



**UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO**

MATHEUS HENRIQUE GUEDES DE OLIVEIRA

**AUTOMAÇÃO DE VERIFICAÇÕES MANUAIS EM PROCESSOS
DE KYC**

CAMPINA GRANDE - PB

2024

MATHEUS HENRIQUE GUEDES DE OLIVEIRA

**AUTOMAÇÃO DE VERIFICAÇÕES MANUAIS EM PROCESSOS
DE KYC**

**Trabalho de Conclusão Curso
apresentado ao Curso Bacharelado em
Ciência da Computação do Centro de
Engenharia Elétrica e Informática da
Universidade Federal de Campina
Grande, como requisito parcial para
obtenção do título de Bacharel em
Ciência da Computação.**

Orientador : João Arthur Brunet Monteiro

CAMPINA GRANDE - PB

2024

MATHEUS HENRIQUE GUEDES DE OLIVEIRA

**AUTOMAÇÃO DE VERIFICAÇÕES MANUAIS EM PROCESSOS
DE KYC**

**Trabalho de Conclusão Curso
apresentado ao Curso Bacharelado em
Ciência da Computação do Centro de
Engenharia Elétrica e Informática da
Universidade Federal de Campina
Grande, como requisito parcial para
obtenção do título de Bacharel em
Ciência da Computação.**

BANCA EXAMINADORA:

**João Arthur Brunet Monteiro
Orientador – UASC/CEEI/UFCG**

**Patrícia Duarte de Lima Machado
Examinador – UASC/CEEI/UFCG**

**Francisco Vilar Brasileiro
Professor da Disciplina TCC – UASC/CEEI/UFCG**

Trabalho aprovado em: 15 de MARÇO de 2024.

CAMPINA GRANDE - PB

RESUMO

Nos últimos anos, as apostas esportivas têm experimentado um crescimento rápido devido à digitalização e à acessibilidade das plataformas online. Contudo, esse aumento também gerou desafios, como a integridade do mercado, o combate a práticas de crime organizado, como lavagem de dinheiro, e a proteção dos consumidores. Nesse cenário, os processos de verificação e validação de documentos, conhecidos como KYC (Know Your Customer - Conheça seu Cliente), tornaram-se essenciais para garantir a conformidade regulatória e a segurança dos usuários.

Este estudo se concentra nos desafios enfrentados pelas casas de apostas ao implementar os processos de KYC em conformidade com as regulamentações, especialmente na necessidade de análise manual para validar os processos de verificação de KYC. Essas verificações manuais são tipicamente demoradas e custosas. Diante desse cenário, foi desenvolvida uma ferramenta de automação que utiliza um modelo LLM (Large Language Model) para reduzir a dependência de intervenções humanas e aprimora a eficiência no processo de verificação de documentos de KYC. A metodologia utilizada abrangeu uma amostra de 163 dados de usuários do período de 26 de Dezembro de 2023 a 26 de Fevereiro de 2024, levando em consideração as principais falhas identificadas durante esse intervalo de tempo.

Os resultados obtidos mostram uma melhoria significativa na performance das verificações de análises manuais. Dessa forma, da amostra realizada com 163 casos o algoritmo forneceu um veredito para 73.6% casos, com precisão de 81,05% na categorização correta do status da análise do usuário. Isso destaca a eficácia geral do algoritmo, tornando o processo mais ágil e eficiente, reduzindo significativamente a carga de trabalho humana e, em contra partida, 26.4% dos casos ainda exigiram análise manual para maior precisão, evidenciando situações em que o algoritmo não conseguiu tomar uma decisão automática.

Contudo, a pesquisa não apenas oferece uma solução prática e tecnológica para os desafios enfrentados pelas casas de apostas, mas também contribui para o avanço na automação de processos de verificação de documentos em ambientes regulamentados.

AUTOMATION OF MANUAL CHECKS IN KYC PROCESSES

ABSTRACT

In recent years, sports betting has experienced rapid growth due to the digitalization and accessibility of online platforms. However, this increase also generated challenges, such as market integrity, combating organized crime practices, such as money laundering, and consumer protection. In this scenario, document verification and validation processes, known as KYC (Know Your Customer), have become essential to ensure regulatory compliance and user safety.

This study focuses on the challenges faced by bookmakers when implementing KYC processes in compliance with regulations, especially the need for manual analysis to validate KYC verification processes. These manual checks are typically time-consuming and costly. Given this scenario, an automation tool was developed that uses an LLM (Large Language Model) model to reduce dependence on human interventions and improve efficiency in the KYC document verification process. The methodology used covered a sample of 163 user data from the period from December 26, 2023 to February 26, 2024, taking into account the main flaws identified during this time interval.

The results obtained show a significant improvement in the performance of manual analysis checks. Thus, from the sample made with 163 cases, the algorithm provided a verdict for 73.6% cases, with an accuracy of 81.05% in correctly categorizing the status of the user's analysis. This highlights the overall effectiveness of the algorithm, making the process more agile and efficient, significantly reducing the human workload and, on the other hand, 26.4% of cases still required manual analysis for greater precision, highlighting situations in which the algorithm was unable to take an automatic decision.

However, the research not only offers a practical and technological solution to the challenges faced by bookmakers, but also contributes to advancing the automation of document verification processes in regulated environments.

AUTOMAÇÃO DE VERIFICAÇÕES MANUAIS EM PROCESSOS DE KYC

Matheus Henrique Guedes de Oliveira
Universidade Federal de Campina Grande Campina
Grande, Paraíba
matheus.oliveira@ccc.ufcg.edu.br

João Arthur Brunet Monteiro
Universidade Federal de Campina Grande Campina
Grande, Paraíba
joao.arthur@computacao.ufcg.edu.br

RESUMO

Nos últimos anos, as apostas esportivas têm experimentado um crescimento rápido devido à digitalização e à acessibilidade das plataformas online. Contudo, esse aumento também gerou desafios, como a integridade do mercado, o combate a práticas de crime organizado, como lavagem de dinheiro, e a proteção dos consumidores. Nesse cenário, os processos de verificação e validação de documentos, conhecidos como KYC (Know Your Customer - Conheça seu Cliente), tornaram-se essenciais para garantir a conformidade regulatória e a segurança dos usuários.

