Trabalho prático 1: ETL

Nuno Filipe Fernandes de Castro

Nº a4944 – Regime Pós-laboral

Ano letivo 2024/2025

Licenciatura em Engenharia de Sistemas Informáticos

Escola Superior de Tecnologia

Instituto Politécnico do Cávado e do Ave

Identificação do Aluno

Nuno Filipe Fernandes de Castro

Aluno número a4944, regime pós-laboral

Licenciatura em Engenharia de Sistemas Informático

RESUMO

Observando o método atualmente utilizado para o tratamento de dados, por parte de uma entidade que realiza rastreios frequentemente, cuja prática envolve a extração manual de informações a partir de ficheiros Excel sempre que há necessidade de consulta ou extração de dados para ações subsequentes. Identificou-se que o processo era suscetível a erros devido a inserção com falhas, resultando em problemas de qualidade e confiabilidade dos dados.

Para resolver estas limitações, foi desenvolvida uma abordagem que automatizou a verificação dos dados essenciais, garantindo que a informação estivesse no formato correto, especialmente em situações que envolviam o envio de mensagem e emails aos clientes. Além disso, os dados foram normalizados para serem integrados numa ferramenta de Power BI, facilitando a análise e visualização

Como resultado, as melhorias implementadas reduziram significativamente o tempo necessário para as tarefas de tratamento de dados associadas aos rastreios, além de simplificarem e aumentarem a precisão na análise e visualização das informações de forma mais eficaz.

ABSTRACT

Observing the current method used for data processing by an entity that frequently conducts screenings, whose practice involves manually extracting information from Excel files whenever there is a need for data consultation or extraction for subsequent actions. It was identified that the process was prone to errors due to faulty data entry, resulting in issues with data quality and reliability.

To address these limitations, an approach was developed to automate the verification of essential data, ensuring that information was in the correct format, particularly in situations involving the sending of messages and emails to clients. Additionally, the data was normalised to be integrated into a Power BI tool, facilitating analysis and visualisation.

As a result, the implemented improvements significantly reduced the time required for data processing tasks related to screenings, while also simplifying and enhancing the accuracy of data analysis and visualisation more effectively.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

Glossário

Business Intelligence – Conjunto de processos, tecnologias e ferramentas que transformam dados brutos em informações úteis para a tomada de decisões estratégicas.

Dashboard – Painel visual que apresenta dados de maneira consolidada, geralmente através de gráficos, tabelas e indicadores.

Excel – Software desenvolvido pela Microsoft, amplamente utilizado para manipulação e análise de dados.

Extract, Transform, Load – Refere-se ao processo de integração de dados que envolve a extração de informação de diversas fontes, a transformação desses dados e o carregamento dos mesmo num sistema de destino.

JavaScript – Linguagem de programação, amplamente utilizada para criar interatividade em páginas web.

Null – Valor especial utilizado em programação e bancos de dados para indicar a ausência de valor ou dado desconhecido.

Power BI – Ferramenta de business intelligence e visualização de dados desenvolvida pela Microsoft. Permite a conexão, transformação e visualização de dados de múltiplas fontes, criando relatórios interativos e dashboards dinâmicos.

String – Tipo de dado utilizado em programação para representar uma sequência de caracteres.

KNIME – Plataforma de análise de dados de código aberto que permite a criação de fluxos de trabalho visuais para extração, transformação e análise de dados.

Workflow – Conjunto de atividades e tarefas organizadas de forma sequencial ou paralela, com o objetivo de completar um processo especifico de forma eficiente e controlada.

Siglas e Acrónimos

BI – Business Intelligence

Cvs – formato de ficheiro de valores separados por vírgulas

xls – formato de ficheiro Excel até à versão Excel 2003

xlsx – formato atual padrão do Excel

SMS – Short Message Service (Serviço de mensagens curtas)

IPCA – Instituto Politécnico do Cávado e do Ave

2FA – Two-Factor Authenticator (Autenticação de Dois Fatores)

1. Introdução

O presente trabalho centra-se na utilização de processos ETL para melhorar o tratamento e integração de dados no contexto de rastreios comunitários de saúde. O ETL é uma metodologia amplamente utilizado por grandes organizações que lidam com grande volume de dados dispersos em múltiplos sistemas, permitindo consolidar essas informações para obter uma visão mais abrangente e uniforme do negócio. Ao realizar estas tarefas torna-se possível preparar e padronizar informações que são essenciais para a tomada de decisões informadas e estratégicas.

