vida de forma geral (e ganhos indiretos de produtividade nos momentos de tempo de trabalho).

7.3 PRÓXIMOS PASSOS

Para evoluir a solução proposta neste texto e ainda continuar aplicando as teorias do método *Lean Startup*, será importante realizar novas interações do MVP, sempre averiguando as respostas dos usuários e as necessidades de alterações/evoluções.

Os ciclos seguintes basicamente seriam compostos pelas seguintes fases:

1. **Rodada de teste do produto e coleta de *feedbacks*:** Essa rodada conta com a participação de usuários, e deve ser realizada para que esses possam testar o produto em suas funcionalidades disponíveis. A coleta subsequente de *feedbacks* é fundamental para entender a percepção desses usuários sobre a ferramenta, incluindo características que poderiam ser inclusos, retrabalhados ou até retirados para as próxima versões.
2. **Análise de *feedbacks*:** Após a coleta dos *feedbacks* é importante retornar à prancheta para entender como o usuário gostaria que o produto fosse modificado. Nesse momento, são definidos os próximos passos em termos de desenvolvimento da próxima versão da ferramenta, considerando o aprendizado da rodada anterior.
3. **Desenvolvimento da nova versão:** Após a definição das *features* a serem incluídas, retrabalhadas ou retiradas para a próxima versão, cabe ao time de projeto implementar e desenvolver o que for necessário para garantir que essas mudanças sejam refletidas ao usuário final.
4. **Relançamento do produto:** Feitas as alterações, é importante que o produto volte ao mercado, para que se reinicie o processo de volta ao passo 1. *Rodada de teste do produto e coleta de feedbacks*.

Algumas das incorporações já mapeadas como possíveis próximos passos são:

¹ Os resultados completos da pesquisa podem ser acessados no Apêndice A

7.3.1 Criação de um arquivo executável

Neste âmbito, seria evoluir ainda mais a experiência do usuário diante do *software*. Para isso, recolheu-se *feedbacks* no sentido de existir uma forma mais simples de aplicar o código na máquina do usuário final. Além disso, uma forma que não fosse obrigatório a instalação das diversas bibliotecas e arquivos (como o próprio Poppler e o Tesseract) no computador de quem for utilizá-la. Uma das formas de se contornar este problema seria a criação de um arquivo executável, ou seja, com a terminação *.EXE*.

Um arquivo deste tipo contém instruções que um computador pode seguir para realizar uma tarefa específica. Esses arquivos são frequentemente utilizados para instalar programas em um computador e executá-los, ou também realizar uma tarefa pré-definida. Quando você clica duas vezes em um arquivo *.EXE*, o sistema operacional do computador carrega as instruções contidas em sua memória e as executa, o que geralmente resulta na instalação ou execução de um programa.

O arquivo *.EXE* contém o código-fonte do programa, bem como outras informações necessárias para que o sistema operacional do computador saiba como executá-lo corretamente.

7.3.2 Criação de integrações com plataformas ou serviços externos

Outra demanda frequente questionada pelos usuários foi com relação à possibilidade de se integrar os resultados gerados pelo programa junto aos sistemas internos ou próprios que as empresas utilizam.

Isto é possível e rondariam o tópico de construção de *API's*, que funcionam basicamente como uma ponte que permite que diferentes *softwares* se comuniquem entre si de forma padronizada e segura, facilitando a integração e a interoperabilidade entre sistemas diferentes.

Ela consiste em, basicamente, regras e protocolos que permite que diferentes softwares se comuniquem entre si. É como se fosse uma ponte que permite que um aplicativo ou sistema solicite e receba informações de outro aplicativo, serviço ou sistema, definindo como os diferentes componentes de *software* devem interagir uns com os outros. Para isso, uma *API* fornece um conjunto de instruções e padrões que os desenvolvedores podem usar para acessar os recursos ou serviços de um sistema, como dados, funcionalidades ou outras operações.

REFERÊNCIAS

FLASH compra ExpenseOn e entra no mercado de despesas corporativas. 2022. Disponível em: <https://startups.com.br/noticias/flash-compra-expenseon-e-entra-no-mercado-de-despesas-corporativas/>. Acesso em: 20 fev. 2024.

POPPLER Documentation. 2023. Disponível em: <https://poppler.freedesktop.org/>. Acesso em: 10 out. 2023.

RIES, E. **The Lean Startup**. [S.l.]: Portfolio Penguin, 2011. 336 p.