Este estudo se concentra nos desafios enfrentados pelas casas de apostas ao implementar os processos de KYC em conformidade com as regulamentações, especialmente na necessidade de análise manual para validar os processos de verificação de KYC. Essas verificações manuais são tipicamente demoradas e custosas. Diante desse cenário, foi desenvolvida uma ferramenta de automação que utiliza um modelo LLM (Large Language Model) para reduzir a dependência de intervenções humanas e aprimora a eficiência no processo de verificação de documentos de KYC. A metodologia utilizada abrangeu uma amostra de 163 dados de usuários do período de 26 de Dezembro de 2023 a 26 de Fevereiro de 2024, levando em consideração as principais falhas identificadas durante esse intervalo de tempo.

Os resultados obtidos mostram uma melhoria significativa na performance das verificações de análises manuais. Dessa forma, da amostra realizada com 163 casos o algoritmo forneceu um veredito para 73.6% casos, com precisão de 81,05% na categorização correta do status da análise do usuário. Isso destaca a eficácia geral do algoritmo, tornando o processo mais ágil e eficiente, reduzindo significativamente a carga de trabalho humana e, em contra partida, 26.4% dos casos ainda exigiram análise manual para maior precisão, evidenciando situações em que o algoritmo não conseguiu tomar uma decisão automática.

Contudo, a pesquisa não apenas oferece uma solução prática e tecnológica para os desafios enfrentados pelas casas de apostas, mas também contribui para o avanço na automação de processos de verificação de documentos em ambientes regulamentados.

REPOSITÓRIOS

<https://github.com/MatheusHG/kyc-automation>

1. INTRODUÇÃO

As apostas esportivas movimentam muito dinheiro e têm regulamentações diferentes em todo o mundo. Isso torna as apostas esportivas um alvo atraente para a lavagem de dinheiro - esconder a origem ilegal do dinheiro para que pareça legal. Essa prática inclui:

- Abrir uma conta de apostas e depositar dinheiro sujo, fazendo parecer que o dinheiro veio de apostas legais. Esse depósito pode inclusive ser feito de forma anônima via boleto.
- Fazer depósitos pequenos em uma conta de apostas sem fazer apostas, fechar a conta e receber o dinheiro de volta de uma forma aparentemente legal.
- Fazer apostas físicas em resultados óbvios para esconder a origem do dinheiro.
- Comprar recibos de bilhetes vencedores e perdedores para provar que o dinheiro veio de apostas esportivas.
- Algumas organizações criminosas até mesmo criam suas próprias casas de apostas online. [1]

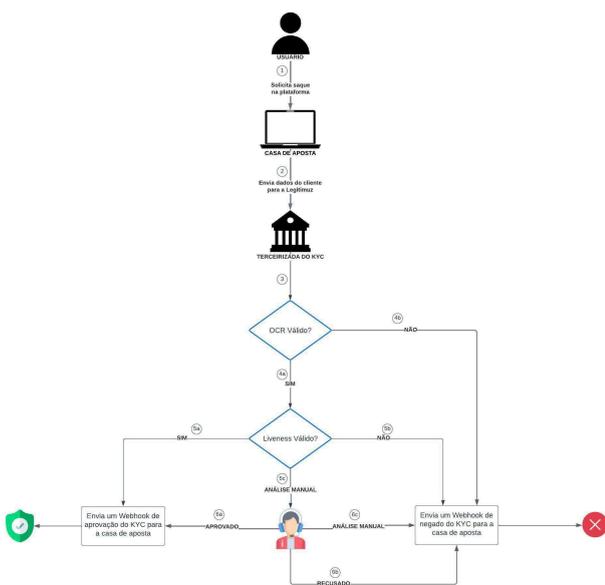
A lavagem de dinheiro e outros tipos de crimes financeiros não só têm um impacto enorme em casas de apostas, mas também na economia em geral. No entanto, no Brasil, o Projeto de Lei 3.626/2023 que rege a Regulamentação de Casas de Apostas, define que as casas de apostas deverão adotar procedimentos de identificação que permitam verificar a validade da identidade dos apostadores, exigindo a utilização da tecnologia de identificação e reconhecimento facial [2].

A implementação dos regulamentos de KYC teve como objetivo principal lidar com os desafios relacionados à lavagem de dinheiro, fraudes e outros crimes financeiros. No entanto, essa iniciativa trouxe consigo a necessidade de estabelecer procedimentos eficazes, enfrentando um dos maiores desafios na análise de documentos: as etapas de OCR e Facematch. Problemas como análises mal sucedidas ou a presença de documentos antigos, rasurados ou capturados por câmeras de baixa qualidade podem dificultar a verificação e resultar na necessidade de uma análise manual, deixando o cliente em espera e correndo o risco de perdê-lo devido à demora no processo.

No cenário da MarjoSports, uma casa de apostas no Brasil que é o estudo de caso deste projeto de TCC, a análise manual de documentos torna-se demorada devido ao elevado número de clientes sujeitos a essa verificação e à limitação de funcionários para lidar com esses casos, o que representa um custo significativo. É importante ressaltar que a contratação de mais funcionários pode não ser uma opção viável devido ao alto

custo associado à hora de trabalho de um humano. Nesse contexto, em um levantamento feito pelo autor deste trabalho entre o período de 26 de Dezembro de 2023 e 26 de Fevereiro de 2024, foram identificadas, em média, cerca de 29,23 validações manuais por dia, consumindo aproximadamente 3 minutos de trabalho humano cada. Uma análise mais aprofundada revelou que 72% dessas verificações poderiam ser automatizadas com soluções de automação, devido à natureza repetitiva e padronizada dessas tarefas, resultando em uma otimização significativa do processo de verificação de documentos.