No contexto dos rastreios de saúde, o volume de dados gerados é pequeno em comparação com grandes corporações, mas, mesmo assim, requer um tratamento cuidadoso para garantir a integridade das informações. Utilizar um processo ETL facilita não apenas o acesso às informações, mas também a normalização, garantindo que a entidade tenha uma visão precisa e coesa dos dados gerados, melhorando a eficácia das análises e a execução de ações futuras.

Este trabalho foca na aplicação de uma abordagem ETL para normalização, correção e integridade dos dados gerados nos rastreios. A metodologia empregada foi essencial para reduzir a margem de erro resultante dos registos manuais e garantir que os dados fossem consistentes e prontos para a análise. A integração desses dados com o Power BI visou melhorar a visualização e análise das informações, permitindo extrair informação valiosa para auxiliar na tomada de decisão de futuras ações.

1.1. Objetivos

- Normalização, correção e filtragem de dados a partir de tabelas, realizadas no Excel, e formato .xls e .xlsx;
- Exportação de listas no formato .csv para posterior importação noutros sistemas;
- Integração de dados em Power BI para posterior análise;

1.2. Contexto

A realização de rastreios na comunidade, gera dados valiosos que podem ser analisados para melhor entender os perfis da população, permitindo a preparação de futuras ações de consciencialização direcionadas a perfis específicos e o planeamento de rastreios futuros focados em áreas específicas.

1.3. Estrutura do documento

Este relatório está divido em várias secções que cobrem cada fase do projeto, desde a identificação do problema até à sua resolução e avaliação dos resultados. A estrutura do documento é a seguinte:

1. Modelo Atual: esta secção apresenta uma análise detalhada do problema enfrentado, descrevendo o estado atual do tratamento dos dados de rastreios realizados pela entidade. São abordadas as limitações dos métodos utilizados, incluindo a dependência de processos manuais que envolvem a correção dos dados. Esta análise evidenciou a necessidade de melhoria na automação, principalmente devido aos erros de inserção manual.
2. Implementação: Esta secção descreve a solução desenvolvida usando a plataforma KNIME para automatizar o processo ETL, desde a entrada dos dados até à preparação para uso no Power BI. A implementação detalhada, a criação de fluxos para filtragem e exportação dos dados, envio automático de emails, e integração dos dados em Power BI. São igualmente justificadas as decisões para garantir um fluxo de dados eficiente e seguro.
3. Análise de resultados e testes: Nesta secção, é feita uma análise dos resultados obtidos com a implementação da solução ETL. É avaliado o impacto da automatização no tempo necessário para tratar os dados, bem como a eficácia na eliminação de erros anteriormente comuns São também discutidos os testes realizados para garantir a qualidade dos dados e da solução desenvolvida.
4. Conclusão: A conclusão resume as melhorias alcançadas, destacando as principais vantagens da automatização do processo ETL. A nova abordagem permitiu reduzir o tempo e esforço necessário para o tratamento de dados,

aumentando a eficiência e precisão dos dados. Foi também destacada a contribuição do projeto para a tomada de decisões mais informadas, demonstrando o valor do ETL como aliado na transformação de dados brutos em informações estratégicas.

2. Modelo atual

A recolha de dados durante os rastreios nem sempre ocorre nas condições ideais para o carregamento direto nas aplicações apropriadas. Dessa forma, torna-se necessário adaptar o processo de recolha aos meios disponíveis.

Muitas vezes, os dados são recolhidos manualmente e posteriormente inseridos numa tabela Excel. Alternativamente, quando há suporte informático disponível durante o evento, a inserção é realizada diretamente na tabela. Em eventos que decorrem ao longo de vários dias consecutivos, podem ser gerados vários ficheiros distintos.

Após a recolha, os dados são unificados, corrigidos, e os elementos necessários são extraídos para criar dois ficheiros específicos: uma lista de contactos e uma lista de emails. Estes ficheiros são utilizados para o envio de informações e agradecimentos pela participação nos rastreios. Além disso, os dados são analisados com o objetivo de apoiar a tomada de decisões sobre ações futuras.

Apesar da dificuldade em controlar as condições de recolha dos dados, o seu tratamento posterior é, frequentemente, um processo demorado, consumindo uma quantidade significativa de tempo do responsável por essas funções. A otimização deste processo é essencial para reduzir a carga de trabalho e melhorar a eficiência geral.

