LISBON
SCHOOL OF
ECONOMICS &
MANAGEMENT
UNIVERSIDADE DE LISBOA

Mestrado

Gestão de Sistemas de Informação

Trabalho Final de Mestrado

Relatório de Estágio

Estágio Curricular nos Departamentos de Data Science e Software Development no Sport Lisboa e Benfica

Rodrigo Heliodoro Vilaça Santos Alcarva

Outubro 2021

LISBON
SCHOOL OF
ECONOMICS &
MANAGEMENT
UNIVERSIDADE DE LISBOA

Mestrado em

Gestão de Sistemas de Informação

Trabalho Final de Mestrado

Relatório de Estágio

Estágio Curricular nos Departamentos de Data Science e Software Development no Sport Lisboa e Benfica

Rodrigo Heliodoro Vilaça Santos Alcarva

Orientação

Prof. Rui Manuel Trigo Pereira Guedes

Outubro 2021

1 Agradecimentos

Foi uma fase da minha vida académica onde tive momentos que me fizeram passar por várias formas de emoção onde tive a sorte de ter excelentes pessoas a apoiar-me.

Quero expressar o maior agradecimento aos meus pais, que são os pilares da minha vida, agradecer-lhes todo o apoio que me deram ao longo deste tempo, mesmo nas fases mais difíceis. Agradeço-lhes, também, por toda a compreensão que tiveram, e por toda a liberdade que me deram, sabendo sempre confiar em mim. Quero agradecer, de forma especial, aos meus avós, com eles aprendi muito e eles estiveram sempre presentes para mim, mesmo perguntando todas as semanas qual era o nome do meu mestrado.

Quero agradecer ao Prof. Rui Guedes pelo seu trabalho incansável durante estes meses, onde ele faria qualquer coisa por mim. Quero agradecer a Gonçalo Ponte, Miguel Nunes e Suds por me integrarem nas suas equipas e me ajudarem a aprender e a saber como é o mundo do trabalho nesta minha primeira experiência profissional.

Aos meus amigos, posso dizer que tenho muita sorte em ter o vosso apoio. Quero agradecer ao Alexandre, à Ana Cláudia, ao André, ao Afonso e ao Bruno pelas horas preciosas que me fizeram relaxar e perceber que devo desfrutar sempre dos momentos. Ao Cabecinha e ao Fraga, por nunca me terem decepcionado, por me fazerem sentir que posso sempre contar com eles, e por me incluírem em tudo o que podem. Tenho, também, de agradecer à Francisca, à Monica e ao Tiago por estarem presentes para mim quando precisei deles. Por ser o meu maior apoio durante este estágio, agradeço afilhado Guilherme. Aprendi muito com ele, aprendi a desfrutar mais da vida, a desfrutar melhor das coisas, a relativizar as coisas, a relativizar melhor, a querer viver todas as experiências que existem neste mundo que me possam vir à cabeça, aprendi a controlar a minha ansiedade e que nada e ninguém é por acaso e tudo tem um significado. Quero agradecer a Ana Leonor por me fazer crescer em todos os sentidos, mas principalmente como pessoa e mentalmente. Ela foi, e durante a maior parte desta fase um grande pilar que não me fez desistir e que me motivou a fazer mais e melhor, que me inspirou a ser uma pessoa melhor, a ser uma pessoa com bom coração, a ser uma pessoa que nunca fica satisfeita com o sucesso e que quer sempre mais.

2 Resumo

Este trabalho final do Mestrado visa detalhar todos os projetos que foram desenvolvidos, durante os meses de outubro a julho, do estágio curricular realizado no Sport Lisboa e Benfica, onde foram objeto de implementação alguns temas e conteúdos abordados no decorrer do Mestrado, mas também onde foi adquirido muito conhecimento no decorrer destes 10 meses.

Inicialmente começamos com uma introdução do desporto em geral e do futebol em particular e reconhecemos que os clubes desportivos cada vez mais inovam nas áreas de *Data Science*, análise e de tecnologia à procura de resultados que lhes permita ter vantagem sobre os rivais. Seguidamente fizemos uma revisão da literatura onde são abordados alguns tópicos que foram posteriormente trabalhados nos vários projectos durante o estágio, tais como *Sports Analytics*, *Data Science* e *Web Scraping*. Em seguida, é feita uma análise e introdução ao grupo Sport Lisboa e Benfica, onde foi apresentado uma comparação entre o que é mencionado na revisão da literatura e o que o SLB anda a realizar. Neste capitulo, existe uma explicação das duas equipas de trabalho que realizaram estes projetos, quais os seus objectivos, o seu trabalho e uma uma comparação entre os métodos de trabalho de cada uma.

A próxima fase é a dos projectos que foram realizados, onde na equipa de *Software Development* o trabalho realizado foi baseado primeiro numa leitura da documentação e depois na parte dos testes. Na equipa de *Data Science* fez-se um projecto de raiz, onde foi criado um algoritmo de extracção de dados a partir de duas fontes de dados, onde foram recolhidas variada informação de jogadores que já passaram pelo SLB, onde posteriormente houve uma junção de dados recolhidos com os dados do SLB. Finalmente passámos à fase de construção de relatórios de *Power BI*, onde procedemos a uma análise profunda dos dados. Posteriormente é feita uma análise crítica do progresso neste estágio curricular, acrescentando igualmente um segmento onde é revisto e mencionado o próprio desenvolvimento pessoal.

Na conclusão há uma análise geral de todo o progresso que houve ao longo destes meses, nomeadamente uma referência a limitações, problemas encontrados e sugestões.

Key Words: Sports Analytics, Data Science, Web Scraping, Business Intelligence

3 Abstract

This final work of the Master aims to detail all the projects that were developed, during the months from October to July, of the curricular internship held at Sport Lisboa e Benfica, where some topics and contents discussed during the Master were subject to implementation, but also where much knowledge was acquired during these 10 months.

Initially we started with an introduction of sports in general and soccer in particular and recognized that sports clubs are increasingly innovating in the areas of Data Science, analysis and technology in search of results that allow them to have an advantage over their rivals. Next we did a literature review where some topics that were later worked on in various projects during the internship, such as Sports Analytics, Data Science and Web Scraping are addressed. Next, an analysis and introduction to the Sport Lisboa e Benfica group is made, where a comparison between what is mentioned in the literature review and what SLB is doing was presented. In this chapter, there is an explanation of the two work teams that carried out these projects, what their objectives are, their work and a comparison between the working methods of each one.

The next phase is the projects that were carried out, where in the Software Development team the work done was based first on a reading of the documentation and then on the testing part. The Data Science team made a project from scratch, where an algorithm was created to extract data from two data sources, where various information of players who have passed through SLB were collected, where later there was a junction of data collected with the SLB data. Finally we passed to the report construction phase of Power BI, where we proceeded to a deep analysis of the data. Afterwards there is a critical analysis of the progress in this curricular internship, also adding a segment where the personal development itself is reviewed and mentioned.

In the conclusion there is an overall analysis of all the progress that has been made over these months, including a reference to limitations, problems encountered and suggestions.

Key Words: Sports Analytics, Data Science, Web Scraping,Business Intelligence

Índice

1 Agradecimentos **1**

2 Resumo **2**

3 Abstract **3**

Índice **4**

Lista de Figuras **5**

4 Introdução **6**

5 Revisão da Literatura **8**

5.1 Sports Analytics 8

5.2 Data Science 9

5.3 Web Scraping 11

5.4 Business Intelligence 13

6 Análise ao Sport Lisboa e Benfica **14**

6.1 Missão, Visão e Valores 17

6.1.1 Missão 17

6.1.2 Visão 17

6.1.3 Valores 18

6.2 Departamentos de Software Development e Data Science do SLB 18

7 Caracterização do estágio **19**

7.1 Projeto Web Scraping 19

7.1.1 As Fases 19

7.1.2 Descrição 20

7.1.3 A prática 22

7.1.4 Dashboard - Power BI 33

7.2 QA Tester 37

8 Análise Crítica **37**

9 Conclusão **40**

9.1 Desenvolvimento Pessoal e Profissional 40

9.2 Problemas e Limitações 41

9.3 Sugestões para o futuro 41

Lista de Figuras

1 Símbolo do Sport Lisboa e Benfica 14

2 Comparação de resultados líquidos entre 19/20 e 20/21 16

3 Mapa com principais projetos do estágio 19

4 Diagrama dos dados 21

5 Diagrama do método 22

6 Página das transferências 26

7 Página do jogador 26

8 Página época a época com código HTML 27

9 Valores de mercado: Website Transfermarkt 30

10 Jogador - Geral 34

11 Jogador - VM 34

12 Jogador - Seleção 34

13 Valorização - VM 34

14 Valorização - Fora do Benfica 34

15 Top 5 Ligas 34

16 Jogadores - Seleção 35

17 Seleção - Comparação Rivais 35

18 Seixal vs Out 35

19 Transferências - Geral 35

20 Transferências Precoces 35

4 Introdução

A exploração e análise de dados associados ao futebol e ao desporto têm gerado um crescente interesse. Quando pensamos em futebol, associamos a um jogo em específico ou a um momento de magia de um determinado jogador ou a felicidade da atuação do nosso clube.

Cada vez mais os clubes têm vontade de inovar em novas áreas, principalmente áreas mais tecnológicas, de modo a ganhar vantagem sobre os seus principais rivais. A relevância dos dados para os clubes são algo central para os seus departamentos e a área de Data Science assume uma relevância central. Existem exemplos de referência e casos de estudo em clubes europeus (por exemplo, o caso do Liverpool FC). O caso Liverpool FC acaba por ser um caso de sucesso, em que desde a chegada do seu treinador Jurgen Kloop, se iniciou um desenvolvimento grande em *Data Science*. Disto resultaram as vitórias da *Premier League* e da Liga dos Campeões, após 30 anos desde a última vez. O seu maior foco acabou por ser direcionado para três áreas, o jogo em sim, onde desenvolveram o conceito de *pitch control*, uma visualização que capta as regiões do espaço controladas por certos jogadores e quais os movimentos certos que devem ser feitos pelos mesmos. Na área de *scouting*, foram usados ingredientes e maneiras corretas para utilização dos dados para a construção de uma identidade de uma equipa. A maneira como o *Liverpool FC*, neste caso, sobressaiu neste campo, foi ter um sentido explícito e quantificável de como o treinador quer que a sua equipa jogue e, ao mesmo tempo, que esteja dentro do orçamento específico, a criação de modelos e métodos para fazer avaliações dos jogadores, compreender qual a dinâmica da equipa e criar modelos para a evolução da mesma com cada jogador adicional recrutado. Por fim, a outra área é a componente económica e também de negócio, que ajudou na tomada de decisão por contratos mais favoráveis para o clube.

A utilização de *Data Science* no mundo do futebol é, atualmente, um fenómeno em ascensão neste momento, querendo obter dados vantajosos em relação aos adversários.

Quando este assunto é mencionado, caímos no erro de pensar que só é utilizado para ajudar o clube de forma desportiva, de forma a ter melhores resultados em relação ao seu desempenho no campo, o que acaba por estar errado. *Data Science* insere-se em vários

departamentos, como por exemplo: *Technical Staff*, *Video Analyst*, *Nutritionist*, *Psychologist*, *Sport Science*, *Scouting* e *Academy*.

Em relação ao nosso país, Portugal é muito centrado no futebol, e segundo um estudo feito pelo programa UEFA GROW, em que se calcula o retorno social do investimento, em Portugal o impacto por atleta atinge os 7,4 mil euros, o que coloca Portugal em terceiro lugar. Este estudo visa compreender o impacto que a participação maciça no futebol tem em todo o espectro de resultados económicos, sanitários e sociais. Mas quantas pessoas conhecem o trabalho que é realizado nos bastidores por várias equipas em várias áreas para colocar o seu clube favorito no topo?