No fluxo atual da casa de apostas, o usuário realiza a validação facial mediante a solicitação de saque ou cadastro, conforme as regras de negócio impostas pela casa de apostas. Em seguida, ele passa pelo processo de OCR para verificar o documento e pelo facematch para confirmar se o rosto corresponde ao documento apresentado. Se aprovado, o usuário pode concluir sua solicitação. Por outro lado, se reprovado, o fluxo é interrompido. Quando há necessidade de análise manual, um funcionário responsável pelo setor de atendimento avalia cada caso individualmente na dashboard, decidindo manualmente se o usuário será aprovado ou reprovado, como ilustrado na figura 1, destacando a necessidade e o potencial para a automação nessa área específica, devido à atividade repetitiva e padronizada das tarefas envolvidas.



[Figura 1 - Fluxo do KYC na casa de aposta]

Diante da necessidade de aprimorar os procedimentos de verificação de documentos em casas de apostas e reduzir os custos associados à análise manual devido ao alto volume de dados, a MarjoSports adquiriu um serviço terceirizado para realizar as verificações de KYC no momento da solicitação de saque na plataforma. Para solucionar as principais causas que levam à espera por análises manuais foi realizado um estudo e desenvolvido um algoritmo de automação, fruto deste trabalho. Onde o algoritmo identifica automaticamente clientes em processo de análise manual e conduz uma reanálise das informações do documento com o uso do modelo LLM (Large Language Model) Gemini do Google para garantir a precisão, consistência dos dados e otimização do processo, reduzindo assim a dependência de análises manuais. Dessa forma, a metodologia incluiu uma amostra de 163 dados de usuários do

período de 26 de dezembro de 2023 a 26 de fevereiro de 2024, considerando as principais falhas identificadas durante esse intervalo. O objetivo é melhorar a eficiência, a precisão e reduzir os custos operacionais associados às verificações de KYC em ambientes regulamentados.

Os resultados demonstram uma melhoria significativa na precisão das verificações manuais, com 57,6% dos usuários sendo aprovados e 16% reprovados pelo algoritmo. Dentre estes, o modelo acertou 62,5% das reprovações necessárias e 99,6% das aprovações corretas. Isso reflete um processo mais otimizado e menos demorado, reduzindo a carga de trabalho humano necessária. No entanto, 26,4% dos casos ainda exigiram análise manual para verificação mais precisa, destacando situações em que o algoritmo não conseguiu definir a solução de forma automática.

A avaliação detalhada do desempenho do algoritmo para extração e validação do CPF, nome completo e data de nascimento mostrou eficácia de 81,3%, 77,5% e 60,7%, respectivamente. Esses resultados evidenciam a eficácia do modelo LLM na automação das verificações de KYC, embora ainda existam desafios a serem superados. A análise completa dos resultados e insights adicionais será apresentada nas seções subsequentes.

Este trabalho está organizado em seis seções. A primeira contextualiza o problema a ser resolvido, enquanto a segunda discute as funcionalidades essenciais e as decisões arquiteturais. A terceira aborda o processo de desenvolvimento do sistema e os desafios enfrentados. A quarta trata da avaliação do sistema por meio de uma análise histórica dos dados submetidos às análises manuais. A quinta apresenta uma conclusão sobre o trabalho realizado, e a sexta oferece sugestões para possíveis direções futuras da aplicação.

2. SOLUÇÃO

Após fazer um levantamento das principais causas pelos quais os usuários iriam para a análise manual. A solução proposta é um algoritmo de automação que refaz a validação de OCR, extraíndo os principais dados presentes em todos os documentos enviados pelos usuários e faz a análise das informações extraídas pelo serviço terceirizado de KYC e pelo algoritmo de otimização, verificando ao máximo de informações presentes, como ilustrado no figura 2.



[Figura 2 - Estrutura da automação]

2.1 Verificações

Como método de checagem das informações presentes no documento, é extraída verificações que validam a existência das informações presentes no documento do usuário.

2.1.1 CPF

É extraída informações do CPF e filtrada para deixar o CPF no padrão de apenas dígitos e analisadas cautelosamente passando por histórico do CPF e verificando se é um CPF válido, verificando se corresponde ao CPF cadastrado na plataforma de casa de aposta e caso não seja possível extrair com precisão o CPF, é seguido se as demais informações, como: Nome completo e data de nascimento do usuário correspondem com as informações do documento.

2.1.2 Nome

Para as verificações de Nome, muitas das causas, acontecem pelo modelo da empresa terceirizada não fazer a correspondência de nomes que possuem abreviações ou não possuem a última parte no nome presente no documento. Com isso, a solução proposta para esse caso, além de refazer a análise para extração das informações no documento é verificar levando em consideração todos esses casos e fazer um matching entre as informações do documento e do usuário.