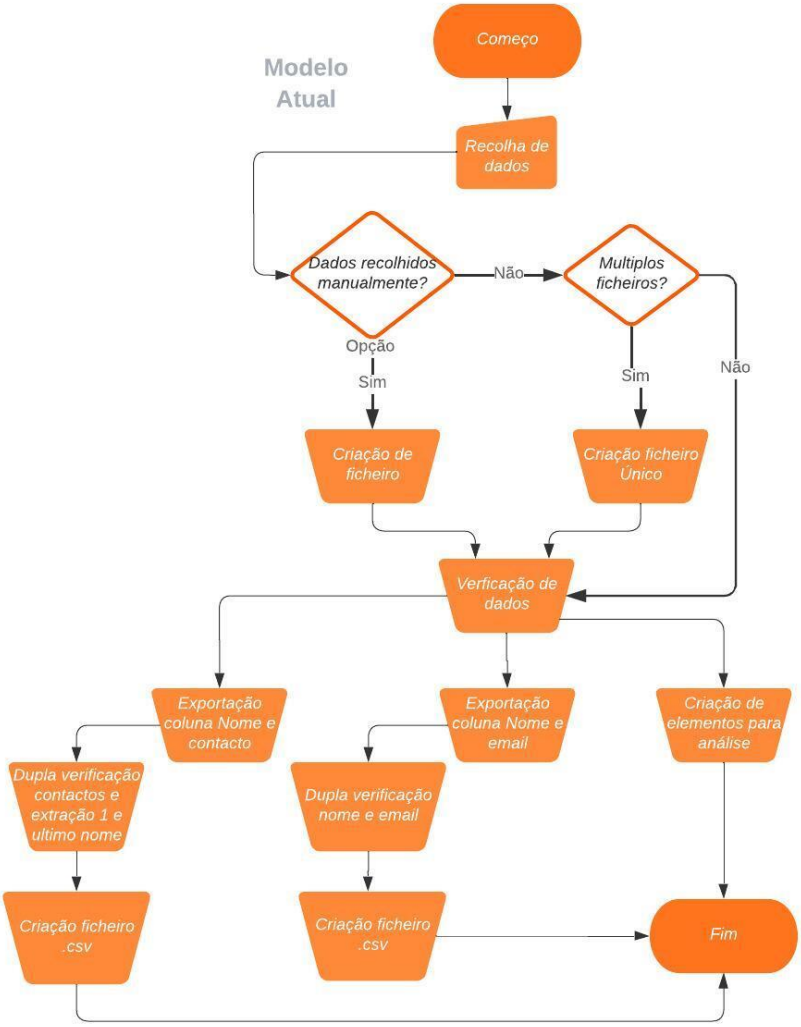

Figura 1 – Diagrama de fluxo do modelo atual

3. Implementação

Recorrendo ao software **KNIME**, uma plataforma de análise de dados, foi desenvolvida uma solução para automatizar todo o processo, desde a receção dos dados recolhidos até ao ponto em que estão prontos para serem utilizados por outras aplicações. A solução inclui a automatização de tarefas fundamentais, como o envio de emails.

Utilizando a linguagem **JavaScript**, adaptada para a aplicação, e fazendo uso de expressões regulares, foi realizado um trabalho completo de normalização, correção de erros, filtragem e exportação dos dados.

