Relatório do trabalho da disciplina de Integração de Sistema Informáticos

Rankings Estatísticos dos Jogadores do Campeonato Brasileiro de Futebol de 2021

Arthur Fellipe Cerqueira Gomes – 24200

Engenharia de Sistemas Informáticos – Pós-Laboral

Outubro de 2024

ESCOLA SUPERIOR DE TECNOLOGIA IPCA

Afirmo por minha honra que não recebi qualquer apoio não autorizado na realização deste trabalho prático. Afirmo igualmente que não copiei qualquer material de livro, artigo, documento web ou de qualquer outra fonte exceto onde a origem estiver expressamente citada.

Arthur Fellipe Cerqueira Gomes – 24200

Lista de Figuras

ÍNDICE

1. Introdução ao Projeto

A integração de sistemas de informação, a utilizar dados como principal ativo, tornou-se uma necessidade crucial para a tomada de decisões estratégicas em empresas e organizações modernas. O processo de ETL (Extração, Transformação e Carga de Dados) desempenha um papel central na consolidação e análise de grandes volumes de dados provenientes de diferentes fontes, de forma a permitir uma visão mais aprofundada e precisa sobre o desempenho e as operações de uma organização.

Neste projeto, foi desenvolvido um fluxo de ETL com foco na obtenção e análise de dados estatísticos dos jogadores do Campeonato Brasileiro de Futebol de 2021. Através da análise dos dados dos jogadores, são geradas métricas e rankings, visualizados em gráficos, o que demonstra o potencial do ETL como ferramenta de suporte à tomada de decisões no contexto desportivo. Ao longo do relatório, será apresentada uma visão detalhada de cada etapa do fluxo de ETL, desde a extração dos dados até à geração dos outputs finais.

A escolha deste tema permite a aplicação prática de conceitos relacionados com a integração de sistemas, proporcionando um cenário ideal para explorar ferramentas de suporte a processos de ETL, bem como novas tecnologias e paradigmas de desenvolvimento.

Neste sentido, foi escolhido o uso do KNIME como a ferramenta principal para a construção do fluxo de ETL. A escolha justifica-se pela versatilidade da plataforma, que permite a integração de dados provenientes de múltiplas fontes, a realização de transformações complexas e a criação de visualizações interativas de forma intuitiva e sem a necessidade de programação extensa. Além disso, o KNIME oferece uma vasta gama de nós pré-configurados, o que facilita a implementação de cálculos e manipulações de dados e torna-o ideal para cenários académicos e profissionais, como o presente projeto, onde a eficiência e a clareza do fluxo são fundamentais.

Todos os ficheiros produzidos pela ferramenta, inclusive outputs do fluxo em diferentes formatos, se encontram anexados ao presente relatório, bem como foram disponibilizados online através do repositório no *GitHub* https://github.com/arthur-fellipe/ISI_TP1_24200.

2. Visão Geral do Fluxo de ETL

2.1. Descrição Geral

O fluxo desenvolvido no KNIME para a análise dos dados dos jogadores do Campeonato Brasileiro de Futebol de 2021 é composto por quatro componentes principais: extração de dados via API, transformação inicial dos dados recebidos, realização de cálculos e classificação, e, finalmente, a criação de gráficos e relatórios para visualização dos resultados. Este fluxo tem como objetivo permitir uma análise robusta e organizada das estatísticas dos jogadores, com saídas em formatos de fácil visualização e exportação.

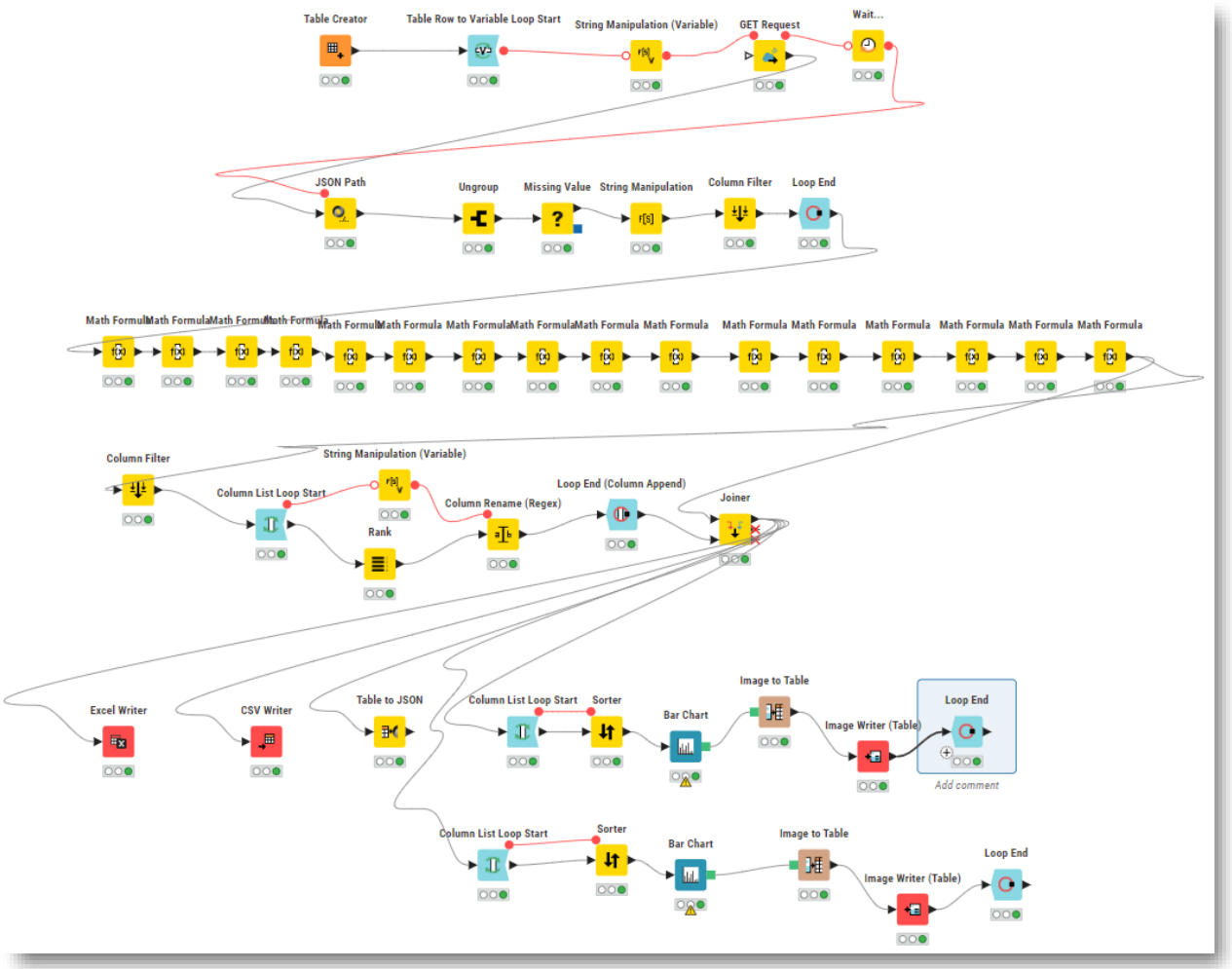

Figura 1 - Visão Geral do Fluxo ETL

O fluxo começa com a extração de dados via API gratuita denominada *API-Football*[1], utilizando o nó GET Request para buscar os dados diretamente da fonte externa. A API em questão fornece diversas informações sobre campeonatos de futebol espalhados pelo mundo, dentre as quais, estatísticas de jogadores que podem ser filtradas por época, clube e liga. Os dados são fornecidos em formato JSON. O plano gratuito é limitado a 100 requisições por dia e, no caso do Campeonato Brasileiro, só poderiam ser acessados os dados referentes às épocas de 2020 a 2022, motivo pelo qual foi escolhida a época de 2021.