Este trabalho acaba por abordar temas e projetos que demonstram como se ajuda o clube a inovar e se desenvolver mais rápido.

O projeto que acabou por durar a maior parte do estágio e que foi devidamente descrito, acaba por estar mais ligado à área de monitorização de antigos jogadores e de possíveis transferências, onde seja também possível haver uma análise de alguns parâmetros.

"I always try to manage the ball in the best way in order to help my teammates to receive and use it"

Kevin De Bruyne

Ao se ler esta frase, pode-se associar De Bruyne como *Data Science*, a bola como os dados e os colegas de equipa como as várias áreas que utilizam esta tecnologia. A área de *Data Science* procura distribuir os dados da melhor forma e com a melhor qualidade possível para serem usados, o que se assemelha a este jogador quando ele está em campo a jogar.

Há que falar também no assunto da proteção de dados, é um assunto que nos últimos tempos tem vindo a ser muito abordado. Como é normal, o SLB tem dados confidenciais e valiosos, dados que nunca foram revelados neste trabalho. De referir que todos os dados que incorporado neste trabalho, são dados públicos e que poderão ser encontrados na internet.

5 Revisão da Literatura

5.1 Sports Analytics

De acordo com Elia Morgule, Ofer H. Azar e Ronnie Lidor, *Sports Analytics* é a investigação e modelação do desempenho desportivo, implementando técnicas científicas. Mais especificamente, à área de *Sports Analytics* refere-se à gestão de dados históricos estruturados, à aplicação de modelos analíticos preditivos que utilizam estes dados, e finalmente, à utilização de sistemas de informação, a fim de informar os decisores e permitir-lhes assistir as suas organizações na obtenção de uma vantagem competitiva no campo de jogo.

A análise e previsão do desporto através destes dados é um campo em rápido crescimento com muitos métodos que podem ser implementados de uma perspectiva diferente para cada situação [11]. Numa equipa, e especificamente para o pessoal técnico e treinadores, o conhecimento das vantagens e desvantagens de cada jogador pode proporcionar valor acrescentado na composição da equipa, em novas transferências, na mudança de ritmo durante um jogo e outros factores qualitativos e quantitativos vitais [14]. A referida Análise de Desempenho é extremamente valiosa para uma equipa, a fim de minimizar os custos orçamentais, maximizar o valor da equipa e melhorar os processos em todos os níveis e segmentos do fluxo [15].

Novas descobertas tecnológicas podem dar a oportunidade de recolher mais dados e exigir novos métodos de análise a serem realizados. Portanto, os novos métodos de análise podem explorar e gerar este valor acrescentado para definir o comportamento dos jogadores de futebol e ajudar o pessoal técnico e os treinadores na melhor tomada de decisões [26].

No seu livro intitulado "Sports Analytics: A Guide for Coaches, Managers, and Other Decision Makers"[1], Benjamin Alamar define o tema como "a gestão de dados históricos estruturados, a aplicação de modelos analíticos preditivos que utilizam esses dados, e a utilização de SI para informar os decisores e permitir-lhes ajudar as suas organizações a obter uma vantagem competitiva no campo do jogo". Utilizando ideias semelhantes, outros livros sobre os tópicos foram publicados durante os últimos cinco anos, tais como "Sport Strategist"de Rein et al [27]. e "Scorecasting: the hidden influences behind how sports are played and games are won"de Moscovitz e Wertheim [24].

Juntos, estes livros fornecem orientações gerais para organizações desportivas profissionais e gestores que procuram confiar na análise desportiva para melhorar o desempenho da sua organização e dos atletas através de uma melhor tomada de decisões.

Entre os mundos do desporto profissional e académico, o *MIT Sloan Sports Analytics Conference* foi estabelecido como um dos eventos mais importantes para a criação e disseminação do conhecimento sobre análise desportiva. Celebrando a sua 9ª edição em Fevereiro de 2015, a conferência baseia-se principalmente em painéis especializados durante os quais as figuras mais dominantes da análise desportiva de organizações desportivas profissionais partilham as suas crenças e experiências sobre iniciativas de análise desportiva. Os trabalhos de investigação e posters são também apresentados durante a conferência, na maioria das vezes por membros do corpo docente ou estudantes de pós-graduação de universidades da América do Norte.

Quanto aos artigos revistos por pares publicados em revistas académicas, o número permanece muito baixo em geral, e praticamente nulo em revistas académicas do SI. A maioria dos artigos publicados provêm de outras disciplinas tais como informática, marketing, matemática, economia e estatística. Em geral, os investigadores tentaram elaborar índices e indicadores-chave de desempenho que possam quantificar a produtividade de um atleta ou de uma equipa numa dada situação. Num dos raros artigos de investigação ligados ao BIA em desportos competitivos por um investigador do SI, Thomas Davenport [6] identifica uma série de potenciais benefícios associados à utilização de análises em desportos profissionais. Ele também fornece exemplos interessantes de aplicações actuais da análise por equipas desportivas profissionais e discute tendências futuras sobre este tema.

5.2 Data Science

Data Science é uma área interdisciplinar para permitir aos peritos em diferentes domínios estudar e trabalhar em conjunto [3]. *Data Science* é um estudo dos dados que envolveu o processamento, análise, interpretação e sentido dos dados [12].

As empresas podem compreender os seus problemas, o seu desempenho empresarial (diário, semanal, mensal e anual) e a previsão do seu desempenho empresarial em questão de minutos a qualquer momento [28].

Em primeiro lugar, *Data Science* permite às empresas recolher e analisar dados sobre as suas operações comerciais, estratégias e desempenho global [22]. Em segundo lugar, as empresas podem melhorar os seus serviços, operações, estratégias e desempenho empresarial com base nos resultados da análise [25]. Em terceiro lugar, as empresas podem melhorar a qualidade da sua modelação preditiva, os decisores da sothat podem planear estratégias adequadas para as suas organizações [8]. Há três grandes benefícios em fazê-lo, no entanto, as formas de executar a *Business Data Science* (BDS) ainda não estão estabelecidas, uma vez que a literatura existente não tem uma orientação conclusiva ou uma abordagem sumária das melhores práticas. Embora haja muitas organizações que se tenham interessado em *Data Science*, elas não sabem como operar e gerir grande volume de dados [22].

Os dados no vácuo são efectivamente inúteis [10], o que requer a capacidade de empregar a capacidade de processamento de informação conhecida como *Data Science* para transformar dados, centrado na aplicação de métodos quantitativos e qualitativos para fornecer conhecimentos a partir de várias formas de dados e resolver problemas relevantes [19]. O interesse pelos dados continua a crescer com o aparecimento de iniciativas na indústria e uma variedade de novas ofertas de programas académicos [29]. *Data Science* depende fortemente das capacidades de programação e análise, no entanto, o conhecimento robusto do domínio é igualmente importante [35]. Como ilustrado em [30], "... a obtenção de um bom algoritmo envolve tanto o código como o contexto, uma mistura dos pontos fortes complementares dos cientistas informáticos e humanistas".

Embora simplificada, esta figura indica que para compreender o impacto da *Data Science*, a investigação deve ilustrar como as competências de programação e de análise se integram com os conhecimentos de domínio para fornecer soluções viáveis a problemas bem definidos. A aplicação de tais análises a problemas actuais não é nova, de facto, a aplicação do OR teve a data do início da Segunda Guerra Mundial [18]. No entanto, o que é novo é um aumento significativo que a programação informática está a jogar na análise moderna. Afirmado por Davenport e Patil [7], a "habilidade mais básica e universal dos cientistas de dados é a capacidade de escrever código". Como tal, isto tem impulsionado avanços nos tipos de técnicas analíticas aplicadas (i.e. aprendizagem de máquina, processamento de linguagem natural), bem como na forma como o software é alavancado para produzir produtos analíticos,

proporcionando às organizações maiores capacidades analíticas.

5.3 Web Scraping

Segundo Ryan Mitchell, *Web Scraping* consiste na prática de recolher dados através de qualquer meio que não seja um programa que interage com um API. Isto é mais comum recorrendo um programa automático que consulta um servidor web, solicita dados (geralmente sob a forma de HTML e outros ficheiros que compõem páginas web), e depois analisa esses dados para extrair a informação necessária [23].

Dado o volume, variedade, velocidade e veracidade da grande disponibilidade de dados disponíveis na Web, a recolha e organização destes dados dificilmente pode ser feita manualmente por investigadores individuais ou mesmo por grandes equipas de investigação [16]. Devido a isto, os investigadores recorrem frequentemente a várias tecnologias e ferramentas para automatizar alguns aspectos da recolha e organização dos dados. Esta prática emergente de utilização de tecnologia para a recolha de dados da Web é frequentemente referida como *Web Scraping* [17].

O *Web Scraping* envolve o desenvolvimento e execução de um guião que navega automaticamente no website e recupera os dados necessários. Estas aplicações de rastreio são frequentemente desenvolvidas utilizando linguagens de programação como o *R* e *Python*. Isto tem a ver com a popularidade geral destas linguagens nas bibliotecas de Data Science e de disponibilidade (por exemplo, pacote *rvest* em *R* ou *Beautiful Soup library* em *Python*) para o rastreio automático e análise de dados da Web.

Os autores de [21] discutem como extrair informação usando mecanismos de extracção a partir de dados não estruturados e de informação adquirida de sítios na web. A extracção automática de dados do HTML do sítio web através da técnica de análise das páginas web utiliza programas especialmente codificados e converte-os noutro formato. O *Web Scraping* analisa e armazena os dados da web em forma estruturada numa base de dados/folha de dados central.

A arquitectura de *scraping* baseada na nuvem para a aquisição de dados não estruturados da web é proposta em [4] utilizando bibliotecas como *BeautifulSoup*, *Scrapy*, *Selenium*, *web driver API*, *HTML Parser library of Python*. *Selenium* e as ferramentas API do driver

da web são aqui seleccionadas para automatizar a extracção de dados da página web. A instância de máquina virtual baseada em *Elastic compute cloud* (EC2) dos serviços web da *Amazon* é utilizada para a implementação desta arquitectura de eliminação de dados baseada em nuvem.

O *Web Crawler* geralmente rasteja o website a partir da primeira página para identificação de ligações em cada página, que é armazenada como estrutura de dados que é utilizada para abrir a respectiva página web repete recursivamente o processo, até que todas as ligações sejam rastreadas. O Extractor de Dados extrai a informação útil e converte-a no formato necessário [20].

Os autores de [13] apresentaram técnicas de raspagem da web para recolher *tweets* históricos dentro de qualquer intervalo de datas utilizando técnicas de raspagem da web contornando as restrições API do *Twitter*. As etiquetas de hipertexto são utilizadas para recuperar informação de texto em quadradinhos. Raspagem de campos de pesquisa no *Twitter* e personalização de consultas, a fim de alargar as capacidades de pesquisa para recolher o histórico dos *tweets* dentro de um intervalo de datas especificado utilizando *Scrapy*, uma estrutura de código aberto para extrair dados de websites escritos em *Python*.

Web Scraping com Naïve Bayes Classification é utilizado para motor de busca de emprego em quatro processos [31]. (1) Criação de modelo de *web scraping*: inserção de modelo de desmantelamento através da definição de documentos HTML a partir de sítios web cuja informação é recolhida. (2) explorando a navegação do sítio, tornando o sistema de exploração de navegação do sítio a partir de sítios cuja informação é recolhida (3) automatizando a navegação e extracção: a partir dos processos número 1 e 2, é realizada a automatização a partir dos dados e da informação adquirida a partir dos sítios web (4) extraindo dados e histórico de pacotes: a informação adquirida a partir do processo número 3 é guardada em tabelas e base de dados.