TESSERACT Documentation. 2018. Disponível em: <https://tesseract-ocr.github.io/>. Acesso em: 3 out. 2023.

APÊNDICES

APÊNDICE A – PESQUISA DE MERCADO LEVANTADA COM ALUNOS DA FACULDADE

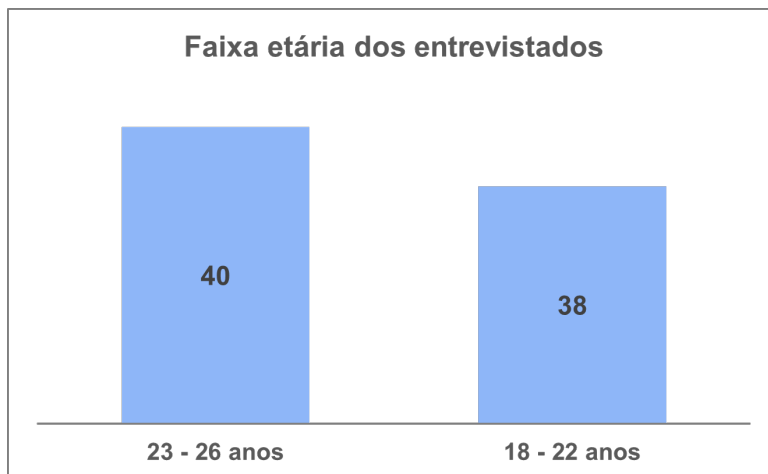


Figura 11 – Faixa etária dos estudantes que responderam a pesquisa de mercado

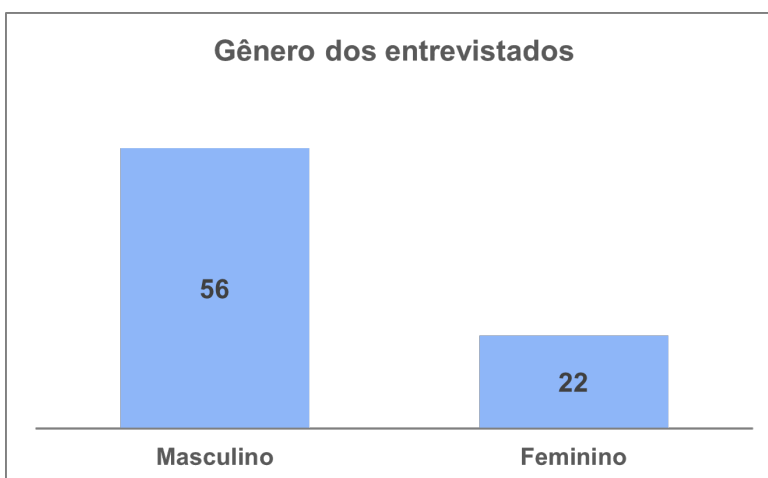


Figura 12 – Gênero dos estudantes que responderam a pesquisa de mercado

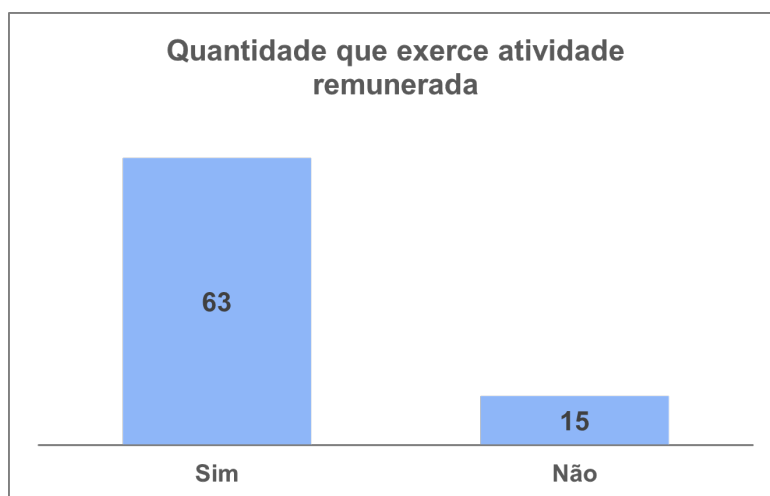


Figura 13 – Estudantes que exercem ou não atividade remunerada que responderam a pesquisa de mercado