Após a obtenção dos dados brutos, é necessário um processo de transformação inicial, que organiza e prepara os dados recebidos para as etapas seguintes. Seguidamente, são realizadas operações de cálculo e classificação, que geram novas métricas a partir dos dados e, em seguida rankings de desempenho dos jogadores. Finalmente, os resultados são apresentados, tanto em sua integralidade em formato *.json*, *.xlsx* e *.csv*, o que possibilita sua transferência, quanto em forma de imagem dos rankings criados para atender ao utilizador final, com auxílio de nós para visualização e exportação.

A seguir, cada um dos componentes principais do fluxo será detalhado, incluindo as funções dos nós mais relevantes em cada etapa.

2.2. Componentes Principais

2.2.1. Extração de Dados via API

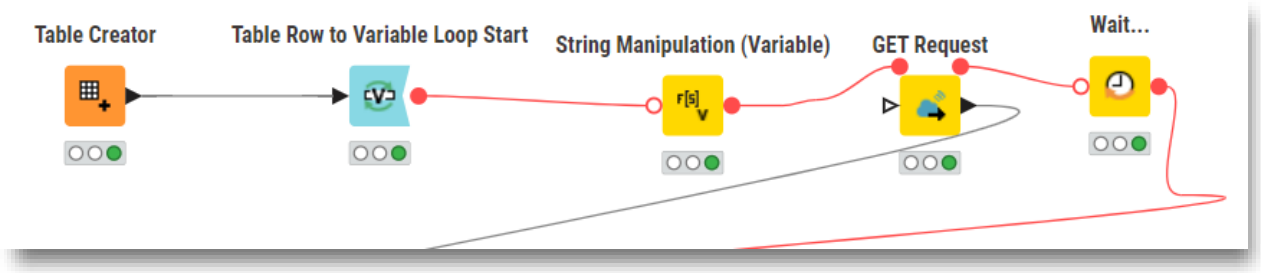

Figura 2 - Extração de Dados via API

A primeira etapa do fluxo consiste na extração de dados diretamente da API, onde são recolhidas informações sobre os jogadores. Para isso, foi utilizado o nó GET Request, que permite realizar pedidos a endpoints de APIs. Contudo, foi verificado que a *API-Football* utiliza a paginação dos dados disponibilizados, agrupando estatísticas de no máximo 20 jogadores a cada página[2], e que

[1] https://www.api-football.com/
[2] https://www.api-football.com/documentation-v3#tag/Players/operation/get-players

para extrair os registos de todos os jogadores seria necessário realizar requisições para cada uma das 47 páginas de dados. Como solução, o nó GET Request foi incluindo em um loop.

O primeiro passo para criação do loop foi a criação de uma tabela com a sequência numérica de 1 a 47, sendo distribuído um número a cada linha. Em seguida, o nó Table Creator foi conectado ao nó Table Row to Variable Loop Start, que transformou os dados contidos em cada linha em uma variável.

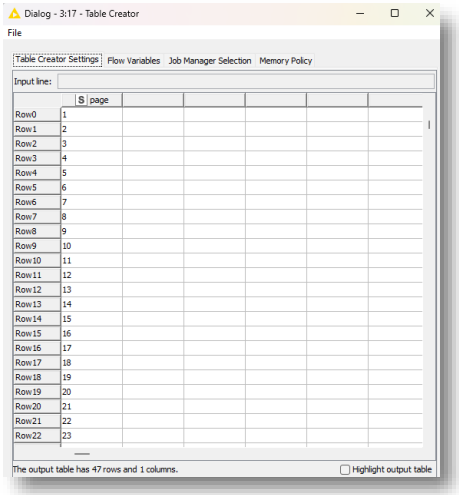

Figura 3 - Tabela de Paginação

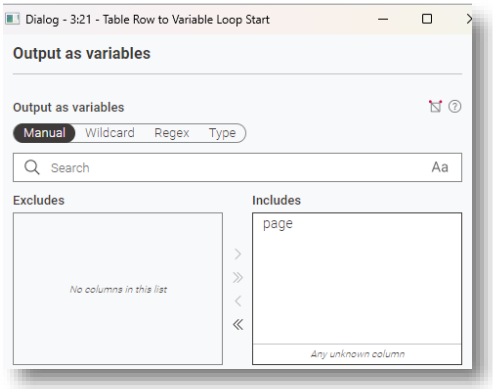

Figura 4 - Table Row to Variable Paginação

O próximo passo foi utilizar o nó String Manipulation (Variable) para concatenar o endpoint adequado para extração das estatísticas desejadas na *API-Football* com a variável gerada no nó anterior, referente à página a ser extraída em cada requisição.

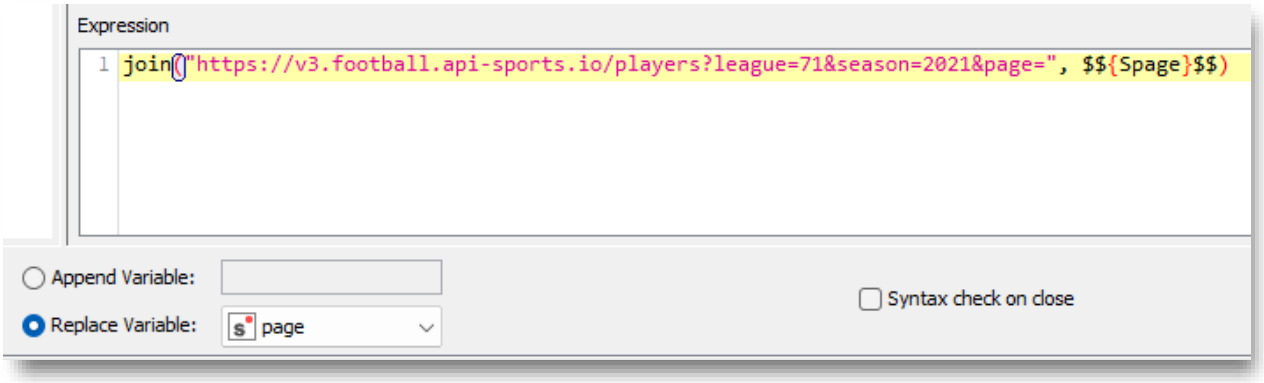

Figura 5 - Concatenção do endpoint com a página

Foi feita então a configuração do nó GET Request, sendo primeiro configurada a seção Request Headers, necessária para autenticação junto à API com a chave de acesso fornecida no momento da subscrição do plano gratuito, e depois a configuração da seção Flow Variables, onde foi estabelecido que a URL seria aquela variável definida na concatenação do nó anterior.