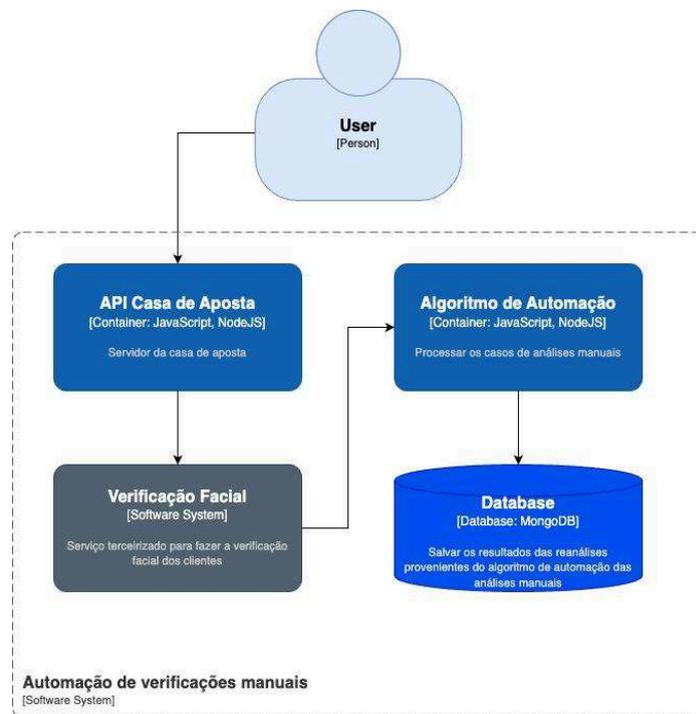
2.1.3 Data de Nascimento

Para verificações de data de nascimento, muitas das causas de análises manuais, acontecem pelo fato do mês estar presente no documento representado por letras e não com números, onde o modelo da empresa terceirizada não era capaz de distinguir essas informações, além de não extrair corretamente números das imagens, fazendo com que fosse para análise manual. Então, a solução para proposta foi preparar o algoritmo para lidar com todo esse tipo de caso e levar em consideração fatores

que levariam a análise manual, como apresentado anteriormente.

2.2 Arquitetura

Nesta seção, é fornecida uma visão geral sobre a arquitetura da solução, apresentando não apenas a estrutura desenvolvida, mas também as decisões tomadas durante o processo de desenvolvimento.



[Figura 3 - Diagrama arquitetural da automação de verificações manuais]

2.2.1 Gemini

Seguindo a arquitetura da aplicação, o algoritmo de automação presente na figura 3, utiliza o Gemini, modelo multimodal do Google DeepMind [3], em sua estrutura no qual desempenha um papel essencial para o funcionamento do algoritmo. Esse modelo tem um propósito fundamental: extrair informações de imagem e converter em JSON (Javascript Object Notation), tais imagens são oriundas de documentos enviados pelos clientes e que estão pendentes de análises manuais.

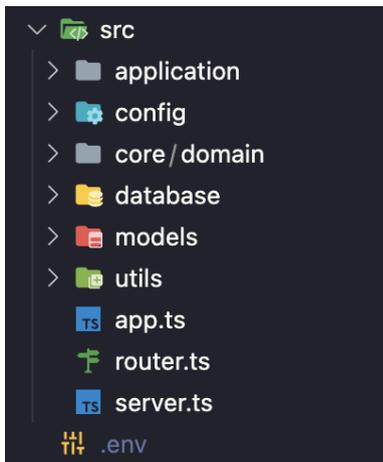


[Figura 4 - Visão abstrata da arquitetura do Gemini]

A configuração desse modelo foi realizada na linguagem de programação JavaScript [4], escolha motivada pela predominância dessa mesma linguagem no código fonte das principais casas de apostas, além de ser uma linguagem base para dezenas de frameworks com alta popularidade e adesão na comunidade de desenvolvimento [5].

2.2.2 Backend

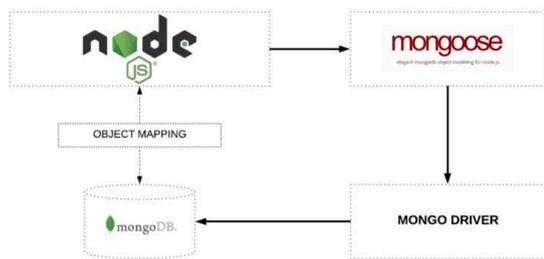
A estrutura no backend foi feita de acordo com as necessidades do sistema, sendo dividido em 4 diretórios principais, apresentados na Figura 5: application, config, core, database domain/entities, models e utils.



[Figura 5 - Estrutura de pastas do algoritmo de verificação automatizada]

A pasta Application é responsável por conter a lógica relacionada à aplicação, como a definição de rotas e as regras de negócio específicas da aplicação, composta de dois arquivos: Controllers e usecases. Os controllers são responsáveis por mapear as rotas da API (como GET, POST, PUT, DELETE), por sua vez o usecases ficam as regras de negócio da aplicação, ou seja, as funcionalidades principais do sistema que serão executadas, além de ser utilizada pelos controllers para realizar operações específicas de acordo com as requisições recebidas. A pasta config ficam as configurações de setup do projeto para garantir sua execução correta. Por sua vez, a pasta core inclui uma classe abstrata das Entidades, fornecendo uma base sólida para a estrutura do sistema.

A pasta database contém o arquivo de configuração e acesso ao MongoDB, o banco de dados utilizado no projeto. Nesse contexto, foi utilizado o mongoose, uma biblioteca de Modelagem de Dados de Objeto para MongoDB e Node.js, que é fundamental para gerenciar o relacionamento entre dados, fornecer validação de esquemas e atuar como tradutor entre objetos no código [10]. Um exemplo de como essa estrutura funciona pode ser visualizado na figura 6.



[Figura 6 - Exemplificação de como o NodeJS trabalha em conjunto com o Mongo]

A pasta models são onde estão os modelos da aplicação, responsáveis por definir a estrutura dos objetos que serão salvos no MongoDB. Por fim, a pasta utils contém arquivos auxiliares para formatação de strings e outras operações utilitárias.

2.2.3 Banco de dados

O banco de dados NoSQL, representado na figura 3, foi utilizado o MongoDB [9], em virtude de ser mais escalonáveis e oferecem desempenho superior. Além disso, a flexibilidade e facilidade de uso dos seus modelos de dados podem acelerar o desenvolvimento em comparação ao modelo relacional, especialmente no ambiente de computação em nuvem. Com isso, os resultados obtidos através do algoritmo de verificação automatizada ficam armazenadas em virtude de não perder os resultados.