A tecnologia de *Web Scraping* utilizada em [9] recolhe dados meteorológicos em tempo real de vários sítios Web e fornece dados meteorológicos actualizados em linha para formar um conjunto de dados meteorológicos. NewsOne - Um sistema de agregação de notícias na plataforma [32] - utiliza o método de *Web Scraping* para extrair o conteúdo de vários sítios web de notícias. Agrega todas as últimas actualizações de notícias de vários recursos

nacionais e internacionais e resume-as para apresentar em palavras curtas e nítidas. Fornece uma interacção orientada para o serviço entre os utilizadores de toda a web.

Sistema automático adaptável de *Web Scraping* extrai informação de páginas web que consistem em blocos repetitivos. Cada bloco representa um objecto de oferta de produto e contém os seus atributos, tais como título da oferta, descrição da oferta, data expiração da oferta, etc., utilizando uma nova abordagem baseada na classificação [33]. Um sistema automático de *Web Scraping*, que se adapta às mudanças estruturais é dado em [5] extraindo informação de páginas web que consistem em blocos repetitivos. A estrutura de *Web Scraping* de [34] oferece uma abordagem fácil e viável através da análise e extracção de dados em grande escala de múltiplos sítios web com intervenção humana mínima para a colheita de objectos de aprendizagem para uma aplicação de *e-Learning*. Conclui-se que o *Web Scraping* é uma técnica muito eficiente para extrair dados de diferentes websites que, além disso, podem ser utilizados para vários fins. A técnica de *Web Scraping* é barata, fácil de implementar, baixa manutenção, maior velocidade e precisão.

5.4 Business Intelligence

Os projectos de *Business Intelligence* desenvolveram uma reputação de serem difíceis, arriscados e caros. O termo *Business Intelligence* (BI) refere-se a tecnologias, aplicações e práticas para a recolha, integração, análise e apresentação de informação empresarial, o que torna o projecto de *Business intelligence* uma mistura de projectos de TI, *Business* e *Analytics*. Segundo estudo realizado pelo grupo Atre, uma grande empresa de serviços de gestão de dados e *Business intelligence*, existem muitas razões para uma elevada taxa de fracasso, a maior é que as empresas tratam os projectos de *Business intelligence* como mais um projecto de TI. No entanto, o *Business intelligence* não é nem um produto nem um sistema, é antes uma estratégia, visão e arquitectura em constante evolução que procura continuamente alinhar as operações e a direcção de uma organização com os seus objectivos estratégicos de negócio. Não há tamanho único quando se trata de metodologias de gestão de projectos de *Business intelligence*.

Power BI é uma evolução dos add-ins anteriormente disponíveis em Excel: *Power Pivot*, *Power Query* e *Power View*. A utilização de *Power BI* com ou sem integração de Excel já

não depende da versão do Microsoft Office instalada na própria organização. Os analistas de dados procuraram simplicidade, novas visualizações e tudo isto está agora disponível no *Power BI*.

6 Análise ao Sport Lisboa e Benfica

Figura 1: Símbolo do Sport Lisboa e Benfica

"Sport Lisboa e Benfica foi fundado a 28 de Fevereiro de 1904, sob o nome Sport Lisboa. Após uma sessão de treino matinal em alguns terrenos em Belém, realizou-se uma reunião à tarde na vizinha Farmácia Franco, à qual assistiram 24 elementos, incluindo os dez da sessão de treino matinal. Eles são considerados os fundadores do Clube"[2].

Tanto a nível desportivo, tanto a nível financeiro na última década, considera-se que o Benfica esteve a um grande nível.

A nível desportivo chegou a um tetra campeonato, chegou a 2 finais europeias, chegou longe vários anos na Liga dos Campeões e ganhou várias taças internas.

A nível financeiro vendeu vários jogadores jovens *Made in* Seixal e outros que evoluíram durante anos no Benfica, fazendo um grande lucro com dezenas de vendas. Para além do número elevado dos valores dessas transferências, fez o Seixal ganhar fama de desenvolvimento de jovens jogadores, fazendo os principais clubes na Europa ficar de olho nas pérolas do Benfica.

Falando dos temas abordados no capítulo anterior, o Benfica é um clube que aos longo de cada época continua a implementar uma estratégia de inovação, digitalização e renovação nas áreas de Tecnologias e Sistemas de Informação.

No relatório de contas de 2020/21, é descrito que na parte de inovação, pretende dispo-

nibilizar ao Grupo e aos seus associados, um conjunto de soluções que contribuem de forma evidente para a simplificação de processos e melhoria contínua de níveis de serviços. Na parte de digitalização, a procura por uma adaptação de métodos de trabalho e novas opções no Comercial e Marketing. Já na parte de renovação, o investimento e a reestruturação de diversos sistemas e infraestruturas críticas para o desempenho desportivo e empresarial/corporativo do Sport Lisboa e Benfica.

Dentro desta estratégia existem 6 projetos:

- A *Player APP* que desempenha um papel fundamental na difusão da informação oficial dentro das equipas, permitindo um contacto próximo entre a equipa técnica e os jogadores;
- No âmbito das soluções de gestão e processos empresariais foi mantido o investimento no *Robotic Process Automation* (RPA), com uma maior adoção nas áreas financeira e de recursos humanos;
- A solução de *reporting* corporativo (RedBI) continuou a sua evolução com a inclusão de novas análises para as áreas de negócio;
- A plataforma de gestão de serviço de IT (*OnePoint*) assegurará ao Benfica a normalização de processos de IT, a gestão centralizada de informação, tal com a centralização de suporte sobre todos os utilizadores, fornecedores e prestadores de serviço;
- A implementação do Benfica *Smart Media Center* (SMC), que tem como principal objetivo a criação de uma plataforma de suporte de gestão e manutenção de conteúdos audiovisuais de forma transversal a todas áreas funcionais do Benfica;
- A infraestrutura de rede e segurança foi reforçada com a renovação de um conjunto de sistemas, aumentando assim a resiliência dos ambientes.

De referir que as soluções tecnológicas continuam a ser uma prioridade no campo desportivo. E a verdade é que o investimento que o clube tem efetuado nesta vertente, tornou a dar frutos com a atribuição de dois prémios esta época:

- Galardão Cosme Damião, na vertente de Inovação, à *Player APP*;

- 1ºLugar internacional, no *MIT Sloan Sports Analytics*, para área de *Sports Data Science*;

Na figura 2 é mostrado uma comparação de os resultados líquidos entre o 1ºsemestre da época 20/21 com a época anterior 19/20.

valores em milhares de euros

	1.º Sem. 20/21 6 meses	1.º Sem. 19/20 6 meses	Variação	%
Resultado líquido do período	8.232	104.153	(95.921)	-92,1%
Resultado operacional (incluindo transações de direitos de atletas)	12.810	116.687	(103.877)	-89,0%
Rendimentos operacionais (excluindo transações de direitos de atletas)	53.546	101.923	(48.377)	-47,5%
Rendimentos com transações de direitos de atletas	77.508	137.033	(59.525)	-43,4%
Rendimentos totais	134.891	244.294	(109.403)	-44,8%

Figura 2: Comparação de resultados líquidos entre 19/20 e 20/21

Desta figura podemos concluir que:

- O resultado líquido no 1.o semestre de 2020/21 ascende a um valor positivo de 8,2 milhões de euros, o que significa que apesar da não presença na fase de grupos da Liga dos Campeões e dos impactos associados à COVID-19, a Benfica SAD conseguiu fechar o semestre com um resultado positivo, correspondendo ao sétimo ano consecutivo em que apresenta lucro nos primeiros seis meses de atividade;

- O resultado operacional ascende a 12,8 milhões de euros, mantendo um valor positivo para o qual muito contribuiu a transferência do jogador Rúben Dias para o *Manchester City*, que apesar de não ter atingido os valores da alienação dos direitos do jogador João Félix no período homólogo, representou uma mais valia significativa;

- Os rendimentos operacionais sem transações de direitos de atleta atingem os 53,5 milhões de euros, o que significa um decréscimo de 47,5% face ao período homólogo,

justificado essencialmente pela inexistência de receitas com *match-day* devido à realização de jogos sem público e pela redução dos rendimentos com prémios distribuídos pela UEFA;

- Os rendimentos com transações de direitos de atletas correspondem a 77,5 milhões de euros e o resultado com transações de direitos de atletas ascende a 69,7 milhões de euros, estando ambos significativamente influenciados pela transferência do jogador Rúben Dias para o Manchester City, à semelhança do semestre homólogo, em que ambos refletiam a alienação dos direitos do jogador João Félix para o Atlético de Madrid;

- Os rendimentos totais no semestre ascendem a 134,9 milhões de euros, o que representa um decréscimo de 44,8% face ao período homólogo, mas que correspondem ao segundo melhor semestre de sempre em termos de rendimentos totais obtidos pela Sociedade SLB.

6.1 Missão, Visão e Valores

6.1.1 Missão

A missão do Grupo Benfica é ser a organização desportiva de maior sucesso em Portugal, tanto no Futebol como noutros desportos, e tanto em termos competitivos como económicos.

6.1.2 Visão

Num dos maiores clubes do país vivemos num ambiente desportivo competitivo, acompanhado das exigências esperadas num dos maiores clubes do mundo. Um anseia e luta para ir mais longe, e o mais distante não pode deixar de ser o "mais alto"nível. Cada jogador, atleta, ou colaborador tem de dar o seu melhor na sua respectiva área. Estas são as premissas que os impulsionam, que os fazem, todos os dias, trabalhar arduamente para alcançar os resultados que o SLB espera alcançar.

6.1.3 Valores

As acções de todos os profissionais do Grupo Benfica devem ser sempre orientadas pela lealdade à organização, sendo honestas, independentes, imparciais, discretas, e não atendendo a quaisquer interesses pessoais.

6.2 Departamentos de Software Development e Data Science do SLB

Neste estágio fui integrado em duas equipas simultaneamente, uma de *Software Development* e a outra de *Sports Data Science*. Eram duas equipas que tinham métodos diferentes uma da outra.

Na equipa *Software Development* é utilizada a metodologia SCRUM, onde se realiza uma reunião diária onde cada um diz o que fez no dia anterior e o que vai fazer no dia, de modo que todos na equipa sabem para onde vai cada um, o que se revela importante porque muitos dependem do trabalho de outro colega. A utilização desta metodologia acaba por ser benéfica para esta equipa, porque o seu objectivo é desenvolver uma plataforma de apoio para toda a estrutura do clube.

A equipa *Sports Data Science* foi criada em 2017 e faz parte do Departamento de Sistemas de Informação do SLB. A sua principal tarefa é a análise e processamento de dados desportivos para ajudar na tomada de decisões dos vários departamentos presentes na estrutura do futebol profissional, treino de futebol e desportos de pavilhão. Este trabalho começa no processo de engenharia de dados, com a extração e manipulação de dados de várias fontes que são posteriormente disponibilizados numa única base de dados. Depois disso, o produto final é a automatização de processos relacionados com esses dados, poupando tempo aos colegas de outros departamentos para se concentrarem noutras tarefas, criação de painéis de controlo em várias áreas do clube ou criação de novos conhecimentos através da utilização de modelos de aprendizagem de máquina ou outras análises matemáticas mais complexas. As suas áreas de intervenção estendem-se a toda a estrutura desportiva do clube: Análise Técnico-Táctica (Futebol e Modalidades), *Scouting* (Futebol), Desempenho Humano (Futebol, Treino e Modalidades) e Gestão de Talento (Futebol).