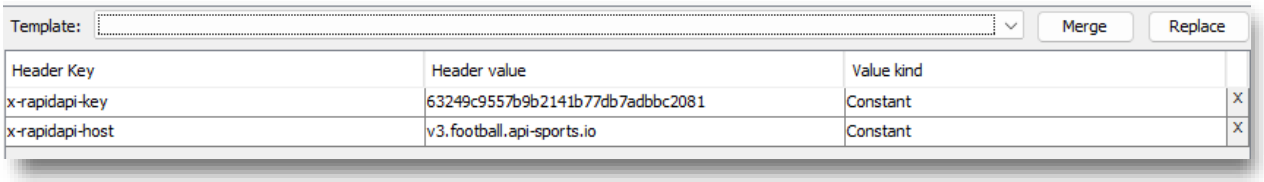

Figura 6 - Request Headers

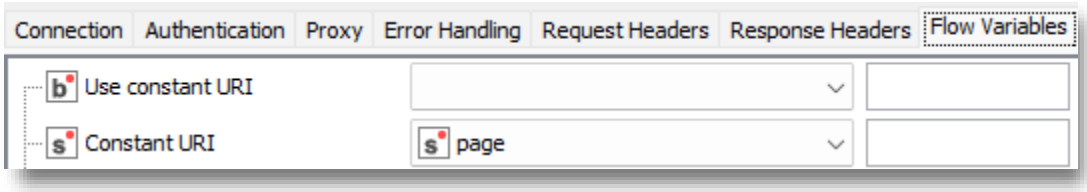

Figura 7 - Configuração URL

Para finalizar a configuração da extração de dados da API um nó Wait... foi inserido para garantir que cada requisição seja feita a cada 6 segundos, o que se faz necessário pois a API tem um limite de 10 requisições por minuto.

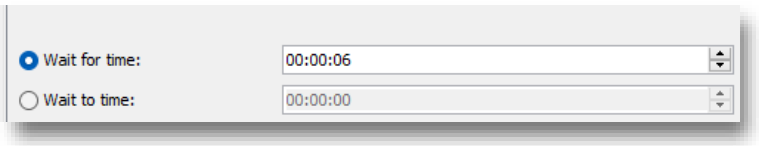

Figura 8 - Configuração Wait

2.2.2. Transformação Inicial dos Dados Recebidos

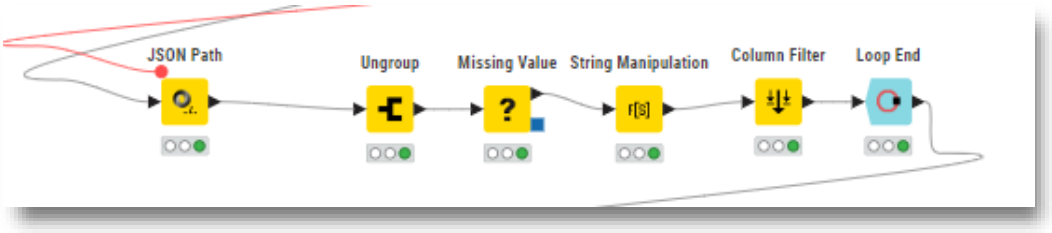

Figura 9 - Transformação Inicial dos Dados Recebidos

Após a extração, ainda dentro do loop criado, os dados ainda estão num formato bruto, e é necessário aplicar transformações iniciais para garantir que estejam prontos para a análise subsequente. Nesse sentido, o nó JSON Path é utilizado para transformar os dados recebidos no formato JSON em uma tabela manipulável dentro do KNIME, bem como para selecionar os tipos de dados que serão analisados, transformando-os em colunas da tabela.

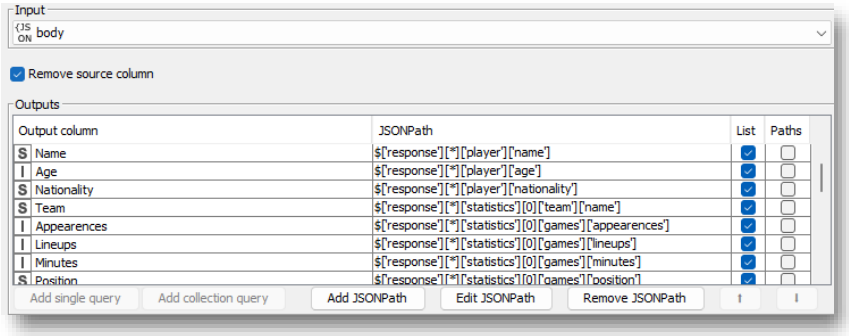

Figura 10 - Configuração JSON Path

A seguir, o nó Ungroup desagrega os dados, permitindo que arrays e objetos complexos sejam transformados em linhas de tabela que podem ser manipuladas de forma mais eficiente.

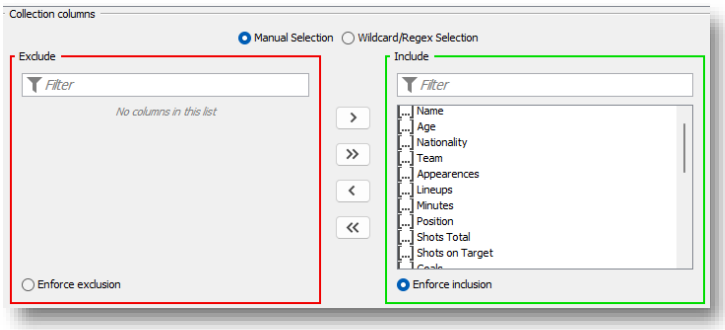

Figura 11 - Configuração Ungroup

Nesta fase, os nós Missing Value e Column Filter são aplicados para limpar os dados, a substituir valores numéricos ausentes por 0 e eliminando colunas irrelevantes para a análise. Adicionalmente, é aplicado o nó String Manipulation para modificar valores textuais dentro das colunas que não se enquadravam à codificação UTF-8, nomeadamente os acentos e caracteres especiais da língua portuguesa, garantindo que os dados estejam consistentes e bem formatados para as próximas etapas.