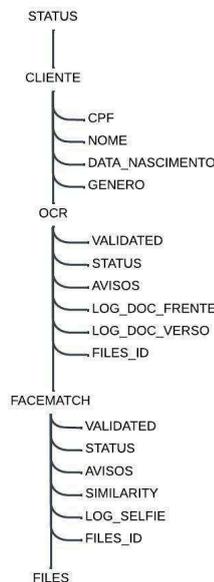
3. PROCESSO DE DESENVOLVIMENTO

3.1 Investigação das principais causas de análise manual

Em um primeiro momento, foi feito um estudo mais aprofundado na plataforma que terceiriza o processo de KYC para a MarjoSports, buscando informações sobre como funciona a mecânica e o fluxo das etapas de verificação. Onde foi selecionado um período de 26 Dezembro de 2023 até 26 Fevereiro de 2024, para análise dos casos que foram julgados como "análise manual" na plataforma, para serem analisados os dados extraídos e a causa da análise manual.

3.2 Extração das informações

Após uma minuciosa análise na primeira etapa e entendendo a estrutura básica gerada nos relatórios dos casos julgados como "análise manual" no formato JSON, ilustrado na figura 7, foi desenvolvido um script fundamentalmente em JavaScript, tendo em vista que o algoritmo de automação para verificação manual é escrita nessa mesma linguagem, para separar e coletar os principais os principais motivos que resultaram em análise manual para o levantamento do estudo.



[Figura 7 - Estrutura do JSON extraído]

3.3 Definição de requisitos

Com base na extração dos dados resultante da fase anterior, combinado com uma análise detalhada das sugestões provenientes dos funcionários de casas de apostas, responsável

pelo setor de atendimento que lidam com as verificações dos casos de clientes pendentes de uma análise manual, foi elaborada uma documentação que descreve os comportamentos e funcionalidades que o algoritmo deverá incorporar. Esse processo foi conduzido com o intuito de estabelecer os requisitos essenciais, os quais servirão como base para captar e comunicar as necessidades, priorizar e alocar tarefas, gerenciar as expectativas, bem como validar o que será possível realizar com base no objeto JSON gerado pela etapa do OCR, além de fazer o gerenciamento do escopo.

A tabela 1 detalha os requisitos identificados:

Requisito	Descrição
R01	Analisar as principais falhas encontradas pelo modelo terceirizado de KYC que resultaram em análise manual.
R02	Realizar a contagem de recorrências das falhas encontradas pela etapa do KYC.
R03	Reanalisar com o modelo Gemini os documentos dos clientes que foram para a análise manual pela empresa terceirizada.
R04	Verificar se as informações extraídas do modelo Gemini e da empresa terceirizada, correspondem com o perfil do usuário cadastrado na plataforma, que deseja completar a etapa exigida do KYC.

[Tabela 1: Requisitos do algoritmo de verificação automatizada de análises manuais]

3.4 Implementação do backend

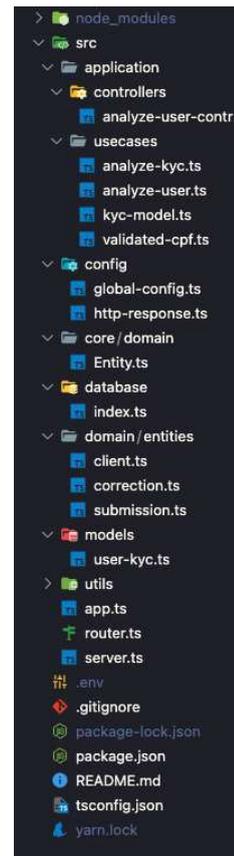
A implementação do backend, no qual a estrutura é mostrada na figura 8, foi desenvolvido usando NodeJS com JavaScript, com a utilização de frameworks para auxiliar no processo de desenvolvimento. Essa escolha foi impulsionada pela busca por um desempenho otimizado e escalável. A abordagem com essa linguagem proporciona uma vasta gama de bibliotecas e ferramentas disponíveis, o que agiliza o desenvolvimento e facilita a manutenção do sistema ao longo do tempo. Além disso, o NodeJS permite uma programação assíncrona que melhora a eficiência na manipulação de solicitações e respostas, tornando-o ideal para aplicativos que exigem alta performance e capacidade de escala.

Para processar os dados exportados dos usuários que passaram por análise manual de forma escalável e eficiente, foi utilizado uma planilha em conjunto com os módulos Stream [6] e csv [7]. Essa abordagem permite lidar com um grande volume de dados de maneira ágil e eficaz, garantindo a integridade e a precisão das informações durante todo o processo de análise e automação.

O Stream é uma interface abstrata para trabalhar com dados de streaming em Node.js e que conseguem lidar com leitura/escrita de arquivos, comunicações em rede, ou qualquer tipo de troca de informação end-to-end de uma forma eficiente. Sua utilização junto ao framework csv, no contexto do algoritmo de otimização de verificações de análises manuais possibilita a leitura pedaço por pedaço, processando o conteúdo sem mantê-lo completo na memória, além da eficiência de memória,

no qual não precisa carregar grandes quantidades de dados em memória antes de ser capaz de processá-los, como também a eficiência temporal, no qual menos tempo é requerido para começar a processar dados, uma vez que você pode começar a processar assim que os tiver, em vez de esperar até que toda carga útil de dados esteja disponível [8]. Dessa forma, ao receber a resposta da reanálise da etapa do OCR, um webhook é acionado para comunicar os resultados ao sistema da casa de apostas, finalizando o processo de análise.

Dessa forma, para os casos nos quais o modelo da Gemini não conseguia ler o documento ou concluir a etapa, os casos se mantinham em análise manual para uma análise de um humano.