7 Caracterização do estágio

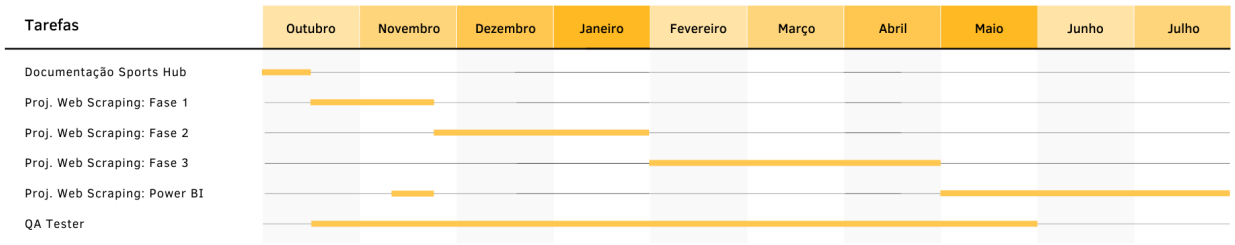

Figura 3: Mapa com principais projetos do estágio

7.1 Projeto Web Scraping

7.1.1 As Fases

Este projeto acabou por se expandir durante todo o estágio, uma vez que a finalização de cada sprint gerava novas iterações com os diferentes feedbacks obtidos.

- **Outubro-Novembro**: Nesta 1ª fase, o principal objetivo era recolher informação variada de todos os jogadores que representaram o SLB entre os seus 12 a 23 anos e assim calcular a percentagem do mecanismo de solidariedade da FIFA.

- **Novembro-Janeiro**: Nesta 2ª fase, o objetivo foi passar de apenas jogadores que representaram o SLB entre as idades dos 12 aos 23 anos para todos os jogadores que representaram o clube e já sairam. Para além disso recolher informação de época a época de cada jogador, como a variação ao longos dos anos do valor de mercado do jogador. Por fim, tornar o algoritmo universal, isto é, gerar a possibilidade de o mesmo ser aplicado ao Benfica, mas analisar qualquer clube.

- **Fevereiro-Abril**: Nesta 3ª fase,.o objetivo foi correr o algoritmo para os grandes rivais do SLB, analisar se os dados estavam todos coerentes e juntar os dados que eu recolhi com os dados que o Benfica tem na sua volumosa base de dados.

- **Maio-Julho**: Nesta 4ª e última fase, o objetivo foi recolher mais dados sobre os atletas nas seleções nacionais e com todos os dados recolhidos criar um *dashboard* em Power BI para responder a várias questões que o clube tinha.

7.1.2 Descrição

Antes de passar à construção do algoritmo, foi necessário descobrir quais eram os dados fundamentais para responder a perguntas-chave, e a partir daí, que ferramentas precisava para exportar esses mesmos dados. As questões-chave que acabaram por ser respondidas neste documento foram:

- Os jogadores formados em clubes são rentáveis no futuro através do mecanismo de solidariedade da FIFA e do seu valor de mercado?
- Existe uma grande diferença entre jogadores formados pelo clube e jogadores que vêm de fora, ao nível do valor de mercado?
- Para os jogadores que permanecem no clube algum tempo, qual é a melhor faixa etária para um jogador estar no clube ? Quando chegam ao clube numa idade muito mais jovem ou quando ele foi recrutado numa idade mais avançada ?
- Os jogadores formados no clube acabam por chegar ao topo das ligas ou ficam abaixo do nível?

Os dados chave para responder a estas perguntas são:

- Em que ano entraram no clube;
- Em que ano saíram do clube;
- Que equipa foi a sua primeira equipa no clube;
- Qual era a sua equipa quando deixaram o clube;
- Para que liga foram depois;
- Quantas transferências tiveram desde que deixaram o clube;

- Valor de mercado no momento da saída do clube;
- Valor de mercado actualmente;
- País e região em que nasceram;
- Posição;
- Percentagem do mecanismo de solidariedade da FIFA.

Com estes dados fundamentais foram então extraídos mais dados para complementar esta investigação, este diagrama que podemos ver na figura 4.

Este projecto é enquadrado num método ETL que contém 3 fases: a parte de extracção, a parte de transformação e a parte de carregamento dos dados.

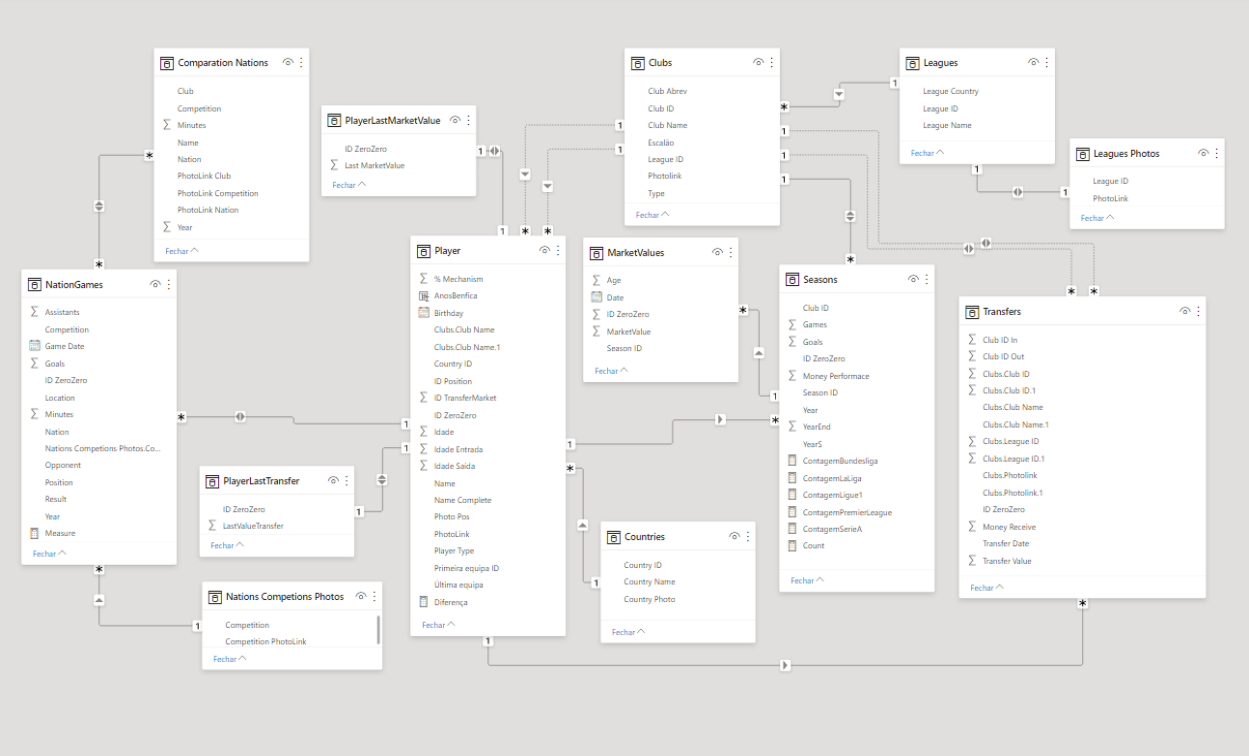

Figura 4: Diagrama dos dados

7.1.3 A prática

7.1.3.1 Extração de dados

O *Web Scraping* pode facilmente recolher dados a partir de um website. Neste projecto, a escolha recaiu sobre dois websites, *ZeroZero* e *Transfermarkt*, porque são dois websites que fornecem muita informação que é importante para este projecto. Informação que está actualizada e correcta, pelo que são duas boas fontes de dados. Na figura 5, pode-se ver um resumo do que acontece, com o *Web Scraping*, onde no fim os dados vão para o ficheiro excel, neste caso.

Figura 5: Diagrama do método

Para começar a exportar dados através do *Web Scraping* usando *python* é necessário instalar e importar duas bibliotecas, neste caso *BeautifulSoup* e *Requests*, através destas duas é possível ligar através do algoritmo *python* a um qualquer sítio web.

BeautifulSoup é uma biblioteca que facilita a raspagem de informação de páginas web. Senta-se no topo de um analisador HTML ou XML, fornecendo expressões pythonicas para iterar, pesquisar e modificar a árvore parse.

Os pedidos fornecer-nos-ão o HTML do nosso alvo, e o *BeautifulSoup* analisará esses dados.

```
1 import requests
2 from bs4 import BeautifulSoup
```

Temos de reconhecer que muitos sítios têm precauções para evitar que os raspadores acedam aos seus dados. A primeira coisa que podemos fazer para contornar isto é falsificar os cabeçalhos que enviamos juntamente com os nossos pedidos para fazer com que o nosso raspador pareça um navegador legítimo:

```
1 headers = {
```

```
    'Access-Control-Allow-Origin': '*',
    'Access-Control-Allow-Methods': 'GET',
    'Access-Control-Allow-Headers': 'Content-Type',
    'Access-Control-Max-Age': '3600',
    'User-Agent': 'Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (
   KHTML, like Gecko) Chrome/47.0.2526.106 Safari/537.36'
}
```

Esta é apenas uma primeira linha de defesa (ou ofensiva, no nosso caso). Há muitas maneiras de os sítios ainda nos poderem manter à distância, mas definir cabeçalhos funciona chocantemente bem para resolver a maioria dos problemas.

Agora vamos buscar uma página e inspeccioná-la com *BeautifulSoup*.

```
page = "https://www.zerozero.pt/team_transfers.php?id=4&epoca_id=150"
pageTree = requests.get(page, headers=headers)
pageSoup = BeautifulSoup(pageTree.content, 'html.parser')
```

Prepara-se as coisas através de um pedido a http://example.com. Cria-se então um objecto *BeautifulSoup* que aceita o conteúdo bruto dessa resposta via req.content. O segundo parâmetro, 'html.parser', é a nossa forma de dizer ao *BeautifulSoup* que este é um documento HTML. Há outros analisadores disponíveis para analisar coisas como XML.

Quando cria-se um objecto *BeautifulSoup* a partir do HTML de uma página, o nosso objecto contém a estrutura HTML dessa página, que pode agora ser facilmente analisado por todos os tipos de métodos.

```
rows = pageSoup.find_all("table",{"class":"zztablestats"})[1].find("tbody
   ").find_all(tr")
transfers = []
linksPlayers = []
for i in rows:
    li = []
    b = i.find_all("td")
    li.append(dataTrans)
    li.append(team)
    for j in b:
        link = j.find("div",{"class":"text"})
        if link:
```

```
            link = j.find("div",{"class":"text"}).find("a").get("href")
            link = "https://www.zerozero.pt/" + link
            if (link not in linksPlayers):
                linksPlayers.append(link)
        li.append(j.text)
        li.append(0)
    transfers.append(li)
```

Assim, dentro da variável *pageSoup* temos todo o código HTML, ou seja, todos os dados que queremos obter. Dito isto, temos duas formas de os extrair com *.find()* e *.find()_all()*

No excerto de código anterior, pode ver os vários usos de *.find()* e *.find_all()*, com vários a serem unidos dentro de uma variável.

Por exemplo, utiliza-se *.find_all("table",{"class": "zztablestats"})* porque dentro desta página existem várias tabelas com esta classe, continuando, acaba-se por escolher a que tem o índice 1. Dentro desta tabela, existe apenas um *tbody*, ou seja, só precisa de usar *.find("tbody")* e dentro do tbody existem várias linhas ("tr"). Assim, as linhas de variáveis acabam por ser uma lista com todos os "tr"na tabela. Note-se que os dados contidos nesta tabela referem-se às transferências que ocorreram numa determinada equipa, numa determinada época. Neste caso, é a equipa com id=4 e epoca_id = 150. Tendo então as linhas variáveis como uma lista, acabamos por iterar esta lista com um *for* e daí extraímos o link para a página do jogador e também a transferência que ocorreu. Neste caso não temos muita informação sobre a transferência, ela é guardada apenas no caso do jogador não ter uma página no *Transfermarkt*. No final, terminamos com uma lista de *links* dos jogadores em que iteramos com um *for*, onde entramos na página de cada jogador.

O website *ZeroZero* coloca um desafio a todos aqueles que pretendem utilizar um *script*. Este website tem um sistema anti-escrito, que detecta se existe um padrão de entradas no website constantemente e em grande número. Assim, se o sistema for activado, o sítio web fica bloqueado durante 20 a 30 minutos. Portanto, é necessário utilizar formas de contornar este sistema.