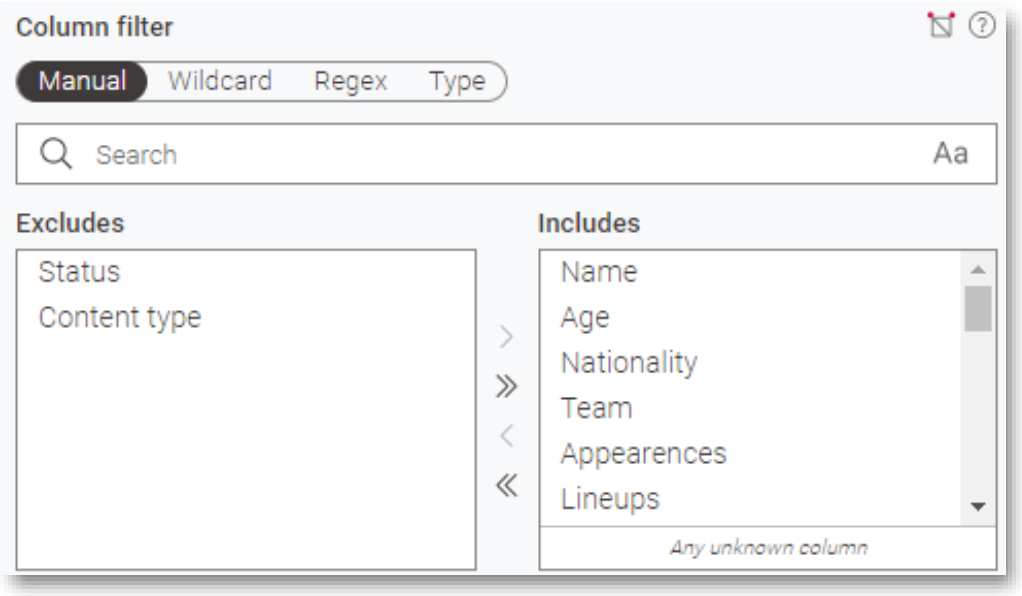

Figura 12 - Colunas Filtradas

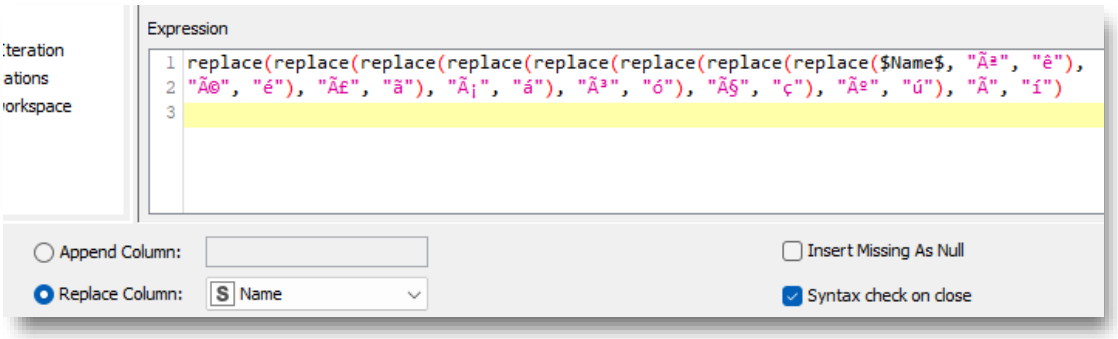

Figura 13 - Correção codificação UTF-8

O último passo realizado é o encerramento do loop de extração e transformação inicial dos dados com utilização do nó Loop End.

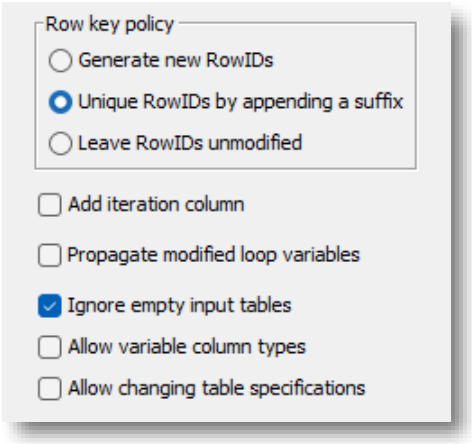

Figura 14 - Configuração Loop End

2.2.3. Realização de Cálculos e Classificação

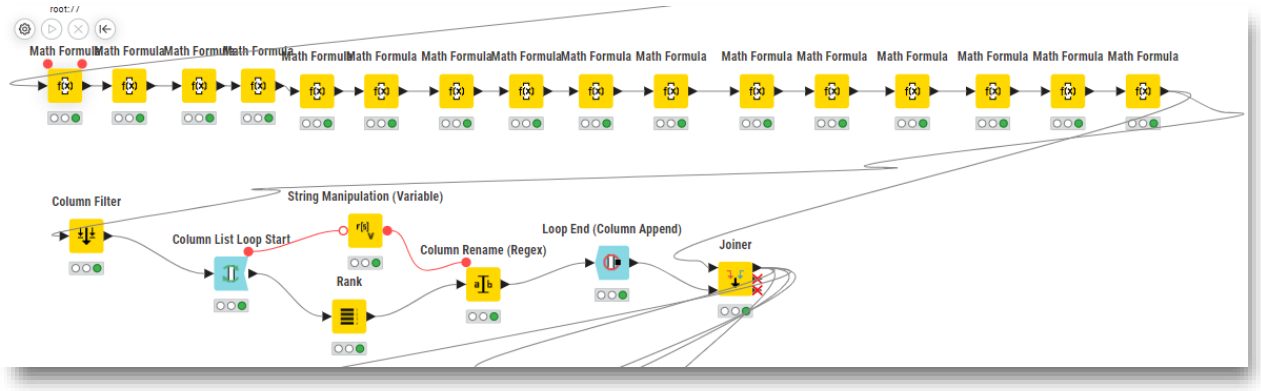

Figura 15 - Realização de Cálculos e Classificação

A terceira fase do fluxo de ETL concentra-se na realização de cálculos e na criação de métricas adicionais, que são fundamentais para a avaliação de desempenho dos jogadores. Nesta etapa, são aplicadas fórmulas matemáticas e operações de agregação para calcular indicadores específicos, a partir dos dados previamente transformados.

A principal ferramenta para estes cálculos é o nó Math Formula, que permite realizar operações aritméticas e lógicas nas colunas de dados. No fluxo desenvolvido, foram configurados 16 nós Math Formula para calcular as seguintes estatísticas:

Participation in Goals	Minutes per Goals	Minutes per Participation in Goals	Minutes per Assists

Penalty Conversion	On Target Conversion per Total Shots	Shots per Goal	Goal Conversion per Total Shots
Goal Conversion per Shots on Target	Percentage of Key Passes per Total Passes	Key Passes per Assist	Total Passes per Assist
Percentage of Successful Dribbles	Percentage of Duels Won	Minutes per Yellow Card	Minutes per Red Card

Figura 16 - Métricas criadas

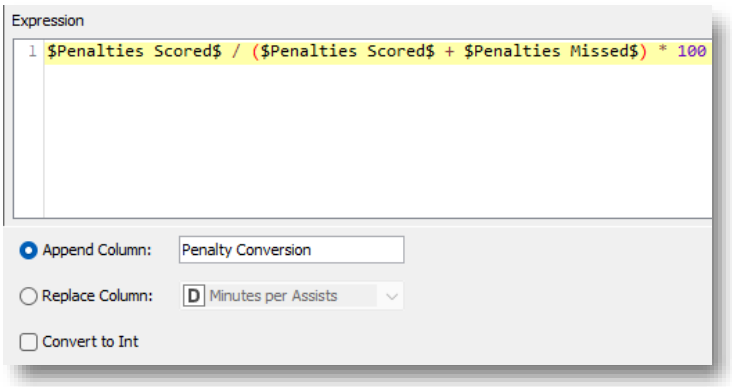

Figura 17 - Exemplo de Configuração do Math Formula

Após a criação destas métricas, é iniciada a etapa de classificação dos jogadores em relação a cada estatística. O primeiro passo é utilizar o nó Column Filter para selecionar as colunas que são rankeadas, nomeadamente aquelas que possuem estatísticas de jogo e não apenas dados de identificação.

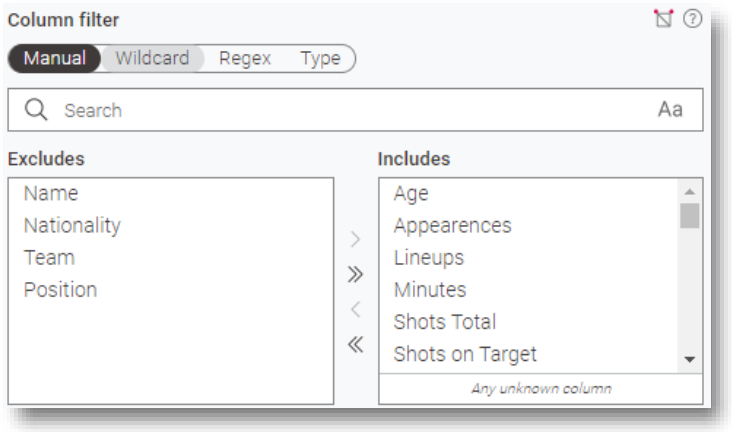

Figura 18 - Colunas a serem rankeadas

Em seguida, utiliza-se o nó Column List Loop Start para iniciar o loop sobre os campos selecionados.