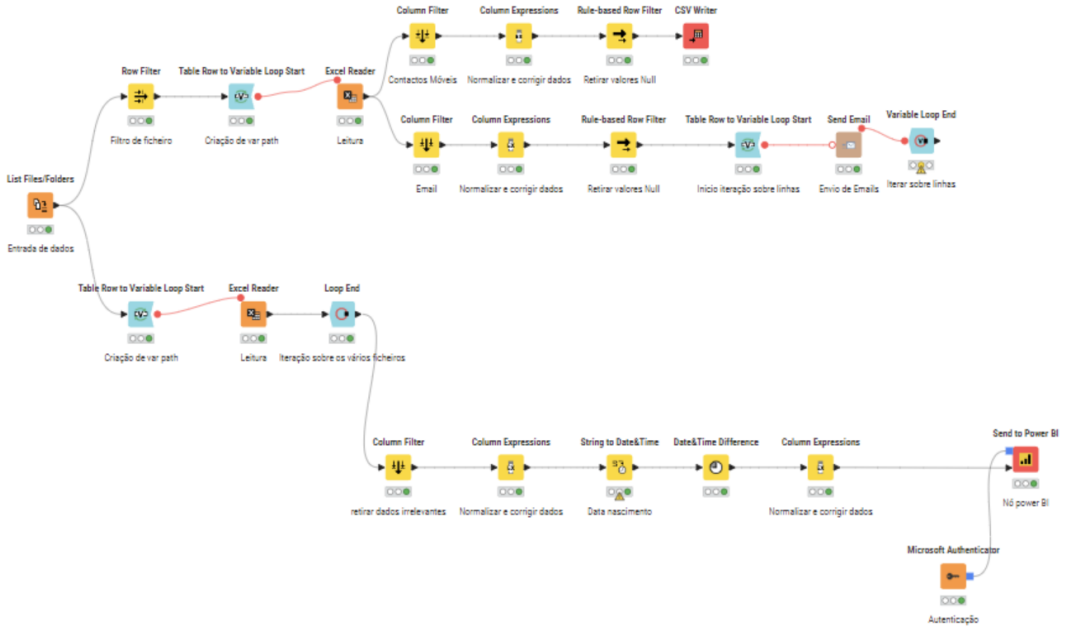

Figura 2 - Visão geral do workflow

3.1. Entrada de dados no fluxo

Como a forma adotada para o armazenamento dos dados são ficheiros **Excel,** foi criada uma pasta dentro do **workflow** da aplicação, denominada "entrada". O caminho desta pasta foi configurado de forma **relativa,** permitindo que o sistema funcione em diferentes dispositivos sem necessidade de ajustes adicionais. Esta pasta serve como um repositório

temporário onde os utilizadores podem depositar os ficheiros à medida que vão sendo recolhidos, para serem posteriormente tratados e integrados.

Na solução existem dois tipos principais de tratamento de dados:

1. **Tratamento individual dos dados de cada ficheiro:** Cada rastreio é tratado de forma isolada, o que levou á criação de um filtro específico para selecionar ficheiros. Assim é possível garantir que apenas os ficheiros relevantes sejam processados.

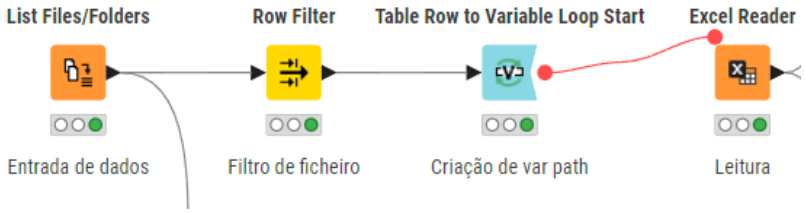

Figura 3 - Leitura de ficheiros com filtro por ficheiro

2. **Análise conjunta de todos os dados:** Os dados de todos os ficheiros são agregados para permitir uma análise global. Este processo é fundamental para obter uma visão abrangente dos rastreios realizados ao longo do tempo. Este tipo de análise possibilita a descoberta de informações demográficas, que são essenciais para a definição de estratégias futuras e o planeamento de ações direcionadas a áreas relevantes.

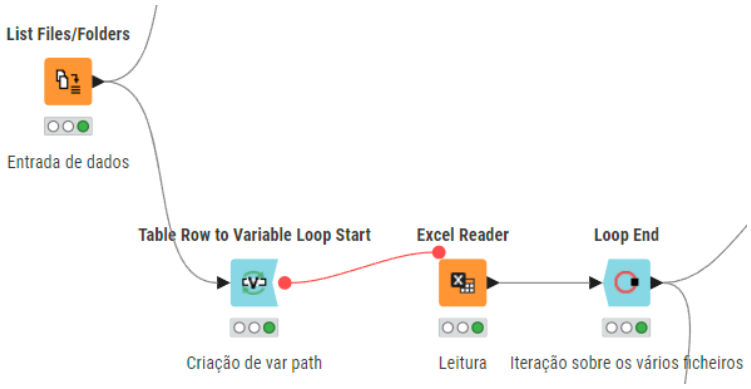

Figura 4 - Leitura de vários ficheiros em simultâneo

3.2. Fluxo para exportação .csv

Um dos requisitos é a capacidade de exportar um ficheiro .csv contendo duas colunas, a primeira com o primeiro e último nome de cada pessoa que participou no rastreio, e a segunda com o respetivo contacto associado. A opção por manter apenas o primeiro e o último nome na respetiva coluna, esta relacionada com a poupança de caracteres no envio dos SMS.