A forma utilizada neste algoritmo foi a de utilizar um *time.sleep()*.

```
for i in range(len(linksPlayers)):
    time.sleep(180)
```

Ao adicionar esta linha de código cada vez que se itera através da lista de ligações do leitor, o algoritmo faz uma pausa de 180 segundos, ou 3 minutos. Houve várias tentativas para baixar o valor dos segundos, mas concluiu-se que abaixo deste valor o *anti-script* irá eventualmente detectá-lo. Pode-se dizer que isto acaba por ser o maior problema porque nos grandes clubes, o número de jogadores é muito elevado e se a lista tiver 600 jogadores, o algoritmo leva 30 horas a correr.

No caso do site *transfermarkt*, não existe um sistema *anti-script*, pelo que não é necessário desperdiçar mais segundos num *time.sleep()*.

Depois desta explicação do que realmente é o *web scraping* e a maneira de o implementá-lo passamos a uma análise mais detalhada dos vários passos deste algoritmo.

Para começar recolhi informações a nível de Ids e nomes do respectivos clubes que eu queria analisar, neste caso seria o SLB e os seus grandes rivais, o Sporting Clube de Portugal e o Futebol Clube do Porto.

```
1 listaEquipasZeroZero = ['Benfica','Porto','Sporting']
2 listaEquipasTransferMarket = ['Sport Lisboa e Benfica','Futebol Clube do
     Porto','Sporting Clube de Portugal']
3 listaEquipasTransferenciasTM = ['SL Benfica','FC Porto','Sporting CP']
4 listaEquipasBs = ['SL Benfica B','FC Porto B','Sporting B']
5 listaIdEquipasZeroZero =
     [['4','3563','228830','6444','6888','25799','7083','25800','64956'],
6 ['6869'],['16']]
```

Um problema encontrado e que acabou por ser resolvido com a criação de várias listas foi o nome dos clubes entre os websites e mesmo em partes diferentes do website iam mudando.

```
1 def equipa(indexEquipa, indexindex, id_season):
2     page = "https://www.zerozero.pt/team_transfers.php?id=" + str(
     listaIdEquipasZeroZero[indexEquipa][indexindex])+ "&epoca_id=" +
     id_season
3     pageTree = requests.get(page, headers=headers)
4     pageSoup = BeautifulSoup(pageTree.content, 'html.parser')
```

Assim é criada a função equipa() que recebe três parâmetros, o *index* da equipa nas listas criadas, ou seja, obriga a quando se for correr o algoritmo na função *main* saber qual o *index* da equipa que queremos analisar, o parâmetro a seguir diz respeito ao numero de id's de

equipas que o clube tem, onde vai receber um número e a função vai buscar o valor daquele *index*, por último este parâmetro diz respeito à época que estamos analisar, sendo que esta análise diz respeito às épocas desde 2020/21 a 2010/11.

Figura 6: Página das transferências

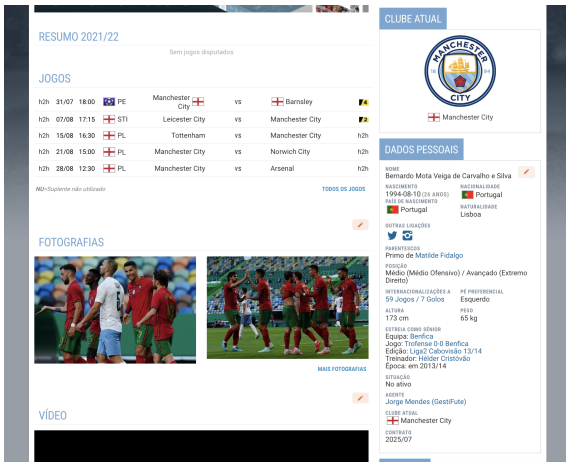

Figura 7: Página do jogador

De seguida, como podemos ver o exemplo da figura 6, esta página contem as entradas e saídas de jogadores daquela equipa na época x, neste caso a época 2021/22 (id=151). Desta página pode-se recolher 2 diferentes dados. O primeiro conjunto são os links de todos os jogadores que se encontram na lista de saída, para no decorrer do algoritmo usar essa lista de links. O segundo conjunto é registar a transferência do jogador para caso o jogador não ter página no *Transfermarket* já haver alguma informação com que se pode ficar.

De seguida com toda a lista de todos os links dos jogadores, iterou-se essa mesma lista.

Começou-se por recolher a informação básica do jogador, temos na figura 7 um exemplo de como é a página de onde recolhi a informação, onde na lateral direita contêm nome, nome completo, data de nascimento, nacionalidade, posição, qual a situação contratual do jogador, se está emprestado pelo Benfica ou se está em definitivo em outra equipa, pois se tivesse emprestado, já não entraria para esta extração de dados.

Caso o jogador não tivesse emprestado, passaria para uma nova página, uma página importante, pois tinha informação de época a época do jogador.

Esta página contêm então informação de quatro momentos do jogador, as épocas do jogador em clubes seniores, em clubes da formação, na seleção A e em seleções pré-sénior. O que acabou por ser uma problema que acabou por ser resolvido facilmente, pois no caso da

figura 8, onde temos o exemplo do jogador Bernardo Silva, como é um jogador de renome internacional acaba por ter os 4 momentos, mas jogadores que não chegaram a um nível elevado podem ter falta de momentos, ou jogadores estrangeiros em que há uma falta de épocas de formação.

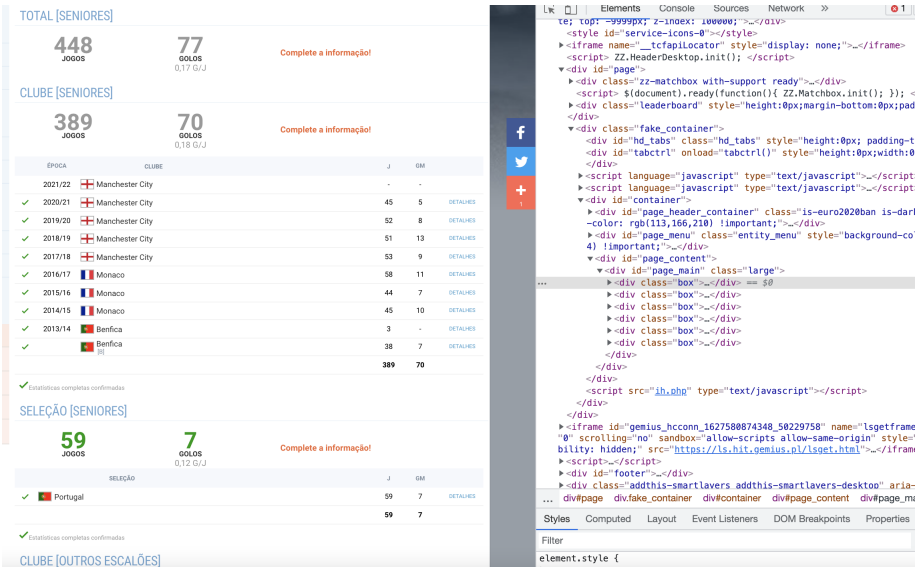

Figura 8: Página época a época com código HTML

Na figura 8 mostra tanto o aspeto da página como o código HTML que tive de utilizar para extrair os dados necessários. Como explicado anteriormente cada momento do jogador está inserido num *div* com a classe *"box"*, assim jogadores que contêm todos os momentos terão um número elevado de estes div's.

Foi dessa maneira que resolvi o problema de jogadores que não tem tanta informação como outros, contar quantos *div's* com a classe *"box"* existem no código. Caso o código conte-se mais de 4, dava a saber que este jogador tinha a informação completa desde épocas em clubes de formação, em clubes profissionais, em seleções nacionais.

Resolvido essa questão, apareceu um novo problema, caso o jogador na mesma época tiver jogado por dois ou mais clubes diferentes essa época terá esse número de entrada, e como podemos ver na figura 8, há uma alínea que não contem o ano da época pois é o mesmo valor da alínea acima. Assim o algoritmo em relação a esta situação está desenvolvido para os diferentes tipos de jogadores, ou seja, consoante a quantidade de *div's* com a classe *"box"*. Depois dessa fase há uma observação sobre o título de cada segmento para assim saber sobre

qual o momento do jogador em que se está a extrair informação, pois há uma distinção entre as épocas em clubes e seleções.

De seguida, o algoritmo coloca em uma lista todas as linhas que encontrou na tabela e há uma iteração dessa mesma lista. Para cada iteração, há uma análise do valor correspondente ao ano da época. Caso o valor não estiver vazio, há uma atribuição desse valor à variável *"season"* e há variável "chave"o valor 0. Caso o valor estiver vazio e o valor da variável "chave"for 0, vai-se buscar as variáveis *"lastSeasonFirstYear"* e *"lastSeasonSecondYear"* e em novas variáveis incrementa-se 1. Assim tem-se o valor da época que estava por conhecer. Caso o valor estiver vazio e variável "chave"for diferente de 0, isto quer dizer que há mais de uma linha seguida com o valor da *"season"* vazio, acaba-se por fazer o mesmo do que em cima, mas sem o incremento de 1.

Passado esta parte do valor da *"season"*, extraiu-se o nome do clube onde o jogador representou e o link do website *zerozero* desse clube. Por fim extraiu-se também os jogos e golos.

Seguidamente, com estes valores recolhidos, dentro de uma lista colocou-se o id do *zerozero* do jogador, os anos da época, o clube, os jogos e os golos, e logo de seguida fez-se *append* dessa lista para a lista final de *seasons*. Para além disso, fez-se *append* do link do clube no *zerozero* na lista *clubsLinks*, onde mais tarde foi utilizada.

Após a conclusão do recolher as épocas do jogador, tanto nos clubes, como nas seleções, entendeu-se que já era possível calcular um valor que era dos principais objetivos no inicío do projeto, o valor do mecanismo de solidariedade da FIFA.

Este mecanismo de solidariedade da FIFA é uma compensação para os clubes formadores. Dos 12 aos 15 anos, cada ano que o jogador está inscrito no clube acresce 0,25%. Dos 16 aos 23 anos, acresce 0,5%. O que no total dá os 5% totais. Este mecanismo foi criado no inicio dos anos 2000 com o objetivo de incentivar os clubes a formarem atletas.

```
finalYears = []
for i in range(len(agesLista)-1):
    if (agesLista[i+1]):
        if (agesLista[i][2]==agesLista[i+1][2]):
            if (agesLista[i][3] > agesLista[i+1][3]):
                if (listaEquipasZeroZero[indexEquipa] in agesLista[i][1]):
```

```
                    if (agesLista[i][0] not in finalYears):
                        finalYears.append(agesLista[i][0])
            else:
                if (listaEquipasZeroZero[indexEquipa] in agesLista[i
    +1][1]):
                    if (agesLista[i+1][0] not in finalYears):
                            finalYears.append(agesLista[i+1][0])
        else:
            if (listaEquipasZeroZero[indexEquipa] in agesLista[i][1]):
                if (agesLista[i][0] not in finalYears):
                    finalYears.append(agesLista[i][0])

value = 0
idades = [12,13,14,15,16,17,18,19,20,21,22,23]
for i in finalYears:
    if i in idades:
        if (i == 12 or i == 13 or i == 14 or i == 15):
            value = value + 0.25
        else:
            value = value + 0.5
```

No excerto de código anterior demonstra como foi calculado o valor. Houve uma iteração à lista com épocas do jogador, onde se analisou que clube o jogador representou nessa época. Dessa análise juntou-se as idades que o jogador tinha quando representou o SLB. Por fim, como já foi explicado anteriormente, o intervalo de idades do mecanismo é entre os 12 aos 23 anos, onde dos 12 aos 15 o valor é de 0.25 por ano e dos 16 aos 23 o valor é de 0.5. Este excerto de código também acabou por ser utilizado para definir o tipo de jogador, se era um jogador *Made in* Seixal ou se era um jogador vindo de fora.