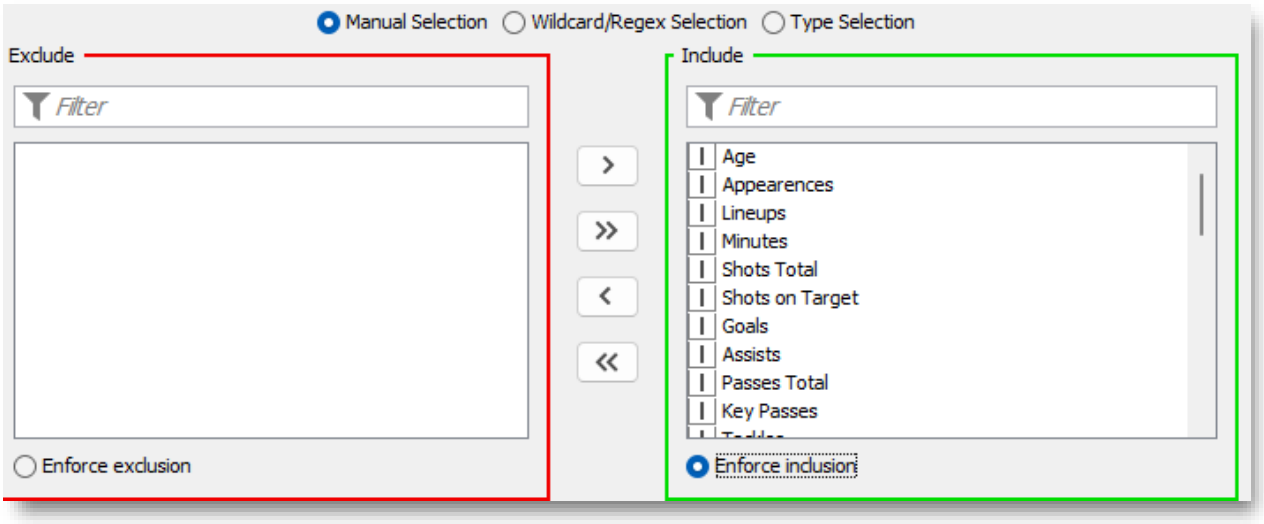

Figura 19 - Column List Loop Start

O último nó gera dois outputs diferentes. O primeiro conecta as variáveis da iteração para o nó String Manipulation (Variable), onde se utiliza concatenação para adicionar “Ranking” ao final do nome da coluna atual e criar nova variável.

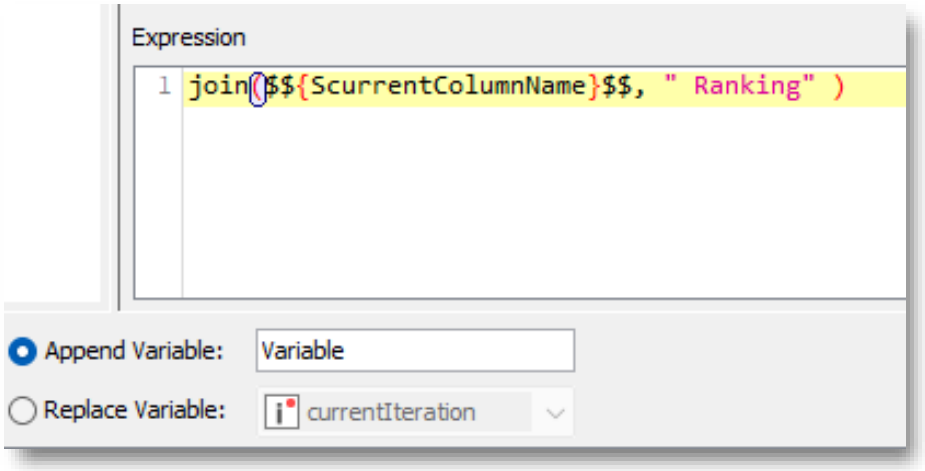

Figura 20 - Concatenação do Nome da Coluna de Ranking

O segundo output leva os dados da coluna atual da iteração ao nó Rank. Este nó é utilizado para gerar rankings que comparam os jogadores de acordo com os diferentes critérios de desempenho. Foi configurado com auxílio à seção Flow Variables que a cada iteração a coluna atual seria rankeada na ordem decrescente e o nome da coluna gerada seria o número da iteração. Além disso, foi configurado que o modo de classificação seria o padrão, no qual 2 jogadores empatados ficam na mesma colocação e a colocação do próximo do ranking não é o número logo a seguir.

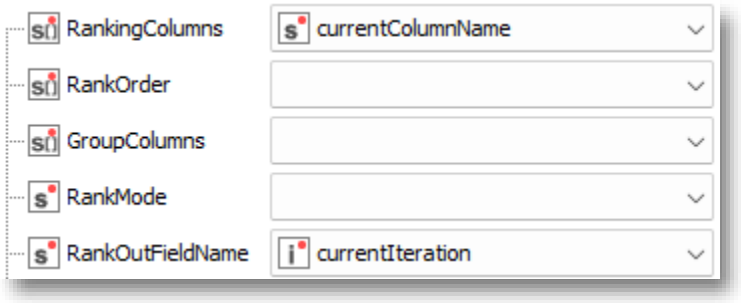

Figura 21 - Flow Variables do Rank

#	RowID	Minutes per Red Card *Number (double)*	38 ↑ *Number (integer)*
575	Row...	851	53
403	Row...	848	54
702	Row...	848	54
417	Row...	806	56

Figura 22 - Exemplo de Classificação com Empate

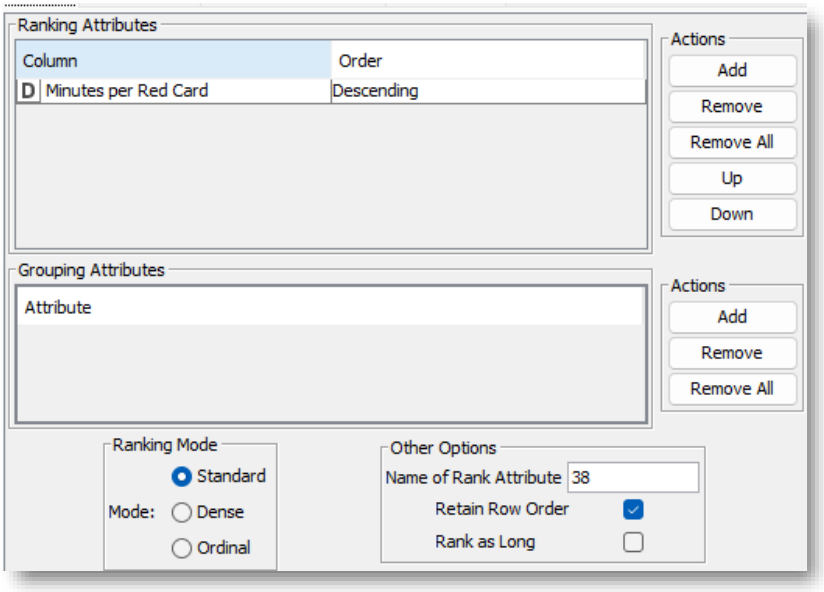

Figura 23 - Configuração do Rank

Entra em ação, então, o nó Column Name (Regex), que a utilizar as variáveis disponíveis no loop muda o nome da coluna de classificação, que deixa de ser o número da iteração atual para ser o resultado da concatenação criada anteriormente com o acréscimo de “Ranking” ao nome da coluna original.