Para garantir que apenas os dados necessários sejam exportados e que a importação nos sistemas subsequentes ocorra sem problemas, foi necessário dividir o processo em várias etapas organizadas e automatizadas:

- **Filtragem das colunas:** Criado um filtro para excluir todas as colunas que desnecessárias, mantendo apenas as colunas referentes ao nome e contacto, evitando assim as informações redundantes.
- **Verificação dos números de telemóvel:** Uma vez que o objetivo final do ficheiro é o envio de SMS, e a aplicação atualmente utilizada para esse fim só suporta números móveis nacionais, foi necessário implementar um processo de rastreio e validação dos números de contacto. Assim, foi utilizada para uma **expressão regular** para garantir que o comprimento do número correspondesse a **nove dígitos** e que os dois primeiros dígitos estivessem de acordo com os códigos dos operadores móveis utilizados em Portugal (91,92,93,96). Os contactos que não correspondam são considerados inválidos.

```
    if (length(column("Contacto")) > 0 && regexMatcher(column("Contacto"),
"^(91|92|93|96)\\d{7}$"))
        {
            column("Contacto");
        } else {
            null;
        }
```

Tabela 1 - Seleção contactos

```
var primeiroNome = substr(column("Nome"), 0, indexOf(column("Nome"), " "));
var ultimoNome = regexReplace(column("Nome"), ".*\\s", "");
join([primeiroNome, " ", ultimoNome])
```

Tabela 2 - Seleção Primeiro e último nome

- **Normalização e correção de dados:** Adicionalmente, foram realizadas etapas de normalização de dados, para corrigir eventuais erros de formatação e garantir que os nomes e contactos seguissem um padrão coerente. Todas as linhas que se encontravam a **null** foram excluídas da exportação para o ficheiro.

```
MISSING $Nome$ => TRUE
MISSING $Contacto$ => TRUE
```

Tabela 3 - Exclusão campos Null

Figura 5 -Exportação csv

3.3. Fluxo para envio automático de email

A configuração do **envio de emails** segue etapas semelhantes às utilizadas anteriormente para o envio de SMS, diferenciando-se apenas nos filtros aplicados para garantir a validade dos dados e evitar falhas durante o processamento. Enquanto o fluxo de SMS filtra números de telemóvel inválidos, o fluxo de envio de emails aplica filtros específicos para garantir que apenas endereços de email válidos sejam considerados, minimizando assim de falhas no processo. que originam erros e leve o processo a parar sem ter

sido executado até ao final. Na fase de preparação dos dados, foram atribuídos valores **null** a todas as linhas que continham campos em branco ou que não correspondiam ao formato habitual de um endereço de email. Isto garante que endereços de email incompletos ou incorretos sejam automaticamente identificados e marcados como inválidos.

```
if (length(column("Email")) > 0 && regexMatcher(column("Email"), "^[^@\\s]+@[^@\\s]+\\.[^@\\s]+$"))
 {
    column("Email");
} else {
    null;
}
```

Tabela 4 - Exclusão emails inválidos

Após a preparação e filtragem dos dados, foi configurado um **loop** que percorre todas as linhas existentes para proceder ao envio de um **email de agradecimento** aos participantes dos rastreios. Para o envio dos emails, foi configurada a ligação a um servidor de email.