[Figura 8 - Estrutura de pastas do algoritmo de verificação automatizada]

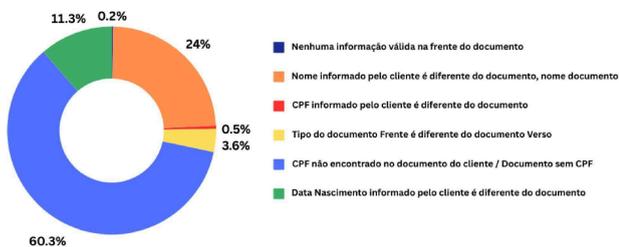
4. AVALIAÇÃO

Nesta seção, foram analisados diversos aspectos cruciais relacionados ao desempenho e a receptividade do algoritmo de automatização de verificações manuais entre os potenciais funcionários de casas de apostas, responsável pelo setor de atendimento. Avaliou-se como os usuários se sentem atualmente, sem o uso do algoritmo de otimização, com o grande volume de dados para serem analisados manualmente nas etapas do KYC, quando estes não são aprovados/reprovados. Ademais, foi avaliado a expectativa e praticidade que o algoritmo poderá oferecer na otimização do tempo de trabalho visando atender as necessidades do usuário. Por fim, também foi avaliado o quão eficiente foi o algoritmo em relação aos dados do período que ficaram como "análise manual" na verificação do KYC.

4.1 Metodologia

Foram analisados 441 casos de usuários que ficaram com o status de "análise manual" entre 26 de dezembro de 2023 e 26 de fevereiro de 2024. Esses dados foram obtidos do painel de KYC fornecido pelo serviço terceirizado da empresa de apostas esportivas MarjoSports, com o objetivo de identificar as principais razões por trás dessas análises manuais. A partir desses dados, uma amostragem foi conduzida para aplicar o algoritmo desenvolvido nesta pesquisa, com o intuito de avaliar sua eficácia e os resultados obtidos.

Nesse aspecto, foi realizado um estudo para identificar as principais razões pelas quais os usuários ficassem em análise manual e os resultados obtidos foram os seguintes: CPF não encontrado no documento do cliente / Documento sem CPF, Nome informado pelo cliente é diferente do documento, Data Nascimento informado pelo cliente é diferente do documento, Tipo do documento Frente é diferente do documento Verso (DOCTYPE), Nenhuma informação válida na frente do documento, CPF informado pelo cliente é diferente do documento, como apresentado na figura 9.



[Figura 9 - Gráfico de rosca sobre os principais erros gerados pelo modelo terceirizado]

Dessa forma, foi disponibilizado um formulário, para avaliar o algoritmo de verificação manual aos atendentes da casa de aposta que tratam os casos de análises manuais, composta de seis perguntas, com todas perguntas sendo objetivas. Para a elaboração das perguntas, elas foram divididas em duas categorias.

A primeira categoria, composta de três perguntas, visa classificar o atual cenário do uso da ferramenta atual de processo de KYC, antes do algoritmo. Para isso foram feitas perguntas diretas sobre o quão eficiente é o processo de verificação de identidade, se ele recomendaria para outra casa de aposta e se é trabalhoso e desgastante o processo de verificações manuais.

A segunda categoria, composta de três perguntas, visa avaliar a qualidade e utilidade do algoritmo de automação dos processos manuais. Para tal objetivo, foram elaboradas as seguintes perguntas: uma como o atendente que faz as análises manuais atualmente avalia os gráficos que contém as estatísticas relacionadas ao relatório, como ele avalia a utilidade do algoritmo para o processo de automação e diminuição do tempo de trabalho e por fim, como ele avalia o quão esse algoritmo irá facilitar em seu dia a dia de trabalho.

4.2 Resultados

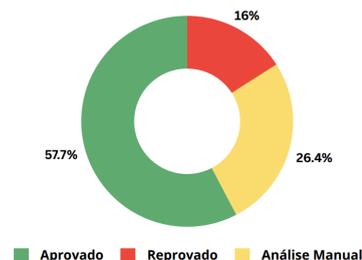
Nesta seção, abordaremos os resultados obtidos da automação da verificação de documentos pendentes de análises manuais em processos de KYC em casas de apostas, utilizando os métodos de análise propostos na seção 2. Com base em uma amostra de 163 dados de usuários do período de 26 de Dezembro de 2023 a 26 de Fevereiro de 2024, levando em

consideração as principais falhas identificadas durante esse período. Além disso, foi feita uma pesquisa para entender a eficácia e obter insights sobre a opinião dos usuários responsáveis pelas verificações manuais em casas de apostas.

O desempenho do algoritmo ilustrado na figura 10, demonstra uma melhoria significativa na performance das verificações de análises manuais, resultando na aprovação de 57,6% e na reprovação de 16% dos usuários. Dessa forma, em relação aos casos julgados como reprovados, o modelo acertou 62,5% das vezes e errou 37,5%. Quanto aos casos julgados como aprovados, o modelo acertou 99,6% e errou 0,4% das vezes. Ou seja, o modelo teve uma eficiência de 81,05% e ineficiência de 18,95%.

Isso torna o processo mais otimizado, menos demorado e reduz a carga de trabalho humano necessária. Porém, 26,4% dos casos ainda demandaram análise manual para uma verificação mais precisa, indicando situações em que o algoritmo não conseguiu definir a solução de forma automática. Diante desse cenário de desempenho, o número médio de análises manuais por dia, que era de 29,23 validações, diminuiu para 12,36 validações manuais por dia.