Depois desta fase de análise e extração de dados do website *zezozero*, ainda com o registo do jogador, começa-se a extração no website *transfermarkt*.

A maneira utilizada para pesquisar o jogador no website do *transfermarkt* foi pelo nome completo, onde acabou por ter uma boa taxa de sucesso, sem contar com os jogadores que já desistiram desta modalidade ou se encontram em escalões amadores, estes acabam por nunca ter uma página no *transfermarkt*. Em casos de insucesso pelo nome do completo, fui

então para o nome e apelido. Para haver mais algum parametro de comparação, foi então feito essa mesma comparação com a data de nascimento que já tinha guardado do website do *zerozero* com esta nova que aparecia num website diferente, caso desse falso, e se houvesse mais opções de jogadores, o algoritmo continuaria a analisar as poucas opções que o website lhe deu, caso não encontrasse nada acabaria só por juntar às listas finais a informação do *zerozero* do jogador, caso a comparação de datas de nascimento desse verdadeiro, o algoritmo começa por perceber se o jogador ainda pertence ao SLB, ou seja, se está emprestado a um clube. Só caso o jogador já não pertencer aos quadros do clube é que o algoritmo avança para uma nova fase, se o jogador estiver emprestado toda a informação anteriormente recolhida não é guardada e o algoritmo parte para um novo jogador.

Como é referido anteriormente, os dados mais importantes para extração no website do *transfermarkt* são referentes às transferências, aos valores de mercado, aos minutos que o jogador jogou no campeonato e nas competições europeias pelo SLB e os jogos feitos pelo jogador ao longos dos anos na seleção nacional.

Assim a primeira coisa feita foi extrair todos os dados relativos a transferências, este que, em uma fase inicial, acaba por também conter informação sobre o mercado de valor do jogador na data da sua transferência.

Este bloco das transferências contêm não só transferências de clube para outro clube, como também subida de escalões dentro de um clube, para além disso, contêm a data em que foi realizada e o valor que o um clube teve de pagar por ele.

Relativo ao valor de mercado, o *transfermarkt* contêm um gráfico dinâmico com os valores de mercado do jogador no decorrer dos anos e em que clube se encontrava.

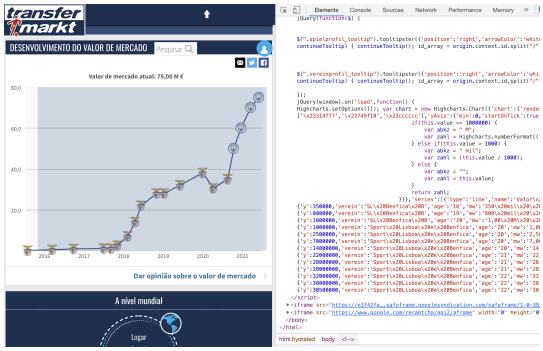

Figura 9: Valores de mercado: Website Transfermarkt

Acabou por ser um obstáculo extrair do gráfico os valores, pois os dados não estavam diretamente no código HTML, são colocados através de um código em *JavaScript*, pode-se ver isso pela figura 9. Depois de várias tentativas, nenhuma teve êxito, aqui tive ajuda de um dos meus superiores, Sudarshan Gopaladesikan,

que tem o cargo de *Head of Sports Data Science*, ele acabou por criar um algoritmo em *R* que retornava uma lista de listas que continha o id do jogador, a idade naquela data, a data, o clube e o valor de mercado. Seguidamente, fui para a página do jogador onde tem o registo por época e competição com jogos, minutos, golos, assistências, cartões. De uma forma menos pormenorizada já existe essa informação extraída do website do *ZeroZero*, mas aqui a diferença é que usei esta informação para outro fim. O relevante aqui passa por ser os número de minutos que o jogador teve em campo no campeonato nacional e nas competições europeias, pois com isso consegui calcular quanto a sua performance no clube deu em termos monetários ao SLB.

Para este calculo, houve uma pesquisa pelos vários documentos da UEFA para saber quanto é que o Benfica ganhou por época, tanto como valor de entrada, como valor de performance na competição.

```
listaReceitasCampeonato = [
    ['2019/20','Sport Lisboa e Benfica',34,9000000],
    ['2018/19','Sport Lisboa e Benfica',34,44000000]]
listaReceitasInternacional =[
    ['2019/20','Champions League','Sport Lisboa e Benfica',6,6300000],
    ['2019/20','Liga Europa','Sport Lisboa e Benfica',2,500000],
    ['2018/19','Champions League','Sport Lisboa e Benfica',6,6300000]]
```

Em cima está um pequeno excerto das duas listas, assim explicando, a *listaReceitasCampeonato*, contêm o valor de entrada da época seguinte, ou seja o Benfica na época 2018/19 foi campeão nacional, ou seja, na época 2019/20 teve entrada direta para a fase de grupos da Liga dos Campeões, o que só disso recebeu 44 milhões, mas nesta lista esse valor foi atribuído à época anterior pois foi a performance dessa época que fez originar este valor.

Já a lista *listaReceitasInternacional*, é referente à performance na época presente. Nas competições internacionais por cada vitória, empate, passar para uma nova fase recebe-se dinheiro e foi esse o calculo. De referir, como podemos ver no excerto de código, um clube numa época pode competir em duas competições diferentes(Por exemplo, Liga dos Campeões e Liga Europa).

```
money = 0
for i in campeonatoReceita:
```

```
    if (i[0] == season):
        if (i[1] == club):
            money =round((minutes/(i[2]*90)) * i[3],2)
```

Assim o dinheiro que o jogador deu ao Benfica a nível da sua performace foi calculado pelos minutos jogados a dividir pelos total de jogos da competição que multiplicou com 90 a multiplicar pelo valor que o Benfica recebeu.

Para finalizar a extração de dados, como o percurso a nível das seleções é importante para o negócio do clube e para uma análise, foi extraído todos os jogos em que cada jogador participou, desde jogos amigáveis a campeonatos do mundo.

Por fim, esta função retorna uma lista com todas as transferências, com todos os jogadores, com todas as épocas dos jogadores, com todos os jogos nas seleções, com todos os clubes e valores de mercado, este último no inicio foi utilizado, mas acabou por ser substituído por outro ficheiro.

Como isto é um algoritmo universal, teve de ser criado um função main(), esta chama a função anterior o número de vezes que o tamanho a lista anos tem. A lista anos contem os valores dos id's de cada época, e vai do valor 150 ao 140.

A maneira utilizada não foi a mais optimizada, mas acabou por juntar toda a informação de todas as épocas.

```
for i in range(lenn):
    if (listaIdEquipasZeroZero[x][i] != None):
        for j in anos:
            lista2020 = equipaA(x,i,j)
            lista1 = lista1 + lista2020[0]
            lista2 = lista2 + lista2020[1]
            lista3 = lista3 + lista2020[2]
            lista4 = lista4 + lista2020[3]
            lista5 = lista5 + lista2020[4]
            lista6 = lista6 + lista2020[5]
            lista7 = lista7 + lista2020[6]
```

Depois de ter as várias listas com a informação ao longo dos vários anos. Foi feita em cada lista uma revisão para ver se havia valores repetidos e caso houvesse eram eliminados.

Posteriormente começou a designação de id's nos vários dados. Isto foi feito para uma

melhor ligação dos dados. Assim cada liga ou clube teria o seu id e vez de dos dados das épocas aparecer o nome do clube, começou a aparecer o id do clube.

Por fim, depois de uma gestão dos dados, foi então gerado vários ficheiros excel's que foram usados nas seguintes fases.

```
with xlsxwriter.Workbook('teste1.xlsx') as workbook:
    worksheet = workbook.add_worksheet()
    for row_num, data in enumerate(a):
        worksheet.write_row(row_num, 0, data)
```

De referir que executar este algoritmo deu imenso problemas, pois no total recolheu informação de 823 jogadores, e sendo que fazia uma pausa de 180 minutos para não ser apanhado pelo sistema *anti-scripts* do *ZeroZero*, pelas contas dá um total de 41 horas, o que acabou por ser um obstáculo pois muitas vezes era encontrado um pequeno erro ou a rede ao qual o computador estava ligado falhava e tinha-se de voltar a repetir tudo do zero.

7.1.3.2 Merge com a BD do Benfica

Esta fase do projeto acabou por demorar mais do que o previsto. Devido a uns problemas técnicos, não consegui utilizar o programa *SQL server management* e durante esta fase acabei por perder uma grande parte dos ficheiros que já tinha feito *merge*. Passando estes problemas, acabei por entregar ao SLB dois ficheiros mais completos, neste caso o dos jogadores e das épocas, e três ficheiros novos, o dos valores de mercado, o dos jogos das seleções e o das transferências.

7.1.4 Dashboard - Power BI

Esta última fase do projeto, o objetivo foi criar um *dashboard* em *Power BI* que atrativamente consolidasse toda a informação que foi extraída e que facilmente desse para analisar e de-se para tirar conclusões.

Este *dashboard* continha um total de 11 relatórios:

Figura 10: Jogador - Geral

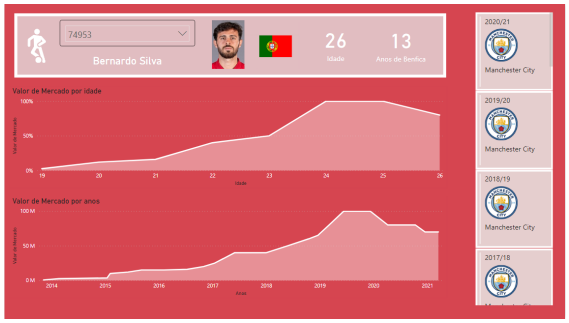

Figura 11: Jogador - VM

Figura 12: Jogador - Seleção

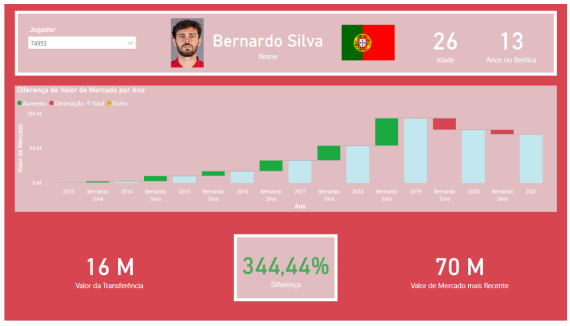

Figura 13: Valorização - VM

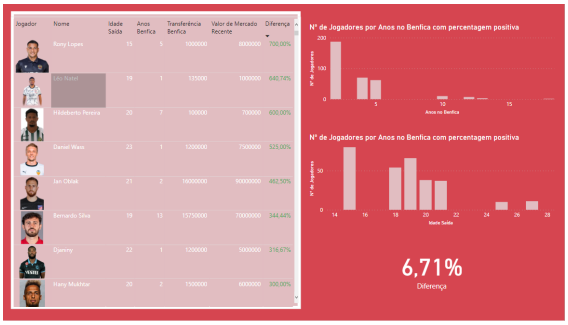

Figura 14: Valorização - Fora do Benfica

Figura 15: Top 5 Ligas

Este *dashboard* aborda vários assuntos, tento uma visualização tanto especifica a um jogador como a um conjunto de jogadores. Acaba por ser notório que todos os dados exportados foram importantes para a construção destes *reports*, dando várias oportunidades de análise importantes. Este *dashboard* foi construído durante várias semanas, com reuniões semana a semana, onde recebia tanto ideias para melhorar o que já estava feito, como novas ideias para implementar. Acabou por ficar um *dashboard* composto, com todos os assuntos importantes e que dão uma visão geral ou especifica do que se quer analisar.