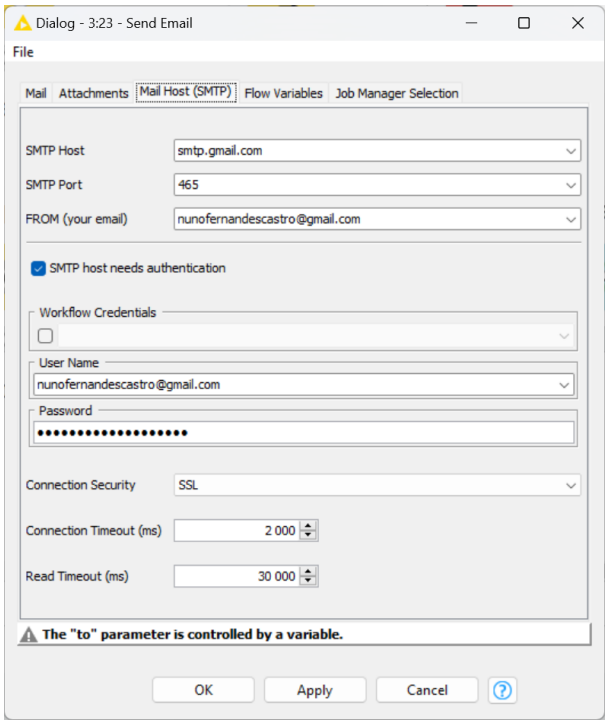

Figura 6 - Configuração servidor email

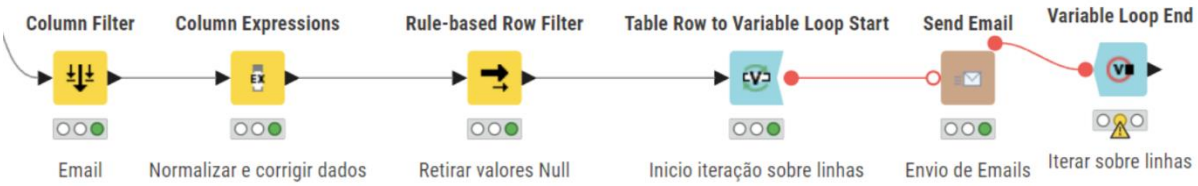

Figura 7 - Envio de emails

3.4. Fluxo para importação de dados em Power BI

Para o tratamento dos dados de todos os rastreios já realizados, o processo inicia-se pela integração de todos os ficheiros numa única tabela, consolidando toda a informação necessária para uma análise abrangente. Como informações como o nome e contactos não são necessários para a análise, o primeiro passo adotado é a exclusão dessa informação ficando só a informação necessária.

Após a exclusão destes dados é necessário começar a normalização e correção da tabela gerada, como estamos a lidar com uma tabela com um maior número de colunas é necessário efetuar mais operações.

3.4.1. Operações sobre coluna data rastreio

Como a coluna **data do rastreio** já se encontra no formato **data** facilita as tarefas a realizar sobre a coluna. O pretendido é que este campo dê origem a duas novas colunas, referentes ao mês e ano, a serem utilizadas posteriormente no tratamento dos dados.

```
regexReplace(column("Data Rastreio"), ".*?(\\d{4}).*", "$1")
regexReplace(column("Data Rastreio"), "\\d{4}-(\\d{2})-\\d{2}", "$1")
```

Tabela 5 - Separação campo data do rastreio

3.4.2. Operações sobre coluna data nascimento

A coluna referente à **data de nascimento** é frequentemente preenchido de forma incorreta e/ou em formatos diferentes, o que pode gerar erros durante a conversão dessa informação para o tipo de dado adequeado. Por isso, é necessário realizar uma correção desses campos, garantindo que todas as datas estejam corretamente formatadas e consistentes.

```
var dataCorrigida = replace(replace(column("Data Nascimento"), "\\", "-"), "/", "-");
if (regexMatcher(dataCorrigida, "^\\d{2}-\\d{2}-\\d{4}$")) {
   join(substr(dataCorrigida, 6, 4), "-", substr(dataCorrigida, 3, 2), "-", substr(dataCorrigida, 0, 2))
} else {
   dataCorrigida
}
```

Tabela 6 - Verificação campo data de nascimento

3.4.3 Operações sobre colunas referentes aos dados rastreados

Como a solução não armazena dados pessoais dos clientes, mas sim informações sobre quais os rastreios foram aceites, todas as colunas destes campos são preenchidas com os valores **"S"** (Sim) e **"N"** (Não). Para garantir uma correta inserção destes dados, é necessário assegurar que os campos são preenchidos no formato adequado e que não ocorrem inserções incorretas de caracteres devido a erros dos utilizadores. Dessa forma, foram aplicadas as mesmas regras de validação em cinco colunas distintas da tabela. Realço que todos os dados inseridos incorretamente são considerados inválidos e, portanto, excluídos de qualquer cálculo estatístico.

```
if (isMissing(column("IMC"))) {
   null
```

```
} else if (regexMatcher(column("IMC"), "^[NS].*")) {
    substr(column("IMC"), 0, 1)
} else {
    null
}
```

Tabela 7 - Tratamento dados dos rastreios

3.4.4 Criação coluna idade

Embora seja um dado relevante para a análise, não é comum solicitar diretamente a **idade** dos participantes nos rastreios. Em vez disso, é guardada apenas a **data de nascimento**, que infelizmente, contém erros tipográficos e inconsistência na formatação, resultando na categorização desse campo como uma **string** em vez de um tipo de dado adequado para operações com data. Para ultrapassar esta limitação, e como já tinha sido feita uma correção do formato destes dados, foi criada uma rotina para converter a data de nascimento para o formato data, possibilitando assim subtrair a data de nascimento e a data atual. Desta forma, tornou-se possível incluir este dado importante nas análises, garantindo maior precisão e qualidade nos resultados obtidos. Como parte do processo de validação e limpeza de dados e para combater **datas de nascimento inválidas** foi criado um filtro adicional para excluir idades superiores a 120 anos.

```
if (column("Idade") > 120) {
    null
} else {
    column("Idade")
}
```