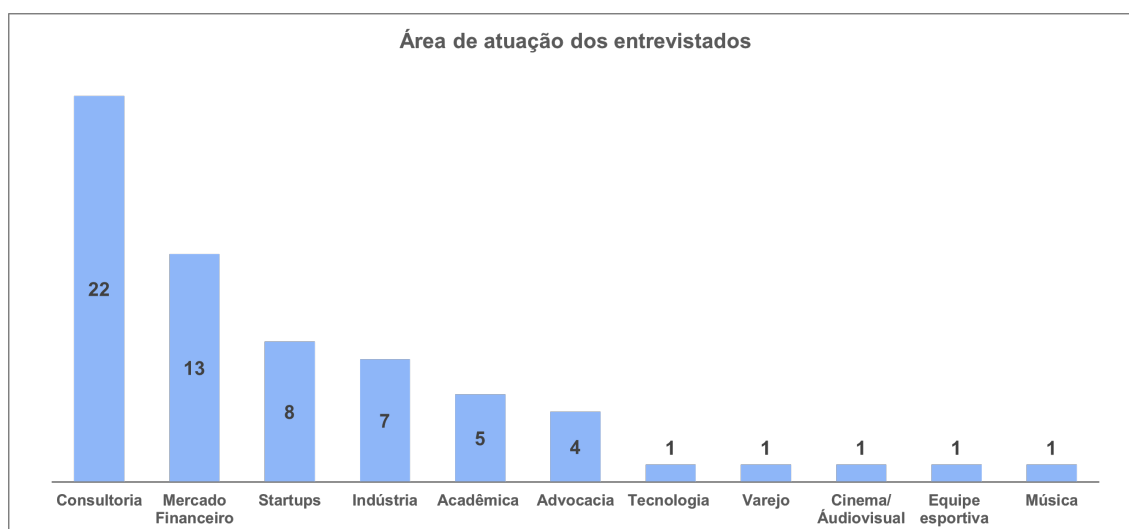


Figura 14 – Distribuição dos estudantes que responderam a pesquisa de acordo com a sua área de atuação no mercado



Figura 15 – Distribuição das finalidades para qual a digitalização ou extração de informações é utilizada

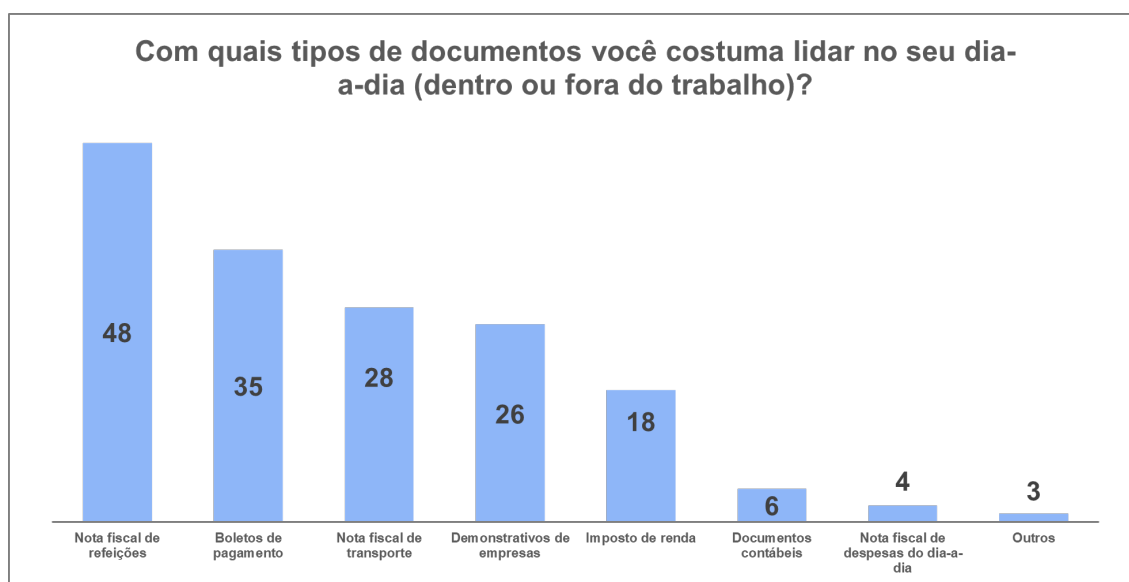


Figura 16 – Distribuição dos tipos de documento que os estudantes presentes na pesquisa costumam lidar no dia-a-dia