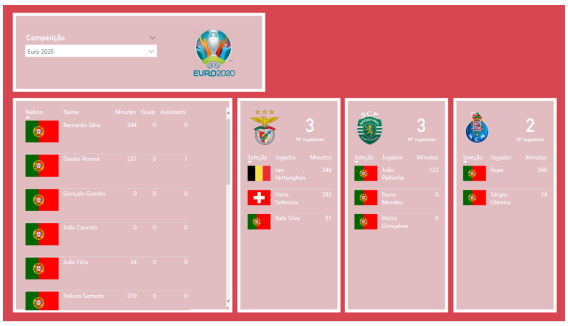

Figura 16: Jogadores - Seleção

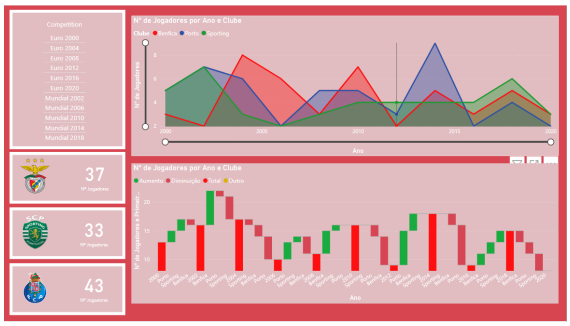

Figura 17: Seleção - Comparação Rivais

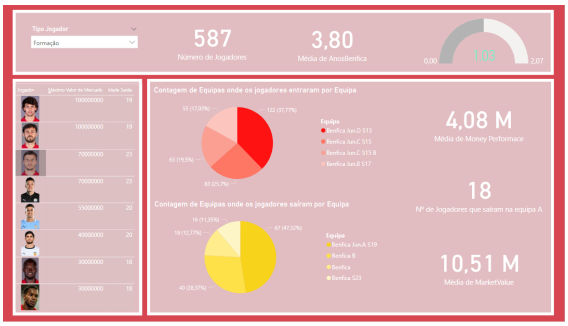

Figura 18: Seixal vs Out

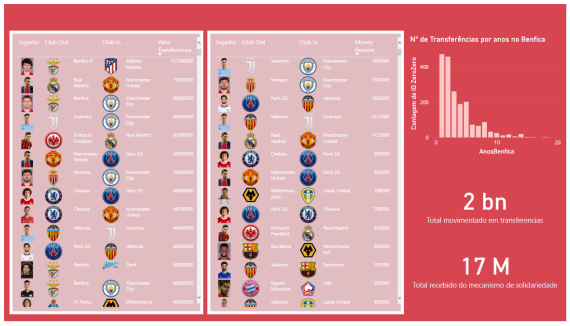

Figura 19: Transferências - Geral

Figura 20: Transferências Precoces

Report - Jogador Geral: Este *report* contêm informação geral do jogador, como nome, idade, posição, anos no Benfica, tipo de jogador, nacionalidade, dinheiro que rendeu ao Benfica pela sua performance e pelo mecanismo de solidariedade, número de transferências, percentagem do mecanismo de solidariedade e os seus jogos e golos por época desportiva.

Report - Jogador VM: Este *report* contêm informação do jogador relativo ao valor de mercado. Tem dois gráficos, um que mostra a evolução do valor de mercado ao longo da idade e outro ao longo dos anos. Para além disso, tem a informação de onde o jogador joga cada época.

Report - Jogador Seleção: Este *report* contêm informação do jogador relativo à seleção. O número de internacionalizações, as competições onde competiu, o número médio de minutos, a posição mais utilizada e o seu valor mercado pelo número de jogos por ano.

Report - Jogador Valorização: Este *report* contêm informação contem um gráfico com o evoluir do valor de mercado ao longos dos anos. Para complementar a informação foi calculado a valorização do jogador relativo ao valor da transferência de quando saiu do Benfica e o seu valor de mercado atual.

Report - Valorização Fora do Benfica: Este *report* contêm informação de uma forma mais geral a informação do *report* anterior, onde tem uma tabela dos jogadores filtrada pela maior valorização, onde tem os campos de idade de saída do Benfica e anos de Benfica. Para além disso, inclui dois gráficos, um com o número de jogadores por anos no Benfica com percentagem positiva e outro com o número de jogador por idade de saída do Benfica com percentagem positiva.

Report - TOP 5 Ligas: Este *report* contêm cinco listas de jogadores, cada um de uma liga diferente. As ligas selecionadas são consideradas as cinco ligas de topo, sendo que a liga portuguesa está em sexto. Aqui têm-se uma visão de quantos jogadores e que jogadores o Benfica pôs nestas ligas.

Report - Jogadores Seleção: Este *report* contêm informação dos jogadores que eram do Benfica e competiram numa competição internacional, mas também tem o extra de incluir uma comparação dos jogadores atuais dos três grandes que participaram na competição. De referir que estes dados dos convocados foram dados extras extraídos, para utilizados neste e no próximo report.

Report - Seleção Comparação Rivais: Este *report* contêm informação do número de jogadores que os três grandes colocaram nas competições desde o Euro 2000 até ao Euro 2020, incluindo os mundiais. Também tem uma análise de qual foi o crescimento ou decréscimo ao longos dos anos.

Report - Seixal vs Out: Este *report* faz uma comparação entre os jogadores que são considerados da formação, ou seja, *Made in* Seixal e os jogadores que vieram de fora. Neste caso, neste algoritmo um jogador é titulado como da formação se jogou num escalão do Benfica até aos 18 anos. Inclui valores como média da percentagem do mecanismo de

solidariedade, média anos no clube, número de jogadores que chegaram à equipa principal, média de valor de mercado, média de performance, e qual as equipas que tem maior número de jogadores em termos de entrada e saída.

Report - Transferências Geral: Este *report* contêm duas tabelas, uma com as maiores transferências que jogadores que representaram o Benfica geraram e outra igualmente com transferências mas filtrada pelas transferências onde o Benfica recebeu mais dinheiro do mecanismo de solidariedade da FIFA. Para além disso inclui um gráfico do número de jogadores por anos de Benfica referente à quantidade de transferências e o dinheiro gerado por estes jogadores, como o total recebido pelo mecanismo de solidariedade.

Report - Transferências Precoces: Este *report* contêm informação dos jogadores que saíram do Benfica mais cedo do que esperado e que acabaram também por chegar a ligas de topo ou gerar transferências com bons valores.

7.2 QA Tester

Este trabalho foi realizado ao longo do estágio, mas em períodos em que o projeto de *web scraping* estava parado. Como é referido anteriormente, eu convivi mais com a equipa de *software development*. Assim, sempre que a equipas de testes tivesse indisponível, e eu tivesse tempo para disponibilizar, era me pedido para testar novas funcionalidades ou *bugs* resolvidos. A maneira de realizar este trabalho era, recebia a informação que páginas testar e davam-me liberdade para testar de tudo, caso encontra-se bugs, colocava na plataforma *Microsoft Azure*, onde descrevia o *Bug* encontrado e assinalava o programador.

8 Análise Crítica

O Benfica desde 2017 tem apostado forte na área de inovação em tecnologias de informação, onde nesse ano começou o primeiro grande projeto de Transformação Digital no desporto, para a Microsoft Portugal. No Caixa Futebol Campus, quase tudo dos jogadores acaba por ser monitorizado, sua qualidade do sono, sua alimentação, o seu estado psicológico, o seu estado físico, o seu esforço nos treinos, no ginásio, nos jogos. Todos estes dados acabam por ser transmitidos para um data lake hospedado pelo Azure. Estas atividades acabam

por serem controladas pelo Laboratório de *Data Science* do Benfica, que é suportado pela Tecnologia Microsoft.

Com o uso de tecnologias de *Machine Learning* há uma otimização dos resultados desportivos no relvado, onde sabe-se qual a informação que irá levar ao sucesso. Esta informação ajuda os jogadores a melhorar a sua performance e faz-los crescer como jogadores.

A nível nacional não há outro clube ao mesmo nível e com uma aposta tão grande nesta área.

Devido a isso, e ao desenvolvimento do meu gosto pela área de Data Science com este mestrado que estava a frequentar, procurei durante o confinamento que nos foi imposto, desenvolver-me e especializar mais na área, para além de dois cursos com certificados que acabei por finalizar, fui procurando conhecer o mundo do trabalho e as várias áreas que poderia gostar. Até que me chegou à frente um workshop com o *Head of Sports Data Science* do Sport Lisboa e Benfica, onde acabou por juntar três segmentos que eu tenho um grande interesse, Data Science, Futebol e o Benfica.

Após ouvir atentamente tudo o que foi dito nesse workshop, acabei por fazer uma pesquisa por vagas no clube, onde acabei por me candidatar expontaneamente a um estágio que poderia nem existir. Acabei por ter a sorte de conseguir uma entrevista e de entrar.

Em relação a projetos, o projeto de *web scraping* acabou por ser desafiante e muito vantajoso para o meu desenvolvimento e que me fez subir as expectativas deste estágio. Não tinha a expectativa ser de ser colocado num projeto sozinho e completamente independente que neste caso foi pedido por órgãos superiores desta organização.

Ter esta responsabilidade tornou o estágio mais interessante e mais motivador, o que por um lado fez-me trabalhar mais para ter resultados positivos.

Os primeiros tempos acabou por ser de estudo, devido a nunca ter utilizado a técnica de *web scraping*. Depois dessa adaptação, foi começar a construção do algoritmo em pequenos passos e foi assim até ao final, em pequenos passos.

Durante este período de tempo, ao estar num projeto com uma grande dimensão e com a uma grande responsabilidade, apareceu-me pela frente vários obstáculos, muitos devido ao meu *know-how* ainda ser naquela altura um bocado baixo para o nível que estava a ser pedido. E isso fez-me crescer como trabalhador pois acabei por nunca pedir por ajuda nestas

dúvidas e tentei de tudo para conseguir resolver as coisas por mim.

Referente à etapa do projeto que implicou a criação de um *dashboard* em *Power BI*, de referir que teve um inicio tardio para o que estava planeado, devido a um atraso na etapa anterior. De referir que era a minha terceira interação com esta tecnologia, onde algumas criações de medidas tiveram de ter muito estudo por trás.

Fazendo uma análise aos gráficos criados, pode-se concluir que o Benfica tem imensos jogadores na formação, onde grande parte entra numa idade muito jovem, entre os 10 e 12 anos, mas há um grande número a sair ainda em equipas de formação, neste caso, nos Juniores e Juvenis (Sub-19 e Sub-17), onde de 587 jogadores apenas 18 saíram pela equipa principal. Este número também pode depender de vários fatores, como o treinador apostar na formação e os jogadores terem espaço para integrar o plantel e jogar regularmente, como ter o jogador ter imenso mercado, etc. Pode-se também referir que a média de percentagem do mecanismo de solidariedade é de 1% em 5, o que traduz que os jogadores acabam por ficar pouco tempo no clube.

Em relação à valorização do jogador após sair do Benfica, acaba por ser muita positiva pois há jogadores que valorizaram até 700% o que acaba por dignificar a formação do Benfica. No total dos jogadores é de referir que a percentagem é positiva e que acabam por ser jogadores que tiveram poucos anos no clube.

Referente às transferências, ao longo destes últimos o Benfica fez transferências de enorme valor, onde muitas estão na lista de maiores transferências feitas no futebol português e mundial. Muitos destes jogadores ainda continuam no clube para onde foram vendidos pelo Benfica o que faz que o valor de mecanismo de solidariedade ainda estar nos 11 milhões, onde este valor acaba por ser um valor excelente para este mecanismo.

Relativo às transferências precoces, de jogadores que acabaram por realizar grandes transferências, mas não no Benfica. Pode-se analisar que o Benfica, quando realizou as transferências apenas em uma delas é que recebeu dinheiro.

9 Conclusão

9.1 Desenvolvimento Pessoal e Profissional

A possibilidade de realizar um estágio para trabalho final de mestrado foi muito importante para o meu desenvolvimento pessoal e profissional uma vez que foi a minha primeira experiência profissional e o meu primeiro contacto com a realidade nesta área. A oportunidade de realizar um estágio no Benfica e de ter acesso a um dos clubes mais famosos do mundo e o meu clube do coração foi, de facto, um enorme privilégio.