Tabela 8 - Excluir idades inválidas

3.4.5 Envio de dados Power BI

No final de todas as verificações e **correções de erros** foi configurado um **nó de integração** que envia as informações tratadas para o **Power BI** da Microsoft. Esta etapa foi fundamental para garantir que os dados pudessem ser facilmente visualizados e utilizados pelos utilizadores finais de forma dinâmica e acessível. O envio destes dados possibilita a criação de **dashboards personalizados** onde os utilizadores podem monitorizar e analisar os resultados dos rastreios de maneira visual e intuitiva.

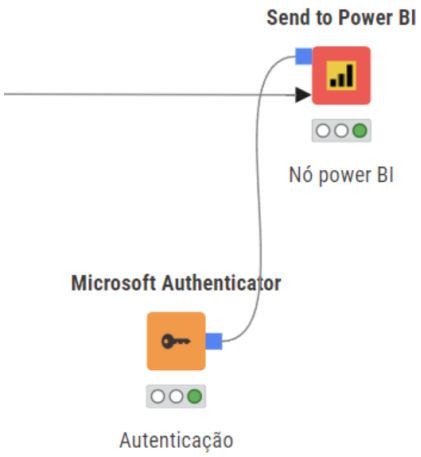

Figura 8 - Nó Power BI

4. Análise de Resultados e testes

Todos os pressupostos do **fluxo criado** foram **plenamente cumpridos** e submetidos a testes, tanto de forma individual quanto integrada. Cada componente do fluxo foi testado isoladamente para garantir o seu funcionamento correto, identificando e corrigindo eventuais falhas antes de prosseguir para a próxima fase.

4.1. Exportação CSV

A funcionalidade foi submetida a testes e a erros forçados para aferir a sua robustez. Foram detetadas falhas, prontamente corrigidas, com a exceção

de anexar a data de criação do ficheiro ao nome do mesmo. Sendo que a solução atual substitui o ficheiro existente.

Diogo Faro,912345688
Antonio Sorte,925546789

Figura 9 - exemplo ficheiro csv

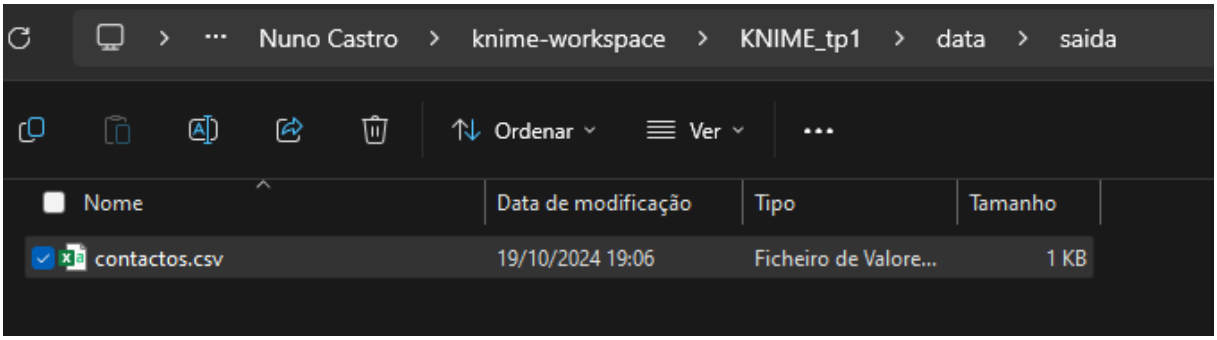

Figura 10 - Ficheiro csv

4.2. Envio de email

Atualmente as contas de email contêm várias restrições ao nível da segurança, incluindo autentificação 2FA, o que fez com que o plano inicial, que era utilizar a conta institucional do IPCA, tivesse de ser alterado e fosse utilizada uma conta de email com os serviços da Google, que permite gerar uma password temporária para uma aplicação especifica e dessa forma fazer a autentificação para o envio dos emails. Todos os testes efetuados foram bem sucedidos, sendo utilizados vários emails para testar o funcionamento do loop para envio das várias linhas da tabela.