RESULTADO DO ALGORITMO NAS ANÁLISES MANUAIS
26 DEZEMBRO DE 2023 - 26 FEVEREIRO 2024



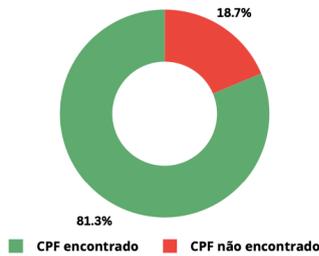
[Figura 10 - Gráfico de rosca sobre o resultado do uso do algoritmo]

Dessa forma, com base nesses resultados, podemos aprofundar nossa análise e examinar minuciosamente o desempenho de cada verificação realizada pelo algoritmo, a fim de analisar os resultados obtidos. Essa análise detalhada será apresentada nas seções a seguir.

4.2.1 CPF

Na avaliação do algoritmo para extração e validação do CPF do usuário em comparação com as informações cadastradas na casa de apostas, foi obtido um ótimo desempenho. A eficácia para validação do CPF foi de 81,3%, com 18,7% dos casos em que não foi possível validar/encontrar o CPF do usuário no documento. Dessa forma, houve uma redução significativa na necessidade de validar manualmente o CPF do usuário, como apresentado na figura 11.

RESULTADO DA VERIFICAÇÃO DO CPF
26 DEZEMBRO DE 2023 - 26 FEVEREIRO 2024

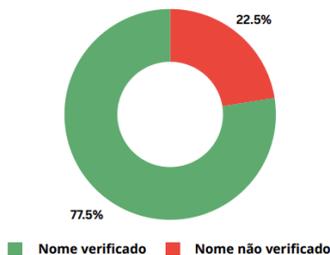


[Figura 11 - Gráfico de rosca sobre o resultado da verificação do CPF]

4.2.2 Nome

Os resultados obtidos com o uso do algoritmo para verificações de nome, demonstrada na figura 12, é evidente que o algoritmo manteve um excelente desempenho ao extrair as informações de nome completo presente no documento, mesmo considerando nomes com abreviações e imagens pouco nítidas. O resultado foi de 77,5% de verificações concluídas com sucesso e 22,5% de verificações em que não foi possível encontrar as informações.

RESULTADO DA VERIFICAÇÃO DO NOME
26 DEZEMBRO DE 2023 - 26 FEVEREIRO 2024

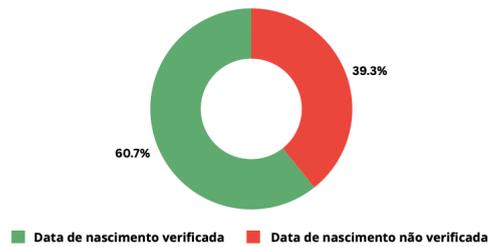


[Figura 12 - Gráfico de rosca sobre o resultado da verificação de nome completo]

4.2.3 Data de Nascimento

Ainda analisando o desempenho do algoritmo em relação a data de nascimento, como apresentado na figura 13, foi possível obter 60,7% de eficácia para extração das informações de data de nascimento presente no documento com a correspondência das informações do usuário e 39,3% para datas de nascimento que não tiveram um bom desempenho. Dessa forma, obtendo um desempenho um pouco abaixo em comparação às outras informações mencionadas e extraídas anteriormente.

RESULTADO DA VERIFICAÇÃO DA DATA DE NASCIMENTO
26 DEZEMBRO DE 2023 - 26 FEVEREIRO 2024



[Figura 13 - Gráfico de rosca sobre o resultado da verificação da data de nascimento]

4.2.4 Avaliações

As avaliações foram altamente positivas, destacando a eficácia e as perspectivas promissoras para o uso contínuo do algoritmo em futuras análises manuais. Os resultados refletem os dados obtidos por meio dos questionamentos realizados com 10 funcionários de casas de apostas, que forneceram insights valiosos. A pesquisa envolveu funcionários de casas de apostas, e ao serem questionados sobre a recomendação do atual processo de KYC para outra casa de aposta, 40% das pessoas recomendaria e 60% não recomendaria. Ademais, ao serem questionados sobre o trabalho atual de análise manuais se seria desgastante e trabalhoso, 10% responderam ser pouco, 10% nem tanto e 80% muito desgastante.

Ao serem questionados sobre o quão eficiente é o processo atual de verificação de identidade da sua casa de aposta, 60% o consideraram pouco eficiente, 40% eficiente e 0% extremamente eficiente. Já na segunda categoria de perguntas, ao serem questionados sobre como avaliavam o desempenho obtido pelo algoritmo no período de 26 de Dezembro de 2023 a 26 de Fevereiro de 2024 para a automação das análises manuais, 0% o consideraram pouco eficiente, 20% eficiente e 80% extremamente eficiente.

Ao serem questionados sobre qual a percepção sobre a otimização do tempo de trabalho com a implementação do algoritmo no dia a dia, 0% disseram ser pouco eficiente, 30% eficiente e 70% extremamente eficiente. Por fim, a última pergunta buscava avaliar quão útil o algoritmo será para o dia a dia de trabalho, e 0% disseram ser pouco eficiente, 20% eficiente e 80% extremamente eficiente.

Dessa forma, os resultados obtidos evidenciam a importância do algoritmo de automação para os casos em que os clientes estão sujeitos a análises manuais, destacando a agilidade do processo de verificação, redução do tempo de espera do cliente e diminuição da intervenção humana.

5. CONCLUSÃO

Durante este estudo, foram explorados de forma aprofundada os desafios enfrentados pelas casas de apostas na implementação de processos de KYC em conformidade com as regulamentações, bem como a solução por meio da automação de verificações manuais. A análise detalhada dos resultados obtidos com a implementação do algoritmo de automação baseado em um modelo LLM (Large Language Model) revelou uma significativa redução na dependência de análises manuais, resultando em uma otimização considerável nos processos de verificação de documentos.