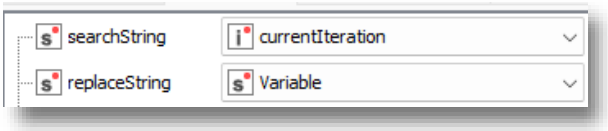

Figura 24 - Flow Variables do Column Name (Regex)

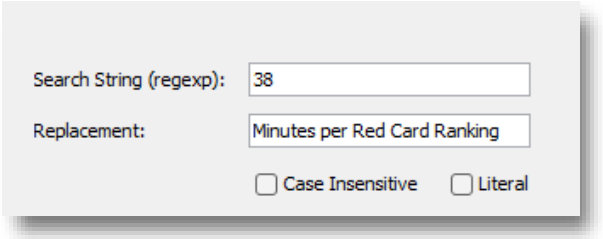

Figura 25 - Configuração do Column Name (Regex)

A iteração é então finalizada pelo nó Loop End (Column Append), que adiciona a nova coluna de classificação ao lado da coluna original. Em seguida, o nó Joiner recebe dois inputs: a tabela que existia após a realização dos cálculos com Math Formula e a tabela gerada pelo loop de classificação. Com essa informação, o Joiner adiciona as colunas de identificação dos jogadores à tabela gerada pelo classificação.

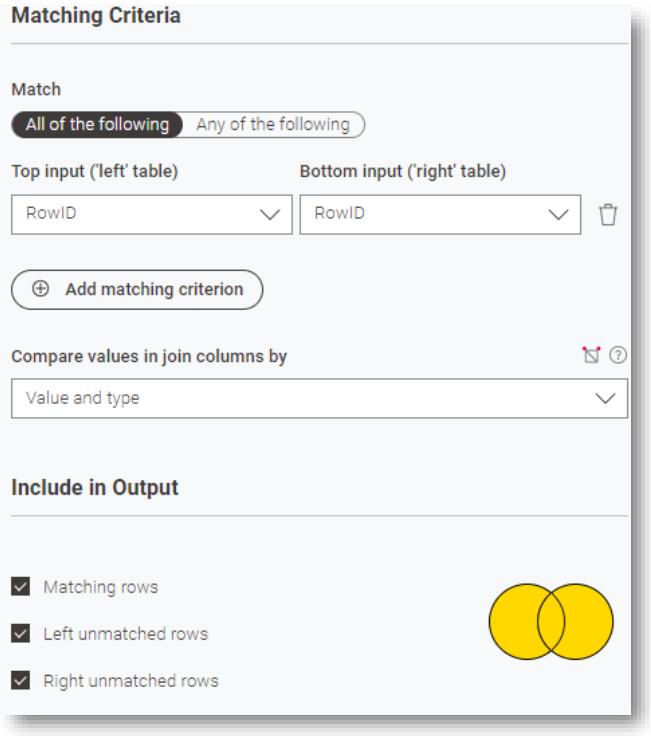

Figura 26 - Joiner pt.1

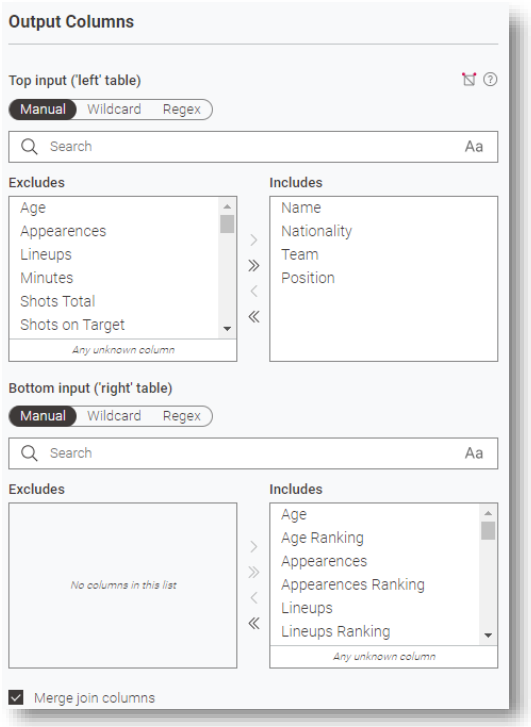

Figura 27 - Joiner pt.2

O resultado desta fase é uma tabela detalhada e final, com os jogadores ordenados de acordo com as métricas de interesse, pronta para ser visualizada e analisada nas próximas etapas.

2.2.4. Criação de Gráficos e Relatórios

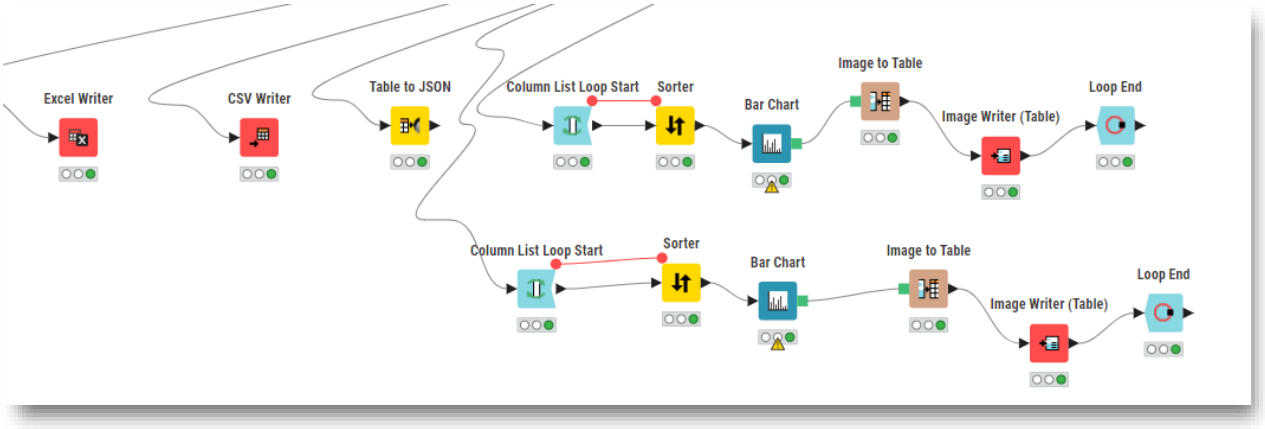

Figura 28 - Criação de Gráficos e Relatórios

A quarta e última fase do fluxo foca na criação de representações gráficas e relatórios, facilitando a interpretação dos resultados das análises realizadas nas etapas anteriores. Para isso, o KNIME oferece uma série de nós voltados para visualização.