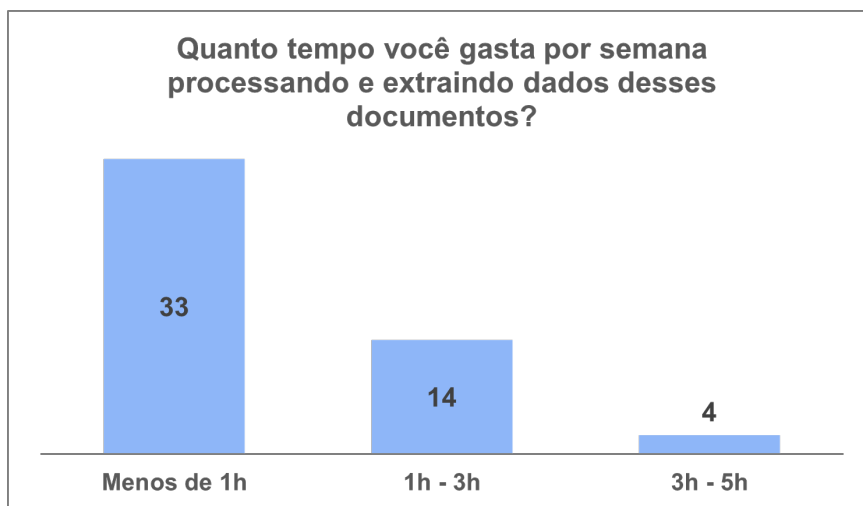


Figura 17 – Tempo gasto para a extração de dados dos documentos

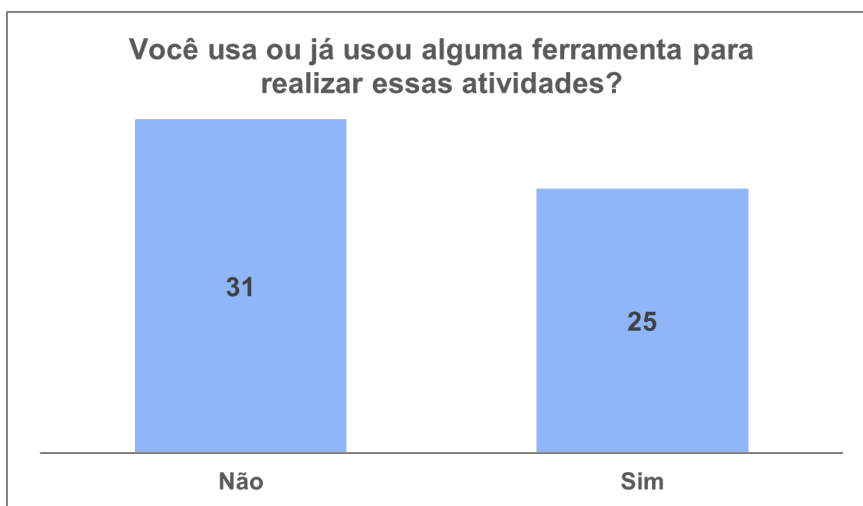


Figura 18 – Quantidade de pessoas presentes na pesquisa que utilizam ou não ferramentas específicas para a extração de dados