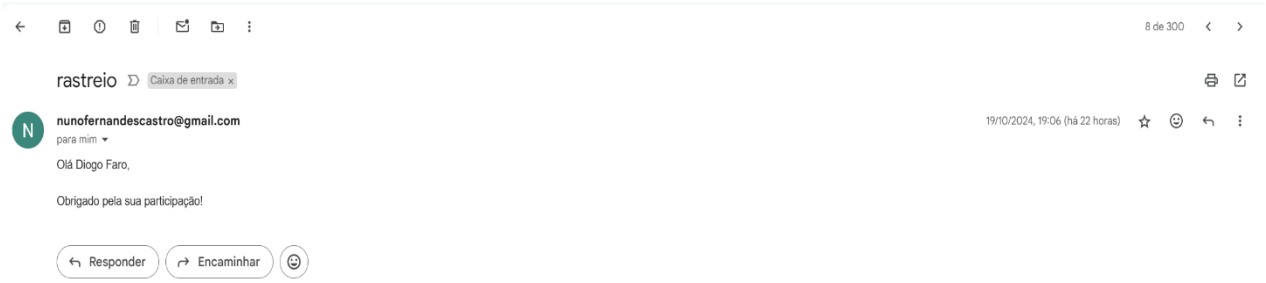

Figura 11 - Envio de email

4.3. Envio de dados para Power BI

A exportação de dados para a aplicação de dados da Microsoft Power BI da Microsoft, ocorreu sem problemas, tendo sido criado um dashboard de exemplo para verificação do carregamento dos dados. Foi detetado que após algum tempo sem ocorrer a exportação de dados é necessário efetuar nova autenticação na aplicação que recebe a informação.

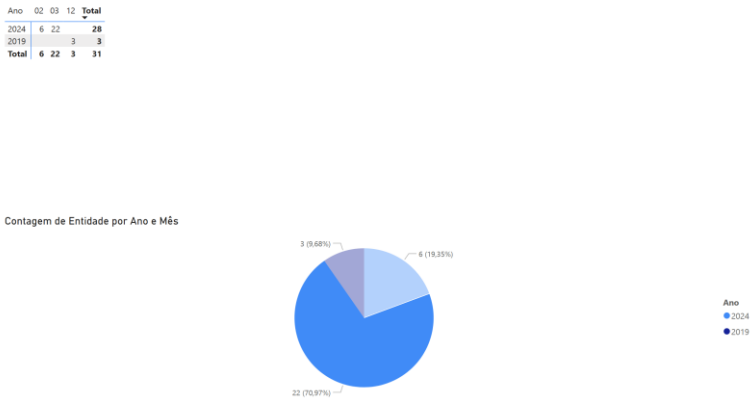

Figura 12 - Exemplo dashboard Power BI

5. Conclusão

Este trabalho prático permitiu explorar a importância da implementação de um processo ETL para melhorar o fluxo de dados dos rastreios realizados. Inicialmente, o método utilizado apresentava diversas limitações, especialmente pela ausência de verificações competentes e a propensão a erros na inserção manual de dados. Através da utilização da plataforma KNIME, foi possível automatizar etapas importantes, incluindo a normalização dos dados e a correção de erros, o que garantiu maior eficiência e precisão ao tratamento dos dados.

Com a nova abordagem, o processo de extração e organização dos dados foi automatizado, reduzindo significativamente o tempo despendido em tarefas manuais e eliminando potenciais falhas. Foi também melhorado o processo de exportação de dados, a criação de um automatismo para envio de email otimizando a comunicação com os participantes nos rastreios, e melhorada a preparação de dados a enviar para análise, adicionado a criação de um nó de envio para o Power BI, permitindo criar dashboards úteis para a tomada de decisão.

Com as melhorias implementadas, foi possível alcançar uma significativa redução de tempo e esforço no tratamento dos dados, além de aumentar a qualidade e confiabilidade dos resultados obtidos, demonstrando como o ETL pode ser um aliado na transformação de dados brutos em informações valiosas para suporte à decisão.

Bibliografia

Change date format. https://forum.knime.com/t/change-date-format/44717/4

Leveraging KNIME and Power BI: Integrating Power BI in KNIME. https://www.phdata.io/blog/leveraging-knime-and-power-bi-integrating-power-bi-in-knime/

How can i eliminate all the missing value from a data set?. https://forum.knime.com/t/how-can-i-eliminate-all-the-missing-value-from-a-data-set/4226

Explore Date&Time in KNIME & Get Best Pratices. https://www.knime.com/blog/explore-date-time-formats-best-practices

“Send Email” node and Gmail account