Nesse sentido, a tabela criada até a fase anterior, a incluir os cálculos e rankings, é exportada em diferentes formatos com os nós Table to JSON, CSV Writer e Excel Writer, a garantir que os resultados finais possam ser analisados externamente em ferramentas como o Excel ou Google Sheets ou integrados em outros sistemas.

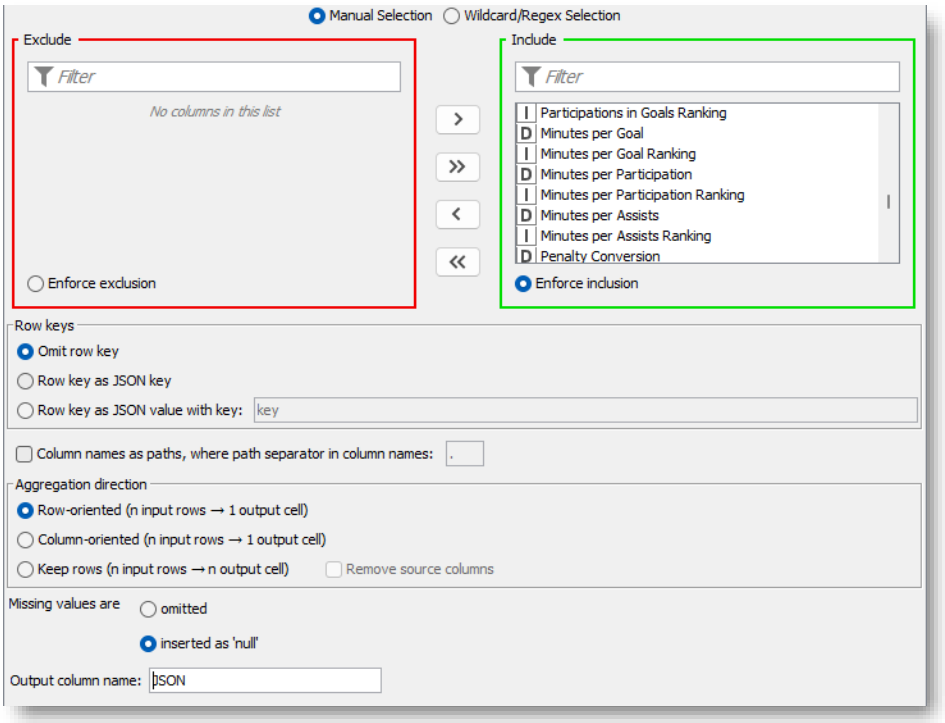

Figura 29 - Table to JSON

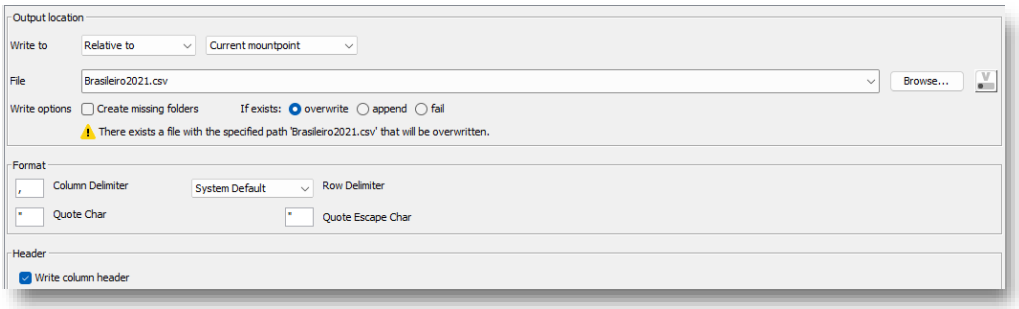

Figura 30 - CSV Writer

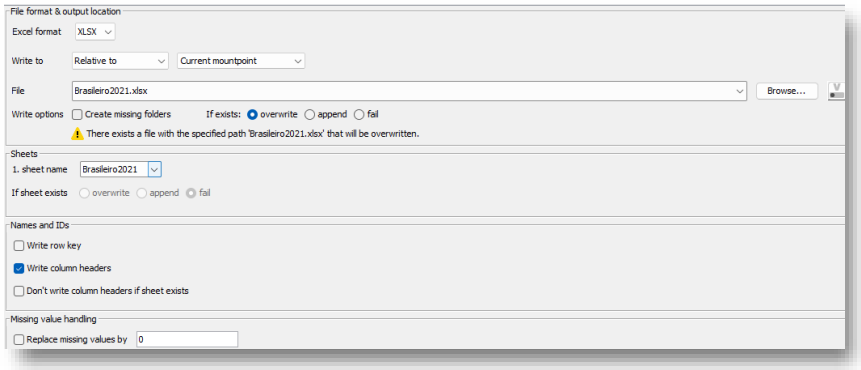

Figura 31 - Excel Writer

Também são criados loops para geração de imagens do ranking de cada métrica de desempenho, tanto em ordem decrescente quanto crescente. Primeiro se utiliza o nó Column List Loop Start para selecionar as colunas que serão representadas em imagens.

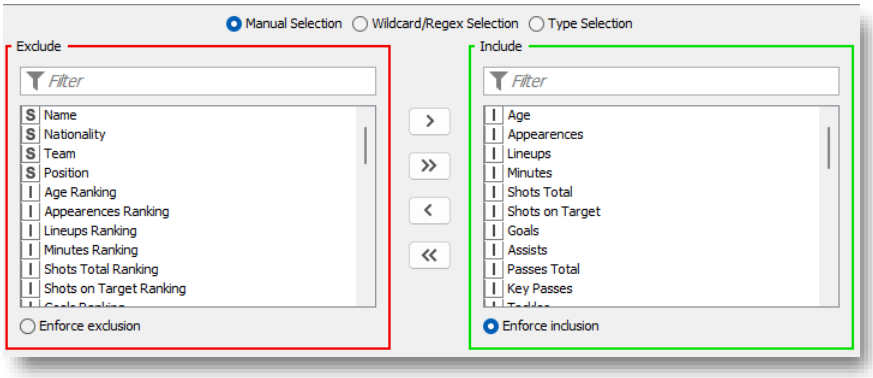

Figura 32 - Seleção de Colunas Representadas em .png

Em seguida o nó Sorter utiliza a variável com o nome da coluna iterada para ordená-la em forma crescente ou decrescente.

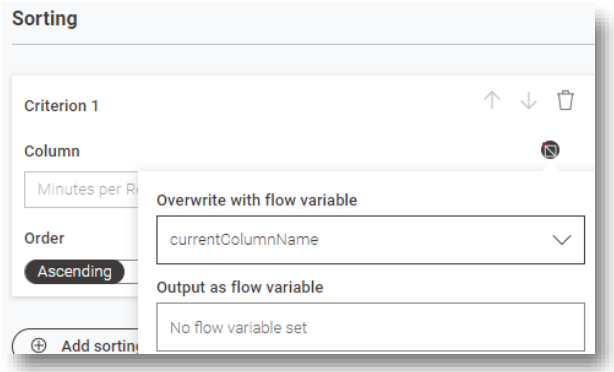

Figura 33 - Sorter

O nó Bar Chart é utilizado para criar gráficos de barras que ilustram as métricas calculadas e os rankings gerados. Os gráficos apresentam os 20 jogadores da liga com melhor ou pior classificação de cada métrica de desempenho selecionada.