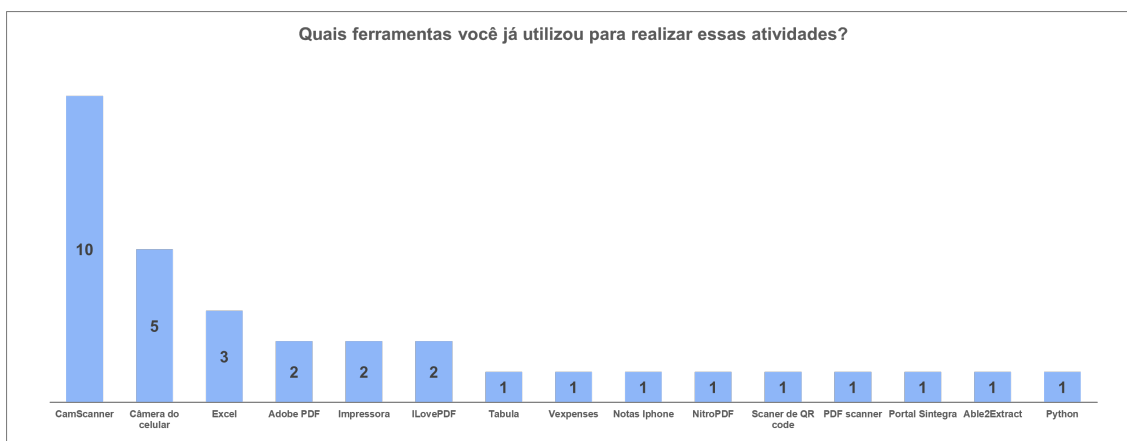


Figura 19 – Distribuição das ferramentas utilizadas

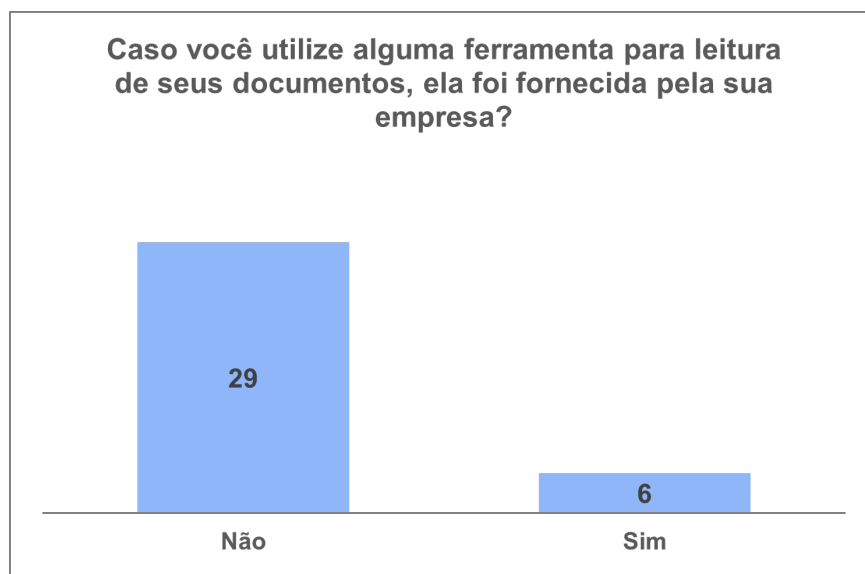


Figura 20 – Quantidade de pessoas que tem a ferramenta oferecida pela empresa em que trabalha para exercer a extração dos dados

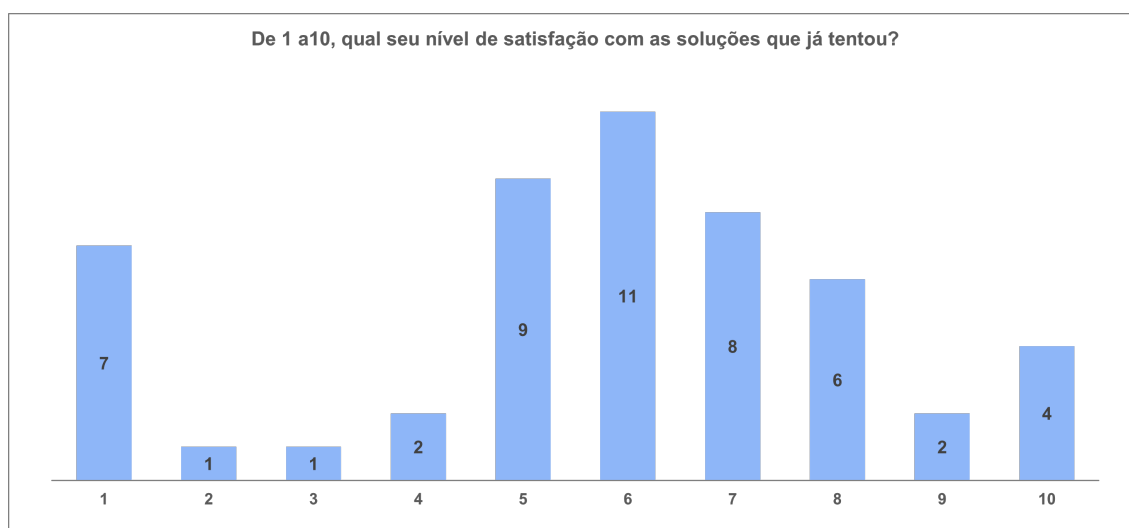


Figura 21 – Distribuição do nível de satisfação daqueles que utilizam ferramentas com a solução já utilizada