Os dados coletados durante o período de pesquisa proporcionaram insights valiosos sobre a eficácia do algoritmo na validação de informações críticas, como CPF, nome completo e data de nascimento dos usuários. Portanto, conclui-se que a implementação da ferramenta de automação representa um avanço significativo no campo da automação de processos de verificação de documentos em ambientes regulamentados. Este estudo não apenas oferece uma solução eficiente para os desafios enfrentados pelas casas de apostas, mas também destaca o potencial da tecnologia para melhorar a eficiência operacional e garantir a conformidade regulatória, contribuindo assim para um ambiente mais seguro e transparente para os usuários.

6. TRABALHOS FUTUROS

Este estudo abre caminho para investigar a eficácia de diferentes modelos LLM (Large Language Models) na otimização dos processos de verificação de documentos e na redução ainda mais significativa das análises manuais. Algumas possíveis abordagens para trabalhos futuros incluem:

6.1 Avaliação de outros modelos LLM

Comparar o desempenho do modelo LLM utilizado neste estudo com outros modelos LLM disponíveis no mercado, como GPT-4 ou T5, para determinar qual oferece a melhor eficiência na automação das verificações manuais.

6.2 Treinamento adicional do modelo LLM

Investigar a possibilidade de realizar treinamento adicional do modelo LLM para aprimorar sua capacidade de reconhecimento e validação de informações críticas nos documentos de KYC.

6.3 Avaliação de métricas adicionais

Além das métricas tradicionais de eficácia, como a taxa de aprovação e reprovação, considerar a avaliação de métricas adicionais, como o tempo médio de processamento, para uma análise mais abrangente do desempenho do modelo LLM.

6.4 Utilização do algoritmo de automação para análises manuais em tempo real

Explorar a aplicação do algoritmo na casa de aposta para processar em tempo real os processos de análises manuais, visando explorar a capacidade de processar um grande volume de dados e fornecer respostas aos usuários de forma instantânea. Com o objetivo de assegurar que o resultado da análise manual seja processado e entregue ao usuário em tempo real, visando uma resposta rápida e eficiente.

Esses trabalhos futuros têm o potencial de aprimorar ainda mais a automação dos processos de verificação de documentos em casas de apostas, contribuindo para uma redução significativa das análises manuais e para a eficiência operacional dessas organizações.

7. AGRADECIMENTOS

Gostaria de expressar meus sinceros agradecimentos a todos que contribuíram para a realização deste trabalho. Em primeiro lugar, agradeço a Deus por Sua orientação e capacitação ao longo de toda a minha jornada acadêmica. Sem Ele, nada disso seria possível.

À minha família, em especial aos meus pais, Francyneide e Manoel, dedico um profundo agradecimento por seu apoio incondicional e por terem trabalhado incansavelmente para me proporcionar educação e oportunidades de crescimento.

Aos meus amigos e familiares, em especial à minha irmã, Brunna, agradeço por todo o incentivo, apoio emocional e encorajamento ao longo desta jornada acadêmica.

Ao professor e orientador, João Arthur Brunet, que sempre foi minha referência desde antes de ingressar na UFCG, expresso minha profunda gratidão por sua orientação, dedicação e paciência ao longo de todo o processo de desenvolvimento deste trabalho. Seus ensinamentos foram fundamentais e contribuíram significativamente para o meu aprendizado.

Por último, mas não menos importante, agradeço ao curso de Ciência da Computação da UFCG por proporcionar um ambiente de aprendizado contínuo, desafiador e enriquecedor, que contribuiu diretamente para o meu crescimento profissional e pessoal.

8. REFERÊNCIAS BIBLIOGRÁFICAS

- [1] SANTOS, Oseias. A Lavagem de Dinheiro no Mundo das Apostas Esportivas. Disponível em: <<https://www.jusbrasil.com.br/artigos/a-lavagem-de-dinheiro-no-mundo-das-apostas-esportivas/2001496757>>. Acesso em: 07 mar. 2024.
- [2] VINHAL, Gabriela. Senado aprova taxaço de apostas esportivas online; cassinos ficam fora. Disponível em: <<https://economia.uol.com.br/noticias/redacao/2023/12/12/senado-aprova-texto-base-de-projeto-que-taxa-mercado-de-apostas-esportivas.htm>>. Acesso em: 02 fev. 2023.
- [3] Visão geral dos modelos multimodais | Google Cloud. Disponível em: <<https://cloud.google.com/vertex-ai/docs/generative-ai/multimodal/overview?hl=pt-br>>. Acesso em: 02 mar. 2024.
- [4] JavaScript. Learn JavaScript Online. Disponível em: <<https://www.javascript.com/>>. Acesso em: 01 abr. 2024.
- [5] BESSA, André. Node.JS: o que é, como funciona esse ambiente de execução JavaScript e um Guia para iniciar. Disponível em: <<https://www.alura.com.br/artigos/node-js>>. Acesso em: 28 mar. 2024.
- [6] Node.js v21.7.2 documentation. Stream. Disponível em: <<https://nodejs.org/api/stream.html>>. Acesso em: 04 abr. 2024.
- [7] NPM. CSV for Node.js and the web. Acesso em: <<https://www.npmjs.com/package/csv>>. Acesso em: 25 mar. 2024.
- [8] NodeJS para iniciantes | Reativa. NodeJS Stream. Disponível em: <<https://nodejs.reativa.dev/0052-nodejs-streams/index>>. Acesso em: 08 abr. 2024.
- [9] MongoDB. The Developer Data Platform | MongoDB. Disponível em: <<https://www.mongodb.com/pt-br>>. Acesso em: 24 mar. 2024.
- [10] KARNIK, Nick. Introdução ao Mongoose para MongoDB. Disponível em: <<https://www.freecodecamp.org/portuguese/news/introducao-ao-mongoose-para-mongodb>>. Acesso em: 09 abr. 2024.