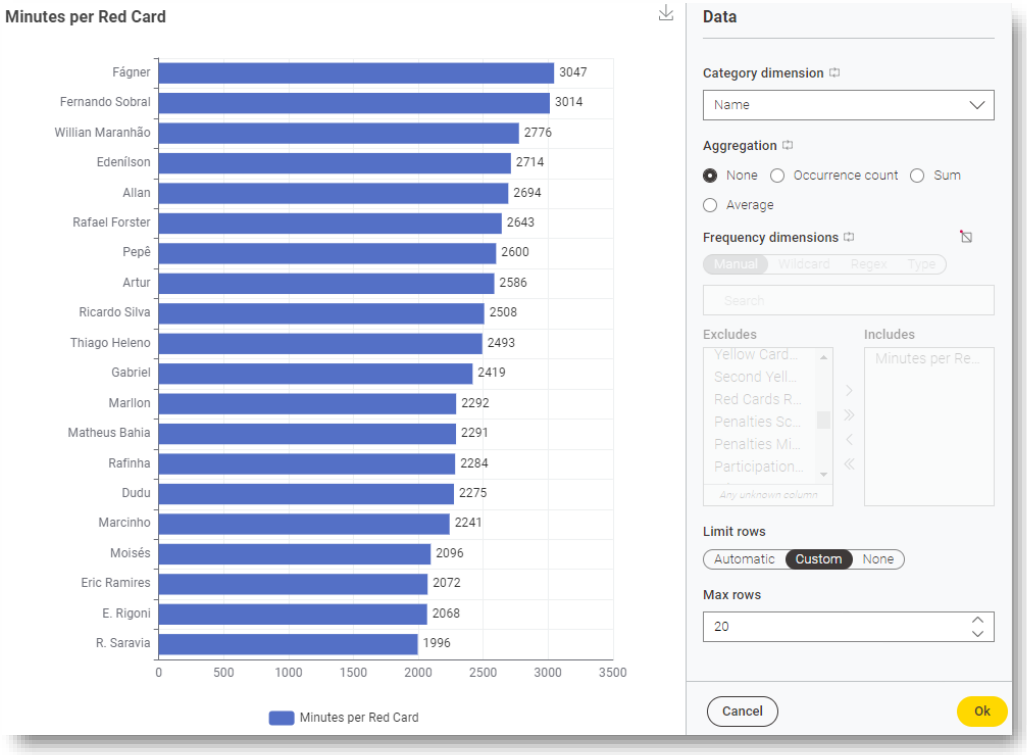

Figura 34 - Criação de Gráfico dos Rankings

O nó Image to Table transforma o gráfico em uma tabela, passo necessário para em seguida o nó Image Writer (Table) exportar os gráficos criados em formato de imagem (.png) para uma pasta "RankingsCrescentes" ou "RankingsDecrescentes", o que facilita a incorporação dessas visualizações em relatórios ou apresentações.

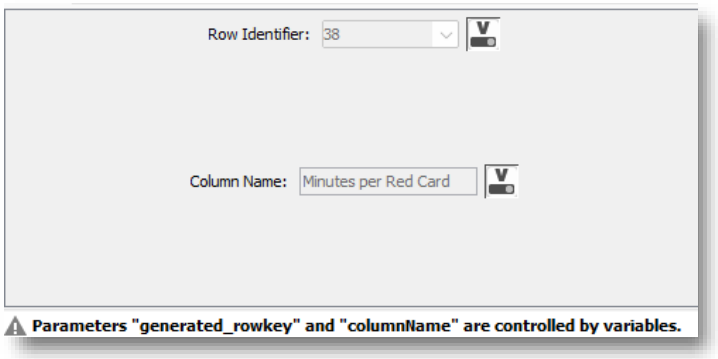

Figura 35 - Image to Table

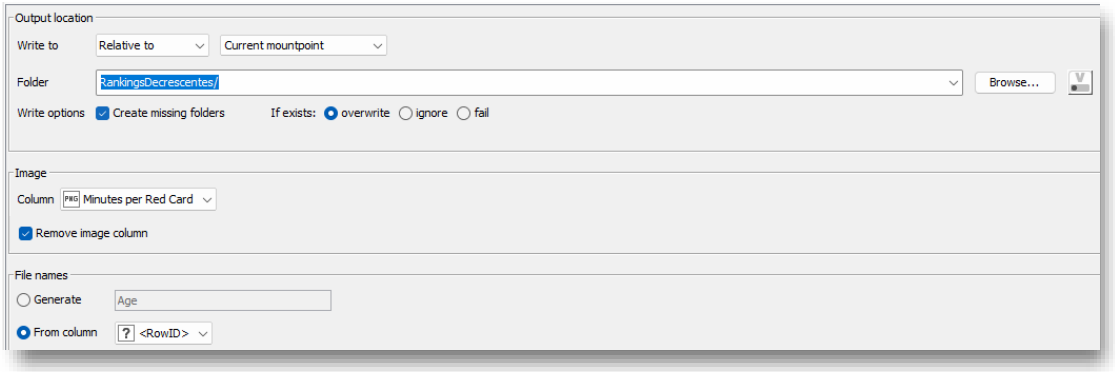

Figura 36 - Image Writer (Table)

Este conjunto de nós para exportação garante que todo o processo de ETL resulte em produtos finais tangíveis, tanto na forma de gráficos quanto de relatórios tabulares, permitindo uma análise clara e precisa do desempenho dos jogadores.

3. Conclusão

O desenvolvimento deste projeto utilizando o KNIME para a criação de um fluxo de ETL dedicado à análise de dados dos jogadores do Campeonato Brasileiro de Futebol de 2021 permitiu a aplicação prática de conceitos-chave de integração de sistemas de informação e análise de dados. Através da estruturação de um processo robusto de extração, transformação, cálculo e visualização, foi possível não só consolidar os conhecimentos teóricos sobre ETL, mas também explorar as capacidades de ferramentas como o KNIME para lidar com dados reais e complexos.

A escolha do KNIME como plataforma de desenvolvimento revelou-se acertada, pela sua flexibilidade e pelo seu suporte a operações diversas, desde a integração com APIs até à visualização dos dados. O uso de nós como Math Formula, Rank, Bar Chart, e outros, demonstrou que é possível construir um fluxo automatizado que não só processa grandes volumes de dados de forma eficiente, mas também permite ajustes dinâmicos conforme os critérios de análise evoluem.

Além disso, o projeto permitiu o aprofundamento na criação de relatórios e dashboards, facilitando a comunicação dos resultados de forma visual e acessível. A estrutura modular do fluxo permite que este seja facilmente adaptado a novos cenários de análise ou para incluir dados de diferentes épocas e competições, o que aumenta o seu potencial de reutilização e escalabilidade

Por fim, o projeto contribuiu significativamente para o desenvolvimento de competências em integração de sistemas, manipulação de dados e visualização de resultados, consolidando conceitos aprendidos na Unidade Curricular e potenciando o uso de tecnologias modernas no contexto de análise de dados desportivos. A partir do que foi construído, há espaço para expandir e refinar a análise, de forma a explorar ainda mais o poder da integração de dados e da automação na tomada de decisões informadas.

4. Bibliografia

API-Football. *Documentation V3*. Disponível em: https://www.api-football.com/documentation-v3. Acesso em: 06 out. 2024.

KNIME AG. *KNIME Documentation*. Disponível em: https://docs.knime.com/. Acesso em: 10 out. 2024.