Este estágio deu-me a possibilidade de pôr em prática conhecimentos adquiridos ao longo do mestrado e vivenciar a diferença entre o mundo académico e o mundo profissional de forma a encaixar ambos neste trabalho.

Acabou ser um desafio para mim estar envolvido em duas equipas diferentes de trabalho, isto deu-me motivação para trabalhar ainda mais e mais eficiente. São desafios destes que gosto de ter pela frente e que me motiva imenso para ser eficiente e perfeccionista.

Realizar este trabalho na área da *Data Science* e puder aprender com pessoas de renome nesta área foi muito gratificante para mim. Conhecer novas maneiras de trabalhar e técnicas que eram desconhecidas para mim foi algo muito importante no meu crescimento. Este estágio acabou por me abrir os horizontes e perceber que esta é a área que quero seguir no futuro, pois é uma área que admiro e que tento sempre aprofundar com leituras de livros ou cursos, onde este estágio acabou por corresponder ao meu entusiasmo.

Estou extremamente grato pela oportunidade que o SLB me deu e estou muito feliz com o resultado final. Acho que com o meu conhecimento prévio de *web scraping*, que não era nenhum , e de *Power BI*, que era o básico, e com o que consegui em apenas poucos meses, o resultado final é realmente satisfatório. A extração de dados foi feita com sucesso e o *dashboard* em *Power BI* ficou completo e atrativo, correspondendo às expectativas. Profissionalmente, esse estágio de 10 meses abriu-me as portas para o mercado de trabalho.

9.2 Problemas e Limitações

Foi um estágio diferente, um estágio em altura de pandemia. A ausência de contacto pessoal acabou ser algo limitativo em termos de desenvolvimento pessoal.

restringindo

Esta pandemia acaba por ser um problema/limitação pois não pude aproveitar este estágio ao máximo, só no último mês é que pude trabalhar localmente no estádio da luz, não tive a oportunidade de ir trabalhar ao centro de estágios no Seixal e não pude ter a 100% a realidade de como é o mundo do trabalho.

Outra limitação foi o acesso às várias tecnologias ou materiais no clube. Ao longo deste estágio houve períodos mortos, devido ao tempo elevado para ter um computador ou ter acesso à VPN (algo que acontece recorrentemente).

9.3 Sugestões para o futuro

De referir, que não posso apontar quase nada, porque as pessoas que semanalmente falava eram atenciosos e sempre que precisa-se de alguma coisa em poucos minutos tinha ajuda.

Posso dar a sugestão de definir um estagiário a apenas uma equipa, estar em duas equipas diferentes ao mesmo tempo para uma pessoa que está a ter o seu primeiro contacto com o mundo do trabalho, por momentos foi algo confuso.

Colocar-me num projeto sozinho foi algo desafiante e gostei porque acabei por evoluir muito, pois tentava descobrir por mim próprio soluções aos obstáculos que encontrava pela frente, onde acabou por durar o tempo todo do estágio, assim uma sugestão era tentar que houvesse mais diversidade de projetos ao longo do periodo de estágio.

Referências

[1] Benjamin C. Alamar. *Sports Analytics: A Guide for Coaches, Managers, and Other Decision Makers*. Columbia University Press, 2013. URL: http://www.jstor.org/stable/10.7312/alam16292.

[2] Sport Lisboa e Benfica. *Como nasceu o nosso Clube*. 2021. URL: https://www.slbenfica.pt/pt-pt/slb/historia/fundacao (acedido em 14/10/2021).

[3] Maura Borrego e Lynita Newswander. "Definitions of Interdisciplinary Research: Toward Graduate-Level Interdisciplinary Learning Outcomes". Em: *The Review of Higher Education* 34 (set. de 2010), pp. 61–84. DOI: 10.1353/rhe.2010.0006.

[4] Ram Sharan Chaulagain et al. "Cloud Based Web Scraping for Big Data Applications". Em: *2017 IEEE International Conference on Smart Cloud (SmartCloud)* (2017), pp. 138–143.

[5] Bhaskar Dastidar, Devanjan Banerjee e Subhabrata Sengupta. "An Intelligent Survey of Personalized Information Retrieval using Web". Em: *I.J. Education and Management Engineering* 5 (jul. de 2016), pp. 24–31.

[6] T. H.s Davenport. "Analytics in sports: The new science of winning". Em: *International Institute for Analytics* (2013), pp. 1–28.

[7] Thomas Davenport e D Patil. "Data Scientist: The Sexiest Job of the 21st Century". Em: *Harvard business review* 90 (out. de 2012), pp. 70–6, 128.

[8] Vasant Dhar. "Data Science and Prediction". Em: *Commun. ACM* 56.12 (dez. de 2013), pp. 64–73. ISSN: 0001-0782. DOI: 10.1145/2500499. URL: https://doi.org/10.1145/2500499.

[9] Fatmasari, Yesi Novaria Kunang e S. D. Purnamasari. "Web Scraping Techniques to Collect Weather Data in South Sumatera". Em: *2018 International Conference on Electrical Engineering and Computer Science (ICECOS)* (2018), pp. 385–390.

[10] Amir Gandomi e Murtaza Haider. "Beyond the hype: Big data concepts, methods, and analytics". Em: *International Journal of Information Management* 35.2 (2015), pp. 137–144. ISSN: 0268-4012. DOI: https://doi.org/10.1016/j.ijinfomgt.2014.10.007. URL: https://www.sciencedirect.com/science/article/pii/S0268401214001066.

[11] Bill Gerrard. "Moneyball and the Role of Sports Analytics: A Decision-Theoretic Perspect". Em: *2016 North American Society for Sport Management Conference (NASSM 2016)* 1 (2016). DOI: https://www.nassm.com/files/conf_abstracts/2016-143.pdf.

[12] Jiawei Han, Micheline Kamber e Jian Pei. "2 - Getting to Know Your Data". Em: *Data Mining (Third Edition)*. Ed. por Jiawei Han, Micheline Kamber e Jian Pei. Third Edition. The Morgan Kaufmann Series in Data Management Systems. Boston: Morgan Kaufmann, 2012, pp. 39–82. ISBN: 978-0-12-381479-1. DOI: https://doi.org/10.1016/B978-0-12-381479-1.00002-2. URL: https://www.sciencedirect.com/science/article/pii/B9780123814791000022.

[13] Aldo Hernandez-Suarez et al. "A Web Scraping Methodology for Bypassing Twitter API Restrictions". Em: *CoRR* abs/1803.09875 (2018). arXiv: 1803.09875. URL: http://arxiv.org/abs/1803.09875.

[14] E. Papalexakis K. Pelechrinis. "Proc. Elev. ACM Int. Conf. Web Search Data Min." Em: *WSDM 18* (2018), pp. 787–788.

[15] Tatti N. Kostakis O. e Gionis A. Springer. "Discovering Recurring Activity in Temporal Networks". Em: 31.6 (2017). DOI: https://doi.org/10.1007/s10618-017-0515-0.

[16] V. Krotov e M. Tennyson. "Research Note: Scraping Financial Data from the Web Using the R Language". Em: *Journal of Emerging Technologies in Accounting* 15 (2018), pp. 169–181.

[17] Richard Landers et al. "A Primer on Theory-Driven Web Scraping: Automatic Extraction of Big Data From the Internet for Use in Psychological Research". Em: *Psychological Methods* 21 (mai. de 2016). DOI: 10.1037/met0000081.

[18] Harold Larnder. “OR Forum—The Origin of Operational Research”. Em: *Operations Research* 32.2 (1984), pp. 465–476. DOI: 10.1287/opre.32.2.465. eprint: https://doi.org/10.1287/opre.32.2.465. URL: https://doi.org/10.1287/opre.32.2.465.

[19] Deanne Larson e Victor Chang. “A review and future direction of agile, business intelligence, analytics and data science”. Em: *International Journal of Information Management* 36.5 (2016), pp. 700–710. ISSN: 0268-4012. DOI: https://doi.org/10.1016/j.ijinfomgt.2016.04.013. URL: https://www.sciencedirect.com/science/article/pii/S026840121630233X.

[20] D. K. Mahto e L. Singh. “A dive into Web Scraper world”. Em: *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)* (2016), pp. 689–693.

[21] S. K. Malik e S. Rizvi. “Information Extraction Using Web Usage Mining, Web Scrapping and Semantic Annotation”. Em: *2011 International Conference on Computational Intelligence and Communication Networks* (2011), pp. 465–469.

[22] Andrew McAfee e Erik Brynjolfsson. “Big Data: The Management Revolution”. Em: *Harvard business review* 90 (out. de 2012), pp. 60–6, 68, 128.

[23] Ryan Mitchell. *Web Scraping with Python: Collecting Data from the Modern Web*. 1st. O’Reilly Media, Inc., 2015. ISBN: 1491910291.

[24] T. Moskowitz e L. J. Wertheim. “Scorecasting: The Hidden Influences Behind How Sports Are Played and Games Are Won”. Em: 2011.

[25] Prithwiraj Nath, Subramanian Nachiappan e Ramakrishnan Ramanathan. “The impact of marketing capability, operations capability and diversification strategy on performance: A resource-based view”. Em: *Industrial Marketing Management* 39.2 (2010), pp. 317–329. ISSN: 0019-8501. DOI: https://doi.org/10.1016/j.indmarman.2008.09.001. URL: https://www.sciencedirect.com/science/article/pii/S0019850108001326.

[26] Ji Ni et al. “A hybrid model for predicting human physical activity status from lifelogging data”. Em: *European Journal of Operational Research* 281.3 (2020), pp. 532–542. DOI: 10.1016/j.ejor.2019.05.03. URL: https://ideas.repec.org/a/eee/ejores/v281y2020i3p532-542.html.

[27] Carlos Portocarrero. *Players are human: On quantifying makeup and personality.* URL: http://wrigleyville.locals.baseballprospectus.com/2015/06/26/players-are-human-on-quantifying-makeup-and-personality/. (accessed: 09.04.2021).

[28] Foster Provost e Tom Fawcett. *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. 1st. O’Reilly Media, Inc., 2013. ISBN: 1449361323.

[29] Tobias Schoenherr e Cheri Speier-Pero. “Data Science, Predictive Analytics, and Big Data in Supply Chain Management: Current State and Future Potential”. Em: *Journal of Business Logistics* 36.1 (2015), pp. 120–132. DOI: https://doi.org/10.1111/jbl.12082. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jbl.12082. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/jbl.12082.

[30] Dayna Simpson et al. “Professional, Research, and Publishing Trends in Operations and Supply Chain Management”. Em: *Journal of Supply Chain Management* 51.3 (2015), pp. 87–100. DOI: https://doi.org/10.1111/jscm.12078. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jscm.12078. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/jscm.12078.

[31] C. Slamet et al. “Web Scraping and Naïve Bayes Classification for Job Search Engine”. Em: 2018.

[32] K. Sundaramoorthy, R. Vijaya Durga e S. Nagadarshini. “NewsOne — An Aggregation System for News Using Web Scraping Method”. Em: *2017 International Conference on Technical Advancements in Computers and Communications (ICTACC)* (2017), pp. 136–140.

[33] B.V.S. Ujwal et al. “Classification-Based Adaptive Web Scraper”. Em: *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. 2017, pp. 125–132. DOI: 10.1109/ICMLA.2017.0-168.

[34] Shreya Upadhyay et al. “Articulating the construction of a web scraper for massive data extraction”. Em: *2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT)*. 2017, pp. 1–4. DOI: 10.1109/ICECCT.2017.8117827.

[35] Matthew A. Waller e Stanley E. Fawcett. “Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management”. Em: *Journal of Business Logistics* 34.2 (2013), pp. 77–84. DOI: https://doi.org/10.1111/jbl.12010. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jbl.12010. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/jbl.12010.